FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2020/2021

Trabajo de Fin de Máster

TÍTULO: Análisis del perfil y predicción sobre la vacunación contra el COVID-19

***Alumno*: Mario González Radillo**

***Tutor*: Aída Calviño Martínez**

Junio (o Septiembre) de 2021

UNIVERSIDAD COMPLUTENSE
MADRID

ÍNDICE

1. INTRODUCCIÓN 1
2. OBJETIVOS 3
3. METODOLOGÍA Y MATERIALES 3
   3.1. Tipos de modelización 3
      3.1.1 kNN o vecino más próximo 3
      3.1.2 Regresión logística 4
      3.1.3 Redes neuronales 4
      3.1.4 *Random forest* y *bagging* 5
      3.1.5 Gradient boosting 6
      3.1.6 Support Vector Machines 6
      3.1.7 Ensamblados 6
   3.2. Técnicas de remuestreo y evaluación 6
      3.2.1 Partición de datos inicial 6
      3.2.2 Validación cruzada 7
      3.2.3 Validación cruzada repetida 7
      3.2.4 Evaluación de la calidad de las variables 7
      3.2.5 Evaluación de los modelos 7
      3.2.6 Evaluación del mejor punto de corte 7
   3.3. Materiales 8
      3.3.1 Barómetros del CIS 8
      3.3.2 *Software* y máquina 8
4. DESCRIPCIÓN DE LOS DATOS 8
5. DEPURACIÓN DE LOS DATOS 10
   5.1. Tratamiento de errores 10
   5.2. Tratamiento de atípicos 15
   5.3. Tratamiento de faltantes 16
   5.1. Estudio y selección de variables preliminar 18
6. MODELIZACIÓN 20
   6.1. KNN o vecino más próximo 20
      6.1.1 Mejor modelo kNN 24
   6.2. Regresión logística 25
      6.2.1 Mejor modelo con regresión logística 28
   6.3. Redes neuronales 30
      6.3.1 Mejor modelo de red neuronal 36

6.4. Random forest y bagging 37
6.4.1 Mejor modelo de *random forest/bagging* 41
6.5. Gradient boosting 41
6.5.1 Mejor modelo *gradient boosting* 51
6.6. Support Vector Machines 51
6.6.1 Mejor modelo de *Support Vector Machines* 55
6.7. Ensamblado 55
7. Selección del mejor modelo 56
7.1. Evaluación del mejor modelo 56
7.1.1 Evaluación del modelo con mayor capacidad predictiva y su mejor punto de corte 56
7.1.2 Evaluación del modelo con mayor equilibrio y su mejor punto de corte 57
8. CONCLUSIONES 58
9. BIBLIOGRAFÍA 60
ANEXOS 62
A. Cuestionarios. 62
B. Modelos. 85
C. Código 93
Código SAS® 93
Código R 128

1. INTRODUCCIÓN

La COVID-19 es una enfermedad infecciosa provocada por un coronavirus descubierta a finales de 2019 en Wuhan, China (OMS, 2020). La rápida extensión de esta enfermedad ha provocado la muerte de más de 2 millones de personas en todo el mundo y el paro de la actividad económica durante meses en 2020. Esta situación ha obligado a las grandes farmacéuticas y agencias de control de medicamentos a acelerar los procesos en busca de una vacuna que pueda hacer retroceder al virus y volver a la normalidad.

Sin embargo, la celeridad con la que se han creado estas vacunas genera dudas en la población sobre la fiabilidad y seguridad de estas. Es por esto que, mediante técnicas de minería de datos en este trabajo, se pretende predecir qué personas estarían dispuestas a que les sea administrada la vacuna y cuáles son las variables que más influyen en esta decisión. No obstante, antes de conocer los objetivos de este proyecto y cómo se llevará a cabo, es conveniente estudiar el estado actual del tema y algunos antecedentes.

En la actualidad, existen cuatro principales vacunas frente a la COVID-19. Dos de ellas, las de Pfizer y Moderna ya han sido autorizadas por el Center for Disease Control estadounidense (CDC, 2020) y tres por la EMA (Agencia Europea del Medicamento). A pesar de estos ensayos y las autorizaciones de las grandes agencias estatales, la población no está del todo convencida. Este sentimiento de rechazo se ha ido reduciendo a medida que han pasado los meses y se han ido inyectando dosis; pero los problemas que se asocian a la vacuna de Astrazeneca podrían provocar un aumento en este rechazo. En la Figura 1, se muestra la evolución de este rechazo.

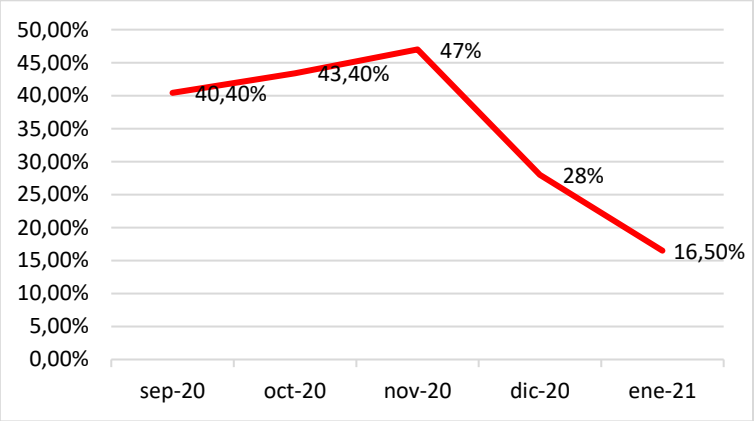


Figura 1. Evolución del rechazo a la vacuna entre septiembre de 2020 y enero de 2021. Fuente: elaboración propia a partir de datos del CIS (2021a).

El diario ABC (2021) asegura que los casos de trombos en algunas personas vacunadas con Astrazeneca han provocado que el porcentaje de personas que consideran esta alternativa como segura haya caído del 59% al 38%.

Aquí entran en juego las técnicas de minería de datos que se tratan en este máster y que permiten predecir determinadas variables objetivo si se cuenta con otras variables. La minería de datos no es nada nuevo en el ámbito sanitario ni en la gestión de la sanidad. El *big data* y su análisis ha permitido a los expertos en sanidad estudiar a distintos niveles: saber qué no está funcionando en la gestión de un hospital, conocer la importancia que tiene un determinado parámetro en un sistema sanitario, predecir la

evolución de los pacientes o todos los posibles resultados de una operación… (Sa, 2018). Tampoco las vacunaciones son ajenas a los procesos de minería de datos como señala Raeven et al. (2019), ya que el tratamiento y análisis de grandes bancos de datos es ya utilizado en el desarrollo de vacunas o incluso en su posterior manufactura (Auschitzky et al., 2014)

No obstante, si bien este proyecto de investigación trata un tema de carácter sanitario como es la vacunación contra la COVID-19; el objeto de este es el análisis del perfil de aquellas personas que rechazan esta vacuna; además de la generación de modelos que permitan predecir si un ciudadano estaría dispuesto o no a vacunarse dadas unas variables. Es, por tanto, un problema u objeto de estudio sociológico o relacionado con las ciencias sociales.

El movimiento antivacunas también se ha tratado desde la perspectiva de la Minería de datos. Por ejemplo, Chan et al. (2020), mediante modelos de geolocalización basados en Minería de datos, llegaron a la conclusión de que existía una correlación entre el porcentaje de vacunación contra el virus de la gripe en determinadas zonas de Estados Unidos y el contenido contra la vacunación generado en redes sociales en las mismas zonas. Aparte de los tweets correspondientes, también se tomó como fuente de datos numerosas encuestas sobre las actitudes de los ciudadanos estadounidenses frente a la vacunación. Así, el estudio de Chan et al. sí guarda una relación con este trabajo de fin de máster, donde los datos se tomarán de los barómetros realizados por el Centro de Investigaciones Sociológicas (CIS) en diciembre de 2020 y enero de 2021. Sin embargo, Chan et al. enfocaron su trabajo hacia la geolocalización, por lo que los métodos distarán de los de este proyecto.

Otro proyecto interesante es Mavragani y Ochoa (2018), que estudiaron cómo la minería de datos y búsquedas online referentes a vacunas afectaron a la hora de controlar el brote de sarampión de 2017 en la Unión Europea; aunque, como ocurría con el anterior, los métodos van más dirigidos al rastreo online de determinados temas.

Pero quizás el artículo más relevante para este proyecto hasta la fecha es Hornsey et al. (2021), que estudia la existencia de pequeños grupos de población muy específicos (“*small pockets”*) muy escépticos frente a las vacunas en general. Estos grupos son tan pequeños que los modelos de regresión habituales no son capaces de detectarlos. Si bien la educación, y sobre todo, la ideología política, son variables que guardan relación con el rechazo a las vacunas; los autores señalan que “es imposible utilizar un único tipo de modelo teórico que arroje luz sobre los matices del escepticismo antivacunas”. No obstante, la muestra de este estudio fue tomada antes de la pandemia, por lo que las percepciones entre la población general pueden haber cambiado.

Por otro lado, IBM (2020) coloca a los modelos predictivos como una herramienta perfecta para conocer el futuro de los clientes y crear campañas especializadas y segmentadas. Aunque esto tiene un enfoque comercial; se podría aplicar perfectamente en una situación en la que se debiera llegar a los colectivos o ciudadanos opuestos a la vacuna. Sin embargo, como bien se puede sacar en claro tras algunas consultas en la web y en el trabajo de Varian (2014), encontrar el modelo perfecto para cada situación es muy complicado si lo que se pretende es hacerlo de antemano. Gracias a la capacidad computacional actual, se puede probar con decenas de modelos para saber cuál funcionará mejor.

Tras este breve repaso del estado del arte, se llega a la conclusión de que si bien se han publicado multitud trabajos y proyectos sobre el uso de la Minería de datos en el desarrollo y logística de las vacunas; aquellas investigaciones de Minería de datos que atajan las vacunas con una perspectiva social lo han hecho hasta hace poco con modelos y herramientas destinadas al rastreo online y en redes sociales. Sin embargo, el estudio de Hornsey et al. sí encuentra algunas tendencias dentro de los perfiles de antivacunas; destacando la importancia del extremismo ideológico como principal generador de antivacunas. Este estudio, no obstante, cuenta con un conjunto de datos previo a la pandemia. La justificación de este proyecto de investigación reside en la gravedad de la pandemia y en la necesidad de reconocer qué perfiles son más propensos a oponerse a la vacuna.

Realizar un estudio de qué ciudadanos quieren o no vacunarse y conocer cuál es su perfil ayudaría para futuras acciones con el fin de cambiar la opinión de estos. Si se conoce con cierta exactitud cuáles son las características que definen a la persona que se opone a esta vacuna contra la COVID-19, será más sencillo llegar a ellos y generar mensajes que les influyan.

2. OBJETIVOS

El objetivo general de este trabajo de fin de máster es predecir qué personas se opondrían a vacunarse contra el COVID nada más estuviera disponible la vacuna.

A pesar de que el objetivo general es el más importante y el eje principal de este trabajo, también se plantean algunos objetivos secundarios:

- Conocer cuáles son las variables que influyen más en la respuesta a "¿Estaría Ud. dispuesto/a a vacunarse inmediatamente cuando se tenga la vacuna?".
- Analizar el perfil del ciudadano que se opone a ser vacunado.

3. METODOLOGÍA Y MATERIALES

3.1. Tipos de modelización

Dado que el objetivo principal de este trabajo es la detección de aquellas personas que se opondrían a ser vacunados; a lo largo de este trabajo se utilizarán distintos métodos de predicción, que se comentan a continuación.

En este apartado se explica brevemente en qué consisten estos métodos y se enumeran sus parámetros modificables. No obstante, se ahonda en las explicaciones sobre parametrización en el punto 6, que recoge la modelización en sí.

3.1.1 kNN o vecino más próximo

El kNN o *k-Nearest Neighbors* (donde *k* es el número de vecinos) es un método de clasificación no parametrizado que, a pesar de su simpleza, es efectivo en muchas ocasiones (Guo et al., 2003). Este modelo clasifica las observaciones en distintas clases dependiendo de cuantos de sus vecinos compartan las características con él.

Este algoritmo parte de la premisa de que los sujetos que se comportan de forma similar están cerca entre sí. Por eso mismo, se basa en el cálculo de la distancia euclídea entre los diferentes sujetos y los vecinos para clasificarlos.

Este método necesita un tratamiento previo de los datos para que puedan funcionar correctamente. Las variables *input* de intervalo necesitan ser estandarizadas o normalizadas y para las variables nominales, se han de crear variables *dummies*. Las variables *dummies* son variables binarias que se generan para cada una de las categorías de cada variable nominal y donde cada observación toma valores 1 y 0, dependiendo de si esa categoría era a la que correspondía la observación o no, respectivamente.

Además, los modelos kNN no son capaces de valorar la importancia de cada uno de las variables; por lo que son muy sensibles a la información que reciben y, sobre todo, al ruido de variables inútiles.

De este modo, son tres los conceptos que se deben tener en cuenta a la hora de generar este tipo de modelos:

- Tipo de estandarización.
- Sets de variables.
- *K* o número de vecinos.

3.1.2 Regresión logística

La regresión logística es uno de los conceptos de modelo más clásicos que aún se utilizan en la modelización predictiva. No obstante, no hay que dejarse engañar por la simplicidad de la regresión, pues para algunos conjuntos de datos resulta muy eficaz.

La regresión logística se basa en una función muy simple que la dota de una gran explicabilidad:

$$P = \frac{1}{1 + e^{-(\beta_1+\beta_1+\cdots+\beta_{\mathrm{mxm}})}}$$

El principal objetivo de la regresión es estimar los parámetros (β0, β1, β2…). De modo que el valor resultante sea la probabilidad de que ocurra un evento, en este caso, la probabilidad de que un sujeto se oponga a vacunarse

Trabajar con distintos números de variables según importancia tiene menos sentido con la regresión que con los kNN; ya que la regresión logística sí es capaz de detectar qué variables tienen mayor relación con la variable objetivo. Es por ello que los distintos elementos modificados para modelizar son otros:

- Tratamiento de las variables
- Inclusión de *dummies*
- Criterio de información
- Criterio de selección
- Inclusión de interacciones

3.1.3 Redes neuronales

Las redes neuronales son modelos de aprendizaje y procesamiento automático que tienen su inspiración en el propio sistema nervioso humano. Estas redes neuronales se

componen de nodos que se conectan entre sí y se organizan en capas. Así, estas conexiones en paralelo también se organizan de modo jerárquico. Como se puede apreciar en la Figura 1, la primera capa es la capa *input*, donde se incluyen las variables con las que se trabaja. Esta capa está unida a la capa oculta mediante unos enlaces que unen cada variable a cada nodo y a los cuales se les es asignado un peso (esto lo hace la función de combinación). Una vez asignados estos pesos, la función de activación es la encargada de agregar y calcular la probabilidad del evento o la salida. El esquema que toma una red neuronal se muestra en la Figura 2:

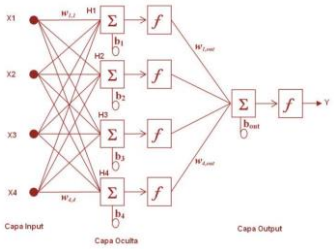


Figura 2. Arquitectura de una red neuronal, donde X son las variables input; H, los nodos; y W, los pesos. Fuente: Portela (2021)

En contraposición con la regresión logística, los resultados de las redes neuronales no son interpretables, pues es el cálculo es intrincado. Las redes neuronales son modelos mucho más complejos que los anteriormente estudiados, con unas posibilidades de parametrización casi infinitas. Así, los parámetros a tener en cuenta son los siguientes:

- Sets de variables
- Número de nodos
- Función de activación
- *Early stopping*
- Algoritmo de aprendizaje

3.1.4 *Random forest* y *bagging*

Los *random forest* son modelos basados en árboles de clasificación que toman distintas muestras del conjunto de datos con reposición y generan predicciones a partir de árboles que más tarde combinan. Los *random forests* se generan con un número de variables prefijado, pero qué variables son utilizadas se sortean de forma aleatoria.

El *bagging* es un tipo de *random forest* en el cual se toman siempre todas las variables *input* posibles; es decir, si un *dataset* contiene 5 variables *input,* los árboles generados por el *bagging* utilizarán 5 variables, mientras que con *random forest,* podrán utilizar 2, 3 o 4.

Los parámetros modificables para este tipo de modelos son los siguientes:

- Sets de variables
- Número de árboles
- Tamaño mínimo de hoja
- Número de variables
- Tamaño de muestra

3.1.5 Gradient boosting

El *gradient boosting* es un tipo de modelo iterativo cuyo algoritmo es capaz de calcular mediante árboles el valor (o predicción) de los residuos e ir ajustando las predicciones de forma iterativa hasta conseguir el menor error en la predicción posible.

Al ser, como *random forest*, un modelo basado en árboles, el *gradient boosting* comparte con él algunos de sus parámetros, que son los siguientes:

- Algoritmo (GBM o XGBoost)
- Sets de variables
- *Shrinkage*
- Número de árboles
- Tamaño mínimo de hoja
- Porcentaje de variables utilizadas (solo con XGBoost)
- Reguladores (solo con XGBoost)

3.1.6 Support Vector Machines

Los *Support Vector Machines* (SVM) son algoritmos de aprendizaje que permite resolver problemas de clasificación mediante la separación de diferentes dimensiones por factores. Existen tres tipos principales de SVM: lineal, polinómico y radial. Para este trabajo, se van a utilizar únicamente el SVM lineal y el SVM radial; excluyendo el SVM polinómico. Esta exclusión viene dada por dos motivos. En primer lugar, el SVM polinómico es el que ofrece mejor resultado en muy pocas ocasiones; ya que, si el conjunto se beneficia de modelos lineales como la regresión logística, el SVM lineal será el mejor; y si no es así; el SVM radial es capaz de detectar las correlaciones no lineales mejor que el polinómico. En segundo lugar, la capacidad computacional necesaria para llevar a cabo el SVM polinómico (el más costoso de los 3) no compensa para generar modelos que probablemente no destaquen frente a SVM lineal o radial.

De este modo, los parámetros con los que se trabaja en SVM:

- Tipo de SVM
- Sets de variables
- C
- *Sigma* (solo con SVM radial)

3.1.7 Ensamblados

En ocasiones, tras la obtención de las predicciones de los mejores modelos; estas puedes combinarse para reducir la varianza del error. En el caso de este trabajo, se tomarán las mejores alternativas de cada uno de los 6 modelos anteriormente citados y se ensamblarán.

3.2. Técnicas de remuestreo y evaluación

3.2.1 Partición de datos inicial

Con el fin de poder evaluar el modelo final sobre una porción de los datos con la que no se haya trabajado hasta el momento, se realiza una partición de datos por la que se toma el 80% de las observaciones como conjunto de entrenamiento o *train* y un 20% como conjunto de *test*.

3.2.2 Validación cruzada

Es una técnica de remuestreo que consiste en la división del conjunto *train* en *k* grupos (en el caso de este trabajo, 5). Con cuatro de ellos se genera el modelo y con el restante (grupo *w*), se evalúa su error. Este proceso se repite de modo que las cinco divisiones hayan sido el grupo *w*.

Esta metodología se utiliza en muy pocas ocasiones en este trabajo, ya que no permite hacer un buen estudio de la varianza. Por este motivo, tan solo se utiliza para hacer estudios preliminares.

3.2.3 Validación cruzada repetida

Esta técnica utiliza la anterior validación cruzada, pero de modo repetido cambiando la semilla aleatoria que genera la partición de los 5 grupos. Para este trabajo se realiza validación cruzada repetida de 50 repeticiones en casi todas las ocasiones; salvo algunas pruebas preliminares con *random forest* (se realizan 10 repeticiones).

Cualquier comparación entre modelos expuesta en este trabajo se hace bajo una idéntica técnica de remuestreo; excepto las comparaciones de paquetes, cuyas peculiaridades se exponen más adelante.

3.2.4 Evaluación de la calidad de las variables

Para estudiar el grado de relación de las variables *input* entre sí y con la variable objetivo, se utiliza el coeficiente V de Cramer aplicado al chi-cuadrado. Este coeficiente toma valores entre 0 y 1; siendo 0 la inexistencia de relación; y 1, relación total.

3.2.5 Evaluación de los modelos

Para evaluar la calidad de los modelos, se utilizará el índice bajo la curva ROC o AUC (*Area Under Roc*). Esta curva representa la sensibilidad (capacidad de detectar los eventos) y especificidad (capacidad de detectar los no-eventos) de un modelo. Así, a diferencia de otros indicadores, el área bajo esta curva no varía dependiendo del punto de corte de probabilidad que se le aplica al modelo (con el que se trabaja una vez escogido el mejor modelo).

A pesar de esto, para los modos de redes neuronales, en el que existe un cambio de *software* como se explicará más adelante, no se puede obtener este indicador; se realizan las comparaciones con tasa de fallos o "*missclasification rate*", que es la proporción de observaciones mal clasificadas.

3.2.6 Evaluación del mejor punto de corte

Al generar modelos para predecir variables objetivo binarias, los modelos llegan a una probabilidad del evento para cada observación. Como norma general, se toma la probabilidad del 0.5 como el punto del corte a partir del cual se considera a esa observación como evento ("1", "Yes"). Utilizar diferentes puntos de corte puede ayudar a tener modelos más acordes a las necesidades de los proyectos. No obstante, estos cambios provocan cambios en la precisión de los modelos; por lo que hay que tener claros algunos conceptos:

- Sensibilidad. Es la capacidad del modelo de detectar los verdaderos positivos o las observaciones que realmente son "evento".

- Especificidad. Es la capacidad del modelo de detectar los verdaderos negativos o las observaciones que no son "evento".
- *Pos Pred*. Es la proporción de verdaderos positivos de aquellas observaciones clasificadas como "evento".
- Índice de Youden. Es un indicador del rendimiento de un modelo de predicción. Cuanto mayor sea, mejor. Sigue la siguiente fórmula:

$$J = Sensibilidad + Especificidad - 1$$

3.3. Materiales

3.3.1 Barómetros del CIS

La base de datos con la que se trabaja en este proyecto se obtiene de la página web del CIS, que permite descargarse el fichero integrado de datos en formato SPSS y corresponde a los barómetros de noviembre y diciembre. Más adelante, en el punto 4 se hace una descripción de esta base de datos.

3.3.2 *Software* y máquina

Para la realización de este trabajo se ha utilizado: SAS® Entreprise *Miner* 14.1 para la depuración y tratamiento previo de los datos; SAS® Base 9.4 para la modelización de redes neuronales; y R para la modelización con el resto de metodologías. En este último, se ha acudido a las librerías de *caret, readxl, writexl, sas7bdat, greybox, dummies, pROC, MASS, reshape, dplyr* y *highcharter*.

A pesar de que la mayoría de este trabajo, sobre todo en cuanto a modelización se refiere, se realiza en R; se opta por utilizar SAS Base en la modelización de redes neuronales; ya que las capacidades de parametrización son mucho mayores en este *software* que en R.

La máquina utilizada para la realización de este trabajo es un *Macbook Air* con procesador Intel i5, 8 GB RAM y *Windows* 10 (*Bootcamp*).

4. DESCRIPCIÓN DE LOS DATOS

El conjunto de datos de este proyecto se obtiene a partir de los barómetros de diciembre de 2020 y enero de 2021 del CIS.

El CIS es un organismo que depende del Ministerio de la Presidencia y cuya principal finalidad es la realización de estudios estadísticos precisos que permitan a los poderes públicos contar con una base sobre la que cimentar sus políticas y administraciones. Las encuestas son la labor más conocida que lleva a cabo el CIS; aunque no es la única, ya que también colabora en apoyando la formación y la investigación en el campo de las ciencias sociales (CIS, 2021b).

Los barómetros del CIS cuentan con una afijación proporcional y se trata de encuestas por ordenador asistidas por vía telefónica que plantean una serie de preguntas básicas recurrentes y algunas otras más actuales (referentes a unas elecciones autonómicas próximas, por ejemplo). El error muestral de estos barómetros se fija en ±1,8% para un nivel de confianza del 95,5%.

Para este trabajo, no se han tomado todas las variables (preguntas) realizadas en los barómetros del CIS por distintos motivos. En primer lugar, a pesar de que los barómetros compartan muchas de sus preguntas; hay determinadas variables que solo se preguntan en meses determinados o pequeñas variaciones que obligan a descartar aquellas variables o preguntas que aparecen en tan solo uno de los dos barómetros utilizados.

En segundo lugar, también se han descartado las preguntas referentes al trato recibido por aquellos entrevistados diagnosticados con COVID, ya que no representaban un porcentaje suficiente de la muestra.

También se han rechazado preguntas muy complejas que se dividían en decenas de variables, y que contaban con preguntas más simples relacionadas. Por ejemplo, una de las preguntas del barómetro aludía a los aspectos en los que la pandemia afectaba a la vida personal de los entrevistados. Esta pregunta se dividía en 13 variables *dummy* con cada uno de los aspectos recogidos por el CIS. Justo antes de esta pregunta, se preguntaba a los encuestados en qué medida había afectado la pandemia a su vida personal. Se opta por tomar solo la segunda de sus variables.

Por último, se rechazan todas las variables que hacen referencia a por qué un entrevistado ha decidido no responder la encuesta; ya que es información que no interesa.

En el Anexo A, se pueden encontrar todas las preguntas y variables de los barómetros de diciembre y enero.

Al unir ambos barómetros, este conjunto cuenta con algo menos de 7700 observaciones y 42 variables. La Tabla 1 resume estas variables.

Tabla 1. Variables: descripción, rol y tipología

Variable	**Descripción**	**Rol - Tipo**
AFE_VIDPER	Nivel de afectación del COVID a la vida personal	Input - Nominal
AFE_VIDSOC	Nivel de afectación del COVID a la vida social	Input - Nominal
CAPITAL	Capital de provincia u otro	Input - Nominal
CCAA	Comunidad autónoma	Input - Nominal
CERCANIA	Partido más cercano	Input - Nominal
CIVIS_COVID	Civismo frente a la situación del COVID	Input - Nominal
CLASESOCIAL	Clase social	Input - Nominal
CNO11	Empleo	Input - Nominal
ECIVIL	Estado civil	Input - Nominal
EDAD	Edad	Input - Intervalo
EFEC_COVID	Señala cuales efectos del COVID le preocupan más	Input - Nominal
ESCIDEOL	Escala ideológica (1 Izquierda – 10 Derecha)	Input - Intervalo
ESCUELA	Indica si ha ido a la escuela o no	Input - Nominal
ESTUDIO	Número del estudio	Input - Nominal
ESTUDIOS	Nivel de estudios	Input - Nominal
GOB_ENCAR	El gobierno que debería encargarse de la gestión de la pandemia	Input - Nominal
INTENCIONG	Intención de voto en elecciones generales	Input - Nominal
INTENCIONGALTER	Intención alternativa de voto en las elecciones generales	Input - Nominal
NACIONALIDAD	Nacionalidad	Input - Nominal

NIVELESTENTREV	Nivel de estudios	Input - Nominal
PARTICIPACIONG	Participación en las últimas generales	Input - Nominal
PRACTICARELIG	Nivel de práctica religiosa	Input - Nominal
PREF_PRES	Preferencia como Presidente del Gobierno	Input - Nominal
PREO_COVID	Nivel de preocupación frente al COVID	Input - Nominal
PRO_PRI	Principal problema actual	Input - Nominal
PRO_SOC	Principal problema social actual	Input - Nominal
PROV	Provincia	Input - Nominal
RECUVOTOG	Votación en las últimas elecciones	Input - Nominal
RELIGION	Religión	Input - Nominal
SEXO	Sexo	Input - Nominal
SIMPATIA	Partido con el que más simpatía tiene	Input - Nominal
SINT_COVID	Indica si ha presentado o no síntomas de COVID	Input - Nominal
SITLAB	Situación laboral	Input - Nominal
TAMUNI	Tamaño del municipio	Input - Nominal
OPOSVACUNA	Indica si el entrevistado se opone o no a vacunarse	Objetivo - Nominal
VAL_ECO	Valoración economía española	Input - Nominal
VAL_ECO_PER	Valoración economía personal	Input - Nominal
VAL_IA	Valoración de Inés Arrimadas	Input - Intervalo
VAL_PC	Valoración de Pablo Casado	Input - Intervalo
VAL_PI	Valoración de Pablo Iglesias	Input - Intervalo
VAL_PS	Valoración de Pedro Sánchez	Input - Intervalo
VAL_SA	Valoración de Santiago Abascal	Input - Intervalo

5. DEPURACIÓN DE LOS DATOS

La primera de las fases antes de trabajar en la modelización predictiva o el análisis del perfil es la depuración de los datos. Este proceso es de suma importancia, ya que evitará futuros problemas o desajustes que puedan empeorar o invalidar las fases posteriores.

Aunque podría realizarse la depuración sin explicar el procedimiento, conviene realizar una memoria de los cambios y acciones llevadas a cabo para el posterior análisis de los resultados. En primer lugar, se estudiará el documento en busca de errores en la toma de los datos o incongruencias. A continuación, se revisarán los datos atípicos del conjunto de datos y valorará si es necesario tratarlos o no. Por último, se deben tratar los datos faltantes o *missing* y decidir cuál es la mejor decisión con respecto a estos.

Casi cualquier software sería válido para la realización de esta fase, la depuración de este conjunto de datos se llevará a cabo en SAS *Miner*; ya que su carácter visual facilita la visualización de los datos y no perjudica en términos de computación; dado que no son procesos muy exigentes.

5.1. Tratamiento de errores

En este apartado, hay que explorar los datos en busca de posibles errores o problemas que solucionar antes de valorar los atípicos o gestionar los faltantes. Estos dos últimos ya son gestiones que alteran de manera considerable los datos; por lo que se realizan más adelante, una vez generados conjuntos de *train* y *test*.

El tratamiento de estos errores es diferente para las variables de intervalo y las variables de clase.

Las variables de intervalo no son muy numerosas en este conjunto de datos, que tan solo cuenta con 7 variables de este tipo (de 42): la edad, la ideología del entrevistado y las valoraciones ideológicas de los 5 principales líderes políticos.

Variables de intervalo

Los errores de este tipo de variables suelen ser valores que no están dentro de los máximos o mínimos esperables de la variable continua o *missings* que el creador del conjunto de datos señala con números como 999 o -1.

La variable "edad" tiene como requisito principal un valor mínimo de 18 años y no tener valores con edades desproporcionadas. Los valores de esta variable están comprendidos entre 18 y 98; por lo que no es necesaria una subsanación de errores.

El resto de las variables de intervalo son las valoraciones de ideología; tanto del propio entrevistado como de los principales líderes políticos españoles. En este caso, sí es necesario abordar los errores. En todas estas variables, los valores deben de estar comprendidos entre 1 y 10. Tras un breve análisis se puede comprobar como en todas ellas existen valores por encima de 10. A continuación, en la Figura 3 se puede apreciar que los máximos de estas variables están por encima de 10.

Variable	Etiqueta	Ausente	N	Mínimo	Máximo
EDAD	EDAD	0	7679	18	98
ESCIDEOL	ESCIDEOL	0	7679	1	999
VAL_IA	VAL_IA	0	7679	1	999
VAL_PC	VAL_PC	0	7679	1	999
VAL_PI	VAL_PI	0	7679	1	999
VAL_PS	VAL_PS	0	7679	1	999
VAL_SA	VAL_SA	0	7679	1	999

Figura 3. Máximos y mínimos de variables de intervalo

Todos estos valores por encima de los límites se sustituyen por valores ausentes; que se tratan más adelante.

Variables de clase

En el caso de las variables de clase, no existen, obviamente, límites numéricos dentro de los cuales tengan que estar comprendidos los valores de las variables. En este apartado, son tres los errores habituales a los que hay que prestar atención: categorías que en realidad representan *missings*, posibles duplicidades o valores que representen lo mismo; y categorías que representen menos de un porcentaje predeterminado de las observaciones. Para este trabajo y este conjunto de datos, se decide establecer este límite entorno al 2%; lo que se traduce en 154 observaciones; un número que se considera lo suficientemente grande.

Como medida general, se han sustituido todas aquellas categorías que hacen referencia a una falta de respuesta (N.C., N.S. O NA, referentes a "No Contesta", "No Sabe" y "Ausente", respectivamente) por ausentes. Esto ocurre en las variables AFE_VIDPERS, AFE_VIDSOC, CERCANIA, CIVIS_COVID, CLASE_SOCIAL, CNO11, ECIVIL, EFECCOVID,

ESCUELA, ESTUDIOS, GOB_ENCAR, INTENCIONG, INTENCIONGALTER, NIVELESENTRE, PARTICIPACIONG, PRACTICARELIG, PRO_PRI, PRO_SOC, RECUVOTOG, RELIGION, SIMPATIA, SINT_COVID, SITLAB, VACUNA, VAL_ECO y VAL_ECO_PER

A continuación, se muestra un resumen de los cambios realizados en aquellas variables de clase más allá de los N.C. o los N.S.

AFE_VIDPERS

En cuanto a esta variable que recoge cómo ha afectado la pandemia a la vida personal de los encuestados, la categoría "Regular" se incluye dentro de la categoría de "algo"; ya que tan solo representaba al 1,5% de las observaciones.

AFE_VIDSOC

Esta variable recoge la afectación del COVID a la vida social de los entrevistados. El error subsanado es exactamente el mismo que en la anterior.

CCAA

En la variable CCAA, existen algunas categorías con menos del 2%: Cantabria, Navarra, La Rioja, Ceuta y Melilla. No obstante, las tres primeras se mantendrán en su propio nivel; ya que todas ellas alcanzan el 1,5% de las observaciones, un número aún suficiente. Esto permite no perder información muy importante. Por tanto, el único error que se subsana es la inclusión de Ceuta y Melilla en Andalucía; por cuestiones meramente geográficas.

CERCANÍA

Esta variable, que hace referencia al partido político con el que el entrevistado se siente más identificado, cuenta con 26 categoría distintas. No obstante, tan solo 11 de estas alcanzan el 2% de observaciones establecido como mínimo. Dos de ellas, además, son N.C y N.S., que se transforma a *missing*. Con el resto de las categorías bajo el umbral del 2%, se ha decidido unirlas a todas bajo una de las categorías ya existente de "Otro Partido", salvo aquellas en cuyas siglas se incluya a Podemos. Estas últimas se incluyen en la categoría "Podemos".

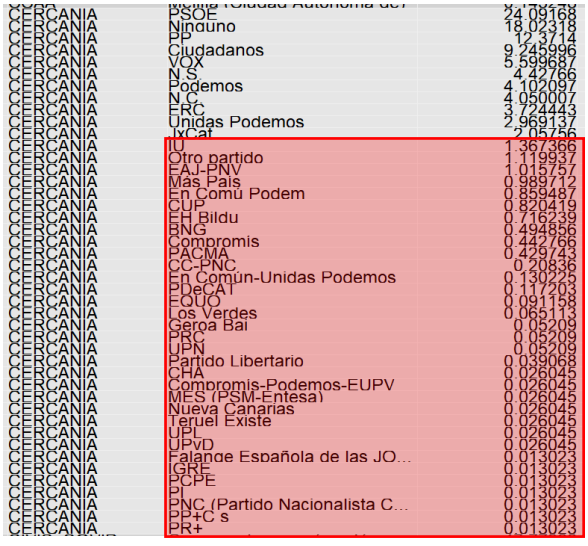


Figura 4. Variable "Cercanía". En rojo, las categorías que no alcanzan el 2% de observaciones

Esta es una buena variable para mostrar cómo se realiza este proceso de tratamiento de errores. En primer lugar, se comprueba si todas las categorías alcanzan el límite preestablecido (2%) y si no es así, cuáles están por debajo. Para esta variable, por ejemplo, todas las categorías señaladas en rojo en la Figura 4 no alcanzan ese umbral.

Todas estas categorías se modifican con los patrones señalados anteriormente para llegar a la siguiente variable, que pasa a tener tan solo 10 categorías (Figura 5)

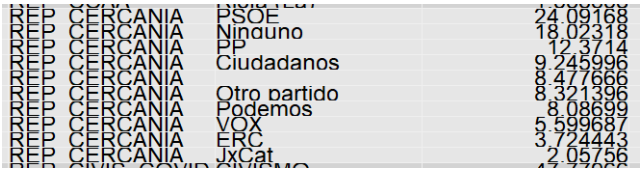


Figura 5. Variable "Cercanía" una vez agrupadas categorías

CIVIS COVID

En esta categoría, solo se unen las categorías “No sabe, duda” y “No lo sabe, duda”; que, obviamente, se refieren a lo mismo.

CLASE SOCIAL

En esta variable, existen algunas categorías por debajo del 2%. “Clase pobre”, “proletariado” y “a los de abajo” se incluyen en “Clase baja”. “Clase alta” se une a “Clase media-alta”. El resto de estas categorías minoritarias (“No cree en las clases”, “Excluidos”, “A la gente común”) se incluyen en una categoría ya existente denominada “Otras”.

CNO11

Esta variable hace referencia a la profesión de los entrevistados. Se corrigen dos errores de duplicidades de categorías que se habían dividido en dos innecesariamente (“Profesionales y científicos…” y “Oficiales, operarios…”). El resto de las categorías bajo el umbral del 2% (Agricultores, Operadores, Militares…) se unen bajo una categoría ya existente: “Otra”.

EFECCOVID

Para esta variable, se unen las categorías de “Ambos por igual” y “Ni unos ni otros”; por ser muy semejantes.

ESCUELA

En la variable que recoge si el entrevistado ha ido a la escuela; se unen las dos categorías que hacen referencia a entrevistados que no han acudido a la escuela. Aun así, estas categorías del “no” no alcanzan ni siquiera el 1,5% de las observaciones. Casi todas las observaciones restantes (un 98,5%) corresponden a “Sí”. Más adelante se valora si es conveniente quedarse con esta variable.

ESTUDIOS

En esta variable, se incluye la categoría “otros” (inferior al 2%) en “Sin Estudios”.

GOB_ENCAR

Esta variable recoge las opiniones de los entrevistados sobre quién debe ser el encargado de llevar el control y gestión de la pandemia. Se recogen bajo la categoría “Otras respuestas”, aquellas respuestas con menos del 2% de las observaciones.

INTENCIONG

Esta variable sigue la tendencia de la variable “CERCANÍA”; con un gran número de categorías, pero no muchas que superen el 2%. Los cambios son muy similares a los de esa variable, aunando en una solo categoría las categorías bajo el umbral del 2%; salvo las que incluyen a Podemos en sus siglas (En Comú, Compromis…)

INTENCIONGALTER

Lo mismo ocurre con esta variable, que recoge la segunda alternativa política de los votantes y se realizan exactamente los mismos cambios.

NIVELESENTRE

Esta variable recoge el nivel de estudio de los entrevistados y el número de categorías es alto, por lo que hay varias por debajo del umbral del 2%. Se unen todas las categorías referentes a posgrado, se unen arquitecturas e ingenierías de todo tipo. El resto de las categorías se recogen en una ya existente denominada “Otro”.

PARTICIPACIONG

Para esta variable, se unen las categorías en dos bloques: “Sí votó” y “No voto”; para poder incluir a aquellas categorías con menos del 2% de representación.

PROV

En esta variable, existen numerosos niveles por debajo del 2%; y dado que estos niveles con pocas observaciones son aun más habituales en aquellas categorías pertenecientes a comunidades autónomas multiprovinciales (donde más interesante podría resultar); se decide eliminar esta variable y trabajar con las CCAA.

PRO_PRI

Esta variable era una pregunta abierta en la encuesta del CIS, por lo que el número de categorías es inmenso. Tan solo se mantienen aquellas con más de un 2% de observaciones. El resto se unen a la categoría ya existente “Otras respuestas”.

PRO_SOC

En la variable sobre problemas sociales, se llevan a cabo los mismos cambios que en la anterior.

RECUVOTOG

Esta variable hace referencia al voto en las anteriores elecciones. El tratamiento de esta variable es exactamente igual al de las variables anteriores que hacen referencia a

partidos políticos. Se unen todas las plataformas de Podemos y los demás partidos que no superan el 2% se incluyen en la categoría “Otro partido”.

SIMPATÍA

Se aúnan todas las categorías con menos del 2% en “Otro partido”.

SITLAB

En esta variable referente a la situación laboral, se unen los dos tipos de entrevistados en paro (los que han trabajado antes y los que no).

OPOSVACUNA

Esta es la variable objetivo de este estudio y, por tanto, se debe ser cauteloso a la hora de hacer cambios. Sin embargo, se necesita tan solo dos categorías (binario); por lo que todas las categorías que representaban a entrevistados que se vacunarían con condiciones se incluyen dentro de la categoría “Sí”. El resto (N.S.; N.C.; Otras respuestas) se convierten en ausentes. Ya que no es recomendable imputar ni tratar de ningún modo los ausentes de una variable objetivo, se filtra la muestra para eliminar todas aquellas observaciones que no respondieran a esta pregunta. Se eliminan, por tanto, en torno a 550 observaciones, que corresponden a un 7,27% de la muestra.

Además, con el fin de adecuarla a los objetivos del trabajo; se decide invertir las categorías de esta variable, para que su significado haga referencia a la oposición a vacunarse. De este modo, los opuestos a la vacuna aparecen en esta variable como “Si”, “Yes” o 1 (evento), dependiendo del *software* en el que se trabaja.

La proporción del evento en la variable objetivo es del 23,46%.

VAL_ECO

Se sustituyen los N.C. y N.S. por *missing*; y se incluye la categoría “Muy buena” (menos del 2% de las observaciones) dentro de “Buena”.

VAL_ECO_PER

Se sustituyen los N.C. y N.S. por *missing*.

Una subsanados estos errores, se comprueba si existe alguna variable cuyo número de *missing* sea superior al 50% de las observaciones. En este caso, solo la variable “SIMPATÍA” supera este umbral. De este modo, esta no cuenta con la información suficiente para ser útil, por lo que se elimina.

A continuación, se realiza una partición de datos de 80% *train* y 20% *test* antes de realizar cambios que de verdad modifiquen el *dataset*.

5.2. Tratamiento de atípicos

La gestión de los valores atípicos en las variables de intervalo es importante para conseguir modelos más estables; ya que algunas clases de estos se ven muy afectadas por los llamados *outliers*. Dependiendo del nivel de asimetría de la variable, se debe utilizar un método de reemplazo para los atípicos u otro. En caso de que la variable se distribuya de forma más o menos simétrica (un índice de asimetría entre -1 y 1), se utiliza

la desviación estándar como método de reemplazo. En el caso de que esta distribución sea asimétrica, se debe acudir al MAD o método de distribución absoluta media.

Como se comentó anteriormente, son 7 las variables de intervalo de este conjunto de datos: la edad, la valoración de la ideología propia y las valoraciones de los 5 principales líderes políticos.

Tras analizar los índices de asimetría de estas variables, se concluye que todas tienen una distribución lo suficientemente simétrica para utilizar la desviación estándar como método de reemplazo de atípicos salvo la variable que indica como los entrevistados valoran a Santiago Abascal. Esta variable tiene una asimetría de 1.78, y por tanto, obliga a utilizar el método MAD.

Sin embargo, *Miner* tan solo detecta atípicos en la variable sobre la valoración a Pablo Casado. En concreto, 18. De este modo, solo se reemplazan los atípicos de esta variable. En la Figura 6, se muestran los resultados de este tratamiento de atípicos:

Variable	Rol	Etiqueta	Entrenamiento
EDAD	INPUT	EDAD	0
REP_ESCIDE	INPUT	Replacement: ESCI...	0
REP_VAL_IA	INPUT	Replacement: VAL_IA	0
REP_VAL_PC	INPUT	Replacement: VAL_PC	18
REP_VAL_PI	INPUT	Replacement: VAL_PI	0
REP_VAL_PS	INPUT	Replacement: VAL_PS	0
REP_VAL_SA	INPUT	Replacement: VAL_SA	0

Figura 6. Variables de intervalo donde se tratan los atípicos

5.3. Tratamiento de faltantes

En primer lugar y antes de trabajar con los datos ausentes, hay que comprobar si existen observaciones que presenten valores *missings* en más del 50% de las variables (42). Para ello, se genera una nueva variable llamada "numMissing" que indica el número de variables con valor ausente de cada observación. En el caso de este conjunto de datos, el máximo es de 20, un valor menor al 50%; por lo que no hace falta descarta observaciones. No obstante, se mantiene esta variable por si pudiera ser de provecho en el futuro.

Tratar los valores *missings* resulta también clave para la consecución de modelos estables y precisos. El tratamiento de estos datos faltantes no es igual en las variables de clase que en las variables de intervalo.

Para las variables de clase, el método que se utiliza es distinto dependiendo si el porcentaje de valores ausentes es mayor o menor del 5%. Para las variables de clase con menos del 5% (307 observaciones) de ausentes, que son la mayoría, se establece como predeterminado el método de distribución. Este método reemplaza los faltantes en las variables de clase de modo que la distribución o el porcentaje de observaciones por categoría de la misma varíe lo menos posible.

A su vez, para aquellas con entre un 5% y un 50% de *missing*, se sustituyen los ausentes por un "No Consta"; ya que interesa mantener esta información por si resultará valiosa durante la construcción de modelos. En la Figura 7, se señalan aquellas variables con más de un 5% de ausentes (rojo) y menos del 5% (verde):

CAPITAL	CAPITAL	C	3	0
NACIONALIDAD	NACIONALIDAD	C	2	0
PREO_COVID	PREO_COVID	C	7	0
PROV	PROV	C	26	0
REP_AFE_VIDPER	Replacement: AFE_VIDPER	C	4	6
REP_AFE_VIDSOC	Replacement: AFE_VIDSOC	C	4	10
REP_CCAA	Replacement: CCAA	C	17	0
REP_CERCANIA	Replacement: CERCANIA	C	9	449
REP_CIVIS_COVID	Replacement: CIVIS_COVID	C	3	31
REP_CLASESOCIAL	Replacement: CLASESOCIAL	C	6	238
REP_CNO11	Replacement: CNO11	C	9	1272
REP_ECIVIL	Replacement: ECIVIL	C	5	20
REP_EFEC_COVID	Replacement: EFEC_COVID	C	3	12
REP_ESCUELA	Replacement: ESCUELA	C	2	9
REP_ESTUDIOS	Replacement: ESTUDIOS	C	6	23
REP_GOB_ENCAR	Replacement: GOB_ENCAR	C	4	192
REP_INTENCIONG	Replacement: INTENCIONG	C	10	196
REP_INTENCIONGALTER	Replacement: INTENCIONGALTER	C	8	2514
REP_NIVELESTENTREV	Replacement: NIVELESTENTREV	C	11	93
REP_PARTICIPACIONG	Replacement: PARTICIPACIONG	C	2	68
REP_PRACTICARELIG	Replacement: PRACTICARELIG	C	6	2177
REP_PREF_PRES	Replacement: PREF_PRES	C	8	558
REP_PRO_PRI	Replacement: PRO_PRI	C	10	131
REP_PRO_SOC	Replacement: PRO_SOC	C	11	125
REP_RECUVOTOG	Replacement: RECUVOTOG	C	10	1076
REP_RELIGION	Replacement: RELIGION	C	6	81
REP_SINT_COVID	Replacement: SINT_COVID	C	2	2
REP_SITLAB	Replacement: SITLAB	C	6	26
REP_VACUNA	Replacement: VACUNA	C	2	0
REP_VAL_ECO	Replacement: VAL_ECO	C	4	129
REP_VAL_ECO_PER	Replacement: VAL_ECO_PER	C	5	35
SEXO	SEXO	C	2	0
TAMUNI	TAMUNI	C	7	0

Figura 7. Variables categóricas. En rojo, aquellas con más de un 5% de ausentes. En verde, aquellos con menos del 5%.

En cuanto a las variables de intervalo, todas aquellas con valores faltantes se imputan con el método de "Distribución", para mantener la distribución de la variable intacta. No obstante, para aquellas que superen el umbral del 5% de valores faltantes, se crea una variable M_, que indica si el dato de dicha observación para esa variable ha sido imputado o no. De este modo, no se pierde información con respecto a la falta de algunos valores que podría ser valiosa en un futuro.

En el caso de este conjunto de datos, todas aquellas variables de intervalo con ausentes (todas salvo "Edad") superan este umbral del 5% (307 observaciones), por lo que se opta por crear estas variables M_. En la Figura 8, se puede ver cómo todas las variables de intervalo salvo "Edad" superan este umbral del 5% y cómo se crean esa M_ variables que indica si se ha imputado ese valor o no.

Variable	Etiqueta	Ausente
REP_EDAD	Replacement: EDAD	0
REP_ESCIDEOL	Replacement: ESCIDEOL	548
REP_REP_ESCIDE	Replacement: Replacement: ESCIDEOL	5566
REP_REP_VAL_IA	Replacement: Replacement: VAL_IA	751
REP_REP_VAL_PC	Replacement: Replacement: VAL_PC	468
REP_REP_VAL_PI	Replacement: Replacement: VAL_PI	415
REP_REP_VAL_PS	Replacement: Replacement: VAL_PS	328
REP_REP_VAL_SA	Replacement: Replacement: VAL_SA	588

M_REP_ESCIDEOL	Imputation Indicator for REP_ESCIDEOL	N	2	0
M_REP_REP_VAL_IA	Imputation Indicator for REP_REP_VAL_IA	N	2	0
M_REP_REP_VAL_PC	Imputation Indicator for REP_REP_VAL_PC	N	2	0
M_REP_REP_VAL_PI	Imputation Indicator for REP_REP_VAL_PI	N	2	0
M_REP_REP_VAL_PS	Imputation Indicator for REP_REP_VAL_PS	N	2	0
M_REP_REP_VAL_SA	Imputation Indicator for REP_REP_VAL_SA	N	2	0

Figura 8. Número de ausentes por cada variable y creación de M_ Variables

5.1. Estudio y selección de variables preliminar

Aunque más adelante se realizará un tratamiento de las variables más profundo y detallado; la cantidad de variables actual es muy grande y con un estudio superficial se pueden rechazar algunas de estas.

Para ello, se crean dos variables aleatorias que otorgan valores entre 0 y 1 para cada una de las observaciones y se estudia el índice V de Cramer. Se rechazan todas aquellas variables con un índice similar o inferior que estas variables aleatorias. La Figura 9 muestra las variables de este conjunto y su V de Cramer.

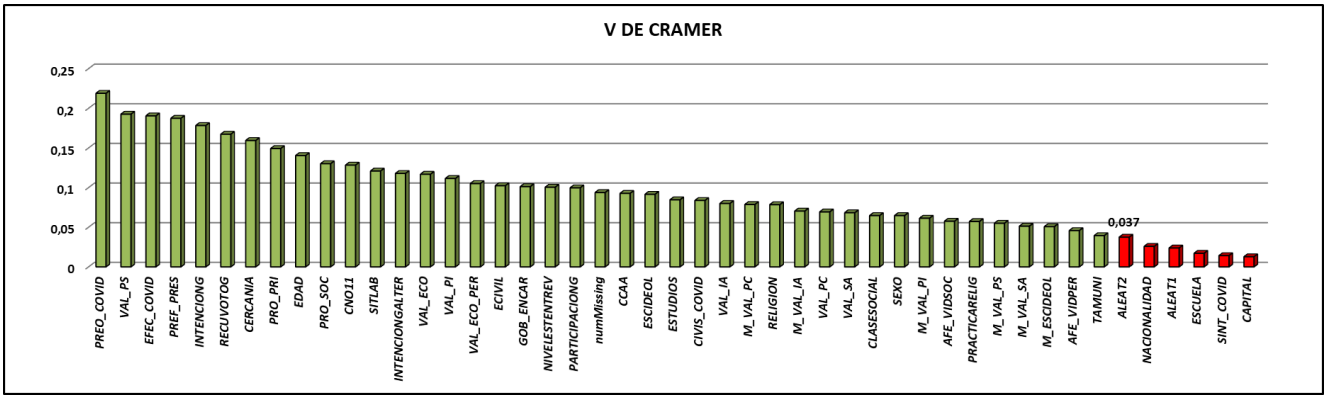


Figura 9. V de Cramer de las distintas variables input en relación con la variable objetivo

Se puede, por tanto, rechazar la variable sobre nacionalidad, la variable que indica si el entrevistado ha ido a la escuela, la variable que indica si el entrevistado ha tenido sintomatología COVID y la variable que indica si el entrevistado vive en la capital de provincia.

Las cinco variables con mayor relación con la variable objetivo son la preocupación por la situación de la pandemia, la valoración al presidente, la preferencia como presidente, la intención de voto y a quién votó el entrevistado en las últimas elecciones generales.

En la primera de estas, se puede apreciar qué diferente se distribuye la variable objetivo dependiendo del nivel de preocupación del entrevistado por la situación sanitaria (Figura 10). Así, como parece lógico, aquellas personas con una mayor preocupación están se oponen menos a vacunarse que aquellas sin preocupación o con muy poca preocupación. En la Figura 10, se muestra en porcentaje esta distribución.

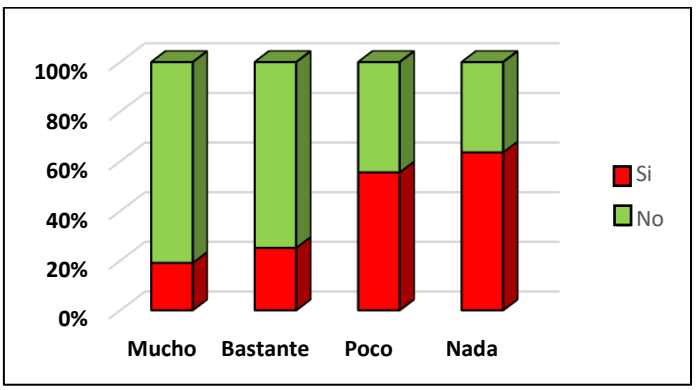


Figura 10. Distribución de la variable objetivo en las distintas categorías de la variable "PREO_COVID". En verde, se representa a aquellos sujetos que no se oponen a la vacuna. En rojo, a los que sí lo hacen.

La segunda variable con mayor V de Cramer es la valoración que el entrevistado hace del presidente del Gobierno, Pedro Sánchez. En este caso, a medida que mejor se valora al presidente, el porcentaje de entrevistados que se opone a ser vacunado disminuye, como se puede comprobar en la Figura 11.

Figura 11. Distribución de la variable objetivo en relación con la variable "VAL_PS"

La siguiente variable con mayor relación a la variable objetivo es la referente al efecto sobre el cual el entrevistado cree que el COVID tiene más efecto. Como se puede comprobar en la Figura 12, aquellos que anteponen la economía a la salud tienden a oponerse a la vacuna más que aquellos que anteponen la salud o que ambos.

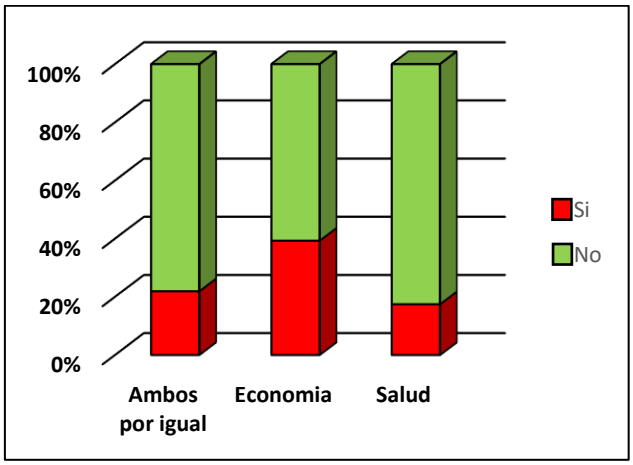


Figura 12. Distribución de la variable objetivo en relación a la variable "EFEC_COVID"

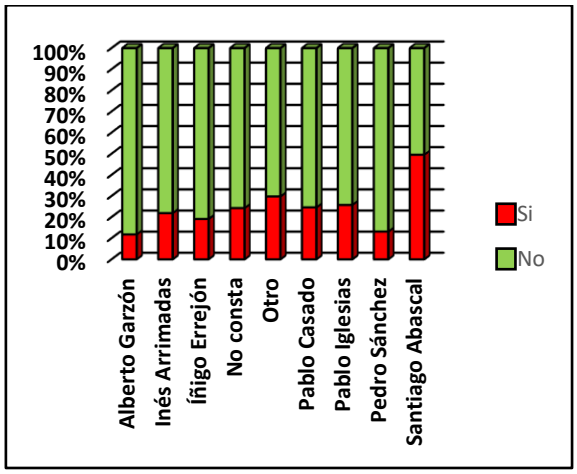


Figura 13. Distribución de la variable objetivo en relación a la variable "PREF_PRES"

La preferencia como Presidente del Gobierno del entrevistado también tiene relación con si el entrevistado se vacunaría o no. Las categorías de “Santiago Abascal” y “Otro” tienen un mayor porcentaje de aversos a la vacuna que el resto; cómo se puede ver en el diagrama de barras de la Figura 13.

Por último, la intención de voto en unas hipotéticas generales también guarda relación con la intención de ponerse o no la vacuna. Los votantes de Vox y aquellos que no votarían tienen una mayor tendencia al “sí”, es decir, a la oposición a la vacuna (Figura 14).

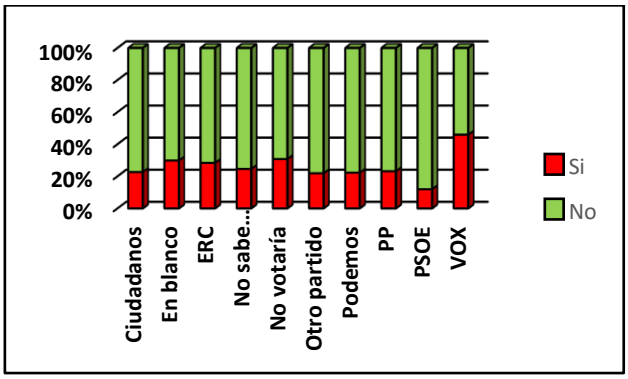


Figura 14. Distribución de la variable objetivo en relación a la variable "INTENCIONG"

6. MODELIZACIÓN

6.1. KNN o vecino más próximo

Como se señalaba en el apartado de metodología, para los kNN se deben tener en cuenta el tipo de estandarización, los sets de variables con los que se trabaja y el número de vecinos.

Para este trabajo, se toma la decisión de estandarizar con los dos métodos ofrecidos por SAS *Miner*: estandarización y rango; para comprobar si alguna de ellas funciona mejor que la otra. La primera de ellas es la forma más habitual de estandarizar, en la cual se resta la media aritmética al valor a estandarizar, y se divide por la desviación típica. La estandarización por rango sigue la siguiente función:

$$x = (x_i - min(x))/(max(x) - min(x))$$

Además, se decide trabajar con tres *sets* de variables diferentes:

- Un *set* que incluya todas las variables salvo las ya eliminadas en el punto anterior.
- Un *set* de variables con una selección suave, llevada a cabo por *Miner*. Con este *software*, se utiliza un nodo de “Selección de variables”, que rechaza aquellas variables con menor importancia y agrupa categorías de las variables nominales que se comportan igual con respecto a la variable objetivo (G_). Para esta selección, las únicas variables de intervalo que se mantienen son “Edad” y “Valoración de Pedro Sánchez”. Se rechazan la mayoría de las variables categóricas, quedando intactas “Participación en las últimas elecciones”, “Efectos adversos más importantes del COVID” y “Civismo COVID”. En el resto de las variables categóricas que se mantienen como *input* (“CNO11”, “ECIVL”, “GOB_ENCAR”, “INTENCIOG”, “INTENCIONGALTER”, “NIVELESENTRE”,

"PREF_PRES", "PERO_COVID", "PRO_PRI", "PRO_SOC", "RECUVOTOG", "SITLAB", "VAL_ECO", "VAL_ECO_PER", "CCAA"), se agrupan aquellas categorías que funcionan de forma similar con respecto a la variable objetivo. En total, este set se reduce a 20 variables *input.*

- Un *set* de variables con una selección de variables agresiva que solo incluya aquellas variables con mayor relevancia según el V de Cramer y que sean las más interesantes para el desarrollo de un análisis del perfil. De este modo se toman, las variables con mayor V de Cramer según la Figura 9. Se incluyen PREO_COVID, VAL_PS, EFEC_COVID, PREF_PRES, EDAD. Es cierto que existían otras variables con importancia similar a PREF_PRES que no han entrado en esta selección, pero todas ellas hacían referencia a los partidos políticos más cercanos o recientes votaciones de los entrevistados. No obstante, estas variables guardan una relación suficientemente fuerte entre sí para poder trabajar con tan solo una de ellas y simplificar las operaciones. En la Tabla 2 se recoge la relación entre variables según el estadístico V de Cramer.

Tabla 2. V de Cramer entre las variables "PREF_PRES", "INTENCIONG", "RECUVOTOG" y "CERCANIA"

	PREF_PRES	INTENCIONG	RECUVOTOG	CERCANIA
PREF_PRES		0.50	0.37	0.43
INTENCIONG			0.52	0.54
RECUVOTOG				0.60
CERCANIA				

Se puede considerar que, a partir de 0.3, las variables guardan relación, y estas variables sobrepasan este valor de forma holgada en la mayoría de las ocasiones. A modo de comparación el V de Cramer entre PREF_PRES y PREO_COVID, por ejemplo, no llega al 0.05.

Todos los modelos kNN que se han probado con validación cruzada repetida se pueden encontrar en la Tabla 25 del Anexo B. Concretamente, para kNN se ha trabajado con 102 modelos, con dos tipos distintos de estandarización, tres *sets* de variables diferentes y 17 números de vecino (k) distintos.

La elección del mejor modelo kNN va a fundamentarse, principalmente, en encontrar el equilibrio entre efectividad, varianza y complejidad/facilidad de aplicación. Si bien es cierto que un aumento en el error no es algo bueno; se pueden hacer pequeñas concesiones a este respecto si esto se traduce en modelos más estables y sencillos. Es por ello que el análisis siguiente no tiene solo en cuenta la efectividad del modelo, sino también qué *set* de variables utiliza el modelo y el número de vecinos. A mayor número de vecinos, existe mayor complejidad y se requiere una mayor capacidad computacional.

El *set* de variables que trabaja con todas las variables es, obviamente, requiere de más tiempo de computación. No obstante, esta complejidad no se traduce en mejores predicciones, ni mejor índice bajo la curva ROC. En la Figura 15, se puede comprobar cómo, al unir los resultados de todos los modelos por el *set* de variables que utilizan. Aquellos modelos que usan todas las variables tienen un índice ROC menor que las otras dos opciones.

Figura 15. Diagrama de cajas que representa el sesgo en índice ROC y varianza de la unión de los modelos con distintos sets de variables

Sin embargo, podría surgir la duda de si los modelos kNN con esta selección de variables pueden llegar a funcionar mejor que los otros con determinado número de vecinos. Para desestimar dicha afirmación solo hay que estudiar la Figura 16.

La Figura 16 representa todos los modelos en términos de error (índice ROC) y varianza. La primera conclusión general que se puede sacar son el mejor desempeño de la estandarización con "método estandarizado" (tres bloques de la izquierda), que es consistentemente mejor que la estandarización con "rango"; salvo con la selección agresiva (azul), donde ambos métodos están muy a la par.

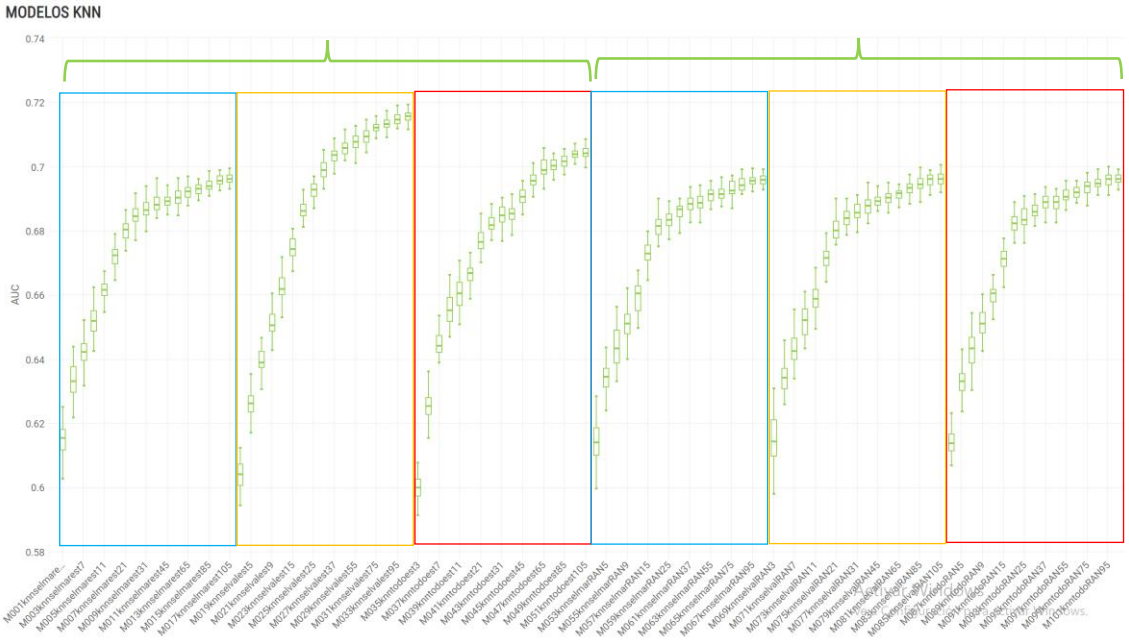


Figura 16. Diagrama de cajas que representa el sesgo y varianza de cada uno de los modelos. Los tres bloques de la izquierda hacen referencia a los modelos con estandarización "estandarizado"; y los tres de la derecha, "rango". En azul, los modelos con selección de variables agresiva; en amarillo, los modelos con selección Miner; y en rojo, modelos con todas las variables.

La siguiente conclusión es, como se señaló anteriormente, el pobre desempeño de la selección con todas las variables (rojo). Sin importar el número de vecinos, esta selección funciona peor que la selección *Miner* y obtiene unos resultados muy similares a los de la selección agresiva. Probablemente esto se deba a sobreajuste. Nada hace

indicar, por tanto, que asumir la complejidad y la dificultad de aplicación que implica utilizar todas las variables.

La otra gran conclusión es que a mayor número de vecinos, mejor funcionan los modelos; aunque a partir de 75, esta mejoría comienza a estabilizarse.

De este modo, se estudiarán más en profundidad los modelos con método "estandarizado" y selección agresiva o selección *Miner*. Al repasar los resultados de estas dos opciones, se observa como la selección de *Miner* funciona mejor que la selección agresiva y otorga al modelo una mayor capacidad predictiva. En la Figura 17, se muestran los modelos más prometedores (k=65 a k=105) con esta selección.

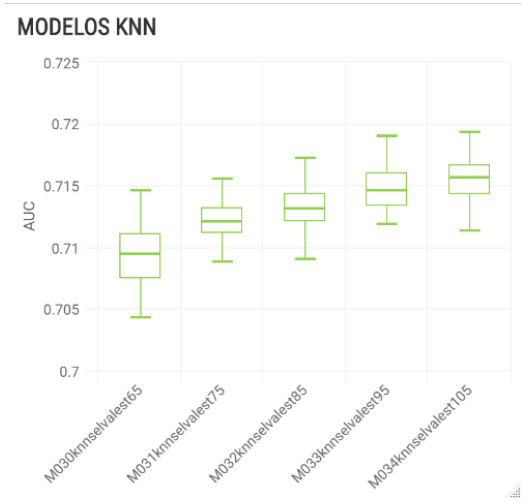


Figura 17. Diagrama de cajas con modelos kNN, selección Miner y "estandarizado" más prometedores.

A medida que aumenta el número de vecinos, el error disminuye. No obstante, al comparar k=95 y k=105, se puede ver claramente como el índice ROC comienza a estancarse. La media de este índice para k=95 es 0.7149; mientras que para k=105, 0.7154. La diferencia es mínima y, además, el modelo de 95 vecinos tiene una varianza algo más pequeña. Por tanto, el modelo kNN con mayor capacidad predictiva es un modelo con método "estandarizado", selección de variables de *Miner* y 95 vecinos.

No obstante, el modelo que predice mejor no tiene por qué tratarse del mejor modelo; ya que la facilidad de aplicación y la simplicidad de los modelos también deben tenerse en cuenta como factores importantes.

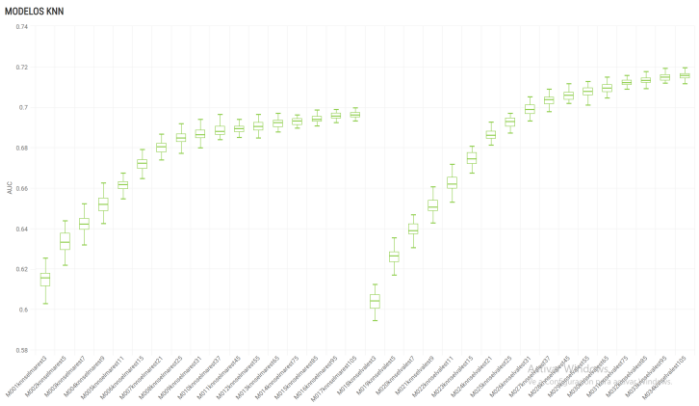


Figura 18. Diagrama de cajas que representa los modelos kNN con "estandarizado" y selección agresiva (bloque de la izquierda); y selección Miner (bloque de la derecha)

Como se señaló anteriormente y como se puede apreciar en la Figura 18, la selección de *Miner* (derecha en la imagen) predice mejor que la selección agresiva (izquierda en la imagen), pero esta última no está tan lejos como para descartarla.

En el gráfico de cajas de la Figura 19, se puede apreciar como la mejoría asociada al aumento de número de vecinos se estabiliza antes de lo que ocurría con la selección de *Miner*. Especialmente interesante resulta el modelo con k=45 (resaltado en rojo en la siguiente figura), que tiene una varianza muy similar a los modelos con k más altos y un índice ROC algo menor (0.6892 de media frente a 0.6961) que se compensa con su simplicidad.

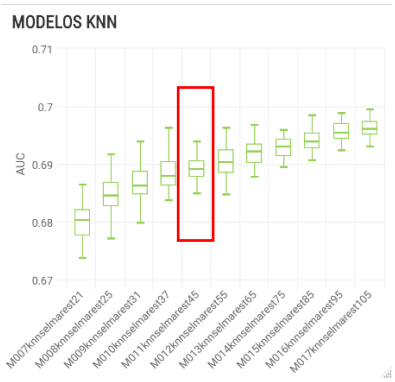


Figura 19. Diagrama de cajas con los modelos kNN con método "estandarizado" y selección agresiva. En rojo, el modelo con k=45.

6.1.1 Mejor modelo kNN

Determinar qué modelo es mejor sin saber con qué finalidad se realiza el mismo es muy complicado. Es por ello que se decide escoger dos modelos ya comentados anteriormente: el modelo con mayor capacidad predictiva y el mejor modelo relación complejidad-error (y varianza).

El modelo que mejor predice es un kNN de 95 con una selección de variables realizada con SAS *Miner* y las variables de intervalo estandarizadas por medio del "método estandarizado" del propio *Miner*. Tiene un índice del área bajo la curva ROC de 0.7149 y una tasa de fallos (*misclassification rate*) de 0.233 con punto de corte en 0.5.

Por otro lado, el modelo kNN con 45 vecinos, método "estandarizado" y selección agresiva de variables es un modelo que predice peor, pero cuya complejidad es mucho menos, lo que facilita tanto la aplicación del modelo como su explicabilidad. En este caso, el índice ROC es de 0.6892 y la tasa de fallos con corte de probabilidad en 0.5 es de 0.225.

Tabla 3. Mejores modelos kNN

	Modelo	**Nº de modelo**	**Nº de variables**	**AUC**	**Tasa de fallos**
Mayor precisión	knnselvalest95	Modelo 33	59 (*dummies*)	0.7149	0.233
Mejor equilibrio	knnselmarest45	Modelo 11	16 (*dummies*)	0.6892	0.225

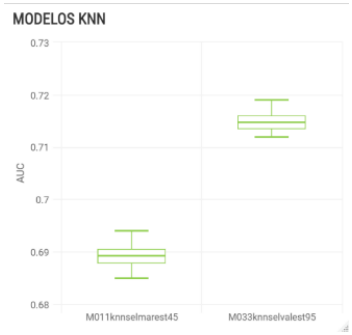


Figura 20. Diagrama de cajas con los mejores modelos kNN.

6.2. Regresión logística

La parametrización para la regresión logística es compleja y sus posibilidades son bastante amplias. A continuación, se comenta con que se ha trabajado para este conjunto de datos:

- El tratamiento de las variables. En ocasiones, realizar transformaciones a las variables de intervalo puede generar mejores modelos. De este modo, uno de los *sets* incluirá tan solo las variables originales; mientras que otro incluirá las variables originales y las variables de intervalo transformadas con un nodo de "Transformar variables" de *Miner* con método "Mejor". Estas transformaciones se recogen en la Tabla 4.

Tabla 4. Variables originales y transformaciones realizadas por Miner

Variable original	Transformación
ESCIDEOL	Discretización (4 tramos)
EDAD	Discretización (3 tramos)
VAL_PS	Discretización (4 tramos)
VAL_PC	Discretización (2 tramos)
VAL_PI	Función inversa
VAL_IA	Discretización (2 tramos)
VAL_SA	Sin transformación
numMissing	Discretización (3 tramos)

- La inclusión de *dummies*. A pesar de que la inclusión de variables *dummies* puede hacer más compleja la interpretación de los parámetros de la regresión, puede generar modelos con un menor número de parámetros que sean igual de buenos o mejores que aquellas regresiones en las que se incluyen las variables categóricas al completo. Así, con ambos sets (originales y originales/transformadas), se prueban modelos tanto con *dummies* como sin ellas.
- El criterio de información. Este criterio es el encargado de valorar hasta qué punto la inclusión de nuevos parámetros genera un aumento en la eficacia que compense la complejidad. En esta regresión, se toman el criterio de información de Akaike (AIC) y el criterio de información Bayesiano (BIC). Este último genera modelos más simples; ya que tiende a penalizar más la complejidad.
- El criterio de selección. La regresión utiliza distintos métodos para seleccionar sus parámetros: *backward*, *stepwise* y *forward*.

- La inclusión de interacciones entre variables. En ocasiones, las interacciones entre variables pueden dar lugar a modelos más complejos, pero más eficaces. No obstante, para esta regresión no conviene utilizar las interacciones, dado que el número elevado de variables, y sobre todo, de variables categóricas generaría modelos muy complejos podrían ser escogidos.

En la Tabla 26 del Anexo B se puede encontrar el listado de todos los modelos de regresión logística (24) llevados a cabo y sus características.

Como ocurría con los modelos kNN, la elección del mejor modelo debe sustentarse en el estudio de precisión, varianza y complejidad de los modelos. Para la regresión logística, es especialmente relevante esta última, pues a menor número de modelos más sencilla será la explicación de los parámetros.

Antes de entrar a valorar y estudiar los modelos, conviene hacer unas aclaraciones sobre algunos problemas o situaciones que se han dado durante la modelización. De los 24 modelos que se plantean; solo se realiza validación cruzada repetida en 15 de ellos; descartando otros 9 por distintos motivos. En primer lugar, se descartan los dos modelos *backward* (AIC y BIC) con las variables originales y transformadas por no llegar a converger. En ocasiones, diferentes criterios de selección concluyen en modelos exactamente iguales, por lo que no tiene sentido trabajar con ambos a la vez y duplicar las operaciones. En el caso de esta modelización, los modelos que se repiten son los siguientes (en negrita el criterio con el que se hace la validación cruzada):

- RegOrigBackAIC, RegOrigStepAIC, **RegOrigForwAIC**
- RegOrigStepBIC, **RegOrigForwBIC**
- RegOrigDumStepBIC, **RegOrigDumForwBIC**
- RegTransStepAIC, **RegTransForwAIC**,
- RegTransStepBIC, **RegTransForwBIC**,
- RegTransDumStepBIC, **RegTransDumStepBIC**

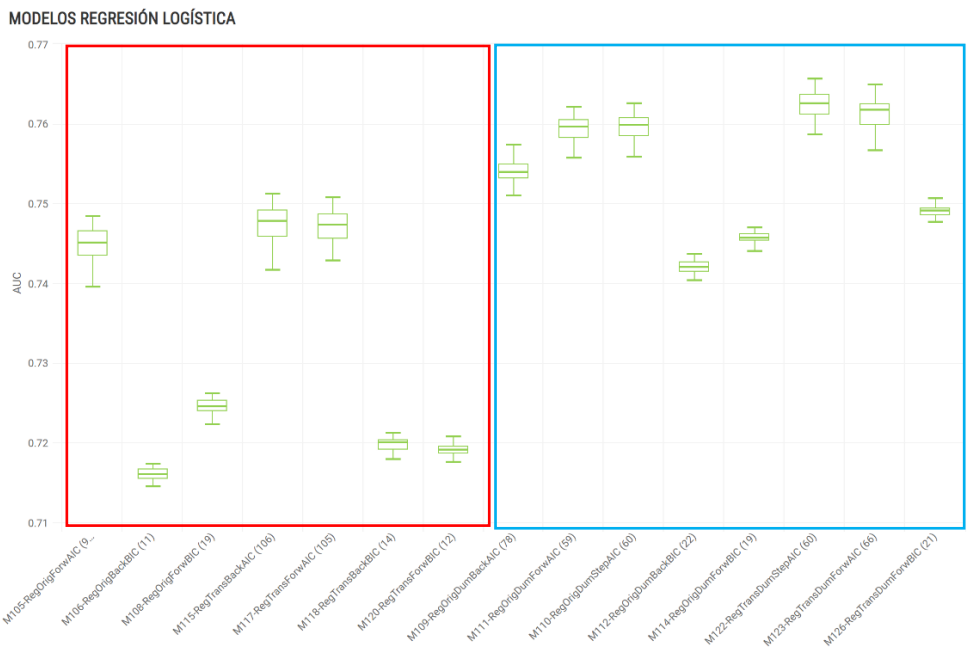


Figura 21. Diagrama de cajas que representa sesgo en índice ROC y varianza. En rojo, modelos sin inclusión de dummies; y en azul, modelos con dummies.

Como es de esperar con un conjunto de datos con tantas variables y sobre todo, variables categóricas, los modelos que utilizan el AIC tienen mayor precisión, pero son excesivamente complejos comparados con sus equivalentes de BIC. A su vez, los modelos que utilizan las variables *dummies* consiguen una mayor precisión con el mismo número de parámetros que sus equivalentes que utilizan las variables categóricas sin "dummificar". En la Figura 21, se muestra una diagrama de cajas que representa el índice ROC y la varianza de estos 15 modelos. Entre paréntesis, se muestra el número de parámetros de cada uno de ellos.

Como se puede apreciar en la Figura 21, los modelos que utilizan *dummies* (resaltados en la zona azul) funcionan mejor que aquellos que nos las utilizan. Como se comentaba anteriormente, todos los modelos que hacen uso del AIC son más precisos, pero su elevado número de parámetros les suma complejidad y generan una mayor varianza. Los modelos con BIC son, en cambio, más simples y mucho más estables.

La idea de este trabajo es seleccionar tanto el modelo que mejor prediga como aquel que encuentre el equilibrio entre precisión, varianza y simplicidad.

Para encontrar el primero de estos, habría que acudir a modelos que utilicen AIC, ya que como se ha señalado, estos tienen un menor sesgo. La Figura 22 muestra con un gráfico de caja y bigotes los más prometedores: los modelos que utilizan las variables originales con *dummies* y *forward* y *stepwise*; y sus equivalentes que utilizan variables originales y transformadas.

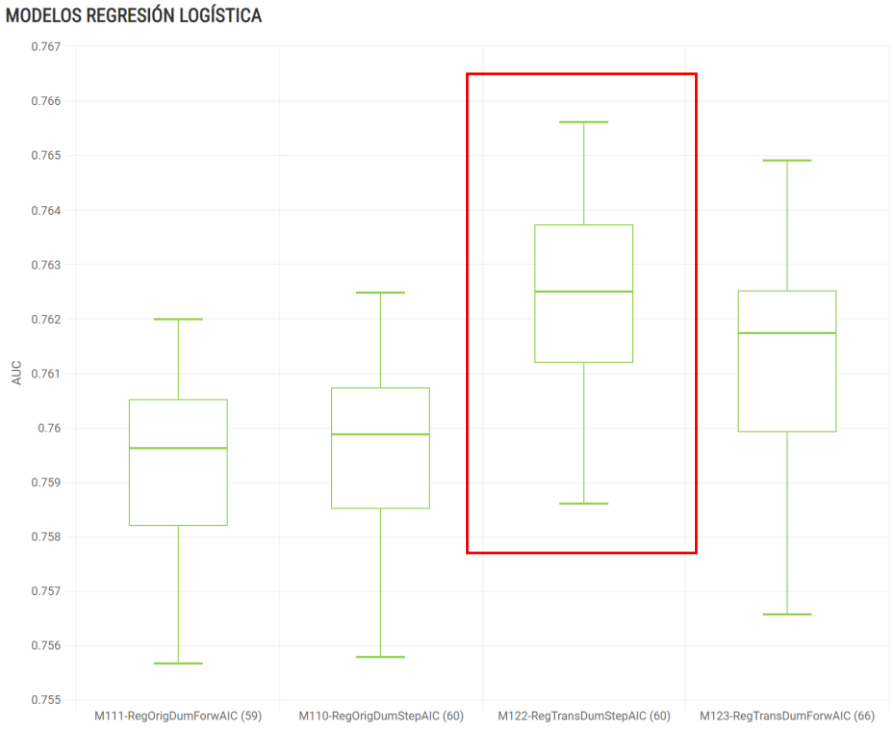


Figura 22. Diagrama de cajas con el sesgo y varianza de los modelos con mayor capacidad predictiva. En rojo, el modelo con variables originales y transformadas, dummies, stepwise y AIC.

Ante modelos tan similares en número de parámetros y varianza; y teniendo en cuenta que se busca el modelo más preciso, se decide tomar como modelo con mayor capacidad predictiva el modelo que utiliza las variables originales y las transformadas con *dummies*, criterio de selección *stepwise* y AIC o el RegTransDumStepAIC (resaltado

en rojo). Este modelo, aunque por poco, también presenta la menor varianza de los cuatro modelos.

En cuanto a los modelos de regresión con más equilibrio, se descartan todos los modelos que no utilizan *dummies,* pues si bien son ligeramente más simples, sacrifican demasiado la precisión. También se descartan los modelos con AIC. De este modo, se valoran los siguientes modelos con *dummies*: los modelos con variables originales, *backward*/*forward* y BIC; y el modelo con variables originales y transformadas que utilizan BIC y *forward.*

Como se puede apreciar en la Figura 23, el único con *backward* (izquierda) puede descartarse sin duda alguna, ya que tiene más parámetros (22) que los otros dos, y además, tiene un mayor error. La elección, por tanto, ha de tomarse entre los otros dos modelos. Dado que la diferencia en índice ROC no es especialmente grande (0.746 frente a 0.749), se decide tomar el modelo con variables originales (remarcado en rojo), que tiene dos parámetros menos (19-21) y no cuenta con variables transformadas en su fórmula, lo que dificultaría la interpretación de los coeficientes.

Figura 23. Diagrama de cajas con el sesgo y varianza con mayor equilibrio. En rojo, el modelo con variables originales, dummies, forward y BIC.

Por lo tanto, el modelo con mejor equilibrio es una regresión logística que utiliza las variables originales del conjunto de datos, *dummies*, criterio de selección *forward* y BIC.

6.2.1 Mejor modelo con regresión logística

El fuerte de la regresión logística es la facilidad que da al analista para interpretar su fórmula y sacar conclusiones sobre la importancia de las variables *input*, por lo que el modelo mejor siempre será aquel que además de ser preciso, sea explicable. No obstante, en este apartado también se incluirán el modelo con mayor capacidad predictiva, en caso de que sea necesaria.

Tabla 5. Mejores modelos de regresión logística.

	Modelo	Nº de modelo	Parámetros	AUC	Tasa de fallos
Mayor precisión	RegTransDumStepAIC	Modelo 122	60	0.7623	0.2152
Mejor equilibrio	RegOrigDumForwBIC	Modelo 114	19	0.7458	0.2168

La fórmula de la regresión logística con mejor equilibrio y, por tanto, más explicabilidad, es la siguiente:

> OPOSVACUNA ~ VAL_PS + EFEC_COVID.Economia + EDAD + PREO_COVID.Poco + SEXO.Hombre + CNO11.N.P. + PREO_COVID.Nada + numMissing + PREF_PRES.Santiago Abascal` + PRO_PRI.Covid + PREF_PRES.Otro + PREO_COVID.Bastante + VAL_ECO_PER.Buena + CCAA.Cataluña + PARTICIPACIONG.No + GOB_ENCAR.Otras respuestas` + CNO11.Personal de apoyo administrativo` + SITLAB.En paro

Si se utilizan los coeficientes de esta fórmula como exponentes del número de Euler (*e*), se obtienen los *odds ratio* que permiten conocer cómo variará la probabilidad del evento dependiendo del valor que toma una variable o la categoría a la que pertenece una observación. En la Tabla 6 se muestran tanto los parámetros como los *odds ratio*:

Tabla 6. Parámetros del mejor modelo de regresión logística.

	Parámetro	e^Parámetro
Intercept	-0.24429349	0.7832577
VAL_PS	-0.14912100	0.8614649
EFEC_COVID.Economia	0.59345962	1.8102403
EDAD	-0.01433713	0.9857652
PREO_COVID.Poco	1.38564585	3.9974068
SEXO.Hombre	-0.48972518	0.6127948
CNO11.N.P.	-0.72505767	0.4842966
PREO_COVID.Nada	1.77786174	5.9171904
numMissing	0.06807238	1.0704428
PREF_PRES.Santiago Abascal	1.10201568	3.0102276
PRO_PRI.Covid	-0.35347439	0.7022440
PREF_PRES.Otro	0.27416339	1.3154297
PREO_COVID.Bastante	0.31021763	1.3637219
VAL_ECO_PER.Buena	-0.24158126	0.7853850
CCAA.Cataluña	0.31781042	1.3741157
PARTICIPACIONG.No	0.37782115	1.4591020
GOB_ENCAR.Otras respuestas	0.52947727	1.6980445
CNO11.Personal de apoyo administrativo	0.48753171	1.6282922
SITLAB.En paro	0.34981700	1.4188079

A pesar de que la interpretación de esta regresión implica aislar cada una de las variables, algo que no es del todo correcto (pues a menudo existen correlaciones entre variables), estudiarla puede resultar interesante para hacerse a la idea del perfil del entrevistado opuesto a la vacuna.

La interpretación de estos parámetros no hace más que reforzar las ideas expuestas en el estudio de las variables. A continuación, se valorarán cuál es la traducción de los *odds ratio* más interesantes a la probabilidad del evento. Hay que tener siempre en cuenta que esta interpretación se basa en la premisa de que el resto de las variables se mantengan constantes. Por ejemplo, por cada punto de mayor valoración que se le dé a Pedro Sánchez, la probabilidad de no querer vacunarse desciende un 14%, o lo que es lo mismo: se multiplica por 0.86.

Si el entrevistado considera que la economía es el aspecto que se ve más afectado por la pandemia, la probabilidad de que se oponga a ser vacunado es un 80% mayor con respecto a los que consideran otras opciones. En cuanto a la edad, por cada año mayor que es el sujeto, se reduce un 1.5% la probabilidad de oposición a la vacuna. Para aquellos entrevistados a los que les preocupa “Poco” la situación de la COVID, la

probabilidad de que no quieran vacunarse se multiplica casi por 4 comparado con aquellos sujetos que contestaron “Mucho” y manteniéndose el resto de las variables intactas.

Otro dato muy interesante y que también sigue la línea del anterior estudio de las variables es el efecto que provoca que el sujeto tenga como presidente preferido a Santiago Abascal. Los entrevistados que lo prefieren tienen una probabilidad de no querer vacunarse 3 veces mayor que los que prefieren a otro candidato. Lo mismo ocurre con los sujetos que prefieren a un candidato minoritario (“PREF_PRES.Otro”), cuya probabilidad de evento es 30% mayor que la de aquellos que prefieren candidatos más populares.

Los hombres son menos reacios a vacunarse que las mujeres, reduciendo su probabilidad de rechazo un 40%. La única comunidad autónoma que incluye la regresión como relevantes es Cataluña. La probabilidad de rechazo de la vacuna es un 30% mayor en los catalanes.

6.3. Redes neuronales

Las redes neuronales son modelos complejos, cuyas posibilidades de parametrización son muy amplias. En este proyecto, se trabaja con los siguientes parámetros:

- *Sets* de variables. Las redes neuronales, como los kNN son muy sensibles a la información que reciben, ya que no son capaces de seleccionar variables como la regresión logística. Además, el número de variables *input*, como se verá más adelante, también afecta al máximo número de nodos de la red; por lo que no se puede trabajar con selecciones de variables grandes. De este modo, una vez estandarizadas y creadas *dummies*, se toman tres *sets* de variables distintos:
    - Selección agresiva con *randomselect*: se trata de la selección agresiva que se utiliza para los modelos kNN, pero aún más reducida, al haber sido sometida a un proceso de *randomselect*, que consiste en la ejecución de 1000 (semillas 12345 a 13344) modelos de regresión logística simple y *stepwise.* Una vez ejecutados, se observa la frecuencia de cada selección de variables. En este caso, en un 60% de estos modelos se tomaba una selección con las siguientes variables *input* (11): PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Igle PREF_PRES_Santiago_A EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD.
    - Selección agresiva (16 variables *input*), que es la misma que en la modelización con kNN.
    - Selección *Miner* (59 variables *input*), que también coincide con la utilizada en kNN.
- Número de nodos. Antes de estimar con qué número de nodos se van a crear estas variables, hay que determinar un número máximo de nodos; ya que un elevado número de estos puede derivar en redes que sobreajusten. Como norma general, se puede decir que son necesarias 20 observaciones por cada parámetro que tenga la red. Para calcular el número de parámetros se debe seguir la siguiente fórmula, donde *h* es el número de nodos y *k*, el número de variables *input*:

$$h \cdot (k+1) + h + 1 = p$$

De este modo, teniendo en cuenta que el número de observaciones es 5500; para cada una de las selecciones, el máximo número de nodos y los nodos utilizados son los que aparecen en la Tabla 7.

Tabla 7. Número máximo de nodos por set de variables

	Nº de variables (con *dummies*)	**Nª máximo de nodos**	**Número de nodos**
Agresiva *randomselect*	11	25	2,3,5,7,9,11,15,19,25
Agresiva	16	17	2,3,5,11,17
Miner	59	4	2,4

- Función de activación. Es otro de los parámetros que SAS permite alterar a la hora de generar redes neuronales. R no permite esta modificación. No obstante, realizar todas las combinaciones posibles con las distintas funciones conllevaría un número de modelos demasiado alto; y una capacidad computacional con la que no se cuenta. De este modo, se decide realizar una exploración previa en la que, con validación cruzada repetida de 5 grupos y 10 repeticiones, se prueban 6 funciones de activación distintas con distinto número de nodos. Como se puede apreciar en la Figura 24, ninguna de las 6 (tangente hiperbólica, logístico, arcotangente, seno, *softmax*, y Gauss) se destaca por encima de las otras; por lo que se decide trabajar con la función de tangente hiperbólica, la de uso más extendido y más fácil aplicación. Además, es la utilizada por el paquete *caret* en R.

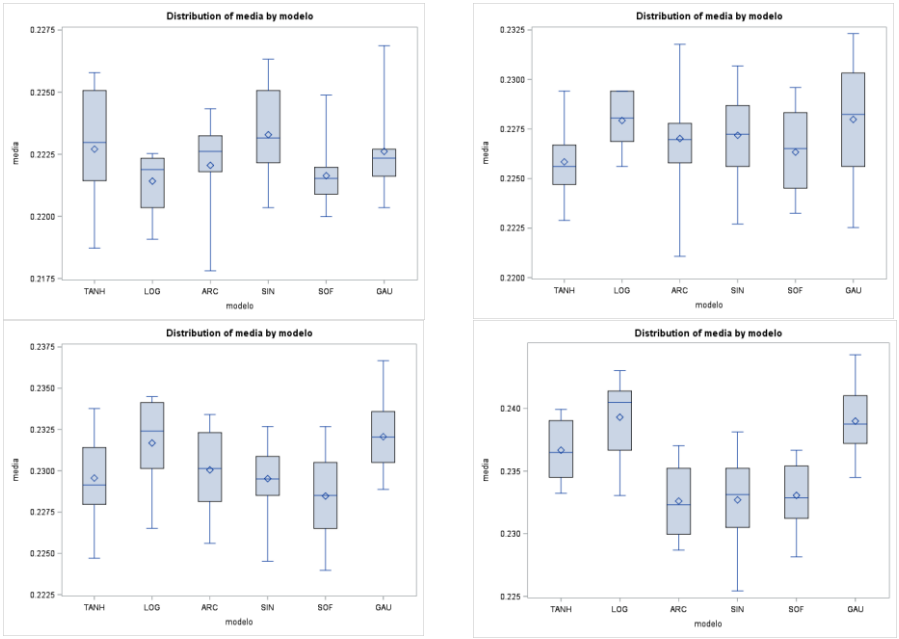


Figura 24. Diagrama de cajas que representa la tasa de fallos de redes con diferentes funciones de activación. De izquierda a derecha y arriba a abajo: 3 nodos, 9 nodos, 15 nodos y 25 nodos.

No obstante, una vez determinada cuál es la mejor red con "targente hiperbólica"; se realizarán pruebas con las distintas funciones; por si interesará utilizarla.

- *Early stopping*. Las redes neuronales realizan una serie de iteraciones para ajustar cada vez más los parámetros, pero este ajuste puede derivar en sobreajuste. Para poder evitar este sobreajuste, cada red que se cree se generará sin *early stopping* y con él. El número de iteración en la que ha de parar la modelización se obtiene con una macro en SAS en donde se visualiza como los valores de la entropía en *train* y validación empiezan a divergir (sobreajuste) a partir de un determinado punto. Para asegurarse, se ejecuta esta macro con dos semillas distintas (12344 - 12345). Para cada red, el *early stopping* será diferente.
  Los valores de *early stopping* para cada selección son los que se muestran en la Tabla 8:

Tabla 8. Valores de early stopping

Selección	Nº de nodos	*Early stopping*
Agresiva *randomselect*	2	15
Agresiva *randomselect*	3	11
Agresiva *randomselect*	5	9
Agresiva *randomselect*	7	9
Agresiva *randomselect*	9	9
Agresiva *randomselect*	11	8
Agresiva *randomselect*	15	8
Agresiva *randomselect*	19	8
Agresiva *randomselect*	25	8
Agresiva	2	17
Agresiva	3	10
Agresiva	5	10
Agresiva	11	9
Agresiva	17	8
Miner	2	9
Miner	3	9

- Algoritmo de aprendizaje. Su parametrización tiene menos influencia en el modelo resultante. Es por esto que se realizarán pruebas con este parámetro una vez determinado cuál es el mejor modelo; por comprobar si alguno de los algoritmos es mejor que el resto. Para las pruebas generales, se utilizará el algoritmo de Levenberg-Marquardt.

Una vez estudiados los parámetros que se pueden modificar con las redes, se presentan en la siguiente figura todos los modelos generados con cada *set* de variables (con y sin *early stopping*), con función de activación "tangente hiperbólica" y algoritmo de aprendizaje de Levenberg-Marquardt. En la Figura 25, se puede comprobar como las selecciones más agresivas (bloques rojo y verde) funcionan mejor que la selección de *Miner*, que es más compleja. Está última sobreajusta de forma clara, presentando una mayor tasa de fallos y una varianza enorme comparada con el resto.

También se puede comprobar como este conjunto de datos se beneficia de redes neuronales más simple con un menor número de nodos. Así, las redes neuronales con selección agresiva y selección agresiva con *randomselect* de 2 nodos son las principales candidatas a ser la mejor red neuronal y se estudiarán aparte.

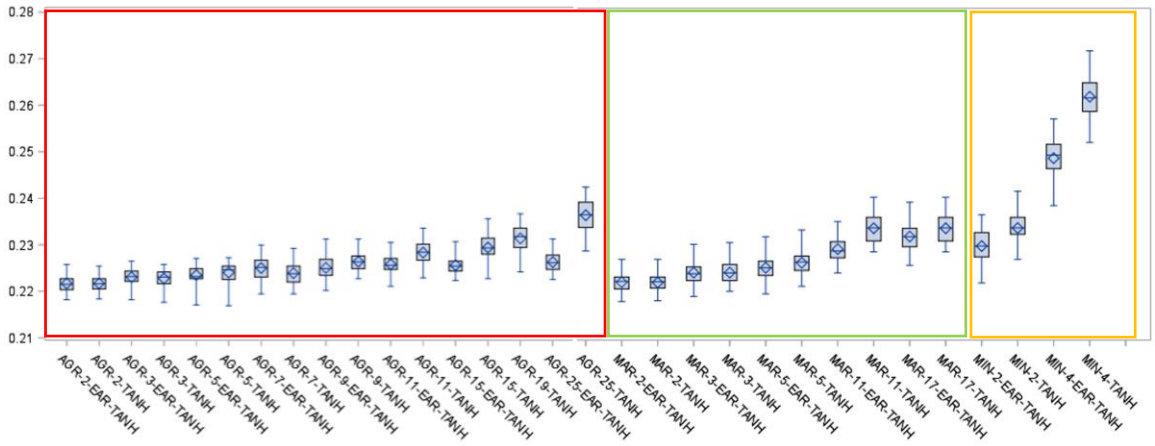


Figura 25. Diagrama de cajas que representa la tasa de fallos de las redes neuronales ejecutadas. En rojo, las redes con selección agresiva randomselect*. En verde, selección agresiva; y en amarillo, selección Miner.*

Una vez descartada la selección de variables *Miner* y los números de nodos altos, conviene comparar las redes neuronales más prometedoras (Figura 26).

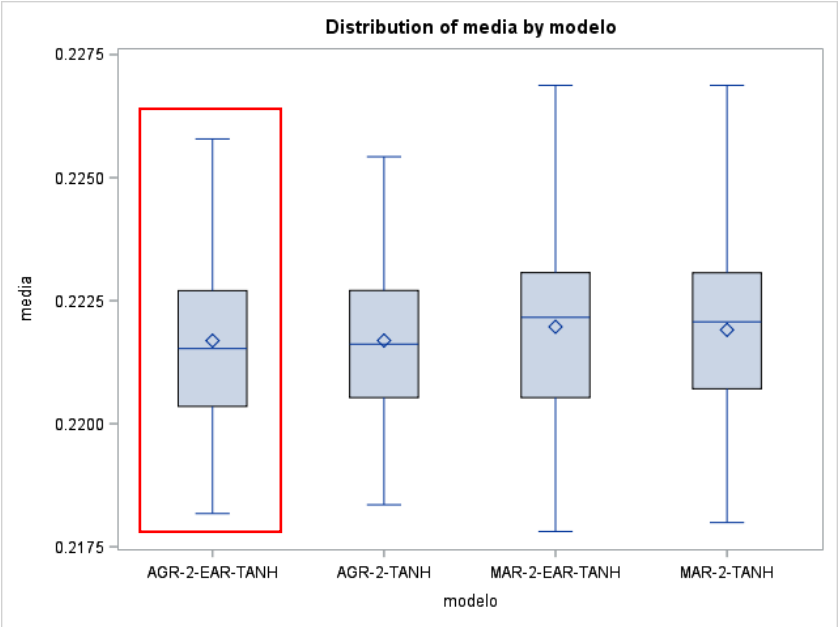


Figura 26. Diagrama de cajas que representa la tasa de fallos de las cuatro mejores redes neuronales.

Como se apreciar en la Figura 26, las redes con una selección de variables algo más complejas (dos de la derecha) tienen una media de tasa de fallos algo más alta y además, mayor varianza que las redes con selección agresiva con *randomselect*. Por lo tanto, la elección ha de tomarse entre las dos redes con dos nodos. Ambas tienen una media de error muy similar; aunque la red sin *early stopping* tiene una varianza algo menor. No obstante, la diferencia es mínima y la facilidad de aplicación de la red con *early stopping* (señalada en rojo) la convierte en una mejor candidata.

Al principio de este apartado, se podía ver cómo ninguna función de activación funcionaba especialmente mejor que las demás con una prueba preliminar. Ahora que ya se ha escogido la mejor red, se prueba con validación cruzada repetida; por si alguna de estas funciones mejorará la red de forma notable (Figura 27).

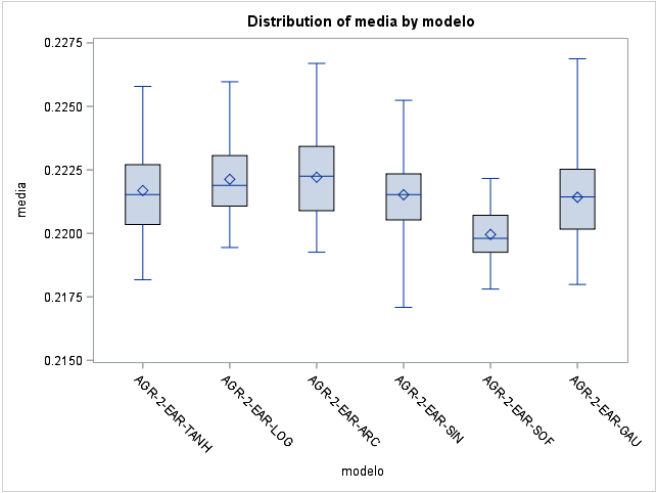


Figura 27. Mejor red con diferentes funciones de activación

Aunque la diferencia entre funciones es muy pequeña; la función de activación *softmax* funciona mejor que el resto; sobre todo, a nivel de varianza. Si bien es cierto que el sesgo es similar en todas ellas (el de la red con *softmax* es también algo menor); la red con *softmax* destaca por su estabilidad y varianza.

El siguiente paso es estudiar cómo afecta a los modelos los cambios en los algoritmos de aprendizaje. Para ello, se toma la mejor de las redes posibles, la red con selección agresiva *randomselect* de 2 nodos con *early stopping* y *softmax*, y se realiza validación cruzada repetida de variaciones de esta con distintos algoritmos.

A pesar de que la red con tangente hiperbólica se ha mostrado peor que la red con *softmax*; también se realizarán pruebas con el cambio de algoritmo en ella, a modo de control.

De este modo, se prueban ambas redes de activación con los algoritmos de Levenberg-Manguardt, Quasi-Newton, Gradiente conjugado, *Back Propagation*, Trureg y DBLDOG. El diagrama de cajas resultante se muestra en la Figura 28.

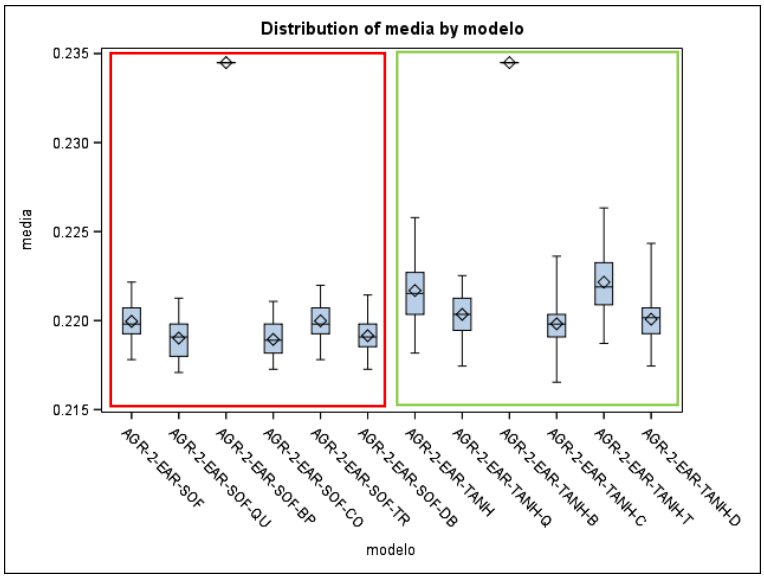


Figura 28. Diagrama de cajas que representa la tasa de fallos de las redes con distintos algoritmos de aprendizaje. En rojo, la red con función "softmax"; en verde, las redes con tangente hiperbólica.

Como se puede apreciar en la Figura 28; la función de activación *softmax* genera modelos más estables, sea cual sea el algoritmo utilizado; por lo que se puede descartar sin duda alguna la función de tangente hiperbólica y estudiar únicamente las redes con *softmax*; representadas en la Figura 29, donde también se ha descartado el algoritmo de *back propagation*, que no funciona perfectamente.

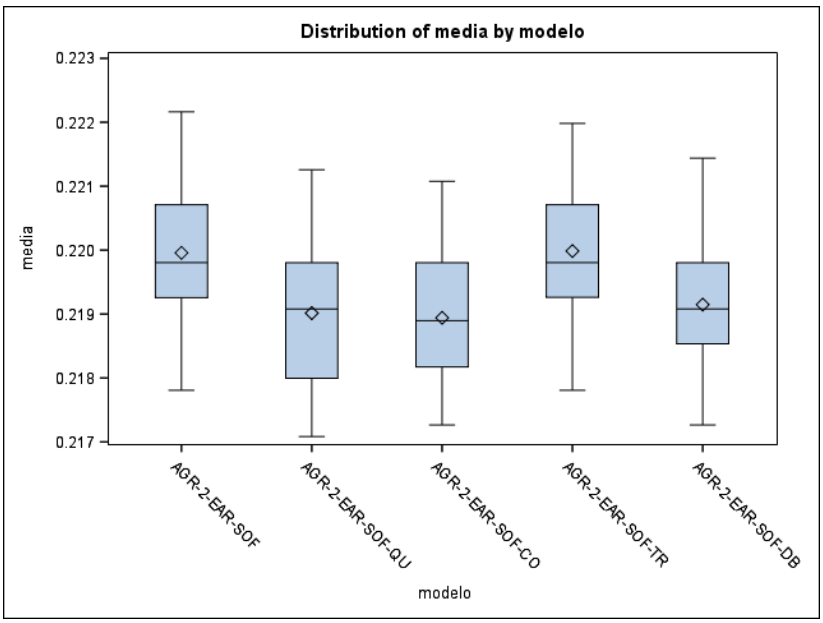


Figura 29. Diagrama de cajas que representa la tasa de fallos de las redes con softmax y diferentes algoritmos de aprendizaje.

Como se puede comprobar en la Figura 29, el cambio de algoritmo genera diferencias mínimas entre las redes. Sin embargo, la red con el algoritmo de "Congra" o "Gradiente conjugado" parece algo más estable que el resto; por lo que se va a estudiar más a fondo junto a la red con algoritmo de Levenberg-Marquandt, que es el más extendido.

Aunque la parametrización en SAS código es mucho mayor que en R; ahora que se conocen las características que hacen mejor a la red neuronal para este conjunto de datos, no está de más realizar algunas pruebas con el paquete *caret* de R para ver si es capaz de mejorar los datos de las redes obtenidas en SAS.

De este modo, se sabe de antemano que las redes que mejor funcionan son aquellas con un número de nodos más bajos; por lo que lo que las pruebas realizadas con R solo contarán con 2 y 3 nodos. Más allá del número de nodos, R sólo permite modificar el *learning rate* de las redes. Así, se trabajará con 2 y 3 nodos y *learning rate* de 0.1, 0.01 y 0.001.

Las comparaciones entre paquetes son complejas, dado que, habitualmente, el remuestreo de las validaciones cruzadas se realiza de forma diferente. Así, solo pueden entenderse como relevantes diferencias muy grandes entre el sesgo o la varianza de los modelos de los distintos paquetes.

En la Figura 30, se pueden ver las tasas de fallos de las redes en R (rojo) y las redes en SAS (verde). A primera vista, estas últimas son redes algo mejores, mostrando que la mayor capacidad de parametrización que permite SAS sí tiene resultado. No obstante, las diferencias no son muy notables. Las redes de SAS tienen menor varianza y menor

sesgo; pero nada muy relevante. Como ya se sabía, la red con algoritmo "gradiente conjugado" es la mejor de todas. En el caso de R, la red con un mejor equilibrio entre sesgo, varianza y complejidad es la de 2 nodos y *learning rate* de 0.01. Sin embargo, como se señaló anteriormente; estas comparaciones están condicionadas por el remuestreo de cada paquete.

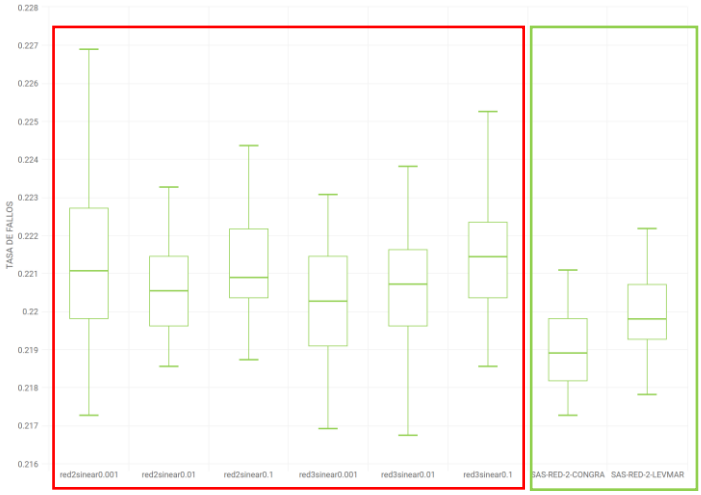


Figura 30. Diagrama de cajas que representa el índice ROC de las redes en R (resaltadas en rojo) y las redes SAS (en verde).

6.3.1 Mejor modelo de red neuronal

Hasta ahora, se han tomado de cada método el modelo con mejor capacidad predictiva y el modelo con un mayor equilibrio entre complejidad, sesgo y varianza. Sin embargo, para las redes neuronales el modelo que mejor predice ya es lo suficientemente simple.

A pesar de esto, también para este caso se van a tomar dos modelos. El primero de ellos es la red generada en SAS, con 2 nodos, función de activación *softmax*, *early stopping* = 15 y algoritmo de aprendizaje de "gradiente conjugado".

También se decide tomar la mejor red de R por un motivo principal: se necesita un modelo de red neuronal con el que, en el futuro, realizar pruebas de ensamblado de modelos. Para este proceso de ensamblado, se necesitan modelos que hayan sido generados con el mismo *software*. Esta elección también viene justificada por la igualdad entre las redes.

Tabla 9. Mejores modelos de red neuronal

	Software	**Sel. variables**	**Nodos**	**F. Activación**	**Algoritmo**	***Early stop***	***Learning rate***	**Nombre de modelo**	**Nº de modelo**	**AUC**	**TASA DE FALLOS**
Mejor modelo	SAS	Agr.*randomselect*	2	Sof	Congra	Sí	-	redagr2earsofcon	Modelo 167	-	0.2189
Modelo R	R	Agr.*randomselect*	2	Tanh	Levmar	No	0.01	red2.0.01	Modelo 170	0.7086	0.2204

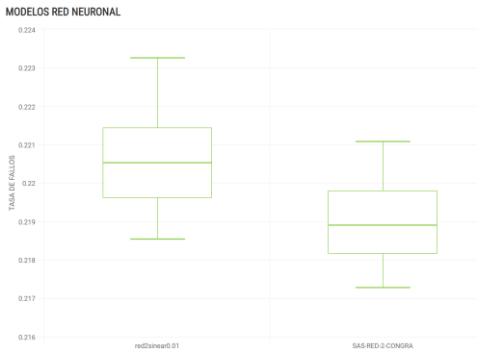


Figura 31. Diagrama de cajas que representa la tasa de fallos de las mejores redes neuronales

6.4. Random forest y bagging

Como se señalaba en la metodología, son varios los parámetros que se pueden modificar en la creación de *random forests*:

- *Sets* de variables. A pesar de que los árboles ya tienen de por sí mecanismos para realizar selección de variables; utilizar distintos grupos de variables puede ayudar a evitar sobreajustes y mejorar la varianza de los modelos. En este caso, se opta por utilizar las mismas tres selecciones que para las redes neuronales: una selección agresiva con *randomselect* (10 *inputs*), la selección agresiva (15 *inputs*) y la selección automática de *Miner* (59 *inputs*).
- Número de árboles. Hace referencia al número de árboles que realiza el modelo. Este número se ha establecido en 1000, ya que se ha podido comprobar que el error tiende a estabilizarse a partir de ese valor. Este valor permitirá agilizar las operaciones computacionales. En la Figura 32, se muestran dos gráficos que representan el error de dos ejemplos de *random forest* con distinto set de variables y características.

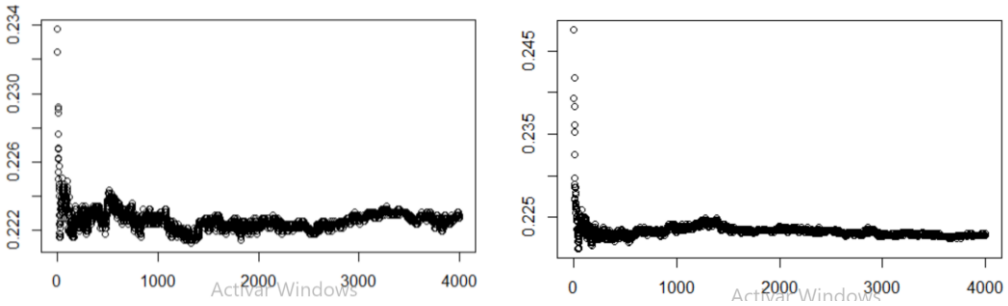


Figura 32. Error con distinto número de árboles.

- Tamaño de hoja. Representa el número mínimo de observaciones que deben caer en cada una de las hojas del árbol. Como forma de protección frente al sobreajuste, se ha establecido en 100 (algo menos del 2% de las observaciones).
- Número de variables. Como se comentaba anteriormente, modificar el número de variables utilizadas en la clasificación del árbol es capaz de generar modelos mejores o peores. Para los distintos sets se prueba con los siguientes números de variables. En la Tabla 10, se resaltan en negrita el número máximo de variables, correspondiente al *bagging*.

Tabla 10. Número de variables utilizadas en el random forest según selección de variables

	Variables *input*	Nº de variables
Agresiva *randomselect*	10	3,4,5,6,7,8,9,**10**
Agresiva	15	3,4,5,6,7,8,9,10,11,12,13,14,**15**
Miner	58	5,10,20,30,40,45,50,55,**58**

- Tamaño de muestra. Para generar los árboles, utilizar distintos tamaños de muestra puede generar mejores modelos y ayudar a combatir el sobreajuste. Para este trabajo, el máximo tamaño de hoja es 4400 (80% del conjunto train – validación cruzada de 5 grupos) y se utilizarán los tamaños 250, 500, 1000, 1500, 2500, 3500 y 4400.

Una vez determinados estos parámetros; se utiliza validación cruzada repetida de 5 grupos y 10 repeticiones para comparar modelos con los tres de variables, los distintos números de variables y tamaños de muestra. Se reduce las repeticiones de validación cruzada repetida de 50 a 10 para agilizar la computación. No obstante, cuando se reduzca el número de modelos a un grupo más prometedor, se realizará vcr de 50 repeticiones; para así poder comparar también con los mejores de otros métodos.

En la Figura 33, se aprecia como los modelos con selección *Miner* (en rojo) funcionan mejor que los modelos con selecciones más agresivas. Además, no solo lo hacen en sesgo; sino también en varianza, con modelos más estables. De este modo, se descartan los modelos con las selecciones más agresivas y estudian más a fondo los que tienen selección *Miner.*

Figura 33. Diagrama de cajas que representa el índice ROC de todos los modelos de random forest y bagging. En azul, los modelos con selección agresiva randomselect; en amarillo, selección agresica; y en rojo, selección Miner.

Las principales tendencias que se pueden apreciar son las siguientes:

- Los modelos que utilizan menos variables tienen una ligera ventaja en sesgo sobre aquellos con más variables; por lo que el *bagging* no parece la mejor opción.
- Los tamaños de muestra mayores funcionan mejor en la selección *Miner*; sobre todo cuando se combinan con números de variables más pequeñas. Para números de variables más alto, la complejidad de la cantidad de variables y grandes tamaños de muestra derivan en sobreajuste; por lo que funcionan mejor los tamaños de muestra intermedios.

Por tanto, se toman los modelos con 5 y 10 variables y máximos tamaños de hoja (3500, 4400) por destacar su desempeño en sesgo. También se tomarán los modelos de 45 variables y muestra 2500; y de 55 y 1000; que destacan en varianza. Con ellos, se realiza validación cruzada repetida de 50 repeticiones. En la Figura 34, se pueden observar las anteriores tendencias y se marcan en rojo los modelos más prometedores con lo que se mostrará la vcr de 50 repeticiones más adelante.

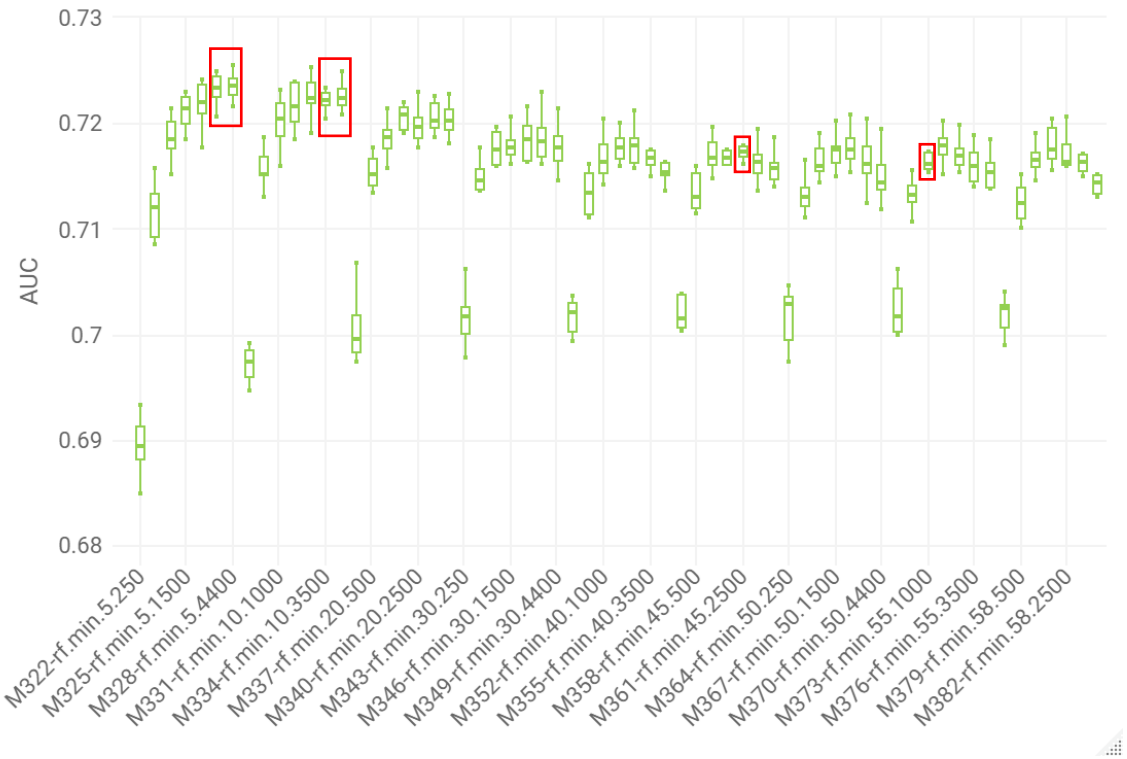


Figura 34. Diagrama de cajas que representa el índice ROC de todos los modelos random forest con selección Miner. En rojo, los modelos más prometedores.

Una vez realizada la validación cruzada repetida de 50 repeticiones, el resultado en forma de diagrama de cajas se muestra en la Figura 35.

Figura 35. Diagrama de cajas que representa el índice ROC de los mejores modelos de random forest

La tendencia es clara: los *random forest* que utilizan un menor número de variables para generar sus árboles de clasificación funcionan mejor; y no solo lo hacen en error; sino también en varianza. Los mejores dos modelos corresponden a modelos que utilizan tan solo 5 variables (de forma aleatoria en cada árbol) de las 58 posibles. Estos modelos se marcan en azul en la Figura 35. Además, también se puede corroborar que cuanto mayor es el tamaño de muestra para números de variables pequeños, mayor es la precisión del modelo. El *random forest* de 5 variables y 4400 observaciones como tamaño de muestra (resaltado en rojo en la Figura 35), es algo mejor en sesgo y varianza que su equivalente con 3500 observaciones.

Anteriormente, se estableció que el tamaño mínimo de hojas fuera del 2% de las observaciones. Sin embargo, al tratarse de un *dataset* relativamente grande, se pretende comprobar si árboles más agresivos y complejos pueden beneficiar al modelo. De este modo, se realiza validación cruzada repetida de 50 repeticiones con tamaños de hoja de 50 (1% de las observaciones) y 20 (0.4%).

Al comparar los diagramas de cajas de estos modelos con tamaño de hoja menor, se puede apreciar como son árboles más precisos e igual de estables. En el diagrama de cajas de la Figura 36, puede comprobarse como el *random forest* de tamaño de hoja máximo 20 (resaltado en rojo) es el mejor sin duda alguna:

6.4.1 Mejor modelo de *random forest/bagging*

Hasta ahora, se estaba concluyendo los apartados con la elección del mejor modelo en capacidad predictiva y modelo más equilibrado. No obstante, con respecto a los *random forest*, el único parámetro que podría simplificar el modelo y su aplicación sería la utilización de sets de variables más sencillos. Sin embargo, se ha podido comprobar que la pérdida en precisión es demasiado alta para justificar su complejidad. Por ese motivo, y dado que de ninguna de las formas un modelo con *random forest* va a permitir explicar los resultados (como si ocurre con regresión logarítmica); se opta por escoger un únicamente el modelo que mejor predice, cuyas características son las siguientes:

Tabla 11. Características del mejor modelo random forest.

	Modelo	Nº de modelo	Selección de variables	Nº de variables	Tamaño de muestra	Tamaño mínimo de hoja	AUC	Tasa de fallos
Mayor precisión	rf.min.5.4400.20	Modelo 386	*Miner*	5	4400	20	0.7319	0.2195

6.5. Gradient boosting

El *gradient boosting* es un tipo de modelo iterativo cuyo algoritmo es capaz de calcular mediante árboles el valor (o predicción) de los residuos e ir añadiéndoselo o restándoselo a la predicción de forma iterativa hasta conseguir el menor error en la predicción posible. Dos son los algoritmos más utilizados a la hora de poner en práctica este método: Gradient Boosting Machine (GBM) y XGBOOST.

GBM

Las opciones de parametrización en los modelos de *gradient boosting*, ya se trate de GBM o XGBoost son muchas; por lo que es imposible estudiar todos los modelos posibles con validación cruzada repetida. Para esquivar estos problemas computacionales; el paquete *caret* permite generar *grids* o rejillas que permiten evaluar las tendencias en error (concretamente *accuracy*) de la combinación de los distintos parámetros. No obstante, antes de entrar a evaluar los parámetros de los que se beneficia este conjunto de datos; conviene señalar cuáles son estos parámetros y qué valores a priori se han escogido para ellos:

- *Sets* de variables. Los algoritmos de *gradient boosting* son sensibles a la información que reciben; por lo que, para este método, se utilizarán la selección agresiva *randomselect*, la selección agresiva y la selección *Miner.*
- *Shrinkage* o constante de regularización. Es el parámetro más importante para obtener buenos modelos de *gradient boosting* y determina la velocidad con la que las nuevas iteraciones se aproximan al valor real en train. A mayor *shrinkage*, más rapidez y menor número de iteraciones. Sin embargo, más rápido no significa mejor. Se debe valorar el *shrinkage* junto al resto de parámetros. Su valor rara vez es mayor de 0.2. Para este conjunto, se realizan pruebas con 0.001, 0.01, 0.03, 0.05, 0.1, 0.2 y 0.25.
- Número de iteraciones o árboles. A diferencia de lo que ocurría con *random forest,* donde a partir de un determinado número de árboles el error se estabilizaba; con *gradient boosting* un número demasiado elevado de árboles

puede derivar en sobreajuste. Para evitar esto, preliminarmente se estudiará el comportamiento de los modelos con 100, 500, 1000, 2000 y 5000 árboles.

- Tamaño mínimo de hoja. Es necesario un control sobre este parámetro y encontrar el punto justo entre árboles muy específicos que sobreajusten y árboles demasiado simples. Como norma general, se suelen toman un 2% de las observaciones como tamaño mínimo. En este caso, 100 observaciones. También, se realizarán pruebas con 50 y 20.

Como se señaló anteriormente, antes de trabajar con validación cruzada repetida; se trabaja con validación cruzada simple de 5 grupos. Para esto, se toma cada una de las selecciones de variables; y se realizan pruebas para todos los anteriores parámetros.

Para la selección con *randomselect;* los resultados de *shrinkage* 0.001, 0.01, 0.03, 0.05, 0.1, 0.2; número de árboles 100, 500, 1000, 2000 y 5000; y tamaños de hoja 20, 50 y 100. La Figura 36 resume el comportamiento del error en estos modelos.

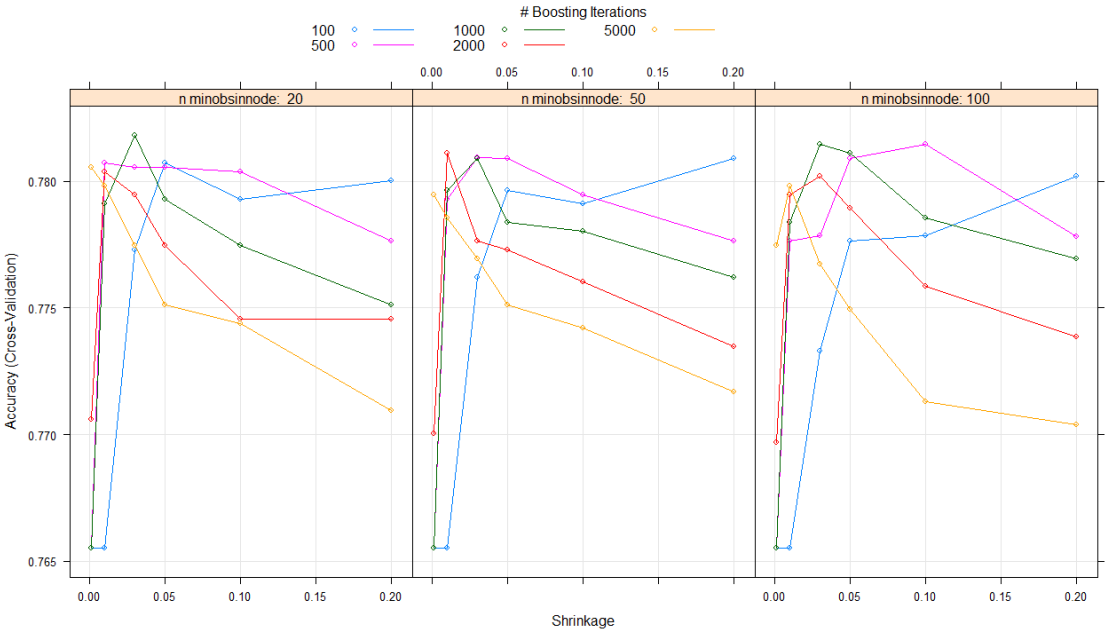


Figura 36. Representación del error con modelos de gradient boosting de diferente parametrización.

Del gráfico de la Figura 36 la principal conclusión a la que se puede llegar es el mejor funcionamiento de modelos con menor número de iteraciones (100, 500, 1000) frente a los modelos con más árboles (en rojo y amarillo). Estos últimos, como es lógico, funcionan mejor con *shrinkages* más pequeños; pero siempre por debajo de los modelos con menos árboles.

De este modo, se vuelve a realizar la misma validación cruzada simple para menos número de árboles (100, 500, 1000 y 1500) y se añade un *shrinkage* más agresivo de 0.25; pues los modelos con 100 árboles (línea azul) podrían beneficiarse de esto. Los resultados aparecen en la Figura 37.

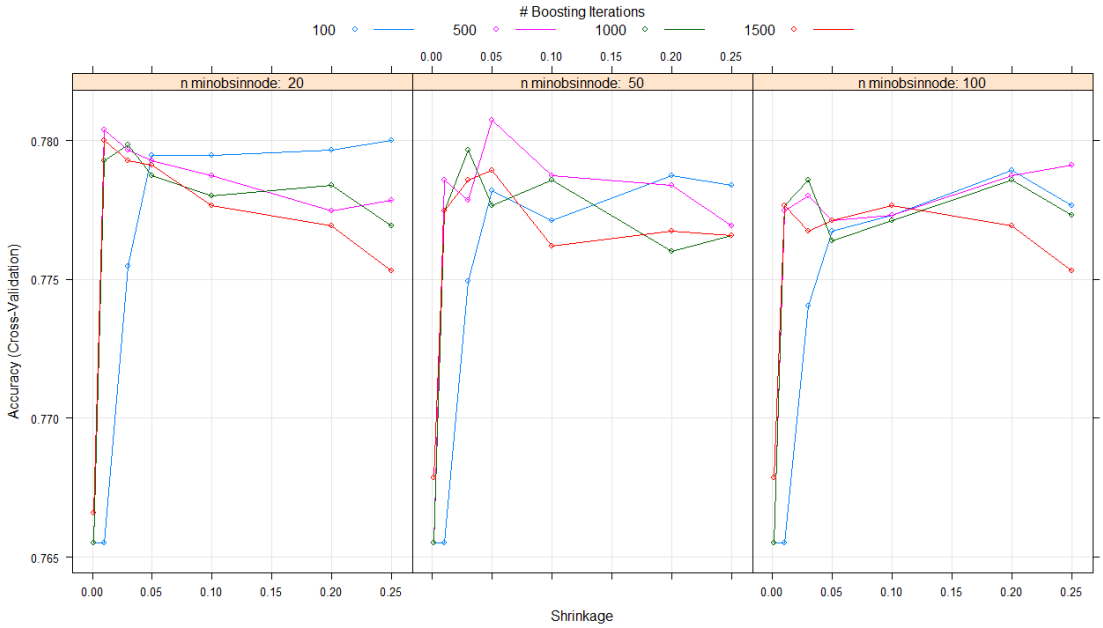


Figura 37. Representación del error de modelos de gradient boosting con selección randomselect.

Dado que las diferencias son bastante bajas, se toma la decisión de para cada uno de los tamaños de hoja y tamaños de árboles, tomar el aparentemente mejor *shrinkage*; con la posibilidad de realizar algunas pruebas más una vez escogido el mejor modelo para ajustarlo, si es posible. Con estas medidas, se realiza validación cruzada repetida de 5 grupos y 50 repeticiones, como hasta ahora.

En el caso de la selección *randomselect;* las características de los modelos más prometedores se resumen en la Tabla 12.

Tabla 12. Modelos con selección randomselect con los que se hace validación cruzada repetida

Tamaño de hoja	Número de árboles	*Shrinkage*
20	100	0.25
	500	0.01
	1000	0.03
	1500	0.01
50	100	0.20
	500	0.05
	1000	0.03
	1500	0.05
100	100	0.20
	500	0.03
	1000	0.03
	1500	0.01

A continuación, se realizará el mismo estudio anterior con los otros sets de variables para valorar con qué parametrizaciones se hace validación cruzada repetida.

Con la selección agresiva, se hace un estudio preliminar exactamente igual al anterior; con el gráfico de la Figura 38 como resultado.

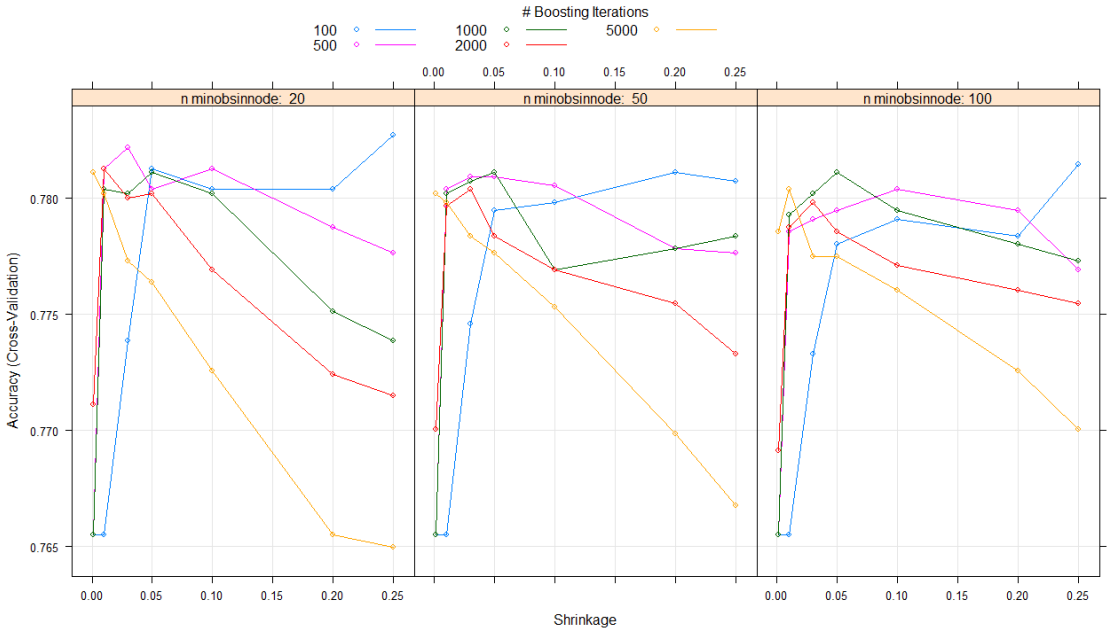


Figura 38. Representación del error de modelos de gradient boosting con selección agresiva.

En este caso, el único número de árboles que se puede descartar de primeras es el de 5000; que funciona peor que todos los demás. En cambio, el modelo con 2000 árboles sí que puede llegar a funcionar; quizá compensado por la mayor complejidad de esta selección de variables. Igual que ocurría anteriormente; se establecen algunas combinaciones de parámetros que parecen prometedores (Tabla 13). Más tarde, se harán retoques y pruebas para ajustar aún mejor los modelos.

Tabla 13. Modelos con selección agresiva con las que se hace validación cruzada repetida.

Tamaño de hoja	Número de árboles	*Shrinkage*
20	100	0.25
	500	0.03
	1000	0.05
	2000	0.03
50	100	0.20
	500	0.03
	1000	0.05
	2000	0.03
100	100	0.25
	500	0.10
	1000	0.05
	2000	0.03

Con respecto a la selección de variables de *Miner*, se puede comprobar en la siguiente figura como los números de árboles mayores sí son igual de competitivos que los números de árboles más bajos; por lo que, para este set de variables, no se descartará el *ntrees* de 5000.

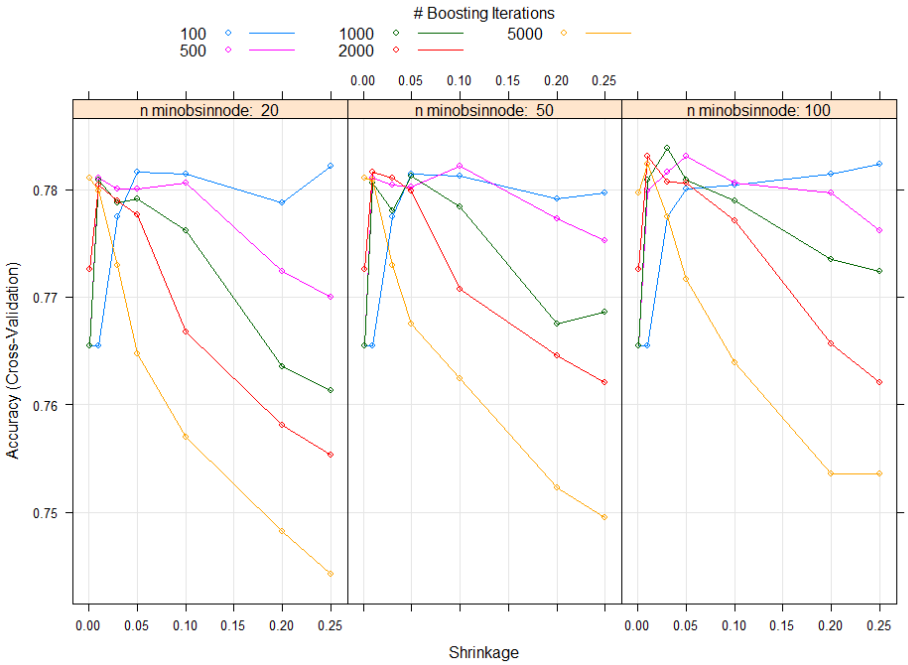


Figura 39. Representación del error de modelos de gradient boosting con selección Miner.

Así, para esta selección *Miner*, en la Tabla 14 se resumen las características de los modelos con los que se hará validación cruzada repetida.

Tabla 14. Modelos con selección agresiva con las que se hace validación cruzada repetida.

Tamaño de hoja	Número de árboles	*Shrinkage*
20	100	0.25
	500	0.1
	1000	0.01
	2000	0.01
	5000	0.001
50	100	0.05
	500	0.1
	1000	0.05
	2000	0.01
	5000	0.001
100	100	0.25
	500	0.05
	1000	0.03
	2000	0.01
	5000	0.01

En total, se realiza un estudio exhaustivo con validación cruzada repetida de 39 modelos; cuya representación gráfica según índice ROC se muestra en la Figura 40.

La primera conclusión que se puede tomar de la Figura 40 es la superioridad de los modelos con selección *Miner* (en rojo) sobre las otras dos selecciones.

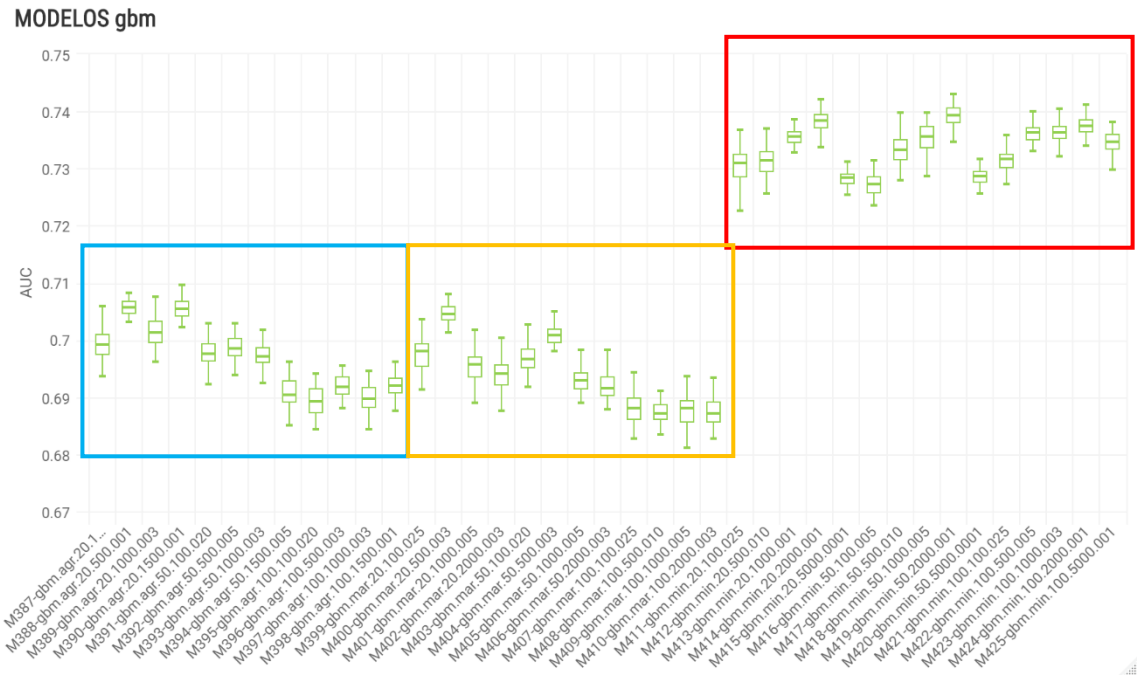


Figura 40. Diagrama de cajas que representa el error de los modelos con GBM. En azul, modelos con selección randomselect; en amarillo, selección agresiva; y en rojo, selección Miner.

La Figura 41 muestra un zoom sobre los modelos con esta selección. En ella, se puede ver cómo es en los modelos con 2000 árboles (resaltados en rojo) donde se alcanza el pico de índice ROC. Los modelos de 5000 árboles tienden a sobreajustar, pues funcionan muy mal para todos los tamaños de hoja. A su vez, los modelos con demasiados pocos árboles (100 y 500 – resaltados en azul; salvo en selección *Miner,* donde el árbol de 500 sí es competitivo) son muy simples y tampoco son capaces de competir con los modelos de 1000 y 2000 árboles.

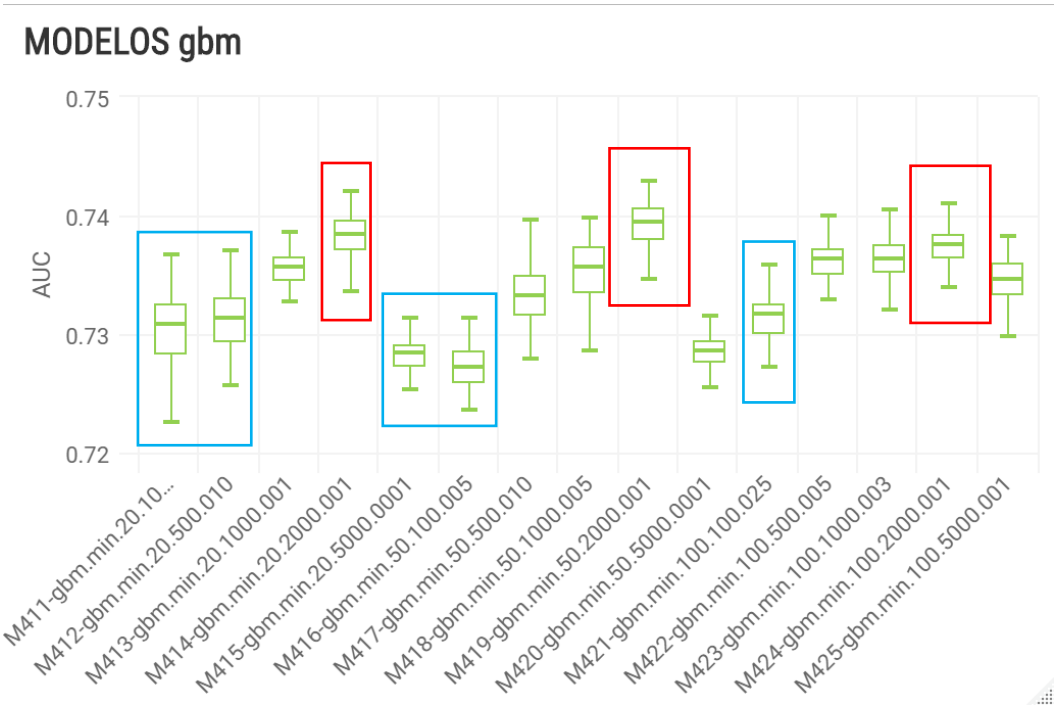


Figura 41. Diagrama de cajas que representa el índice ROC de los modelos gbm con selección Miner. En azul, los modelos con 100 y 500 como número de árboles; y en rojo, los de 2000.

Entre los mejores modelos destacados en rojo (2000 árboles); se opta por el modelo de la derecha; cuya varianza es más baja que la de los otros. Otro motivo para escogerlo es su tamaño mínimo de hoja (el más grande), que ronda el 2% de las observaciones; y permite evitar de forma más sencilla el sobreajuste.

Se toma este modelo y se vuelve a ejecutar validación cruzada repetida modificando ligeramente el número de árboles, para ver si se puede ajustar aún más el modelo. La Figura 42 muestra el ROC del hasta ahora el mejor modelo junto a sus equivalentes de 1500 y 2500 árboles.

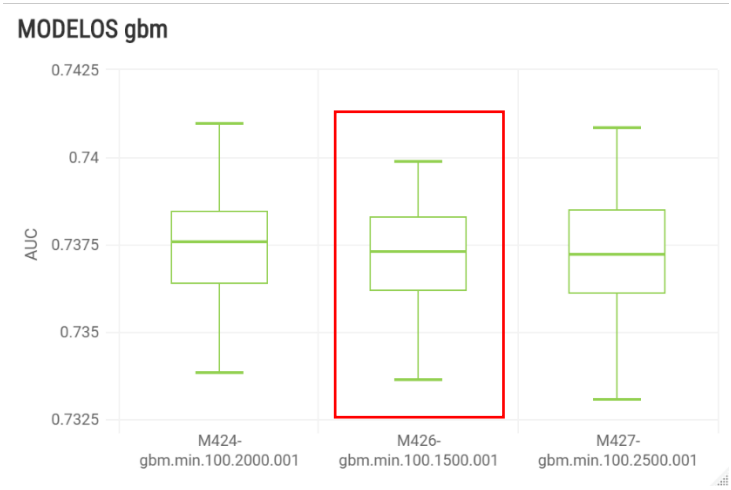


Figura 42. Diagrama de cajas que representa el índice ROC de los modelos gbm con 2000, 1500, 2500 árboles.

Como se puede apreciar; una reducción en el número de árboles se resuelve en una muy ligera subida del error; pero compensada por la menor varianza. Por este motivo, se decide tomar el modelo con 1500 árboles (resaltado en rojo).

Ahora que ya se cuenta con el mejor modelo en cuanto a número de árboles; se quiere observar si ligeros cambios en la constante de regularización o *shrinkage* tienen algún efecto positivo en el modelo. Por este motivo, se realiza validación cruzada repetida modificando este parámetro. El del modelo original era 0.01; y las nuevas pruebas se realizan con 0.02 y 0.03 (0.03 era el mejor *shrinkage* para su modelo equivalente con 100 árboles); ya que lo normal es aumentar esta constante a medida que se cuenta con un menor número de árboles.

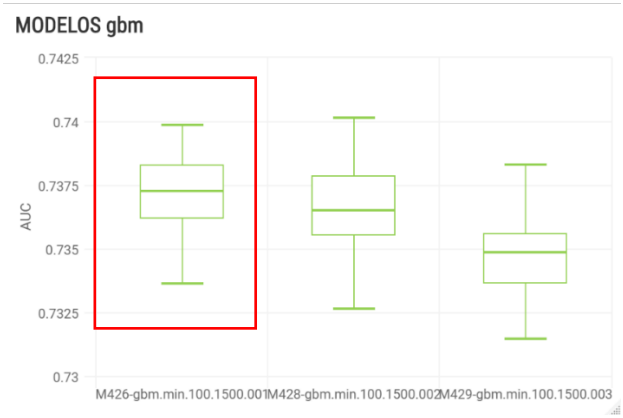


Figura 43. Diagrama de cajas que representa el índice ROC de modelos con distintos valores de shrinkage

Tras realizar estas pruebas, se comprueba que ninguno de los nuevos valores de *shrinkage* mejora al 0.01; ya que, como se puede comprobar en la Figura 43, el modelo con 0.01 (resaltado en rojo) presenta mejores resultados tanto en error como en varianza.

Por tanto, el mejor modelo de GBM es un modelo con selección *Miner*, 1500 iteraciones, 100 observaciones como tamaño mínimo de hoja y *shrinkage* de 0.01. A continuación, se realizarán pruebas con el otro algoritmo disponible en *gradient boosting* (XGBoost) y se compararán los mejores modelos de ambos.

XGBoost

El XGBoost es un algoritmo de *gradient boosting* que comparte la mayoría de sus características y procedimientos; pero que añade una regularización que permite tener un mayor control sobre la varianza del error. En teoría, este paquete debería conseguir modelos más estables que los modelos con GBM. Sin embargo, no siempre es así. Aunque las técnicas son similares, el ajuste de los modelos se realiza con distintos algoritmos y esto puede provocar que los modelos con GBM sean mejores.

Tanto GBM como XGBoost son modelos de *gradient boosting*; por lo que XGBootst puede parametrizarse con las mismas opciones que GBM. En este caso, todos los parámetros para realizar las pruebas preliminares serán los mismos; salvo la selección de variables, pues solo se utilizará la selección *Miner*, que se mostró como la mejor opción con GBM de forma muy clara.

Además de los parámetros que se tuneaban con GBM (número de árboles, tamaño mínimo de hoja y *shrinkage*), con XGBoost se pueden modificar dos parámetros más:

- Porcentaje de variables utilizadas. Es un parámetro similar al número de variables de *random forest*. Permite variar el porcentaje de variables que se utiliza para generar el modelo. De este modo, se puede evitar el sobreajuste. Una vez obtenidos los modelos más prometedores, se variará este parámetro en busca de mejoras.
- *Gamma/alpha* o reguladores. Son constantes que, en raras ocasiones, consiguen mejorar el modelo.

Como ocurría con GBM con la siguiente rejilla o *grid* se realiza una validación cruzada simple de 5 grupos para observar patrones que ayuden a reducir las opciones y poder llevar a cabo validación cruzada repetida de 50 repeticiones con unos pocos modelos.

```
xgbmgrid<-expand.grid(
  min_child_weight=c(20,50,100),
  eta=c(0.001,0.01,0.03,0.05,0.1,0.2,0.25),
  nrounds=c(100,500,1000,2000,5000),
  max_depth=2,gamma=0,colsample_bytree=1,subsample=1)
```

La Figura 44 resume la precisión de cada una de las distintas combinaciones.

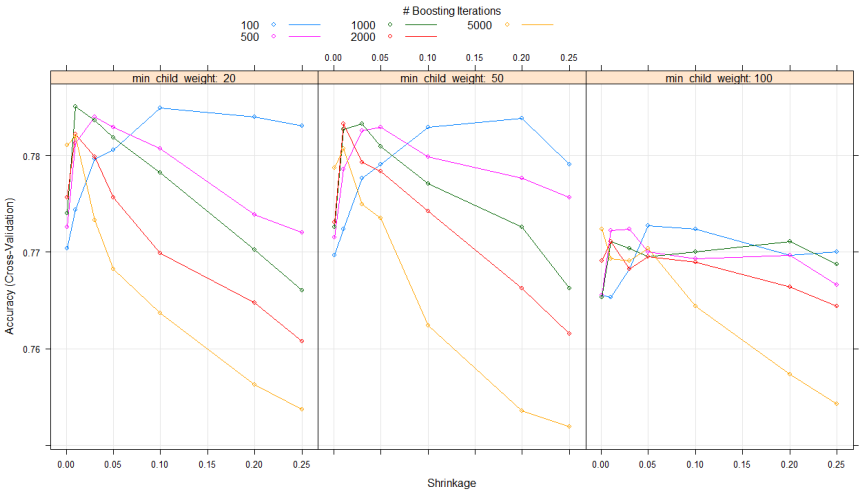


Figura 44. Representación de la accuracy de los modelos XGBoost

El gráfico de la Figura 44 permite rechazar el tamaño de hoja de 100 (correspondiente al 2% de las observaciones), representado a la derecha; ya que presenta una *accuracy* bastante inferior a las otras dos. También se puede rechazar el tamaño de árbol 5000 (línea amarilla), que está bastante por debajo del resto de tamaños de árboles.

Por tanto, como se hizo con GBM, se escoge el *shrinkage* adecuado para cada número de árboles y se realiza validación cruzada repetida de 50 repeticiones con las combinaciones de parámetros de la Tabla 15.

Tabla 15. Combinaciones de parámetros de los modelos con los que se hace validación cruzada repetida

Tamaño de hoja	Número de árboles	*Shrinkage*
20	100	0.1
	500	0.03
	1000	0.01
	2000	0.01
50	100	0.2
	500	0.05
	1000	0.03
	2000	0.01

El índice ROC de los anteriores es modelos está representado en la Figura 45.

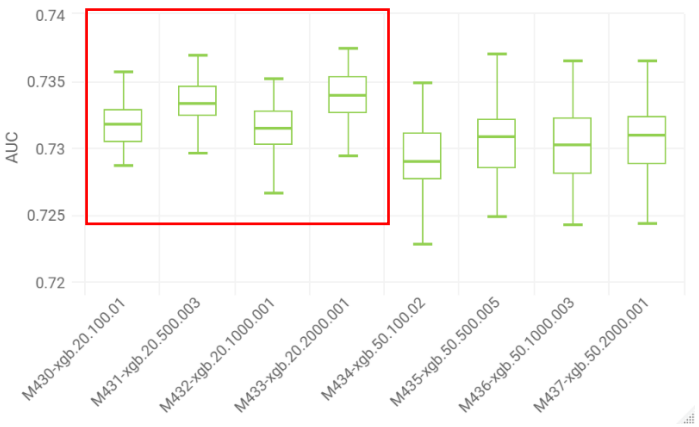


Figura 45. Diagrama de cajas que representa el índice ROC de los modelos con XGBoost

A diferencia de lo que ocurría con GBM, los modelos con XGBoost se benefician de un tamaño de hoja más pequeño (tamaño de hoja 20 resaltado en rojo en la figura); en concreto destacan los modelos de 500 y 2000 iteraciones. Por este motivo, estos dos modelos son los escogidos para realizar pruebas con el parámetro de porcentaje de variables utilizadas. En este caso, se utilizan 5%, 12.5%, 25%, 50% y 75%.

En la Figura 46, se puede observar cómo los modelos con aleatorización de las variables utilizadas (en azul) funcionan mejor que los modelos sin ese tratamiento. No son diferencias drásticas, pero sí lo suficientemente grandes para justificar su uso.

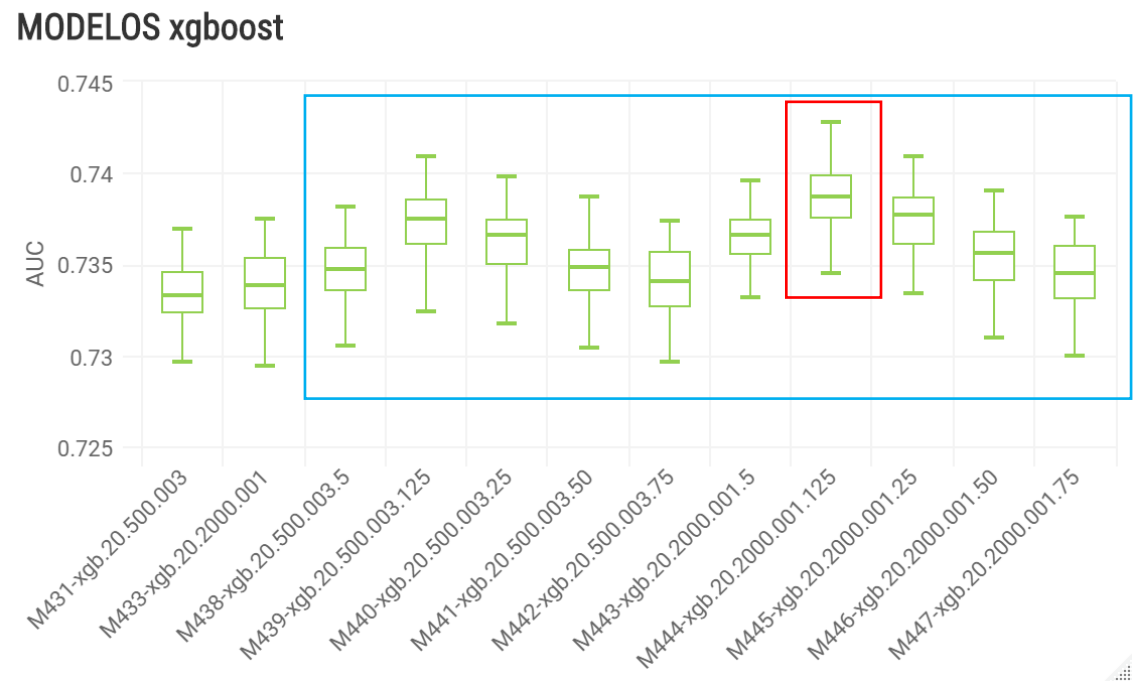


Figura 46. Diagrama de cajas que representa el índice ROC de modelos con distintos porcentajes de variables utilizadas. En azul, los modelos que no utilizan el 100% de las variables. En rojo, el mejor modelo.

De forma clara, se puede observar cómo el mejor modelo (en rojo) es un modelo *XGBoost* con 2000 iteraciones y que utiliza un 12.5% de las variables de forma aleatoria; o lo que es lo mismo, 7 variables.

Comparación entre GBM y XGBoost

Ahora que ya se dispone de los dos mejores modelos para cada uno de los algoritmos, se comparan en la Figura 47.

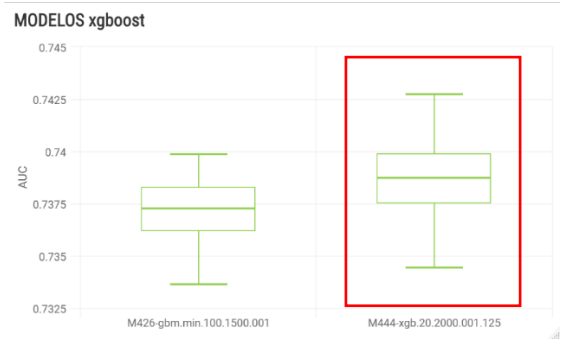


Figura 47. Diagrama de cajas que representa el índice ROC de los mejores modelos con GBM y XGBoost. En rojo, el mejor modelo.

Como se aprecia en la Figura 47, el modelo con XGBoost es un modelo ligeramente más preciso y que sacrifica muy poco en cuestión de varianza si se compara con el modelo GBM.

6.5.1 Mejor modelo *gradient boosting*

Como hasta ahora, se plantean dos modelos: el modelo con la mayor capacidad predictiva y el modelo que tenga el mayor equilibrio entre sesgo, varianza y simplicidad. En el caso del *gradient boosting*, la simplicidad solo se puede medir con el tiempo de ejecución de los modelos, que viene dado, principalmente, por el número de iteraciones o árboles del modelo.

Por ese motivo, además de tomar el modelo que mejor predice; se escoge un modelo que sea más simple y que se pueda ejecutar de forma más rápida. Con esta idea en mente, se toma un modelo GBM con 500 árboles y 100 de tamaño de hoja mínimo. Este modelo es cuatro veces más rápido que el modelo con mejor capacidad predictiva.

Las características de estos modelos aparecen la Tabla 16.

Tabla 16. Características de los mejores modelos de gradient boosting

	Algoritmo	**Nº árboles**	**Tamaño de hoja**	***Shrinkage***	**% de variables**	**Nombre del modelo**	**Nº del modelo**	**AUC**	**TASA DE FALLOS**
Mayor precisión	XGBoost	2000	20	0.01	12,5%	xgb.20.2000.001.125	Modelo 444	0.7387	0.2156
Mejor equilibrio	GBM	500	100	0.05	-	gbm.min.100.500.005	Modelo 422	0.7361	0.2185

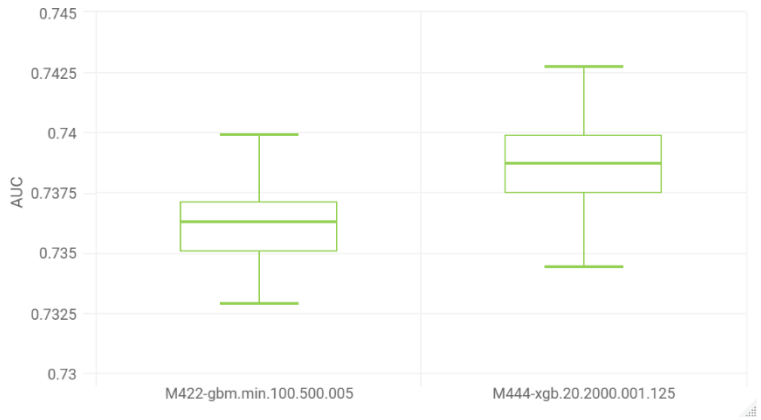


Figura 48. Diagrama de caja que representa el índice ROC de los mejores modelos de gradient boosting

6.6. Support Vector Machines

Los SPV son sensibles a la información que reciben y no tienen mecanismos de discriminación de variables integrados; por lo que se utilizan los tres *sets* de variables que se han venido usando hasta ahora: selección *randomselect*, selección agresiva y selección *Miner*.

A continuación, se valorarán los modelos de SVM lineal y SVM radial por separado, para más tarde comparar los mejores de cada uno de ellos.

Support Vector Machines: Lineal

La parametrización en el SVM Lineal es bastante reducida; ya que solo hay un parámetro modificable: el parámetro C. Este parámetro no tiene valores predeterminados; y el adecuado para cada conjunto de datos puede estar entre 0.0001 y 10000. Además, la optimización en SVM es lenta y costosa computacionalmente. Por este motivo, se realiza validación cruzada simple (5 grupos) para determinar en qué región se encuentra el parámetro C idóneo para este conjunto. Los valores de C utilizados para esta prueba son 0.0001, 0.001, 1, 100, 1000 y 10000. Se obvian los valores entre 0.001 y 1 porque se ha podido observar que no suponen cambios con respecto a C=1.

Los resultados de esta prueba para cada selección se presentan en términos de *accuracy* en la Tabla 17.

Tabla 17. Resumen del error de los modelos SVM lineal preliminares. En verde, los modelos más prometedores.

Selección	Parámetro C	Accuracy
Selección *randomselect*	0.0001	0.7655
	0.001	0.7660
	1	0.7728
	100	0.7728
	1000	0.7728
	10000	0.7728
Selección agresiva	0.0001	0.7655
	0.001	0.7668
	1	0.7728
	100	0.7728
	1000	0.7728
	10000	0.7713
Selección *Miner*	0.0001	0.7655
	0.001	0.7655
	1	0.7728
	100	0.7728
	1000	0.7410
	10000	0.7064

Como se puede ver en la anterior tabla, hay que ser muy agresivo con el parámetro C (valores muy altos o muy bajos) para obtener cambios; ya que los valores usuales entre 0 y 1 optimizan de la misma forma y llegan al mismo modelo.

De este modo; para cada *set* de las variables se toma únicamente el parámetro C=1 para realizar validación cruzada repetida de 50 repeticiones. La representación gráfica del error se muestra en la Figura 49.

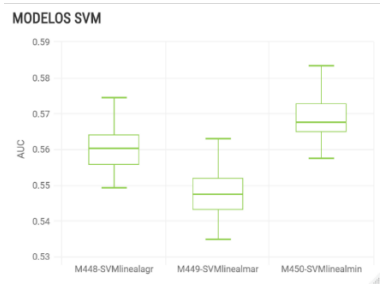


Figura 49. Diagrama de cajas que representa el índice ROC de los mejores modelos con SVM lineal

Como se puede apreciar de forma clara, el mejor de estos modelos es el que utiliza la selección *Miner* (a la derecha). Sin embargo, desde ya se puede asegurar que de ningún modo puede competir con otros modelos; pues su índice ROC no llega al 0.6.

Support Vector Machines: RBF o Radial

Este tipo de SVM tiene dos parámetros que se puede modificar: el parámetro C (el mismo que en lineal) y el parámetro *sigma*; que es un parámetro de escala. Ambos parámetros están correlacionados; y su comportamiento ha de observarse en conjunto, Por ese motivo, se realiza validación cruzada simple de 5 grupos para observar las tendencias de estos parámetros. Los valores utilizados son los siguientes:

- Parámetro C: 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 25, 100, 1000, 10000 y 100000.
- *Sigma*: 0.0001, 0.005, 0.01 y 0.05

La Figura 50 representa la precisión en *accuracy* de estas combinaciones de parámetros con las distintas selecciones.

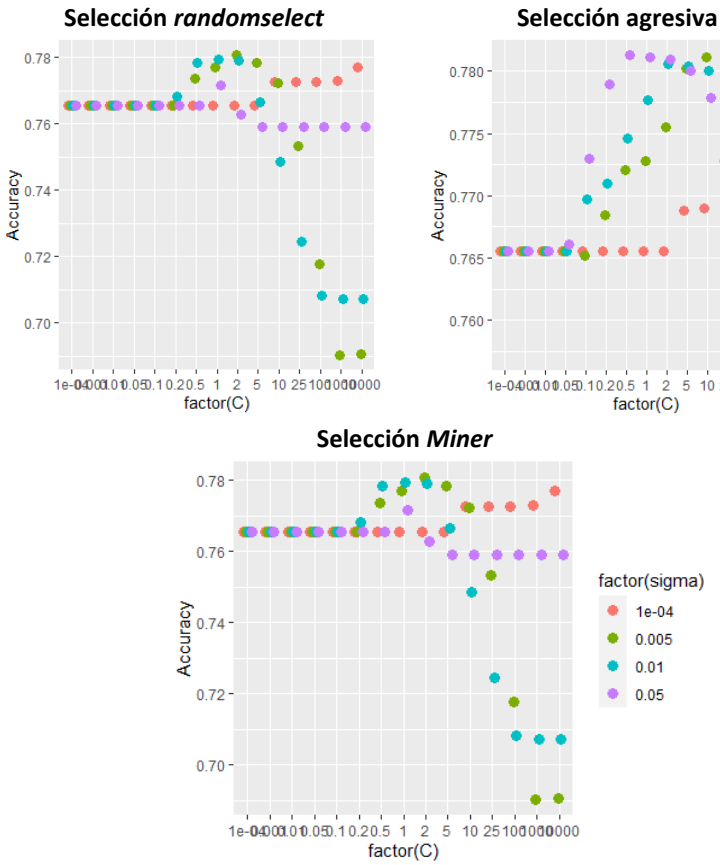


Figura 50. Representación de la precisión de los modelos con SVM radial

En la Figura 50, se puede ver cómo evoluciona el acierto (eje Y) de los modelos con cada *sigma* (se representa cada valor con puntos de diferentes colores) unido a cada valor de C (eje X). Para cada *sigma*, se toma el valor del parámetro C que alcanza el mayor *accuracy*. Las tendencias son claras y se ve cómo a menor *sigma*, mayor debe ser el C. La Tabla 18 resume las combinaciones de *sigma* y C con las que se realizará validación cruzada repetida.

Tabla 18. Combinaciones de SVM radial con los que se realiza validación cruzada repetida

Selección de variables	*Sigma*	C
Selección *randomselect*	0.05	1
	0.01	1
	0.005	2
	0.0001	10000
Selección agresiva	0.05	0.5
	0.01	2
	0.005	10
	0.0001	10000
Selección *Miner*	0.05	1
	0.01	1
	0.005	2
	0.0001	10000

La Figura 51 resume el índice ROC de los anteriores modelos.

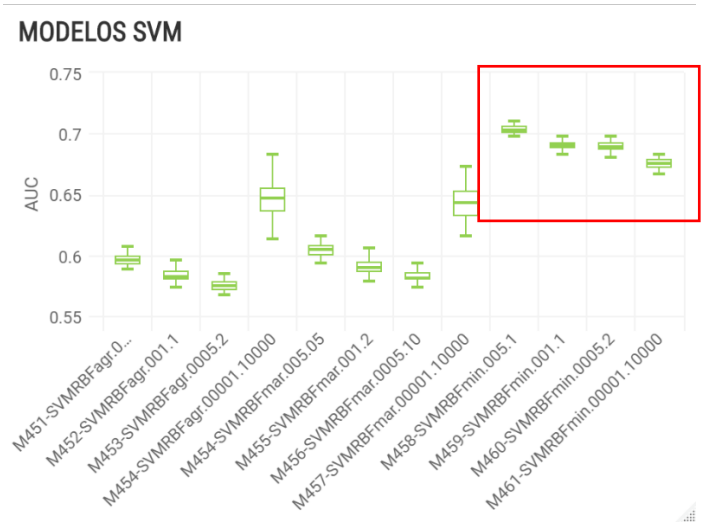


Figura 51. Diagrama de cajas que representa el índice ROC de los mejores modelos con SVM RBF. En rojo, los modelos con selección Miner.

En el gráfico de la Figura 51, se puede comprobar como los modelos con selección *Miner* (en rojo) son bastante más competitivos que sus equivalentes con otras selecciones. Entre los modelos con esa selección, destaca el modelo con el *sigma* más alto (0.05) y C=1. Aunque se aprecia que *sigma* es el más alto probado; probar con otros valores aún más alto no es posible.

Comparación entre SVM Lineal y SVM RBF

La diferencia entre ambos tipos de SVM es muy notable (Figura 52).

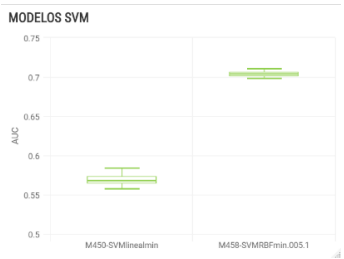


Figura 52. Comparación en índice ROC entre los mejores modelos de SVM lineal y RBF.

El SVM radial (a la derecha) es mucho mejor en error y bastante mejor en varianza; por lo que no hay duda de cúal de los dos es mejor.

6.6.1 Mejor modelo de *Support Vector Machines*

Aunque hasta ahora se han elegido dos modelos, uno con mayor precisión y otro con mayor equilibrio, con SVM solo se selecciona el modelo con mayor capacidad predictiva; pues ninguna característica de los SVM puede hacerlos más simples o con mayor equilibrio.

Las características del mejor modelo SVM se muestran en la Tabla 19.

Tabla 19. Características del mejor modelo con SVM

Selección de variables	Tipo SVM	C	Sigma	Nombre de modelo	Número de modelo	AUC	Tasa de fallos
Selección *Miner*	RBF	1	0.05	SVMRBFmin.005.1	Modelo 458	0.7034	0.2265

6.7. Ensamblado

Una vez se conocen los mejores modelos para todos los métodos de predicción, puede resultar interesante combinar las predicciones de probabilidad de evento de los distintos mejores modelos para cada observación, con el fin último de reducir, principalmente, la varianza.

De este modo, con la validación cruzada repetida de 50 repeticiones, se obtiene la probabilidad de oposición a la vacuna para cada observación y cada modelo; y se realiza la media aritmética en combinaciones de 2, 3, 4, 5 y 6 (total de los mejores modelos).

La Figura 53 resume el índice ROC de estos ensamblados.

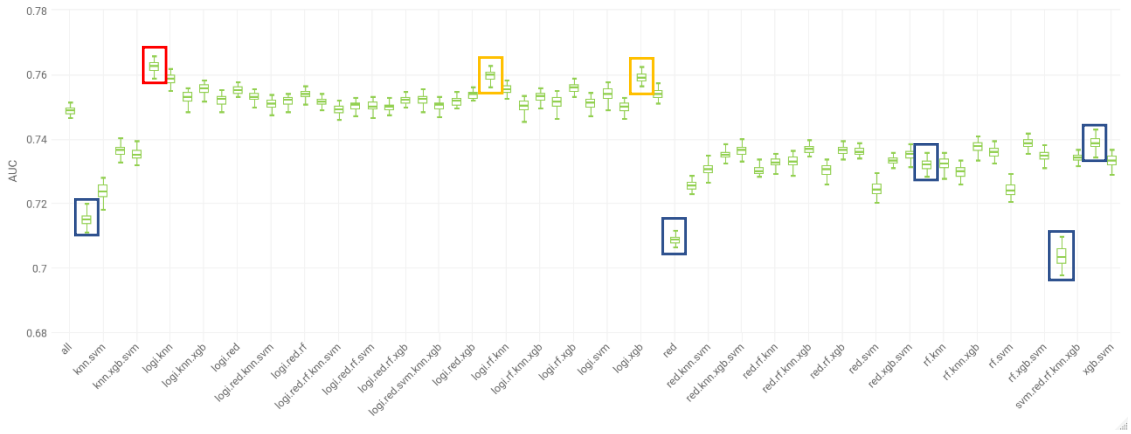


Figura 53. Diagrama de cajas que representa el error de los mejores modelos de cada método y sus ensamblados.

En azul, se representan los modelos simples sin ensamblar. Como norma general. El ensamblado tiende a reducir ligeramente la varianza, pero este conjunto de datos se beneficia enormemente del uso de la regresión logística (resaltada en rojo) y ninguna combinación es capaz de competir con ella.

Algunas combinaciones simples, como logística-*random forest* o logística-*gradient boosting* (resaltadas en amarillo en la Figura 53), sí reducen ligeramente la varianza; pero la mejora es tan pequeña que no justifican la complejidad que entrañan.

7. Selección del mejor modelo

Tras analizar los mejores modelos posibles y comprobar que las técnicas de ensamblado no generan modelos más interesantes, se puede asegurar que el modelo con la mayor capacidad predictiva es una regresión logística con las siguientes características (Tabla 20).

Tabla 20. Características del modelo que mejor predice

	Variables	Criterio de selección	Criterio de información	Modelo	Nº de modelo	Parámetros	AUC	Tasa de fallos
Mayor precisión	Originales y Transformadas	*Stepwise*	AIC	RegTransDumStepAIC	Modelo 122	60	0.7623	0.2152

Además, si lo que se pretende es tener un modelo con buena capacidad predictiva (no la mejor), pero una mayor explicabilidad con la que obtener conclusiones más claras; se debe acudir a otra regresión logística con distintas características (Tabla 21).

Tabla 21. Características del modelo con mayor equilibrio entre precisión y explicabilidad.

	Variables	Criterio de selección	Criterio de información	Modelo	Nº de modelo	Parámetros	AUC	Tasa de fallos
Mayor equilibrio	Originales	*Forward*	BIC	RegOrigDumForwBIC	Modelo 114	19	0.7458	0.2168

7.1. Evaluación del mejor modelo

Una vez escogidos los modelos con mayor capacidad predictiva y mayor equilibrio entre sesgo, varianza y simplicidad, conviene ver qué tal funcionan con el conjunto de datos *test* que se obtuvo al comienzo de este trabajo con una partición de datos 80%-20%. Por tanto, se cuenta con 1380 observaciones con las que no se ha generado el modelo y qué pueden dilucidar si el obtenido es un modelo útil en la práctica o no.

7.1.1 Evaluación del modelo con mayor capacidad predictiva y su mejor punto de corte

En cuanto al área bajo la curva ROC, este modelo tiene un índice de 0.7217, una diferencia notable con el índice ROC en validación, que era de 0.7623. Como ya se sabía, es un modelo complejo, lo que lo convierte en más proclive al sobreajuste.

En cuanto a la tasa de fallos, este modelo falla en un 23.39% de las ocasiones, comparado con el 21.52% que lo hacía en validación. Esta tasa de fallos se obtiene colocando el punto de corte de la probabilidad del evento en 0.5, lo que otorga el mínimo valor de *missclasification rate* o tasa de fallos.

No obstante, en ocasiones conviene cambiar el ese punto de corte, de modo que el modelo arroje unas predicciones más acordes al fin último del proyecto. Hay que tener claro que cualquier cambio en este punto de corte derivarán en una tasa de fallos más alta. Sin embargo, determinados puntos de corte pueden aumentar la capacidad de detectar los eventos, es decir, saber que sujetos sí se oponen a la vacuna. Como

contrapartida, estos puntos de corte incluirían como opositores a la vacuna a bastantes sujetos que no lo son. En resumen, un aumento de la sensibilidad del modelo deriva necesariamente en un descenso de la especificidad.

Tabla 22. Índices de precisión con diferentes puntos de corte

	Punto de corte	Tasa de fallos	Sensibilidad	Especificidad	Pos Pred
Punto corte 0.5	0.5	23.39%	19.70%	94.13%	50.80%
Igual sensibilidad y especificidad	0.197	33.09%	66.77%	66.95%	38.34%
Máximo Índice de Youden	0.192	33.24%	69.85%	65.85%	38.69%
Proporción del evento	0.2353	30.41%	58.46%	73.01%	40%

La Tabla 22 representa las tasas de precisión para diferentes puntos de corte. "Pos Pred" hace referencia a cuántos de los sujetos que el modelo clasifica como evento son verdaderos positivos.

Para comparar con el punto de corte 0.5, se han utilizado: el valor que iguala sensibilidad y especificidad, el valor que maximiza el índice de Youden y la proporción del evento. El punto de corte 0.5 destaca por su tasa de fallos, que es menor que la proporción del evento, una buena guía en la que basarse; pero su sensibilidad es muy baja, y solo detecta a 1 de cada 5 opuestos a la vacuna. Este problema concreto necesita de una sensibilidad razonable, pues es primordial detectar a los reacios a vacunarse. Asimismo, colocar a no eventos como opositores tampoco es un problema muy relevante, aunque una mínima precisión es necesaria.

De este modo, se decide tomar el punto de corte de la proporción del evento, pues otorga una sensibilidad aceptable (que roza el 60%) y una tasa de verdaderos positivos sobre el total de positivos (40%) también aceptable.

7.1.2 Evaluación del modelo con mayor equilibrio y su mejor punto de corte

Además de trabajar con el modelo más preciso, también es interesante estudiar cómo funciona este modelo más simple y si esta simplicidad es beneficiosa en términos de sobreajuste.

El índice ROC de este modelo en *test* es de 0.7238, algo más bajo que su ROC en validación, que era de 0.7458. Además, en *test,* este ROC es incluso mayor que el del modelo con mejor capacidad predictiva. Así, el descenso de este índice es menos pronunciado que el del modelo más preciso. Además, su tasa de fallos con punto de corte 0.5 en *test* (21.65%) también se encuentra mucho más cerca de su equivalente en validación (21.58%), al contrario de lo que ocurría con el anterior modelo.

Al igual que en el anterior apartado, se valoran diferentes puntos de corte.

Tabla 23. Índices de precisión del modelo con mayor equilibrio

	Punto de corte	Tasa de fallos	Sensibilidad	Especificidad	Pos Pred
Punto corte 0.5	0.5	21.65%	21.84%	95.74%	61.21%

Igual sensibilidad y especificidad	0.221	33.53%	66.77%	66.38%	37.94%
Máximo Índice de Youden	0.285	26.43%	54.15%	79.55%	44.9%
Proporción del evento	0.2353	31.93%	64%	69.32%	39.1%

Igual que con el anterior modelo se busca un grado de sensibilidad aceptable sin dejar de lado la precisión. Con este modelo concreto, el punto de corte que mejor consigue eso es el que maximiza el índice de Youden. Con respecto al punto de corte de 0.5, la sensibilidad sube del 21.84% al 54.15%; resintiéndose la tasa de fallos mucho menos que el resto de los puntos de corte.

8. CONCLUSIONES

Al comienzo de este proyecto, se planteaba el objetivo principal de generar modelos de predicción que permitieran detectar qué sujetos se opondrían a ser vacunados y saber si existen características comunes a estos. A pesar de que sí se han obtenido modelos que reducen la tasa de fallos de las predicciones en *test*; estos modelos no son perfectos y están lejos de explicar completamente en el movimiento antivacunas.

Para este trabajo, se han generado más de 450 modelos (y más de 50 ensamblados); pero los resultados no han sido tan satisfactorios como se esperaba. Por más pruebas con los algoritmos más punteros que se han realizado, no se han conseguido bajadas en la tasa de fallos drásticas. Es más, el modelo más preciso ha resultado ser una regresión logística; probablemente, la metodología menos compleja y más clásica de todas las realizadas.

Como se señalaba en la introducción, en el estudio de Hornsey et al. (2021) ya se recalcaba lo difícil que resulta utilizar un único modelo teórico que consiga identificar cómo son aquellas personas que rechazan la vacuna; en gran medida porque determinaron que los sujetos antivacunas formaban parte de pequeños grupos o "*small pockets*" con diferentes características.

En el caso de este proyecto, la regresión logística sí permite detectar algunas tendencias. Por ejemplo, los votantes de VOX o aquellos sujetos que dan peor valoración al Presidente del Gobierno tienen una mayor probabilidad de rechazar la vacuna. Ocurre lo mismo con aquellas personas más preocupadas por la situación económica que la situación sanitaria.

No obstante, se debe tener en cuenta que los resultados de variación de probabilidad del evento que arroja la regresión logística no tienen en cuenta posibles correlaciones entre variables. No es capaz de detectar esos "*small pockets*" de los que hablaban Hornsey et al. Esto provoca que no se puedan tomar estos resultados como 100% precisos, pues en el análisis del efecto de cada una de las variables, se parte de la base de que el resto permanezcan constantes.

Sin embargo, sí hay tendencias y sí se detectan. Todo el trabajo realizado no es en vano, ya que más allá de aprender a trabajar con un conjunto de datos complejo y metodologías de predicción muy distintas, se ha generado un modelo que es capaz de

clasificar de forma correcta a 4 de cada 5 sujetos en *test.* Además, un modelo así permite obtener probabilidad de evento a cada persona con pocas y siempre relevantes variables *input*.

Para ilustrar esto, se toma al sujeto del conjunto *test* que tienen una probabilidad predicha del evento mayor y al sujeto con la menor probabilidad del evento. Sus características (algunas de las variables *input* que toma la regresión) se recogen en la Tabla 24. Se pueden ver las diferencias entre ambos sujetos de forma clara. Así, el entrevistado con mayor probabilidad de evento es una mujer joven votante de VOX poco preocupada por el COVID y que antepone la economía a la situación sanitaria. Por otro lado, el sujeto con menor probabilidad de oponerse a la vacuna es un hombre jubilado con buena situación económica y buena valoración del Presidente; que se preocupa más por la situación sanitaria que la económica.

Tabla 24. Características de los sujetos con mayor y menor probabilidad del evento

	En contra de la vacuna	**A favor de la vacuna**
Sexo	Mujer	Hombre
¿Catalán?	Sí	No
Edad	24	67
¿En paro?	No	No
¿Santiago Abascal como presidente?	Sí	No
COVID: ¿Economía o Salud?	Economía	Salud
¿Situación económica personal buena?	No	Sí
¿Preocupación por el COVID?	Poco	Mucho
¿Problema principal en España: COVID?	No	Sí
Valoración (1-10) de Pedro Sánchez	3	8
Probabilidad de evento	**0.9287**	**0.0118**
¿Opuesto a la vacuna?	**Sí**	**No**

Puede que las variables *input* de este conjunto de datos no sean las mejores, pero sí que contienen información, y mediante uno de los métodos de modelización más simples, se han conseguido determinar tendencias que permitirían conocer la probabilidad de que una persona se oponga a ser vacunado y se actúe en consecuencia.

9. BIBLIOGRAFÍA

ABC (2021, 25 de marzo). ¿Puedo rechazar la vacuna de Astrazeneca?. *ABC*. Recuperado el 25 de abril aquí.

Auschitzky, E., Hammer, M., & Rajagopaul, A. (2014). How big data can improve manufacturing. *McKinsey Consulting*. Recuperado el 26 de diciembre aquí.

CDC (2020). Diferentes vacunas contra el COVID-19. *CDC.* Recuperado el 26 de diciembre aquí.

Chan, M. S., Jamieson, K. H., & Albarracin, D. (2020). Prospective associations of regional social media messages with attitudes and actual vaccination: A big data and survey study of the influenza vaccine in the United States. *Vaccine*, *38*(40), 6236-6247. https://doi.org/10.1016/j.vaccine.2020.07.054

CIS (2020a). *Barómetro de diciembre de 2020.* Recuperado el 26 de diciembre aquí.

CIS (2020b). *Barómetro de enero de 2021.* Recuperado el 26 de diciembre aquí.

CIS (2021a). *Barómetros.* Recuperado el 26 de diciembre aquí.

CIS (2021b). *Quiénes somos.* Recuperado el 26 de diciembre aquí.

Guo, G.; Wang, H.; Bell, D.; Bi, Y. y Greer, K. (2003). KNN Model-Based Approach in Classification. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003.* Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62

Hornsey J. M.; Edwards M., Lobera J.; Díaz-Catalán, C. y Barlow, F. C. (2021). Resolving the small-pockets problem helps clarify the role of education and political ideology in shaping vaccine scepticism. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12500

IBM (2020). *A business guide into modern predictive analytics.* Recuperado el 28 de diciembre aquí.

Mavragani, A., & Ochoa, G. (2018). The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak. *Big Data and Cognitive Computing*, *2*(1), 2. https://doi.org/10.3390/bdcc2010002

OMS (2020). Brote de enfermedad por coronavirus. *WHO*. Recuperado el 26 de diciembre aquí.

Portela, J. (2021). *Apuntes de la asignatura Machine Learning*.

Raeven, R. H. M., van Riet, E., Meiring, H. D., Metz, B., & Kersten, G. F. A. (2019). Systems vaccinology and big data in the vaccine development chain. *Immunology*, *156*(1), 33-46. https://doi.org/10.1111/imm.13012

Sa, S. (2018). Big Data in Healthcare Management: A Review of Literature. *American Journal of Theoretical and Applied Business*, *4*(2), 57.

https://doi.org/10.11648/j.ajtab.20180402.14

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28. https://doi.org/10.1257/jep.28.2.3

ANEXOS

A. Cuestionarios.

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

«Información sujeta a secreto estadístico (Ley 12/89, de 9 de mayo, de la Función Estadística Pública) y al Reglamento General de Protección de Datos y la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales.» Plan Estadístico Nacional 2017-2020. RD 410/2016, 31 oct. y RD 1043/2017 de 22 dic.

«Buenos días/tardes, soy (nombre propio) y estoy realizando una encuesta telefónica para el Centro de Investigaciones Sociológicas (CIS) sobre temas de interés general. Por este motivo solicitamos su colaboración y se la agradecemos anticipadamente. Este teléfono ha sido obtenido al azar. Esta conversación será grabada para supervisar la calidad y después se borrará en un plazo inferior a un mes, le garantizamos el absoluto anonimato y secreto de sus respuestas en el más estricto cumplimiento de las leyes sobre secreto estadístico y protección de datos personales. Tras la realización de la encuesta su número de teléfono será disociado de las respuestas que pueda dar, que a su vez serán anonimizadas para que en ningún caso puedan ser asociadas a usted. Si desea conocer sus derechos de protección de datos y ampliar esta información puede consultar la página web www.cis.es ¿Ha comprendido la información leída? ¿Sería tan amable de contestar a unas preguntas, algunas de ellas sobre datos de carácter sensible como la intención de voto? No está obligado a contestar todas las preguntas. Muchas gracias.»

PC1. Pregunta contacto 1. ¿Me puede decir a qué provincia y municipio estoy llamando...?

[TIPO_TEL]
FIJO *1*
MÓVIL *2*

[CCAA]
_____

[PROVINCIA]
_____

[MUNICIPIO]
_____

[ENTREV]
_____

ENTREVISTADOR/A: SI LA PERSONA QUE CONTESTA ES DIFERENTE DE LA QUE COGIÓ EL TELÉFONO PRESENTARSE:

Buenos días/tardes, mi nombre es… y le llamo del Centro de Investigaciones Sociológicas porque estamos realizando una encuesta de opinión sobre temas de interés general. Dura de 12 a 15 minutos aproximadamente. ¿Sería tan amable de colaborar con nosotros?

[SEXO]
Hombre *1*
Mujer *2*

[EDADEXACTA]
_____

[EDAD]
de 18 a 24 ..... *1*
de 25 a 34 ..... *2*
de 35 a 44 ..... *3*
de 45 a 54 ..... *4*
de 55 a 64 ..... *5*
65 y más *6*

P.0 En primer lugar quisiera preguntarle si tiene Ud....

[P0]
La nacionalidad española *1*
La nacionalidad española y otra *2*
Otra nacionalidad *3*

Saltos:
Si P0=3 ir a fin cuestionario.

P.1 Me gustaría hacerle algunas preguntas sobre la crisis del coronavirus. Pensando en todos los efectos de esta pandemia, ¿diría Ud. que la crisis del coronavirus le preocupa mucho, bastante, poco o nada?

[P1]
Mucho *1*
Bastante *2*
Poco *3*
(NO LEER) Regular *4*
Nada *5*
N.S. *8*
N.C. *9*

P.2 En estos momentos, ¿qué le preocupa a Ud. más, los efectos de esta crisis sobre la salud, o los efectos de la crisis sobre la economía y el empleo?

[P2]
Los efectos sobre la salud *1*
Los efectos sobre la economía y el empleo *2*
(NO LEER) Ambos por igual *3*
(NO LEER) Ni unos ni otros *4*
N.S. *8*
N.C. *9*

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

P.3 Tal como está evolucionando la situación del coronavirus en España ¿cree Ud. que era necesario que el Parlamento español tomara medidas de control y aislamiento más exigentes, como se ha hecho, o que se podía continuar como se estaba? (NO LEER ÚNICAMENTE LAS SEÑALADAS CON NO LEER)

[P3]

- *Había que tomar medidas más exigentes 1*
- *Podíamos continuar como estábamos 2*
- *(NO LEER) Medidas dependiendo de cada zona 3*
- *(NO LEER) Medidas menos exigentes 4*
- *(NO LEER) Otras medidas anticipatorias, más eficaces y mejor planteadas 5*
- *(NO LEER) Medidas de prevención, más medios técnicos y humanos (haciendo más pruebas, más rastraedores/as...) 6*
- *(NO LEER) Depende de cada situación 7*
- *(NO LEER) Cumplimiento de las medidas, más control y penalizaciones 8*
- *(NO LEER) Mejor gestión y coordinación política 9*
- *(NO LEER) Medidas de concienciación ciudadana y responsabilidad 10*
- *(NO LEER) No tomar ninguna medida 11*
- *(NO LEER) Medidas tomadas por un equipo de expertos 12*
- *Otras respuestas 96*
- *(NO LEER) No lo sabe, duda 98*
- *N.C. 99*

P.4 En temas de salud, ¿ha tenido Ud. que contactar con los servicios sanitarios por pensar que tenía síntomas relacionados con el coronavirus?

[P4]

- *Sí 1*
- *No ..... 2.*
- *N.C. 9*

Filtros:
P4a si P4=1.

P.4a ¿A qué servicios recurrió Ud.? (RESPUESTA MÚLTIPLE)

[P4A]

- *A mi médico/a de atención primaria 1*
- *Al servicio de urgencias de atención primaria 2*
- *Al servicio de urgencias del hospital 3*
- *Al 061/112 ..... 4*
- *A un teléfono 900 que la comunidad puso para estos casos 5*
- *Otras respuestas (especificar) 6*
- *N.C. 9*

[P4A_COD_1]

_____

Filtros:
P4b si P4=1

P.4b ¿Cómo fue la atención que recibió? (LEER).

[P4B]

- *Muy buena 1*
- *Buena 2*
- *(NO LEER) Regular 3*
- *Mala 4*
- *Muy mala 5*
- *N.S. 8*
- *N.C. 9*

Filtros:
P4c si P4=1.

P.4c ¿Le indicaron que debía guardar medidas de aislamiento?

[P4C]

- *Sí 1*
- *No 2*
- *N.C. 9*

Filtros:
P4d si P4=1 y si P4c=1.

P.4d. ¿Le hicieron la prueba del coronavirus?

[P4D]

- *Sí 1*
- *No 2*
- *N.S. 8*
- *N.C. 9*

Filtros:
P4e si 1 en P4, 1 en P4c y 1 en P4d.

P.4e. ¿Y le diagnosticaron finalmente infección por coronavirus?

[P4E]

- *Sí 1*
- *No 2*
- *N.C. 9*

Filtros:
P4f si 1 en P4, 1 en P4c, 1 en P4d y 1 en P4e.

P.4f. ¿Y cómo evolucionó su enfermedad? (LEER).

[P4F]

- *Tuve síntomas leves y la pasé en casa 1*
- *Tuve síntomas importantes, pero la pasé en casa 2*
- *Tuve que ingresar en el hospital 3*
- *N.C. 9*

Filtros:
P4g si 1 en P4, 1 en P4c, 1 en P4d ,1 en P4e y 1 o 2 en P4f.

P.4g. ¿Quién le atendió sanitariamente durante la enfermedad? (RESPUESTA MÚLTIPLE, LEER SI ES PRECISO).

[P4G]

- *El/la médico/a de atención primaria 1*
- *La/el enfermera/o de atención primaria 2*
- *El 061/112 ..... 3*
- *Otros/as profesionales sanitarios/as 4*
- *No tuve seguimiento por ningún/a profesional sanitario/a 5*
- *N.C. 9*

Filtros:
P4h si 1 en P4, 1 en P4c, 1 en P4d ,1 en P4e, 1 o 2 en P4f y 1,2,3 o 4 en p4g.

P.4h Y esa atención se llevó a cabo de forma… (MÚLTIPLE) (LEER).

[P4H]

- *Telefónica 1*
- *A través de Internet 2*
- *Vinieron a visitarme a casa 3*
- *Acudí al centro de salud 4*
- *En algún momento acudí a urgencias del hospital 5*
- *Otras respuestas (especificar) 6*
- *N.C. 9*

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

[P4H_COD]
_____

Filtros:
P4i si 1 en P4, 1 en P4c, 1 en P4d ,1 en P4e, 1 o 2 en P4f y 1,2,3 o 4 en p4g.

P4i ¿Y cómo valora, en su conjunto, la atención que ha recibido? (LEER).

[P4I]

- *Muy buena* 1
- *Buena* 2
- *(NO LEER) Regular* 3
- *Mala* 4
- *Muy mala* 5
- *N.S.* 8
- *N.C.* 9

P.5. Después de estos últimos meses vividos con la pandemia del coronavirus, ¿cree Ud. que son convenientes reformas en la sanidad española?

[P5]

- *Sí* 1
- *No* 2
- *N.S.* 8
- *N.C.* 9

Filtros:
P5a si P5=1.

P.5a. ¿Qué reformas piensa Ud. que son necesarias…?

[P5A]

	Sí	*No*	*N.S.*	*N.C.*
Dedicar más recursos económicos	1	2	8	9
Aumentar las plantillas	1	2	8	9
Aumentar la coordinación entre las CC.AA.	1	2	8	9
Aumentar las instalaciones y los recursos dedicados a prevenir y abordar las pandemias	1	2	8	9

Filtros:
P5b si P5=1 y P5a=1.

P.5b.¿Y con qué grado de urgencia piensa Ud. que habría que abordar estas reformas: con mucha, bastante, regular, poca o apenas sin urgencia?

[P5B]

	Mucha urgencia	*Bastante urgencia*	*Regular*	*Poca urgencia*	*Apenas sin urgencia*	*N.S.*	*N.C.*
Dedicar más recursos económicos	1	2	3	4	5	8	9
Aumentar las plantillas	1	2	3	4	5	8	9
Aumentar la coordinación entre las CC.AA.	1	2	3	4	5	8	9
Aumentar las instalaciones y los recursos dedicados a prevenir y abordar las pandemias	1	2	3	4	5	8	9

P.6 ¿Estaría Ud. dispuesto/a a vacunarse inmediatamente cuando se tenga la vacuna?

[P6]

- *Sí* 1
- *No* 2
- *Sí, si tiene garantías, si está probada, si es fiable* 3
- *Sí, según el origen de la vacuna* 4
- *Sí, si hay información suficiente* 5
- *Sí, por consejo de autoridades, científicos/as, o sanitarios/as* 6
- *Otras respuestas* 7
- *(NO LEER) No sabe, duda* 8
- *N.C.* 9

P.7 ¿Quién le gustaría a Ud. que se hiciera cargo de la lucha contra la pandemia primordialmente: el Gobierno de España, el Gobierno de las comunidades autónomas, o ambos en colaboración?

[P7]

- *El Gobierno de España* 1
- *El Gobierno de las comunidades autónomas* 2
- *Ambos en colaboración* 3
- *Otras respuestas (especificar)* 6
- *(NO LEER) No lo sabe, duda* 8
- *N.C.* 9

[P7_COD]
_____

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

P.8 Por lo que Ud. está viendo, ¿cree que la mayoría de los/as españoles/as estamos dando un ejemplo de civismo y solidaridad en la forma de afrontar las medidas contra la COVID-19, o piensa que la mayoría está siendo poco cívica e indisciplinada?

[P8]

- *Cree que la mayoría está reaccionando con civismo y solidaridad 1*
- *Cree que la mayoría está siendo poco cívica e indisciplinada 2*
- *(NO LEER) No lo sabe, duda 8*
- *N.C. 9*

P.9 Considerando lo que está ocurriendo con la pandemia, ¿todo lo que sucede le está afectando a Ud. mucho, bastante, algo, nada o casi nada en su vida personal?

P.10 ¿Y le está afectando a Ud. mucho, bastante, algo o nada o casi nada en su vida social y de relaciones?

P9. Vida personal
P10. Vida social

[P9]

- *Me está afectando mucho 1*
- *Me está afectando bastante 2*
- *(NO LEER) Regular 3*
- *Me está afectando algo 4*
- *No me está afectando nada o casi nada 6*
- *N.S. 8*
- *N.C. 9*

[P10]

- *Me está afectando mucho 1*
- *Me está afectando bastante 2*
- *(NO LEER) Regular 3*
- *Me está afectando algo 4*
- *No me está afectando nada o casi nada 6*
- *N.S. 8*
- *N.C. 9*

Filtros:
P11a si 1, 2, 3 o 4 en P9

P.11A ¿En qué aspecto o aspectos le está afectando a Ud. en su vida personal? (REPUESTA ESPONTÁNEA).

[P11A]

- *Por el distanciamiento con los seres queridos 1*
- *Por el trabajo y/o economía personal 2*
- *Por el estado anímico negativo (ansiedad, tristeza...) 3*
- *Las restricciones y libertad de movimientos 4*
- *Por miedo al contagio suyo o de otras personas 5*
- *Por el cambio de las condiciones laborales 6*
- *Por el aislamiento y el confinamiento 7*
- *Por los cambios en la vida cotidiana 8*
- *Por pérdida de trabajo (despidos, cierres, no encontrar...) 9*
- *Por los cambios de rutina en clases 10*
- *Otras respuestas 97*
- *N.S. 98*
- *N.C. 99*

Filtros:
P11b si 1, 2, 3 o 4 en P10

P11.B ¿En qué aspecto o aspectos le está afectando en su vida social? (RESPUESTA ESPONTÁNEA).

[P11B]

- *Por la distancia respecto a la gente 1*
- *Por la distancia respecto a los amigos/as 2*
- *Por el asislamiento, confinamiento y no poder salir 3*
- *Por la distancia respecto a la familia 4*
- *Por el cese o limitación de actividades culturales, de ocio y deportivas 5*
- *Por el miedo a que se produzcan contagios en las relaciones sociales y familiares 6*
- *Aspectos económicos, laborales y/o profesionales 7*
- *Aspectos psicológicos, anímicos y emocionales 8*
- *Por el cese o limitación de actividades de hostelería y restauración 9*
- *Por la ausencia de contacto físico en las relaciones sociales, pérdida de calidad en las relaciones 10*
- *Otras respuestas 97*
- *N.S 98*
- *N.C 99*

P.12 Al margen de las medidas oficiales para el control de la COVID tomadas por las autoridades desde que estamos bajo la pandemia, ¿qué medidas propias de control ha estado siguiendo Ud.?

[P12]

- *Ninguna, hace vida normal 1*
- *Tiene cuidado con las cosas que toca o por dónde va, pero en lo demás hace vida normal 2*
- *Permanece prácticamente en aislamiento, saliendo de casa sólo para adquirir alimentos y para ir a consultas médicas 3*
- *No sale de casa para nada que no resulte imprescindible y le traen los suministros y medicinas 4*
- *Otras respuestas (especificar) 6*
- *N.S., duda. 8*
- *N.C. 9*

[P12_COD]

_____

P.13 Refiriéndonos a la situación económica general de España actualmente, ¿cómo la calificaría Ud.: muy buena, buena, mala o muy mala?

[P13]

- *Muy buena 1*
- *Buena 2*
- *(NO LEER) Regular 3*
- *Mala 4*
- *Muy mala 5*
- *N.S. 8*
- *N.C. 9*

P.14 ¿Cómo calificaría Ud. su situación económica personal en la actualidad: muy buena, buena, mala o muy mala?

[P14]

- *Muy buena 1*
- *Buena 2*
- *(NO LEER) Regular 3*
- *Mala 4*
- *Muy mala 5*
- *N.S. 8*
- *N.C. 9*

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

P.15 ¿Cuál es, a su juicio, el principal problema que existe actualmente en España? ¿Y el segundo? ¿Y el tercero? (ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O COMPLETAR "OTRO, ¿CUÁL?").

1er lugar

[P15_1]

Los peligros para la salud: COVID-19	*53*
La crisis económica, los problemas de índole económica	*8*
El paro	*1*
Los problemas políticos en general	*51*
Lo que hacen los partidos políticos	*50*
El mal comportamiento de los/as políticos/as	*13*
La corrupción y el fraude	*11*
La sanidad	*6*
Los problemas de índole social	*16*
La inmigración	*18*
Los problemas relacionados con la calidad del empleo	*9*
La independencia de Cataluña	*45*
Las pensiones	*12*
La violencia de género	*19*
La falta de acuerdos, unidad y capacidad de colaboración.	*46*
Poca conciencia ciudadana (falta de civismo, de sentido espíritu cívico)	*55*
Otro	*96*
Ninguno	*97*
N.S.	*98*
N.C.	*99*

[P151_OTRO]

_____

2º lugar

[P15_2]

Los peligros para la salud: COVID-19	*53*
La crisis económica, los problemas de índole económica	*8*
El paro	*1*
Los problemas políticos en general	*51*
Lo que hacen los partidos políticos	*50*
El mal comportamiento de los/as políticos/as	*13*
La corrupción y el fraude	*11*
La sanidad	*6*
Los problemas de índole social	*16*
La inmigración	*18*
Los problemas relacionados con la calidad del empleo	*9*
La independencia de Cataluña	*45*
Las pensiones	*12*
La violencia de género	*19*
La falta de acuerdos, unidad y capacidad de colaboración.	*46*
Poca conciencia ciudadana (falta de civismo, de sentido espíritu cívico)	*55*
Otro	*96*
Ninguno	*97*
N.S.	*98*
N.C.	*99*

[P152_OTRO]

_____

3er lugar

[P15_3]

Los peligros para la salud: COVID-19	*53*
La crisis económica, los problemas de índole económica	*8*
El paro	*1*
Los problemas políticos en general	*51*
Lo que hacen los partidos políticos	*50*
El mal comportamiento de los/as políticos/as	*13*
La corrupción y el fraude	*11*
La sanidad	*6*
Los problemas de índole social	*16*
La inmigración	*18*
Los problemas relacionados con la calidad del empleo	*9*
La independencia de Cataluña	*45*
Las pensiones	*12*
La violencia de género	*19*
La falta de acuerdos, unidad y capacidad de colaboración.	*46*
Poca conciencia ciudadana (falta de civismo, de sentido espíritu cívico)	*55*
Otro	*96*
Ninguno	*97*
N.S.	*98*
N.C.	*99*

[P153_OTRO]

_____

P.16 ¿Y cuál es el problema que a Ud., personalmente, le afecta más? ¿Y el segundo? ¿Y el tercero? (ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O COMPLETAR "OTRO, ¿CUÁL?").

1er lugar

[P16_1]

Los peligros para la salud: COVID-19	*53*
La crisis económica, los problemas de índole económica	*8*
Avituallamiento de víveres en el hogar	*60*
Tener que estar enclaustrado/a en casa	*59*
El paro	*1*
La sanidad	*6*
Las pensiones	*12*
Los problemas relacionados con la calidad del empleo	*9*
Los problemas políticos en general	*51*
Lo que hacen los partidos políticos	*50*
El mal comportamiento de los/as políticos/as	*13*
La educación	*22*
La corrupción y el fraude	*11*
Los problemas de índole social	*16*
Las preocupaciones y situaciones personales	*29*
La vivienda	*7*
La independencia de Cataluña	*45*
Los problemas relacionados con la juventud	*20*
Otro	*96*
Ninguno	*97*
N.S.	*98*
N.C.	*99*

[P161_OTRO]

_____

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

2º lugar
[P16_2]

- *Los peligros para la salud: COVID-19 ... 53*
- *La crisis económica, los problemas de índole económica ... 8*
- *Avituallamiento de víveres en el hogar ... 60*
- *Tener que estar enclaustrado/a en casa ... 59*
- *El paro ... 1*
- *La sanidad ... 6*
- *Las pensiones ... 12*
- *Los problemas relacionados con la calidad del empleo ... 9*
- *Los problemas políticos en general ... 51*
- *Lo que hacen los partidos políticos ... 50*
- *El mal comportamiento de los/as políticos/as ... 13*
- *La educación ... 22*
- *La corrupción y el fraude ... 11*
- *Los problemas de índole social ... 16*
- *Las preocupaciones y situaciones personales ... 29*
- *La vivienda ... 7*
- *La independencia de Cataluña ... 45*
- *Los problemas relacionados con la juventud ... 20*
- *Otro ... 96*
- *Ninguno ... 97*
- *N.S. ... 98*
- *N.C. ... 99*

[P162_OTRO]
_____

3er lugar
[P16_3]

- *Los peligros para la salud: COVID-19 ... 53*
- *La crisis económica, los problemas de índole económica ... 8*
- *Avituallamiento de víveres en el hogar ... 60*
- *Tener que estar enclaustrado/a en casa ... 59*
- *El paro ... 1*
- *La sanidad ... 6*
- *Las pensiones ... 12*
- *Los problemas relacionados con la calidad del empleo ... 9*
- *Los problemas políticos en general ... 51*
- *Lo que hacen los partidos políticos ... 50*
- *El mal comportamiento de los/as políticos/as ... 13*
- *La educación ... 22*
- *La corrupción y el fraude ... 11*
- *Los problemas de índole social ... 16*
- *Las preocupaciones y situaciones personales ... 29*
- *La vivienda ... 7*
- *La independencia de Cataluña ... 45*
- *Los problemas relacionados con la juventud ... 20*
- *Otro ... 96*
- *Ninguno ... 97*
- *N.S. ... 98*
- *N.C. ... 99*

[P163_OTRO]
_____

P.17 A continuación, voy a mencionarle los nombres de algunos/as líderes políticos/as y nos gustaría que, en relación a lo que cada uno/a está diciendo y haciendo sobre la COVID-19 en estos momentos, lo/a califique de 1 a 10, siendo el 1 "muy mal" y el 10 "muy bien". (SI NO CONOCE ALGUNO/A ME LO DICE).

[P17]

	NO CONOCE	Muy mal 1	2	3	4	5	6	7	8	9	Muy bien 10	N.S.	N.C.
Pedro Sánchez	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Casado	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Iglesias	97	1	2	3	4	5	6	7	8	9	10	98	99
Santiago Abascal	97	1	2	3	4	5	6	7	8	9	10	98	99
Inés Arrimadas	97	1	2	3	4	5	6	7	8	9	10	98	99

P.18 El presidente del Gobierno, Pedro Sánchez, ¿le inspira a Ud., personalmente, mucha confianza, bastante confianza, poca o ninguna confianza?

[P18]

- *Mucha confianza ... 1*
- *Bastante confianza ... 2*
- *Poca confianza ... 3*
- *Ninguna confianza ... 4*
- *N.S. ... 8*
- *N.C. ... 9*

P.19 ¿Y el líder del principal partido de la oposición, Pablo Casado, le inspira, personalmente, mucha confianza, bastante confianza, poca o ninguna confianza?

[P19]

- *Mucha confianza ... 1*
- *Bastante confianza ... 2*
- *Poca confianza ... 3*
- *Ninguna confianza ... 4*
- *N.S. ... 8*
- *N.C. ... 9*

P.20 De los/as principales líderes políticos/as, ¿quién preferiría que fuese el/la presidente/a del Gobierno en estos momentos? (SOLO UN NOMBRE).

[P20]

- *Pedro Sánchez ... 1*
- *Pablo Casado ... 2*
- *Santiago Abascal ... 3*
- *Pablo Iglesias ... 4*
- *Alberto Garzón ... 5*
- *Inés Arrimadas ... 6*
- *Íñigo Errejón ... 7*
- *(NO LEER) Ninguno/a de ellos/as ... 97*
- *(NO LEER) Otro/a, ¿quién? ... 96*
- *N.S. ... 98*
- *N.C. ... 99*

[P20_COD]
_____

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

P.21 Suponiendo que mañana se celebrasen nuevamente elecciones generales, es decir, al Parlamento español, ¿a qué partido votaría Ud.? (RESPUESTA ESPONTÁNEA).

[INTENCIONG]

- *PSOE (Partido Socialista Obrero Español)* 2
- *PP (Partido Popular)* 1
- *VOX* 18
- *Podemos* 3
- *IU (Izquierda Unida)* 5
- *Unidas Podemos* 21
- *En Comú Podem* 6
- *En Común - Unidas Podemos* 67
- *Ciudadanos* 4
- *Más País* 50
- *ERC (Esquerra Republicana de Catalunya)* 8
- *JxCat (Junts per Catalunya)* 9
- *CUP* 19
- *EAJ-PNV (Partido Nacionalista Vasco)* 11
- *EH Bildu (Euskal Herria Bildu)* 12
- *CC-PNC (Coalición Canaria – Partido Nacionalista Canario)* 13
- *Nueva Canarias* 16
- *UPN (Unión del Pueblo Navarro)* 14
- *Compromís* 7
- *BNG (Bloque Nacionalista Galego)* 24
- *PRC (Partido Regionalista de Cantabria)* 43
- *Teruel Existe* 68
- *PACMA (Partido Animalista)* 17
- *FAC (Foro Asturias)* 15
- *Otro partido, ¿cuál?* 95
- *Voto nulo* 77
- *En blanco* 96
- *No votaría* 97
- *No sabe todavía* 98
- *N.C.* 99

[INTENCIONG_COD]

Filtros:
P21a si P21 no es 77, 96, 97, 98 o 99.

P.21a En el caso de que por cualquier razón finalmente no votase por el partido que me ha dicho, ¿a qué otro partido votaría Ud.? (RESPUESTA ESPONTÁNEA).

[INTENCIONGALTER]

- *PSOE (Partido Socialista Obrero Español)* 2
- *PP (Partido Popular)* 1
- *VOX* 18
- *Podemos* 3
- *IU (Izquierda Unida)* 5
- *Unidas Podemos* 21
- *En Comú Podem* 6
- *En Común - Unidas Podemos* 67
- *Ciudadanos* 4
- *Más País* 50
- *ERC (Esquerra Republicana de Catalunya)* 8
- *JxCat (Junts per Catalunya)* 9
- *CUP* 19
- *EAJ-PNV (Partido Nacionalista Vasco)* 11
- *EH Bildu (Euskal Herria Bildu)* 12
- *CC-PNC (Coalición Canaria – Partido Nacionalista Canario)* 13
- *Nueva Canarias* 16
- *UPN (Unión del Pueblo Navarro)* 14
- *Compromís* 7
- *BNG (Bloque Nacionalista Galego)* 24
- *PRC (Partido Regionalista de Cantabria)* 43
- *Teruel Existe* 68
- *PACMA (Partido Animalista)* 17
- *FAC (Foro Asturias)* 15
- *(NO LEER) No votaría a ningún otro partido (cita el mismo partido)* 93
- *Otro partido, ¿cuál?* 95
- *En blanco* 96
- *No votaría* 97
- *N.S.* 98
- *N.C.* 99

[INTENCIONGALTER_COD]

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

Filtros:
P22 si P21=77, 96, 97, 98 o 99.

P.22 Sin ningún compromiso por su parte, ¿me podría decir por qué partido siente Ud. más simpatía? (RESPUESTA ESPONTÁNEA).

[SIMPATIA]

- *PSOE (Partido Socialista Obrero Español)* *2*
- *PP (Partido Popular)* *1*
- *VOX* *18*
- *Podemos* *3*
- *IU (Izquierda Unida)* *5*
- *Unidas Podemos* *21*
- *En Comú Podem* *6*
- *En Común - Unidas Podemos* *67*
- *Ciudadanos* *4*
- *Más País* *50*
- *ERC (Esquerra Republicana de Catalunya)* *8*
- *JxCat (Junts per Catalunya)* *9*
- *CUP* *19*
- *EAJ-PNV (Partido Nacionalista Vasco)* *11*
- *EH Bildu (Euskal Herria Bildu)* *12*
- *CC-PNC (Coalición Canaria – Partido Nacionalista Canario)* *13*
- *Nueva Canarias* *16*
- *UPN (Unión del Pueblo Navarro)* *14*
- *Compromís* *7*
- *BNG (Bloque Nacionalista Galego)* *24*
- *PRC (Partido Regionalista de Cantabria)* *43*
- *Teruel Existe* *68*
- *PACMA (Partido Animalista)* *17*
- *FAC (Foro Asturias)* *15*
- *Otro partido, ¿cuál?* *95*
- *Ninguno* *97*
- *N.S.* *98*
- *N.C.* *99*

[SIMPATIA_COD]

_____

P.23 Cuando se habla de política se utilizan normalmente las expresiones izquierda y derecha. Situándonos en una escala de 10 casillas, como un termómetro, que van del 1 al 10, en la que 1 significa "lo más a la izquierda" y 10 "lo más a la derecha", ¿en qué casilla se colocaría Ud.?

[ESCIDEOL]

- *1 Izda.* *1*
- *2* *2*
- *3* *3*
- *4* *4*
- *5* *5*
- *6* *6*
- *7* *7*
- *8* *8*
- *9* *9*
- *10 Dcha.* *10*
- *N.S.* *98*
- *N.C.* *99*

P.24 Y siguiendo este mismo criterio, a los/as líderes de los principales partidos políticos, ¿dónde los/as ubicaría? (ENTREVISTADOR/A: PEDIR A LA PERSONA ENTREVISTADA QUE INDIQUE LA CASILLA EN LA QUE COLOCARÍA CADA LÍDER).

[ESCIDEOLPOLI]

	NO CONOCE	*Izquierda 1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Derecha 10*	*N.S.*	*N.C.*
Pedro Sánchez	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Casado	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Iglesias	97	1	2	3	4	5	6	7	8	9	10	98	99
Santiago Abascal	97	1	2	3	4	5	6	7	8	9	10	98	99
Inés Arrimadas	97	1	2	3	4	5	6	7	8	9	10	98	99

P.25 ¿Y qué valoración le merece cada uno/a de los/as siguientes políticos/as? Puntúelos/as de 1 a 10, sabiendo que el 1 significa que lo/a valora "muy mal" y el 10 que lo/a valora "muy bien".

[P25]

	NO CONOCE	*Muy mal 1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Muy bien 10*	*N.S.*	*N.C.*
Pedro Sánchez	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Casado	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Iglesias	97	1	2	3	4	5	6	7	8	9	10	98	99
Santiago Abascal	97	1	2	3	4	5	6	7	8	9	10	98	99
Inés Arrimadas	97	1	2	3	4	5	6	7	8	9	10	98	99

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

P.26 ¿Me podría decir si en las elecciones generales del 10 de noviembre de 2019...? (LEER RESPUESTAS).

[PARTICIPACIONG]

- *Fue a votar y votó 1*
- *Votó por correo 7*
- *No tenía edad para votar 2*
- *Fue a votar pero no pudo hacerlo 3*
- *No fue a votar porque no pudo 4*
- *Prefirió no votar 5*
- *No tenía derecho a voto 6*
- *No recuerda 8*
- *N.C. 9*

P.26a ¿Y podría decirme a qué partido o coalición votó? (RESPUESTA ESPONTÁNEA).

Filtros:
P26a si P26=1 o 7

[RECUVOTOG]

- *PSOE 2*
- *PP 1*
- *VOX 18*
- *Unidas Podemos 21*
- *En Comú Podem 6*
- *En Común - Unidas Podemos 67*
- *Ciudadanos 4*
- *Más País 50*
- *ERC 8*
- *JxCat 9*
- *CUP 19*
- *EAJ-PNV 11*
- *EH Bildu 12*
- *CCa-PNC-NC 13*
- *Navarra Suma (UPN) 14*
- *Més Compromís 7*
- *BNG (Bloque Nacionalista Galego) 24*
- *PRC (Partido Regionalista de Cantabria) 43*
- *Teruel Existe 68*
- *PACMA (Partido Animalista) 17*
- *Otros partidos 95*
- *En blanco 96*
- *Voto nulo 77*
- *No recuerda 98*
- *N.C. 99*

P.27 En todo caso, ¿qué partido considera más cercano a sus propias ideas? (RESPUESTA ESPONTÁNEA).

[CERCANIA]

- *PSOE (Partido Socialista Obrero Español) 2*
- *PP (Partido Popular) 1*
- *VOX 18*
- *Podemos 3*
- *IU (Izquierda Unida) 5*
- *Unidas Podemos 21*
- *En Comú Podem 6*
- *En Común-Unidas Podemos 67*
- *Ciudadanos 4*
- *Más País 50*
- *ERC (Esquerra Republicana de Catalunya) 8*
- *JxCat (Junts Per Catalunya) 9*
- *CUP 19*
- *EAJ-PNV (Partido Nacionalista Vasco) 11*
- *EH Bildu (Euskal Herria Bildu) 12*
- *CC-PNC (Coalición Canaria – Partido Nacionalista Canario) 13*
- *Nueva Canarias 16*
- *UPN (Unión del Pueblo Navarro) 14*
- *Compromís 7*
- *BNG (Bloque Nacionalista Galego) 24*
- *PRC (Partido Regionalista de Cantabria) 43*
- *Teruel Existe 68*
- *PACMA (Partido Animalista) 17*
- *FAC (Foro Asturias) 15*
- *Otro partido, ¿cuál? 95*
- *Ninguno 97*
- *N.S. 98*
- *N.C. 99*

[CERCANIA_COD]

PREGUNTAS SOLO PARA CATALUÑA

CX.1 Por favor, ¿cómo calificaría la situación económica de Cataluña, muy buena, buena, regular, mala o muy mala?

[CX1]

- *Muy buena 1*
- *Buena 2*
- *Regular 3*
- *Mala 4*
- *Muy mala 5*
- *N.S. 8*
- *N.C. 9*

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

CX.2 ¿Cuál es, a su juicio, el principal problema que tiene Cataluña en la actualidad? ¿Y el segundo? (ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O COMPLETAR "OTRO, ¿CUÁL?").

1er lugar

[CX2_1]

- *Los peligros para la salud: COVID-19 ..... 53*
- *La crisis económica, los problemas de índole económica 8*
- *El paro 1*
- *Los problemas políticos en general 51*
- *Lo que hacen los partidos políticos 50*
- *El mal comportamiento de los/as políticos/as 13*
- *La corrupción y el fraude 11*
- *La sanidad 6*
- *Los problemas de índole social 16*
- *La inmigración 18*
- *Los problemas relacionados con la calidad del empleo 9*
- *La independencia de Cataluña 45*
- *Las pensiones 12*
- *La violencia de género 19*
- *La falta de acuerdos, unidad y capacidad de colaboración. 46*
- *Poca conciencia ciudadana (falta de civismo, de sentido espíritu cívico) 55*
- *Otro 96*
- *Ninguno 97*
- *N.S. 98*
- *N.C. 99*

[CX21_OTRO]

_____

2º lugar

[CX2_2]

- *Los peligros para la salud: COVID-19 ..... 53*
- *La crisis económica, los problemas de índole económica 8*
- *El paro 1*
- *Los problemas políticos en general 51*
- *Lo que hacen los partidos políticos 50*
- *El mal comportamiento de los/as políticos/as 13*
- *La corrupción y el fraude 11*
- *La sanidad 6*
- *Los problemas de índole social 16*
- *La inmigración 18*
- *Los problemas relacionados con la calidad del empleo 9*
- *La independencia de Cataluña 45*
- *Las pensiones 12*
- *La violencia de género 19*
- *La falta de acuerdos, unidad y capacidad de colaboración. 46*
- *Poca conciencia ciudadana (falta de civismo, de sentido espíritu cívico) 55*
- *Otro 96*
- *Ninguno 97*
- *N.S. 98*
- *N.C. 99*

[CX22_OTRO]

_____

CX.3 En términos generales, ¿cómo calificaría Ud. la gestión que ha realizado en estos últimos años la Generalitat de Cataluña, muy buena, buena, regular, mala o muy mala?

[CX3]

- *Muy buena 1*
- *Buena 2*
- *Regular 3*
- *Mala 4*
- *Muy mala 5*
- *N.S. 8*
- *N.C. 9*

CX.4 ¿Cómo calificaría Ud. la gestión realizada por Pere Aragonés como president en funciones de la Generalitat desde octubre de 2020: muy buena, buena, regular, mala o muy mala?

[CX4]

- *Muy buena 1*
- *Buena 2*
- *Regular 3*
- *Mala 4*
- *Muy mala 5*
- *N.S. 8*
- *N.C. 9*

CX.5 De los siguientes partidos que voy a mencionarle, dígame por favor, ¿cuál es a su juicio el que en la actualidad...?

[CX5]

	PSC	*PP*	*Ciudadanos*	*ERC*	*JxCat*	*CUP*	*En Comú Podem*	*PDeCAT*	*PNC*	*(NO LEER) Ninguno*	*N.S.*	*N.C.*
Mejor defiende los intereses de Cataluña	1	2	3	4	5	6	7	8	9	97	98	99
Mejor representa las ideas de la gente como Ud.	1	2	3	4	5	6	7	8	9	97	98	99
Le inspira más confianza	1	2	3	4	5	6	7	8	9	97	98	99
Está más unido	1	2	3	4	5	6	7	8	9	97	98	99
Tiene mejores líderes en Cataluña	1	2	3	4	5	6	7	8	9	97	98	99
Está más capacitado para gobernar Cataluña	1	2	3	4	5	6	7	8	9	97	98	99
Tiene más capacidad para llegar a acuerdos con el Gobierno de España	1	2	3	4	5	6	7	8	9	97	98	99

CX.6 Suponiendo que mañana se celebrasen elecciones al Parlamento catalán, ¿a qué partido votaría Ud.? (RESPUESTA ESPONTÁNEA).

[INTENCIONCAT]

- *JxCat 9*
- *ERC (Esquerra Republicana de Catalunya) 8*
- *PSC (Partido Socialista de Cataluña) 2*
- *En Comú Podem 6*
- *CUP 19*
- *Ciudadanos 4*
- *PP 1*
- *VOX 18*
- *PDeCAT 20*
- *PNC (Partido Nacionalista Catalán) 21*
- *Otro partido, ¿cuál? 95*
- *Voto nulo 77*
- *En blanco 96*
- *No votaría 97*
- *No sabe todavía 98*
- *N.C. 99*

[INTENCIONCAT_COD]

Filtros:
CX7 si CX6=77, 96, 97, 98 o 99.

CX.7 Sin ningún compromiso por su parte, ¿me podría decir por qué partido siente Ud. más simpatía? (RESPUESTA ESPONTÁNEA).

[SIMPATIACAT]

- *JxCat 9*
- *ERC (Esquerra Republicana de Catalunya) 8*
- *PSC (Partido Socialista de Cataluña) 2*
- *En Comú Podem 6*
- *CUP 19*
- *Ciudadanos 4*
- *PP 1*
- *VOX 18*
- *PDeCAT 71*
- *PNC (Partido Nacionalista Catalán) 72*
- *Otro partido, ¿cuál? 95*
- *Ninguno 97*
- *N.S. 98*
- *N.C. 99*

[SIMPATIACAT_COD]

CX.8 En todo caso, ¿qué partido de Cataluña considera más cercano a sus propias ideas? (RESPUESTA ESPONTÁNEA).

[CERCANIACAT]

- *JxCat 9*
- *ERC (Esquerra Republicana de Catalunya) 8*
- *PSC (Partido Socialista de Cataluña) 2*
- *En Comú Podem 6*
- *CUP 19*
- *Ciudadanos 4*
- *PP 1*
- *VOX 18*
- *PDeCAT 71*
- *PNC (Partido Nacionalista Catalán) 72*
- *Otro partido, ¿cuál? 95*
- *Ninguno 97*
- *N.S. 98*
- *N.C. 99*

[CERCANIACAT_COD]

CX.9 ¿Y qué partido preferiría que gobernase en Cataluña? (RESPUESTA ESPONTÁNEA).

[PREPARTIGOBCAT]

- *JxCat 9*
- *ERC (Esquerra Republicana de Catalunya) 8*
- *PSC (Partido Socialista de Cataluña) 2*
- *En Comú Podem 6*
- *CUP 19*
- *Ciudadanos 4*
- *PP 1*
- *VOX 18*
- *PDeCAT 71*
- *PNC (Partido Nacionalista Catalán) 72*
- *Otro partido, ¿cuál? 95*
- *Ninguno 97*
- *N.S. 98*
- *N.C. 99*

[PREPARTIGOBCAT_COD]

CX.10 Le agradecería que me indicara si conoce a cada uno/a de los/as siguientes líderes políticos/as y qué valoración le merece. Puntúelos/as de 1 a 10, sabiendo que el 1 significa que lo/a valora "muy mal" y el 10 que lo/a valora "muy bien".

[CX10]

	NO CONOCE	*Muy mal 1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Muy bien 10*	*N.S.*	*N.C.*
Carles Puigdemont	97	1	2	3	4	5	6	7	8	9	10	98	99
Oriol Junqueras	97	1	2	3	4	5	6	7	8	9	10	98	99
Quim Torra	97	1	2	3	4	5	6	7	8	9	10	98	99
Miquel Iceta	97	1	2	3	4	5	6	7	8	9	10	98	99
Marta Rovira	97	1	2	3	4	5	6	7	8	9	10	98	99
Salvador Illa	97	1	2	3	4	5	6	7	8	9	10	98	99
Carlos Carrizosa	97	1	2	3	4	5	6	7	8	9	10	98	99
Lorena Roldán	97	1	2	3	4	5	6	7	8	9	10	98	99
Carles Riera	97	1	2	3	4	5	6	7	8	9	10	98	99
Jaume Asens	97	1	2	3	4	5	6	7	8	9	10	98	99
Jèssica Albiach	97	1	2	3	4	5	6	7	8	9	10	98	99
Alejandro Fernández	97	1	2	3	4	5	6	7	8	9	10	98	99

CX.11 De los/as líderes políticos/as que le he mencionado, ¿quién preferiría que fuese el/la president/a de la Generalitat de Cataluña en estos momentos?

[CX11]

- *Carles Puigdemont* 1
- *Oriol Junqueras* 2
- *Quim Torra* 3
- *Miquel Iceta* 4
- *Marta Rovira* 5
- *Salvador Illa* 6
- *Carlos Carrizosa* 7
- *Lorena Roldán* 8
- *Carles Riera* 9
- *Jaume Asens* 10
- *Jèssica Albiach* 11
- *Alejandro Fernández* 12
- *(NO LEER) Ninguno/a de ellos/as* 97
- *(NO LEER) Otro/a, ¿quién?* 96
- *N.S.* 98
- *N.C.* 99

[CX11_COD]
_____

PREGUNTAS PARA EL TOTAL NACIONAL

P.28 ¿Ha ido Ud. a la escuela o cursado algún tipo de estudios? (ENTREVISTADOR/A: en caso negativo, preguntar si sabe leer y escribir).

[ESCUELA]

- *No, es analfabeto/a* 1
- *No, pero sabe leer y escribir* 2
- *Sí, ha ido a la escuela* 3
- *N.C.* 9

Filtros:
P28a si P28=3

P.28a ¿Cuáles son los estudios de más alto nivel oficial que Ud. ha cursado (con independencia de que los haya terminado o no)? Por favor, especifique lo más posible, diciéndome el curso en que estaba cuando los terminó (o los interrumpió) y también el nombre que tenían entonces esos estudios (ej.: 3 años de estudios primarios, primaria, 5º de bachillerato, Maestría Industrial, preuniversitario, 4º de EGB, licenciatura, doctorado, FP1, etc.). (ENTREVISTADOR/A: si aún está estudiando, anotar el último curso que haya completado y el ciclo correcto en las opciones de respuesta. Si no ha completado la primaria, anotar nº de años que asistió a la escuela, diferenciando entre menos de 5 y más de 5).

[CURSOENTREV]
CURSO _____ *N.S. - N.R. = 98 N.C. = 99*

[NOMBREESTENTREV]
NOMBRE DE ESTUDIOS _____ *N.S. - N.R. = 98 N.C. = 99*

[NIVELESTENTREV]

- *01.Menos de 5 años de escolarización* 1
- *02.Educación primaria (Educación primaria de LOGSE, 5º Curso de EGB, Enseñanza primaria antigua)* 2
- *03.Cualificación profesional grado inicial (FP grado inicial). PCPI (Programas de Cualificación Profesional Inicial, que no precisan de titulación académica de la primera etapa de secundaria para su realización). Programas de garantía social* 3
- *04.Educación secundaria (ESO, EGB. Graduado Escolar. Certificado de Escolaridad, Bachillerato Elemental)* 4
- *05.FP de grado medio (Ciclo/módulo formativo de FP (grado medio), de Artes Plásticas y Diseño, Música y danza, Enseñanzas deportivas, FP I, Bachiller laboral elemental. Oficialía Industrial; Bachillerato Comercial).* 5
- *06.Bachillerato (Bachillerato LOGSE, BUP, Bachillerato superior (6º), Bachillerato universitario (7º), Incluidos COU y PREU)* 6
- *07.FP de grado superior (Ciclo/módulo formativo de FP (grado superior) de Artes Plásticas, Diseño, Música y danza, Deporte, FP II, Bach. Laboral Sup., Maestría industrial, Perito Mercantil; Secretariado de 2º grado; Grado Medio conservatorio)* 7
- *08.Arquitectura-ingeniería técnica (Arquitectura/ingeniería técnica, aparejador/a; peritos/a)* 8
- *09.Diplomatura (ATENCIÓN: solo Diplomaturas oficiales, no codificar aquí los tres primeros años de una licenciatura o grado con mayor duración)* 9
- *10.Grado (Estudios de grado, Enseñanzas Artísticas equivalentes (desde 2006))* 10
- *11.Licenciatura (Titulaciones con equivalencia oficial: 2º ciclo INEF; Danza y arte dramático (desde 1992); Grado superior de música)* 11
- *12.Arquitectura/ingeniería* 12
- *13.Máster oficial universitario (Especialidades médicas o equivalente)* 13
- *14.Doctorado* 14
- *15.Títulos propios de posgrado (máster no oficial, etc.)* 15
- *16. Otros estudios* 16
- *N.S.* 98
- *N.C.* 99

P.29 ¿Cómo se define Ud. en materia religiosa: católico/a practicante, católico/a no practicante, creyente de otra religión, agnóstico/a, indiferente o no creyente, o ateo/a?

[RELIGION]

- *Católico/a practicante* 1
- *Católico/a no practicante* 2
- *Creyente de otra religión* 3
- *Agnóstico/a (no niegan la existencia de Dios pero tampoco la descartan)* 4
- *Indiferente, no creyente* 5
- *Ateo/a (niegan la existencia de Dios)* 6
- *N.C.* 9

Estudio: Barómetro de diciembre de 2020
Clave: CIS3303

Filtros:
P29a si 1, 2 o 3 en P29.

P.29a. ¿Con qué frecuencia asiste Ud. a misa u otros oficios religiosos, sin contar las ocasiones relacionadas con ceremonias de tipo social, por ejemplo, bodas, comuniones o funerales?

[PRACTICARELIG]

- *Casi nunca* 1
- *Varias veces al año* 2
- *Alguna vez al mes* 3
- *Casi todos los domingos o festivos* 4
- *Varias veces a la semana* 5
- *N.C.* 9

P.30 ¿Cuál es su estado civil?

[ECIVIL]

- *Casado/a* 1
- *Soltero/a* 2
- *Viudo/a* 3
- *Separado/a* 4
- *Divorciado/a* 5
- *N.C.* 9

P.31 ¿En qué situación laboral se encuentra Ud. actualmente?

[SITLAB]

- *Trabaja* 1
- *Jubilado/a o pensionista (anteriormente ha trabajado)* 2
- *Pensionista (anteriormente no ha trabajado)* 3
- *En paro y ha trabajado antes* 4
- *En paro y busca su primer empleo* 5
- *Estudiante* 6
- *Trabajo doméstico no remunerado* 7
- *Otra situación* 8
- *N.C.* 9

[SITLAB_COD]

_____

P.31a ¿Me puede decir cuál es su ocupación actual?

Filtros:
P31a si P31=1.

[CNO11]

- *Directores/as y gerentes* 1
- *Profesionales y científicos/as e intelectuales* 2
- *Técnicos/as y profesionales de nivel medio* 3
- *Personal de apoyo administrativo* 4
- *Trabajadores/as de los servicios y vendedores/as de comercios y mercados* 5
- *Agricultores/as y trabajadores/as cualificados/as agropecuarios/as, forestales y pesqueros/as* 6
- *Oficiales/as, operarios/as y artesanos/as de artes mecánicas y de otros oficios* 7
- *Operadores/as de instalaciones y máquinas y ensambladores/as* 8
- *Ocupaciones elementales* 9
- *Ocupaciones militares y cuerpos policiales* 10
- *Otra/o* 11
- *N.C.* 99

P.32 ¿A qué clase social diría Ud. que pertenece? (RESPUESTA ESPONTÁNEA).

[CLASESOCIAL]

- *Clase alta* 1
- *Clase media-alta* 2
- *Clase media-media* 3
- *Clase media-baja* 4
- *Clase trabajadora/obrera* 5
- *Clase baja* 12
- *Clase pobre* 6
- *Infraclase* 7
- *Proletariado* 8
- *A los/as de abajo* 9
- *Excluidos/as* 10
- *A la gente común* 11
- *Otra (especificar)* 96
- *No cree en las clases* 97
- *No sabe, duda* 98
- *N.C.* 99

[CLASESOCIAL_COD]

_____

HEMOS TERMINADO. MUCHAS GRACIAS POR SU AMABILIDAD
Y POR EL TIEMPO QUE NOS HA DEDICADO.

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

«Información sujeta a secreto estadístico (Ley 12/89, de 9 de mayo, de la Función Estadística Pública) y al Reglamento General de Protección de Datos y la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales.» Plan Estadístico Nacional 2017-2020. RD 410/2016, 31 oct. y RD 1043/2017 de 22 dic.

«Buenos días/tardes, soy (nombre propio) y estoy realizando una encuesta telefónica para el Centro de Investigaciones Sociológicas (CIS) sobre temas de interés general. Por este motivo solicitamos su colaboración y se la agradecemos anticipadamente. Este teléfono ha sido obtenido al azar. Esta conversación será grabada para supervisar la calidad y después se borrará en un plazo inferior a un mes, le garantizamos el absoluto anonimato y secreto de sus respuestas en el más estricto cumplimiento de las leyes sobre secreto estadístico y protección de datos personales. Tras la realización de la encuesta su número de teléfono será disociado de las respuestas que pueda dar, que a su vez serán anonimizadas para que en ningún caso puedan ser asociadas a usted. Si desea conocer sus derechos de protección de datos y ampliar esta información puede consultar la página web www.cis.es ¿Ha comprendido la información leída? ¿Sería tan amable de contestar a unas preguntas, algunas de ellas sobre datos de carácter sensible como la intención de voto? No está obligado a contestar todas las preguntas. Muchas gracias.»

PC1. Pregunta contacto 1. ¿Me puede decir a qué provincia y municipio estoy llamando...?

[TIPO_TEL]
- ***FIJO*** *1*
- ***MÓVIL*** *2*

[CCAA]

[PROVINCIA]

[MUNICIPIO]
_____

[ENTREV]
_____

ENTREVISTADOR/A: SI LA PERSONA QUE CONTESTA ES DIFERENTE DE LA QUE COGIÓ EL TELÉFONO PRESENTARSE:

Buenos días/tardes, mi nombre es… y le llamo del Centro de Investigaciones Sociológicas porque estamos realizando una encuesta de opinión sobre temas de interés general. Dura de 12 a 15 minutos aproximadamente. ¿Sería tan amable de colaborar con nosotros?

[SEXO]
- ***Hombre*** *1*
- ***Mujer*** *2*

[EDADEXACTA]
_____

[EDAD]
- ***de 18 a 24*** *1*
- ***de 25 a 34*** *2*
- ***de 35 a 44*** *3*
- ***de 45 a 54*** *4*
- ***de 55 a 64*** *5*
- ***65 y más*** *6*

P.0 En primer lugar quisiera preguntarle si tiene Ud....

[P0]
- ***La nacionalidad española*** *1*
- ***La nacionalidad española y otra*** *2*
- ***Otra nacionalidad*** *3*

Saltos:
Si P0=3 ir a fin cuestionario.

P.1 Me gustaría hacerle algunas preguntas sobre la crisis del coronavirus. Pensando en todos los efectos de esta pandemia, ¿diría Ud. que la crisis del coronavirus le preocupa mucho, bastante, poco o nada?

[P1]
- ***Mucho*** *1*
- ***Bastante*** *2*
- ***Poco*** *3*
- ***(NO LEER) Regular*** *4*
- ***Nada*** *5*
- ***N.S.*** *8*
- ***N.C.*** *9*

P.2 En estos momentos, ¿qué le preocupa a Ud. más, los efectos de esta crisis sobre la salud, o los efectos de la crisis sobre la economía y el empleo?

[P2]
- ***Los efectos sobre la salud*** *1*
- ***Los efectos sobre la economía y el empleo*** *2*
- ***(NO LEER) Ambos por igual*** *3*
- ***(NO LEER) Ni unos ni otros*** *4*
- ***N.S.*** *8*
- ***N.C.*** *9*

P.3 En relación con la situación sanitaria generada por el coronavirus en España , diría Ud. que…

[P3]
- ***Lo peor ha pasado ya*** *1*
- ***Seguimos en el peor momento*** *2*
- ***Lo peor está por llegar*** *3*
- ***(NO LEER) No lo sabe, duda*** *98*
- ***N.C.*** *99*

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

P.4 En temas de salud, ¿ha tenido Ud. que contactar con los servicios sanitarios por pensar que tenía síntomas relacionados con el coronavirus?</span

[P4]

- *Sí* 1
- *No* 2
- *N.C.* 9

Filtros:
P4a si P4=1.

P.4a ¿A qué servicios recurrió Ud.? (RESPUESTA MÚLTIPLE)

[P4A]

- *A mi médico/a de atención primaria* 1
- *Al servicio de urgencias de atención primaria* 2
- *Al servicio de urgencias del hospital* 3
- *Al 061/112* 4
- *A un teléfono 900 que la comunidad puso para estos casos* 5
- *Otras respuestas (especificar)* 6
- *N.C.* 9

[P4A_COD_1]

_____

Filtros:
P4b si P4=1.

P.4b ¿Cómo fue la atención que recibió? (LEER)

[P4B]

- *Muy buena* 1
- *Buena* 2
- *(NO LEER) Regular* 3
- *Mala* 4
- *Muy mala* 5
- *N.S.* 8
- *N.C.* 9

Filtros:
P4c si P4=1.

P.4c ¿Le indicaron que debía guardar medidas de aislamiento?

[P4C]

- *Sí* 1
- *No* 2
- *N.C.* 9

Filtros:
P4d si P4=1 y si P4c=1.

P.4d. ¿Le hicieron la prueba del coronavirus?

[P4D]

- *Sí* 1
- *No* 2
- *N.S.* 8
- *N.C.* 9

Filtros:
P4e si 1 en P4, 1 en P4c y 1 en P4d.

P.4e. ¿Y le diagnosticaron finalmente infección por coronavirus?

[P4E]

- *Sí* 1
- *No* 2
- *N.C.* 9

Filtros:
P4f si 1 en P4, 1 en P4c, 1 en P4d y 1 en P4e.

P.4f. ¿Y cómo evolucionó su enfermedad? (LEER)

[P4F]

- *Tuve síntomas leves y la pasé en casa* 1
- *Tuve síntomas importantes, pero la pasé en casa* 2
- *Tuve que ingresar en el hospital* 3
- *N.C.* 9

Filtros:
P4g si 1 en P4, 1 en P4c, 1 en P4d, 1 en P4e y 1 o 2 en P4f.

P.4g. ¿Quién le atendió sanitariamente durante la enfermedad? (RESPUESTA MÚLTIPLE, LEER SI ES PRECISO)

[P4G]

- *El/la médico/a de atención primaria* 1
- *La/el enfermera/o de atención primaria* 2
- *El 061/112* 3
- *Otros/as profesionales sanitarios/as* 4
- *No tuve seguimiento por ningún/a profesional sanitario/a* 5
- *N.C.* 9

Filtros:
P4h si 1 en P4, 1 en P4c, 1 en P4d, 1 en P4e, 1 o 2 en P4f y 1, 2, 3 o 4 en P4g.

P.4h Y esa atención se llevó a cabo de forma… (RESPUESTA MÚLTIPLE) (LEER)

[P4H]

- *Telefónica* 1
- *A través de Internet* 2
- *Vinieron a visitarme a casa* 3
- *Acudí al centro de salud* 4
- *En algún momento acudí a urgencias del hospital* 5
- *Otras respuestas (especificar)* 6
- *N.C.* 9

[P4H_COD]

_____

Filtros:
P4i si 1 en P4, 1 en P4c, 1 en P4d, 1 en P4e, 1 o 2 en P4f y 1, 2, 3 o 4 en P4g.

P4i ¿Y cómo valora, en su conjunto, la atención que ha recibido? (LEER)

[P4I]

- *Muy buena* 1
- *Buena* 2
- *(NO LEER) Regular* 3
- *Mala* 4
- *Muy mala* 5
- *N.S.* 8
- *N.C.* 9

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

P.5 ¿Está Ud. dispuesto/a a vacunarse del COVID-19 inmediatamente? (UNA SOLA RESPUESTA)
(ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O MARCAR "OTRAS RESPUESTAS") </span

[P5]

Sí	*1*
No	*2*
Sí, si tiene garantías, si está probada, si es fiable	*3*
Sí, según el origen de la vacuna	*4*
Sí, si hay información suficiente	*5*
Sí, por consejo de autoridades, científicos/as, o sanitarios/as	*6*
Otras respuestas	*7*
(NO LEER) No sabe, duda	*8*
N.C.	*9*

Filtros:
P5a si P5 =2.

P.5a ¿Cuál es la razón principal por la que no se vacunaría inmediatamente? (UNA SOLA RESPUESTA).
(ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O COMPLETAR "OTRAS RAZONES, ¿CUÁL?").

[P5A]

No se fía de estas vacunas	*1*
No cree que sean eficaces	*2*
Por miedo a que tengan riesgos para la salud/efectos secundarios-colaterales	*3*
Por tener pocas probabilidades de contagio	*4*
Por haber pasado la COVID-19	*5*
Prefiere esperar para ver cómo funcionan	*6*
Por otra razón, ¿cuál?	*7*
(NO LEER) No lo sabe, duda	*98*
N.C.	*99*

[P5A_COD]
_____

P.6 ¿Quién le gustaría a Ud. que se hiciera cargo de la lucha contra la pandemia primordialmente: el Gobierno de España, el Gobierno de las comunidades autónomas, o ambos en colaboración?

[P6]

El Gobierno de España	*1*
El Gobierno de las comunidades autónomas	*2*
Ambos en colaboración	*3*
Otras respuestas (especificar)	*6*
(NO LEER) No lo sabe, duda	*8*
N.C.	*9*

[P6_COD]
_____

P.7 Por lo que Ud. está viendo, ¿cree que la mayoría de los/as españoles/as estamos dando un ejemplo de civismo y solidaridad en la forma de afrontar las medidas contra el COVID-19, o piensa que la mayoría está siendo poco cívica e indisciplinada?

[P7]

Cree que la mayoría está reaccionando con civismo y solidaridad	*1*
Cree que la mayoría está siendo poco cívica e indisciplinada	*2*
(NO LEER) No lo sabe, duda	*8*
N.C.	*9*

P.8 Considerando lo que está ocurriendo con la pandemia, ¿todo lo que sucede le está afectando a Ud. mucho, bastante, algo, nada o casi nada en su vida personal?

P.9 ¿Y le está afectando a Ud. mucho, bastante, algo o nada o casi nada en su vida social y de relaciones?
P8. Vida personal
P9. Vida social

[P8]

Me está afectando mucho	*1*
Me está afectando bastante	*2*
(NO LEER) Regular	*3*
Me está afectando algo	*4*
No me está afectando nada o casi nada	*6*
N.S.	*8*
N.C.	*9*

[P9]

Me está afectando mucho	*1*
Me está afectando bastante	*2*
(NO LEER) Regular	*3*
Me está afectando algo	*4*
No me está afectando nada o casi nada	*6*
N.S.	*8*
N.C.	*9*

Filtros:
P10a si P8=1, 2, 3 o 4.

P.10A ¿En qué aspecto o aspectos le está afectando a Ud. en su vida personal?
(ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O MARCAR "OTRAS RESPUESTAS").

[P10A]

Por el distanciamiento con los seres queridos	*1*
Por el trabajo y/o economía personal	*2*
Por el estado anímico negativo (ansiedad, tristeza...)	*3*
Las restricciones y libertad de movimientos	*4*
Por miedo al contagio suyo o de otras personas	*5*
Por el cambio de las condiciones laborales	*6*
Por el aislamiento y el confinamiento	*7*
Por los cambios en la vida cotidiana	*8*
Por pérdida de trabajo (despidos, cierres, no encontrar...)	*9*
Por los cambios de rutina en clases	*10*
Otras respuestas	*97*
N.S.	*98*
N.C	*99*

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

Filtros:
P10b si P9=1, 2, 3 o 4.

P10.B ¿En qué aspecto o aspectos le está afectando en su vida social?
(ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O MARCAR "OTRAS RESPUESTAS").

[P10B]

- *Por la distancia respecto a la gente 1*
- *Por la distancia respecto a los amigos/as 2*
- *Por el asislamiento, confinamiento y no poder salir 3*
- *Por la distancia respecto a la familia 4*
- *Por el cese o limitación de actividades culturales, de ocio y deportivas 5*
- *Por el miedo a que se produzcan contagios en las relaciones sociales y familiares 6*
- *Aspectos económicos, laborales y/o profesionales 7*
- *Aspectos psicológicos, anímicos y emocionales 8*
- *Por el cese o limitación de actividades de hostelería y restauración 9*
- *Por la ausencia de contacto físico en las relaciones sociales, pérdida de calidad en las relaciones 10*
- *Otras respuestas 97*
- *N.S. 98*
- *N.C. 99*

P.11 Al margen de las medidas oficiales para el control de la COVID tomadas por las autoridades desde que estamos bajo la pandemia, ¿qué medidas propias de control ha estado siguiendo Ud.?

[P11]

- *Ninguna, hace vida normal 1*
- *Tiene cuidado con las cosas que toca o por dónde va, pero en lo demás hace vida normal 2*
- *Permanece prácticamente en aislamiento, saliendo de casa sólo para adquirir alimentos y para ir a consultas médicas 3*
- *No sale de casa para nada que no resulte imprescindible y le traen los suministros y medicinas 4*
- *Otras respuestas (especificar) 6*
- *N.S., duda 8*
- *N.C. 9*

[P11_COD]

_____

P.12 Refiriéndonos a la situación económica general de España actualmente, ¿cómo la calificaría Ud.: muy buena, buena, mala o muy mala?

[P12]

- *Muy buena 1*
- *Buena 2*
- *(NO LEER) Regular 3*
- *Mala 4*
- *Muy mala 5*
- *N.S. 8*
- *N.C. 9*

P.13 ¿Cómo calificaría Ud. su situación económica personal en la actualidad: muy buena, buena, mala o muy mala?

[P13]

- *Muy buena 1*
- *Buena 2*
- *(NO LEER) Regular 3*
- *Mala 4*
- *Muy mala 5*
- *N.S. 8*
- *N.C. 9*

P.14 ¿Cuál es, a su juicio, el principal problema que existe actualmente en España? ¿Y el segundo? ¿Y el tercero? (ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O COMPLETAR "OTRO, ¿CUÁL?")

1er lugar

[P14_1]

- *Los peligros para la salud: COVID-19 ..... 53*
- *La crisis económica, los problemas de índole económica 8*
- *El paro 1*
- *Los problemas políticos en general 51*
- *Lo que hacen los partidos políticos 50*
- *El mal comportamiento de los/as políticos/as 13*
- *La corrupción y el fraude 11*
- *La sanidad 6*
- *Los problemas de índole social 16*
- *La inmigración 18*
- *Los problemas relacionados con la calidad del empleo 9*
- *La independencia de Cataluña 45*
- *Las pensiones 12*
- *La violencia de género 19*
- *La falta de acuerdos, unidad y capacidad de colaboración. 46*
- *Poca conciencia ciudadana (falta de civismo, de sentido espíritu cívico) 55*
- *Otro 96*
- *Ninguno 97*
- *N.S. 98*
- *N.C. 99*

[P141_OTRO]

_____

2º lugar

[P14_2]

- *Los peligros para la salud: COVID-19 ..... 53*
- *La crisis económica, los problemas de índole económica 8*
- *El paro 1*
- *Los problemas políticos en general 51*
- *Lo que hacen los partidos políticos 50*
- *El mal comportamiento de los/as políticos/as 13*
- *La corrupción y el fraude 11*
- *La sanidad 6*
- *Los problemas de índole social 16*
- *La inmigración 18*
- *Los problemas relacionados con la calidad del empleo 9*
- *La independencia de Cataluña 45*
- *Las pensiones 12*
- *La violencia de género 19*
- *La falta de acuerdos, unidad y capacidad de colaboración. 46*
- *Poca conciencia ciudadana (falta de civismo, de sentido espíritu cívico) 55*
- *Otro 96*
- *Ninguno 97*
- *N.S. 98*
- *N.C. 99*

[P142_OTRO]

_____

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

3er lugar
[P14_3]

- *Los peligros para la salud: COVID-19.....53*
- *La crisis económica, los problemas de índole económica.....8*
- *El paro.....1*
- *Los problemas políticos en general.....51*
- *Lo que hacen los partidos políticos.....50*
- *El mal comportamiento de los/as políticos/as.....13*
- *La corrupción y el fraude.....11*
- *La sanidad.....6*
- *Los problemas de índole social.....16*
- *La inmigración.....18*
- *Los problemas relacionados con la calidad del empleo.....9*
- *La independencia de Cataluña.....45*
- *Las pensiones.....12*
- *La violencia de género.....19*
- *La falta de acuerdos, unidad y capacidad de colaboración. 46*
- *Poca conciencia ciudadana (falta de civismo, de sentido espíritu cívico).....55*
- *Otro.....96*
- *Ninguno.....97*
- *N.S.....98*
- *N.C.....99*

[P143_OTRO]
_____

P.15 ¿Y cuál es el problema que a Ud., personalmente, le afecta más? ¿Y el segundo? ¿Y el tercero? (ENTREVISTADOR/A: NO LEER – MARCAR LA OPCIÓN QUE MÁS SE APROXIME A LO DICHO TEXTUALMENTE O COMPLETAR "OTRO, ¿CUÁL?"). </span

1er lugar </span
[P15_1]

- *Los peligros para la salud: COVID-19.....53*
- *La crisis económica, los problemas de índole económica.....8*
- *Avituallamiento de víveres en el hogar.....60*
- *Tener que estar enclaustrado/a en casa.....59*
- *El paro.....1*
- *La sanidad.....6*
- *Las pensiones.....12*
- *Los problemas relacionados con la calidad del empleo.....9*
- *Los problemas políticos en general.....51*
- *Lo que hacen los partidos políticos.....50*
- *El mal comportamiento de los/as políticos/as.....13*
- *La educación.....22*
- *La corrupción y el fraude.....11*
- *Los problemas de índole social.....16*
- *Las preocupaciones y situaciones personales.....29*
- *La vivienda.....7*
- *La independencia de Cataluña.....45*
- *Los problemas relacionados con la juventud.....20*
- *Otro.....96*
- *Ninguno.....97*
- *N.S.....98*
- *N.C.....99*

[P151_OTRO]
_____

2º lugar</span
[P15_2]

- *Los peligros para la salud: COVID-19.....53*
- *La crisis económica, los problemas de índole económica 8*
- *Avituallamiento de víveres en el hogar.....60*
- *Tener que estar enclaustrado/a en casa.....59*
- *El paro.....1*
- *La sanidad.....6*
- *Las pensiones.....12*
- *Los problemas relacionados con la calidad del empleo.....9*
- *Los problemas políticos en general.....51*
- *Lo que hacen los partidos políticos.....50*
- *El mal comportamiento de los/as políticos/as.....13*
- *La educación.....22*
- *La corrupción y el fraude.....11*
- *Los problemas de índole social.....16*
- *Las preocupaciones y situaciones personales.....29*
- *La vivienda.....7*
- *La independencia de Cataluña.....45*
- *Los problemas relacionados con la juventud.....20*
- *Otro.....96*
- *Ninguno.....97*
- *N.S.....98*
- *N.C.....99*

[P152_OTRO]
_____

3er lugar</span
[P15_3]

- *Los peligros para la salud: COVID-19.....53*
- *La crisis económica, los problemas de índole económica 8*
- *Avituallamiento de víveres en el hogar.....60*
- *Tener que estar enclaustrado/a en casa.....59*
- *El paro.....1*
- *La sanidad.....6*
- *Las pensiones.....12*
- *Los problemas relacionados con la calidad del empleo.....9*
- *Los problemas políticos en general.....51*
- *Lo que hacen los partidos políticos.....50*
- *El mal comportamiento de los/as políticos/as.....13*
- *La educación.....22*
- *La corrupción y el fraude.....11*
- *Los problemas de índole social.....16*
- *Las preocupaciones y situaciones personales.....29*
- *La vivienda.....7*
- *La independencia de Cataluña.....45*
- *Los problemas relacionados con la juventud.....20*
- *Otro.....96*
- *Ninguno.....97*
- *N.S.....98*
- *N.C.....99*

[P153_OTRO]
_____

P.16 A continuación, voy a mencionarle los nombres de algunos/as líderes políticos/as y nos gustaría que, en relación a lo que cada uno/a está diciendo y haciendo sobre la COVID-19 en estos momentos, lo/a califique de 1 a 10, siendo el 1 "muy mal" y el 10 "muy bien". (SI NO CONOCE ALGUNO/A ME LO DICE).
[P16]

	NO CONOCE	*Muy mal 1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Muy bien 10*	*N.S.*	*N.C.*
Pedro Sánchez	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Casado	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Iglesias	97	1	2	3	4	5	6	7	8	9	10	98	99
Santiago Abascal	97	1	2	3	4	5	6	7	8	9	10	98	99
Inés Arrimadas	97	1	2	3	4	5	6	7	8	9	10	98	99

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

P.17 El presidente del Gobierno, Pedro Sánchez, ¿le inspira a Ud., personalmente, mucha confianza, bastante confianza, poca o ninguna confianza?

[P17]

Mucha confianza 1
Bastante confianza 2
Poca confianza 3
Ninguna confianza 4
N.S. 8
N.C. 9

P.18 ¿Y el líder del principal partido de la oposición, Pablo Casado, le inspira, personalmente, mucha confianza, bastante confianza, poca o ninguna confianza?

[P18]

Mucha confianza 1
Bastante confianza 2
Poca confianza 3
Ninguna confianza 4
N.S. 8
N.C. 9

P.19 De los/as principales líderes políticos/as, ¿quién preferiría que fuese el/la presidente/a del Gobierno en estos momentos? (SOLO UN NOMBRE).

[P19]

Pedro Sánchez 1
Pablo Casado 2
Santiago Abascal 3
Pablo Iglesias 4
Alberto Garzón 5
Inés Arrimadas 6
Íñigo Errejón 7
(NO LEER) Ninguno/a de ellos/as 97
(NO LEER) Otro/a, ¿quién? 96
N.S. 98
N.C. 99

[P19_COD]

P.20 A continuación voy a leerle la lista de los/as ministros y ministras que forman el Gobierno. Dígame, por favor, para cada uno/a de ellos/as si lo/a conoce y cómo lo/a valoraría en una escala de 1 a 10, sabiendo que el 1 significa que lo/a valora "muy mal" y el 10 que lo/a valora "muy bien". (ENTREVISTADOR/A: UTILIZAR EL SCROLL PARA COMPLETAR TODAS LAS VALORACIONES).

[P20]

	NO CONOCE	Muy mal 1	2	3	4	5	6	7	8	9	Muy bien 10	N.S.	N.C.
José Luis Ábalos	97	1	2	3	4	5	6	7	8	9	10	98	99
Nadia Calviño	97	1	2	3	4	5	6	7	8	9	10	98	99
Carmen Calvo	97	1	2	3	4	5	6	7	8	9	10	98	99
Juan Carlos Campo	97	1	2	3	4	5	6	7	8	9	10	98	99
Manuel Castells	97	1	2	3	4	5	6	7	8	9	10	98	99
Isabel Celaá	97	1	2	3	4	5	6	7	8	9	10	98	99
Carolina Darias	97	1	2	3	4	5	6	7	8	9	10	98	99
Yolanda Díaz	97	1	2	3	4	5	6	7	8	9	10	98	99
Pedro Duque	97	1	2	3	4	5	6	7	8	9	10	98	99
José Luis Escrivá	97	1	2	3	4	5	6	7	8	9	10	98	99
Alberto Garzón	97	1	2	3	4	5	6	7	8	9	10	98	99
Arancha González Laya	97	1	2	3	4	5	6	7	8	9	10	98	99
Fernando Grande-Marlaska	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Iglesias	97	1	2	3	4	5	6	7	8	9	10	98	99
Salvador Illa	97	1	2	3	4	5	6	7	8	9	10	98	99
Reyes Maroto	97	1	2	3	4	5	6	7	8	9	10	98	99
María Jesús Montero	97	1	2	3	4	5	6	7	8	9	10	98	99
Irene Montero	97	1	2	3	4	5	6	7	8	9	10	98	99
Luis Planas	97	1	2	3	4	5	6	7	8	9	10	98	99
Teresa Ribera	97	1	2	3	4	5	6	7	8	9	10	98	99
Margarita Robles	97	1	2	3	4	5	6	7	8	9	10	98	99
José Manuel Rodríguez Uribes	97	1	2	3	4	5	6	7	8	9	10	98	99

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

P.21 Suponiendo que mañana se celebrasen nuevamente elecciones generales, es decir, al Parlamento español, ¿a qué partido votaría Ud.? (RESPUESTA ESPONTÁNEA).

[INTENCIONG]

PSOE (Partido Socialista Obrero Español)	*2*
PP (Partido Popular)	*1*
VOX	*18*
Podemos	*3*
IU (Izquierda Unida)	*5*
Unidas Podemos	*21*
En Comú Podem	*6*
En Común - Unidas Podemos	*67*
Ciudadanos	*4*
Más País	*50*
ERC (Esquerra Republicana de Catalunya)	*8*
JxCat (Junts per Catalunya)	*9*
CUP	*19*
EAJ-PNV (Partido Nacionalista Vasco)	*11*
EH Bildu (Euskal Herria Bildu)	*12*
CC-PNC (Coalición Canaria – Partido Nacionalista Canario)	*13*
Nueva Canarias	*16*
UPN (Unión del Pueblo Navarro)	*14*
Compromís	*7*
BNG (Bloque Nacionalista Galego)	*24*
PRC (Partido Regionalista de Cantabria)	*43*
Teruel Existe	*68*
PACMA (Partido Animalista)	*17*
FAC (Foro Asturias)	*15*
Otro partido, ¿cuál?	*95*
Voto nulo	*77*
En blanco	*96*
No votaría	*97*
No sabe todavía	*98*
N.C.	*99*

[INTENCIONG_COD]

Filtros:
P21a si P21 no es 77, 96, 97, 98 o 99.

P.21a En el caso de que por cualquier razón finalmente no votase por el partido que me ha dicho, ¿a qué otro partido votaría Ud.? (RESPUESTA ESPONTÁNEA).

[INTENCIONGALTER]

PSOE (Partido Socialista Obrero Español)	*2*
PP (Partido Popular)	*1*
VOX	*18*
Podemos	*3*
IU (Izquierda Unida)	*5*
Unidas Podemos	*21*
En Comú Podem	*6*
En Común - Unidas Podemos	*67*
Ciudadanos	*4*
Más País	*50*
ERC (Esquerra Republicana de Catalunya)	*8*
JxCat (Junts per Catalunya)	*9*
CUP	*19*
EAJ-PNV (Partido Nacionalista Vasco)	*11*
EH Bildu (Euskal Herria Bildu)	*12*
CC-PNC (Coalición Canaria – Partido Nacionalista Canario)	*13*
Nueva Canarias	*16*
UPN (Unión del Pueblo Navarro)	*14*
Compromís	*7*
BNG (Bloque Nacionalista Galego)	*24*
PRC (Partido Regionalista de Cantabria)	*43*
Teruel Existe	*68*
PACMA (Partido Animalista)	*17*
FAC (Foro Asturias)	*15*
(NO LEER) No votaría a ningún otro partido (cita el mismo partido)	*93*
Otro partido, ¿cuál?	*95*
En blanco	*96*
No votaría	*97*
N.S.	*98*
N.C.	*99*

[INTENCIONGALTER_COD]

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

Filtros:
P22 si P21 = 77, 96, 97, 98 o 99.

P.22 Sin ningún compromiso por su parte, ¿me podría decir por qué partido siente Ud. más simpatía? (RESPUESTA ESPONTÁNEA).

[SIMPATIA]

PSOE (Partido Socialista Obrero Español) ***2***
PP (Partido Popular) ***1***
VOX ***18***
Podemos ***3***
IU (Izquierda Unida) ***5***
Unidas Podemos ***21***
En Comú Podem ***6***
En Común - Unidas Podemos ***67***
Ciudadanos ***4***
Más País ***50***
ERC (Esquerra Republicana de Catalunya) ***8***
JxCat (Junts per Catalunya) ***9***
CUP ***19***
EAJ-PNV (Partido Nacionalista Vasco) ***11***
EH Bildu (Euskal Herria Bildu) ***12***
CC-PNC (Coalición Canaria – Partido Nacionalista Canario) ***13***
Nueva Canarias ***16***
UPN (Unión del Pueblo Navarro) ***14***
Compromís ***7***
BNG (Bloque Nacionalista Galego) ***24***
PRC (Partido Regionalista de Cantabria) ***43***
Teruel Existe ***68***
PACMA (Partido Animalista) ***17***
FAC (Foro Asturias) ***15***
Otro partido, ¿cuál? ***95***
Ninguno ***97***
N.S. ***98***
N.C. ***99***

[SIMPATIA_COD]
_____

P.23 Cuando se habla de política se utilizan normalmente las expresiones izquierda y derecha. Situándonos en una escala de 10 casillas, como un termómetro, que van del 1 al 10, en la que 1 significa "lo más a la izquierda" y 10 "lo más a la derecha", ¿en qué casilla se colocaría Ud.?

[ESCIDEOL]

1 Izda. ***1***
2 ..... ***2***
3 ..... ***3***
4 ..... ***4***
5 ..... ***5***
6 ..... ***6***
7 ..... ***7***
8 ..... ***8***
9 ..... ***9***
10 Dcha. ***10***
N.S. ***98***
N.C. ***99***

P.24 Y siguiendo este mismo criterio, a los/as líderes de los principales partidos políticos, ¿dónde los/as ubicaría? (ENTREVISTADOR/A: PEDIR A LA PERSONA ENTREVISTADA QUE INDIQUE LA CASILLA EN LA QUE COLOCARÍA CADA LÍDER).

[ESCIDEOLPOLI]

	NO CONOCE	Izquierda 1	2	3	4	5	6	7	8	9	Derecha 10	N.S.	N.C.
Pedro Sánchez	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Casado	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Iglesias	97	1	2	3	4	5	6	7	8	9	10	98	99
Santiago Abascal	97	1	2	3	4	5	6	7	8	9	10	98	99
Inés Arrimadas	97	1	2	3	4	5	6	7	8	9	10	98	99

P.25 ¿Y qué valoración le merece cada uno/a de los/as siguientes políticos/as? Puntúelos/as de 1 a 10, sabiendo que el 1 significa que lo/a valora "muy mal" y el 10 que lo/a valora "muy bien".

[P25]

	NO CONOCE	Muy mal 1	2	3	4	5	6	7	8	9	Muy bien 10	N.S.	N.C.
Pedro Sánchez	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Casado	97	1	2	3	4	5	6	7	8	9	10	98	99
Pablo Iglesias	97	1	2	3	4	5	6	7	8	9	10	98	99
Santiago Abascal	97	1	2	3	4	5	6	7	8	9	10	98	99
Inés Arrimadas	97	1	2	3	4	5	6	7	8	9	10	98	99

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

P.26 Cambiando de tema ¿sabe Ud. que el pasado mes de diciembre el Congreso de los Diputados aprobó la Ley de Regulación de la Eutanasia?

[P26]

- *Sí* 1
- *No* 2
- *(NO LEER) No sabe qué es/qué significa "eutanasia"* 3
- *N.C.* 9

Filtros:
P26a si P26 = 1, 2 o 9.

P.26a Y, por lo que Ud. sabe, ¿está de acuerdo o en desacuerdo con la eutanasia?

[P26A]

- *Totalmente de acuerdo* 1
- *De acuerdo* 2
- *En desacuerdo* 4
- *Totalmente en desacuerdo* 5
- *(NO LEER) Ni de acuerdo ni en desacuerdo* 3
- *(NO LEER) Duda, no sabe* 8
- *N.C.* 9

Filtros:
P26a si P26 = 1, 2 o 9 y P26a=3 u 8.

P.26b ¿Por qué razón duda en su opinión sobre la eutanasia?
(PREGUNTA ABIERTA, ANOTAR TODO LO QUE MENCIONE LA PERSONA ENTREVISTADA).

[P26B]

N.S. = 998 N.C. = 999

P.27 ¿Me podría decir si en las elecciones generales del 10 de noviembre de 2019...? (LEER RESPUESTAS).

[PARTICIPACIONG]

- *Fue a votar y votó* 1
- *Voto por correo* 7
- *No tenía edad para votar* 2
- *Fue a votar pero no pudo hacerlo* 3
- *No fue a votar porque no pudo* 4
- *Prefirió no votar* 5
- *No tenía derecho a voto* 6
- *No recuerda* 8
- *N.C.* 9

Filtros:
P27a si P27=1 o 7

P.27a ¿Y podría decirme a qué partido o coalición votó? (RESPUESTA ESPONTÁNEA).

[RECUVOTOG]

- *PSOE* 2
- *PP* 1
- *VOX* 18
- *Unidas Podemos* 21
- *En Comú Podem* 6
- *En Común - Unidas Podemos* 67
- *Ciudadanos* 4
- *Más País* 50
- *ERC* 8
- *JxCat* 9
- *CUP* 19
- *EAJ-PNV* 11
- *EH Bildu* 12
- *CCa-PNC-NC* 13
- *Navarra Suma (UPN)* 14
- *Més Compromís* 7
- *BNG (Bloque Nacionalista Galego)* 24
- *PRC (Partido Regionalista de Cantabria)* 43
- *Teruel Existe* 68
- *PACMA (Partido Animalista)* 17
- *Otros partidos* 95
- *En blanco* 96
- *Voto nulo* 77
- *No recuerda* 98
- *N.C.* 99

P.28 En todo caso, ¿qué partido considera más cercano a sus propias ideas? (RESPUESTA ESPONTÁNEA).

[CERCANIA]

- *PSOE (Partido Socialista Obrero Español)* 2
- *PP (Partido Popular)* 1
- *VOX* 18
- *Podemos* 3
- *IU (Izquierda Unida)* 5
- *Unidas Podemos* 21
- *En Comú Podem* 6
- *En Común-Unidas Podemos* 67
- *Ciudadanos* 4
- *Más País* 50
- *ERC (Esquerra Republicana de Catalunya)* 8
- *JxCat (Junts Per Catalunya)* 9
- *CUP* 19
- *EAJ-PNV (Partido Nacionalista Vasco)* 11
- *EH Bildu (Euskal Herria Bildu)* 12
- *CC-PNC (Coalición Canaria – Partido Nacionalista Canario)* 13
- *Nueva Canarias* 16
- *UPN (Unión del Pueblo Navarro)* 14
- *Compromís* 7
- *BNG (Bloque Nacionalista Galego)* 24
- *PRC (Partido Regionalista de Cantabria)* 43
- *Teruel Existe* 68
- *PACMA (Partido Animalista)* 17
- *FAC (Foro Asturias)* 15
- *Otro partido, ¿cuál?* 95
- *Ninguno* 97
- *N.S.* 98
- *N.C.* 99

[CERCANIA_COD]

Estudio: Barómetro de enero de 2021
Clave: ECIS3307

P.29 ¿Ha ido Ud. a la escuela o cursado algún tipo de estudios? (ENTREVISTADOR/A: en caso negativo, preguntar si sabe leer y escribir).

[ESCUELA]

No, es analfabeto/a ***1***
No, pero sabe leer y escribir ***2***
Sí, ha ido a la escuela ***3***
N.C. ***9***

Filtros:
P29a si P29=3.

P.29a ¿Cuáles son los estudios de más alto nivel oficial que Ud. ha cursado (con independencia de que los haya terminado o no)? Por favor, especifique lo más posible, diciéndome el curso en que estaba cuando los terminó (o los interrumpió) y también el nombre que tenían entonces esos estudios (ej.: 3 años de estudios primarios, primaria, 5º de bachillerato, Maestría Industrial, preuniversitario, 4º de EGB, licenciatura, doctorado, FP1, etc.). (ENTREVISTADOR/A: si aún está estudiando, anotar el último curso que haya completado y el ciclo correcto en las opciones de respuesta. Si no ha completado la primaria, anotar nº de años que asistió a la escuela, diferenciando entre menos de 5 y más de 5).

[CURSOENTREV]

CURSO _____ *N.S. - N.R. = 98 N.C. = 99*

[NOMBREESTENTREV]

NOMBRE DE ESTUDIOS _____ *N.S. - N.R. = 98 N.C. = 99*

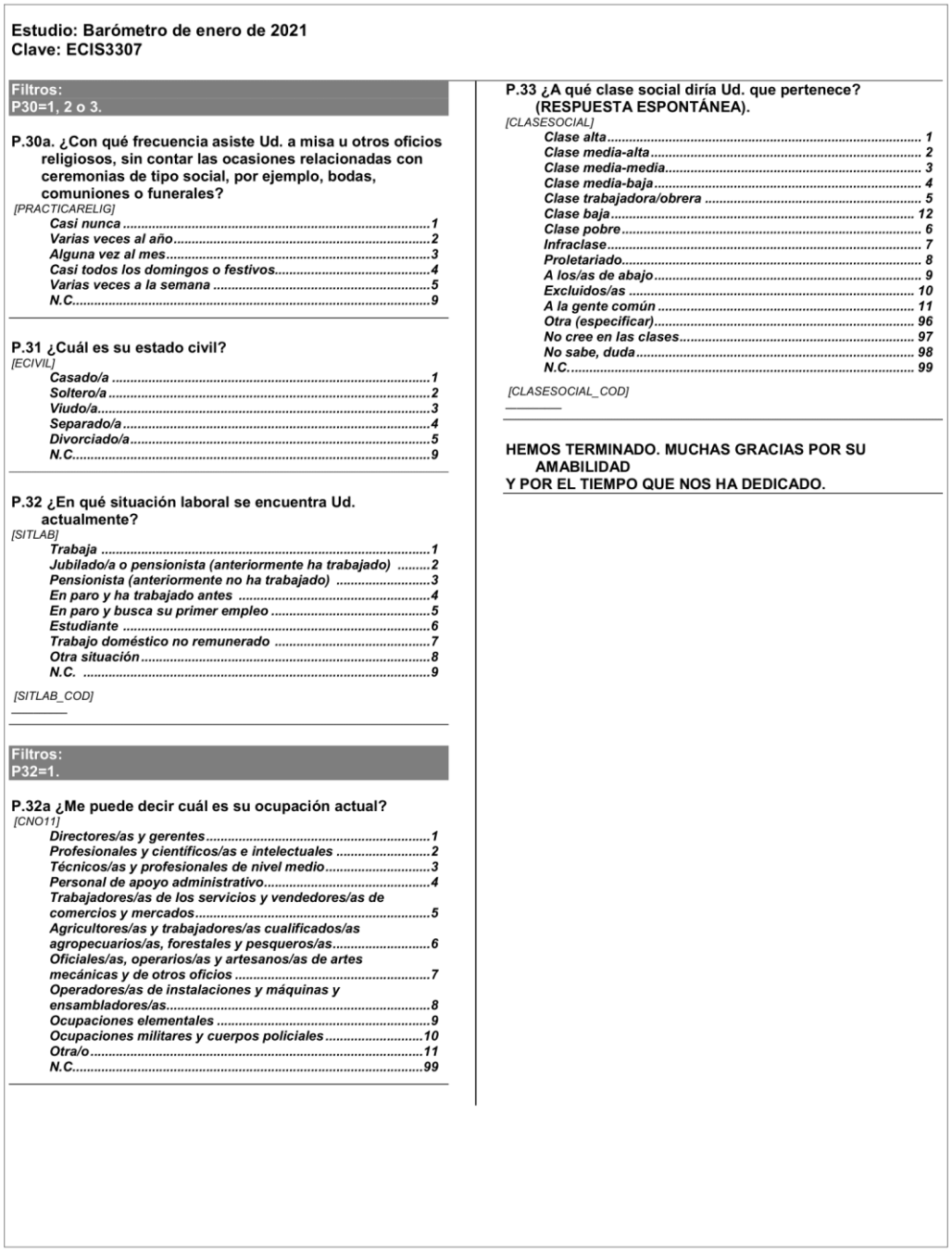
[NIVELESTENTREV]

01.Menos de 5 años de escolarización ***1***
02.Educación primaria (Educación primaria de LOGSE, 5º Curso de EGB, Enseñanza primaria antigua) ***2***
03.Cualificación profesional grado inicial (FP grado inicial). PCPI (Programas de Cualificación Profesional Inicial, que no precisan de titulación académica de la primera etapa de secundaria para su realización). Programas de garantía social ***3***
04.Educación secundaria (ESO, EGB. Graduado Escolar. Certificado de Escolaridad, Bachillerato Elemental) ***4***
05.FP de grado medio (Ciclo/módulo formativo de FP (grado medio), de Artes Plásticas y Diseño, Música y danza, Enseñanzas deportivas, FP I, Bachiller laboral elemental. Oficialía Industrial; Bachillerato Comercial). ***5***
06.Bachillerato (Bachillerato LOGSE, BUP, Bachillerato superior (6º), Bachillerato universitario (7º), Incluidos COU y PREU) ***6***
07.FP de grado superior (Ciclo/módulo formativo de FP (grado superior) de Artes Plásticas, Diseño, Música y danza, Deporte, FP II, Bach. Laboral Sup., Maestría industrial, Perito Mercantil; Secretariado de 2º grado; Grado Medio conservatorio) ***7***
08.Arquitectura-ingeniería técnica (Arquitectura/ingeniería técnica, aparejador/a; peritos/a) ***8***
09.Diplomatura (ATENCIÓN: solo Diplomaturas oficiales, no codificar aquí los tres primeros años de una licenciatura o grado con mayor duración) ***9***
10.Grado (Estudios de grado, Enseñanzas Artísticas equivalentes (desde 2006)) ***10***
11.Licenciatura (Titulaciones con equivalencia oficial: 2º ciclo INEF; Danza y arte dramático (desde 1992); Grado superior de música) ***11***
12.Arquitectura/ingeniería ***12***
13.Máster oficial universitario (Especialidades médicas o equivalente) ***13***
14.Doctorado ***14***
15.Títulos propios de posgrado (máster no oficial, etc.) ***15***
16. Otros estudios ***16***
N.S. ***98***
N.C. ***99***

P.30 ¿Cómo se define Ud. en materia religiosa: católico/a practicante, católico/a no practicante, creyente de otra religión, agnóstico/a, indiferente o no creyente, o ateo/a?

[RELIGION]

Católico/a practicante ***1***
Católico/a no practicante ***2***
Creyente de otra religión ***3***
Agnóstico/a (no niegan la existencia de Dios pero tampoco la descartan) ***4***
Indiferente, no creyente ***5***
Ateo/a (niegan la existencia de Dios) ***6***
N.C. ***9***



Estudio: Barómetro de enero de 2021
Clave: ECIS3307

Filtros:
P30=1, 2 o 3.

P.30a. ¿Con qué frecuencia asiste Ud. a misa u otros oficios religiosos, sin contar las ocasiones relacionadas con ceremonias de tipo social, por ejemplo, bodas, comuniones o funerales?
[PRACTICARELIG]
- Casi nunca 1
- Varias veces al año 2
- Alguna vez al mes 3
- Casi todos los domingos o festivos 4
- Varias veces a la semana 5
- N.C. 9

P.31 ¿Cuál es su estado civil?
[ECIVIL]
- Casado/a 1
- Soltero/a 2
- Viudo/a 3
- Separado/a 4
- Divorciado/a 5
- N.C. 9

P.32 ¿En qué situación laboral se encuentra Ud. actualmente?
[SITLAB]
- Trabaja 1
- Jubilado/a o pensionista (anteriormente ha trabajado) 2
- Pensionista (anteriormente no ha trabajado) 3
- En paro y ha trabajado antes 4
- En paro y busca su primer empleo 5
- Estudiante 6
- Trabajo doméstico no remunerado 7
- Otra situación 8
- N.C. 9

[SITLAB_COD]
_____

Filtros:
P32=1.

P.32a ¿Me puede decir cuál es su ocupación actual?
[CNO11]
- Directores/as y gerentes 1
- Profesionales y científicos/as e intelectuales 2
- Técnicos/as y profesionales de nivel medio 3
- Personal de apoyo administrativo 4
- Trabajadores/as de los servicios y vendedores/as de comercios y mercados 5
- Agricultores/as y trabajadores/as cualificados/as agropecuarios/as, forestales y pesqueros/as 6
- Oficiales/as, operarios/as y artesanos/as de artes mecánicas y de otros oficios 7
- Operadores/as de instalaciones y máquinas y ensambladores/as 8
- Ocupaciones elementales 9
- Ocupaciones militares y cuerpos policiales 10
- Otra/o 11
- N.C. 99

P.33 ¿A qué clase social diría Ud. que pertenece? (RESPUESTA ESPONTÁNEA).
[CLASESOCIAL]
- Clase alta 1
- Clase media-alta 2
- Clase media-media 3
- Clase media-baja 4
- Clase trabajadora/obrera 5
- Clase baja 12
- Clase pobre 6
- Infraclase 7
- Proletariado 8
- A los/as de abajo 9
- Excluidos/as 10
- A la gente común 11
- Otra (especificar) 96
- No cree en las clases 97
- No sabe, duda 98
- N.C. 99

[CLASESOCIAL_COD]
_____

HEMOS TERMINADO. MUCHAS GRACIAS POR SU AMABILIDAD
Y POR EL TIEMPO QUE NOS HA DEDICADO.

B. Modelos.

Modelos kNN

Tabla 25. Modelos kNN.

Método de estandarización	Selección de variables	k	Nombre de modelo	Número de modelo
Estandarizado	Selección agresiva	3	knnselmarest3	Modelo 1
		5	knnselmarest5	Modelo 2
		7	knnselmarest7	Modelo 3
		9	knnselmarest9	Modelo 4
		11	knnselmarest11	Modelo 5
		15	knnselmarest15	Modelo 6

		21	knnselmarest21	Modelo 7
		25	knnselmarest25	Modelo 8
		31	knnselmarest31	Modelo 9
		37	knnselmarest37	Modelo 10
		45	knnselmarest45	Modelo 11
		55	knnselmarest55	Modelo 12
		65	knnselmarest65	Modelo 13
		75	knnselmarest75	Modelo 14
		85	knnselmarest85	Modelo 15
		95	knnselmarest95	Modelo 16
		105	knnselmarest105	Modelo 17
	Selección *Miner*	3	knnselvalest3	Modelo 18
		5	knnselvalest5	Modelo 19
		7	knnselvalest7	Modelo 20
		9	knnselvalest9	Modelo 21
		11	knnselvalest11	Modelo 22
		15	knnselvalest15	Modelo 23
		21	knnselvalest21	Modelo 24
		25	knnselvalest25	Modelo 25
		31	knnselvalest31	Modelo 26
		37	knnselvalest37	Modelo 27
		45	knnselvalest45	Modelo 28
		55	knnselvalest55	Modelo 29
		65	knnselvalest65	Modelo 30
		75	knnselvalest75	Modelo 31
		85	knnselvalest85	Modelo 32
		95	knnselvalest95	Modelo 33
		105	knnselvalest105	Modelo 34
	Todas las variables	3	knntodoest3	Modelo 35
		5	knntodoest5	Modelo 36
		7	knntodoest7	Modelo 37
		9	knntodoest9	Modelo 38
		11	knntodoest11	Modelo 39
		15	knntodoest15	Modelo 40
		21	knntodoest21	Modelo 41
		25	knntodoest25	Modelo 42
		31	knntodoest31	Modelo 43
		37	knntodoest37	Modelo 44
		45	knntodoest45	Modelo 45
		55	knntodoest55	Modelo 46
		65	knntodoest65	Modelo 47
		75	knntodoest75	Modelo 48
		85	knntodoest85	Modelo 49
		95	knntodoest95	Modelo 50
		105	knntodoest105	Modelo 51
Rango	Selección agresiva	3	knnselmarran3	Modelo 52
		5	knnselmarran5	Modelo 53
		7	knnselmarran7	Modelo 54
		9	knnselmarran9	Modelo 55
		11	knnselmarran11	Modelo 56
		15	knnselmarran15	Modelo 57
		21	knnselmarran21	Modelo 58
		25	knnselmarran25	Modelo 59
		31	knnselmarran31	Modelo 60
		37	knnselmarran37	Modelo 61
		45	knnselmarran45	Modelo 62
		55	knnselmarran55	Modelo 63
		65	knnselmarran65	Modelo 64
		75	knnselmarran75	Modelo 65
		85	knnselmarran85	Modelo 66
		95	knnselmarran95	Modelo 67
		105	knnselmarran105	Modelo 68
	Selección *Miner*	3	knnselvalran3	Modelo 69
		5	knnselvalran5	Modelo 70
		7	knnselvalran7	Modelo 71
		9	knnselvalran9	Modelo 72
		11	knnselvalran11	Modelo 73
		15	knnselvalran15	Modelo 74

		21	knnselvalran21	Modelo 75
		25	knnselvalran25	Modelo 76
		31	knnselvalran31	Modelo 77
		37	knnselvalran37	Modelo 78
		45	knnselvalran45	Modelo 79
		55	knnselvalran55	Modelo 80
		65	knnselvalran65	Modelo 81
		75	knnselvalran75	Modelo 82
		85	knnselvalran85	Modelo 83
		95	knnselvalran95	Modelo 84
		105	knnselvalran105	Modelo 85
	Todas las variables	3	knntodoran3	Modelo 86
		5	knntodoran5	Modelo 87
		7	knntodoran7	Modelo 88
		9	knntodoran9	Modelo 89
		11	knntodoran11	Modelo 90
		15	knntodoran15	Modelo 91
		21	knntodoran21	Modelo 92
		25	knntodoran25	Modelo 93
		31	knntodoran31	Modelo 94
		37	knntodoran37	Modelo 95
		45	knntodoran45	Modelo 96
		55	knntodoran55	Modelo 97
		65	knntodoran65	Modelo 98
		75	knntodoran75	Modelo 99
		85	knntodoran85	Modelo 100
		95	knntodoran95	Modelo 101
		105	knntodoran105	Modelo 102

Modelos Regresión Logística

Tabla 26. Modelos de regresión logística

Variables	¿*Dummies*?	Criterio de información	Criterio de selección	Modelo	Nº modelo
Originales	Sin *dummies*	AIC	*Backward*	RegOrigBackAIC	Modelo 103
			Stepwise	RegOrigStepAIC	Modelo 104
			Forward	RegOrigForwAIC	Modelo 105
		BIC	*Backward*	RegOrigBackBIC	Modelo 106
			Stepwise	RegOrigStepBIC	Modelo 107
			Forward	RegOrigForwBIC	Modelo 108
	Con *dummies*	AIC	*Backward*	RegOrigDumBackAIC	Modelo 109
			Stepwise	RegOrigDumStepAIC	Modelo 110
			Forward	RegOrigDumForwAIC	Modelo 111
		BIC	*Backward*	RegOrigDumBackBIC	Modelo 112
			Stepwise	RegOrigDumStepBIC	Modelo 113
			Forward	RegOrigDumForwBIC	Modelo 114
Originales y transformadas	Sin *dummies*	AIC	*Backward*	RegTransBackAIC	Modelo 115
			Stepwise	RegTransStepAIC	Modelo 116
			Forward	RegTransForwAIC	Modelo 117
		BIC	*Backward*	RegTransBackBIC	Modelo 118
			Stepwise	RegTransStepBIC	Modelo 119
			Forward	RegTransForwBIC	Modelo 120
	Con *dummies*	AIC	*Backward*	RegTransDumBackAIC	Modelo 121
			Stepwise	RegTransDumStepAIC	Modelo 122
			Forward	RegTransDumForwAIC	Modelo 123
		BIC	*Backward*	RegTransDumBackBIC	Modelo 124
			Stepwise	RegTransDumStepBIC	Modelo 125
			Forward	RegTransDumForwBIC	Modelo 126

Modelos Red Neuronal

Tabla 27. Modelos de red neuronal

Software	Sel. variables	Nodos	F. Activación	Algoritmo	*Early stop*	Learning rate	Nombre de modelo	Nº de modelo
SAS	*Miner*	2	Tanh	Levmar	Sí	-	redmin2ear	Modelo 127
					No	-	redmin2	Modelo 128
		4			Sí	-	redmin4ear	Modelo 129
					No	-	redmin4	Modelo 130
	Agresiva	2	Tanh	Levmar	Sí	-	redmar2ear	Modelo 131
					No	-	redmar2	Modelo 132
		3			Sí	-	redmar3ear	Modelo 133
					No	-	redmar3	Modelo 134
		5			Sí	-	redmar5ear	Modelo 135
					No	-	redmar5	Modelo 136
		11			Sí	-	redmar11ear	Modelo 137
					No	-	redmar11	Modelo 138
		17			Sí	-	redmar17ear	Modelo 139
					No	-	redmar17	Modelo 140
	Agr. *randomselect*	2	Tanh	Levmar	Sí	-	redagr2ear	Modelo 141
					No	-	redagr2	Modelo 142
		3			Sí	-	redagr3ear	Modelo 143
					No	-	redagr3	Modelo 144
		5			Sí	-	redagr5ear	Modelo 145
					No	-	redagr5	Modelo 146
		7			Sí	-	redagr7ear	Modelo 147
					No	-	redagr7	Modelo 148
		9			Sí	-	redagr9ear	Modelo 149
					No	-	redagr9	Modelo 150
		11			Sí	-	redagr11ear	Modelo 151
					No	-	redagr11	Modelo 152
		15			Sí	-	redagr15ear	Modelo 153
					No	-	redagr15	Modelo 154
		19			Sí	-	redagr19ear	Modelo 155
					No	-	redagr19	Modelo 156
		25			Sí	-	redagr25ear	Modelo 157
					No	-	redagr25	Modelo 158
		2	Log	Levmar	Sí	-	redagr2earlog	Modelo 159
			Arc			-	redagr2eararc	Modelo 160
			Sin			-	redagr2earsin	Modelo 161
			Sof			-	redagr2earsof	Modelo 162
			Gau			-	redagr2eargau	Modelo 163
			Sof	Qua		-	redagr2earsofqua	Modelo 164
				Bprop		-	redagr2earsofcon	Modelo 165
				Con		-	redagr2earsoftru	Modelo 166
				Trur		-	redagr2earsofcon	Modelo 167
				Dbl		-	redagr2earsofdbl	Modelo 168
R	Agr.*randomselect*	2	Tanh	Levmar		0.1	red2.0.1	Modelo 169
						0.01	red2.0.01	Modelo 170
						0.001	red2.0.001	Modelo 171
		3				0.1	red3.0.1	Modelo 172
						0.01	red3.0.01	Modelo 173
						0.001	red3.0.001	Modelo 174

Modelos Random Forest

Tabla 28. Modelos de random forest

Selección	Tamaño de hoja	Nº de variables	Tamaño de muestra	Nombre de modelo	Nº de modelo
Agresiva *radnomselect*	100	3	250	rf.agr.3.250	Modelo 175
			500	rf.agr.3.500	Modelo 176
			1000	rf.agr.3.1000	Modelo 177
			1500	rf.agr.3.1500	Modelo 178
			2500	rf.agr.3.2500	Modelo 179
			3500	rf.agr.3.3500	Modelo 180
			4400	rf.agr.3.4400	Modelo 181
		4	250	rf.agr.4.250	Modelo 182

			500	rf.agr.4.500	Modelo 183
			1000	rf.agr.4.1000	Modelo 184
			1500	rf.agr.4.1500	Modelo 185
			2500	rf.agr.4.2500	Modelo 186
			3500	rf.agr.4.3500	Modelo 187
			4400	rf.agr.4.4400	Modelo 188
		5	250	rf.agr.5.250	Modelo 189
			500	rf.agr.5.500	Modelo 190
			1000	rf.agr.5.1000	Modelo 191
			1500	rf.agr.5.1500	Modelo 192
			2500	rf.agr.5.2500	Modelo 193
			3500	rf.agr.5.3500	Modelo 194
			4400	rf.agr.5.4400	Modelo 195
		6	250	rf.agr.6.250	Modelo 196
			500	rf.agr.6.500	Modelo 197
			1000	rf.agr.6.1000	Modelo 198
			1500	rf.agr.6.1500	Modelo 199
			2500	rf.agr.6.2500	Modelo 200
			3500	rf.agr.6.3500	Modelo 201
			4400	rf.agr.6.4400	Modelo 202
		7	250	rf.agr.7.250	Modelo 203
			500	rf.agr.7.500	Modelo 204
			1000	rf.agr.7.1000	Modelo 205
			1500	rf.agr.7.1500	Modelo 206
			2500	rf.agr.7.2500	Modelo 207
			3500	rf.agr.7.3500	Modelo 208
			4400	rf.agr.7.4400	Modelo 209
		8	250	rf.agr.8.250	Modelo 210
			500	rf.agr.8.500	Modelo 211
			1000	rf.agr.8.1000	Modelo 212
			1500	rf.agr.8.1500	Modelo 213
			2500	rf.agr.8.2500	Modelo 214
			3500	rf.agr.8.3500	Modelo 215
			4400	rf.agr.8.4400	Modelo 216
		9	250	rf.agr.9.250	Modelo 217
			500	rf.agr.9.500	Modelo 218
			1000	rf.agr.9.1000	Modelo 219
			1500	rf.agr.9.1500	Modelo 220
			2500	rf.agr.9.2500	Modelo 221
			3500	rf.agr.9.3500	Modelo 222
			4400	rf.agr.9.4400	Modelo 223
		10 (*bagging*)	250	rf.agr.10.250	Modelo 224
			500	rf.agr.10.500	Modelo 225
			1000	rf.agr.10.1000	Modelo 226
			1500	rf.agr.10.1500	Modelo 227
			2500	rf.agr.10.2500	Modelo 228
			3500	rf.agr.10.3500	Modelo 229
			4400	rf.agr.10.4400	Modelo 230
Agresiva	100	3	250	rf.mar.3.250	Modelo 231
			500	rf.mar.3.500	Modelo 232
			1000	rf.mar.3.1000	Modelo 233
			1500	rf.mar.3.1500	Modelo 234
			2500	rf.mar.3.2500	Modelo 235
			3500	rf.mar.3.3500	Modelo 236
			4400	rf.mar.3.4400	Modelo 237
		4	250	rf.mar.4.250	Modelo 238
			500	rf.mar.4.500	Modelo 239
			1000	rf.mar.4.1000	Modelo 240
			1500	rf.mar.4.1500	Modelo 241
			2500	rf.mar.4.2500	Modelo 242
			3500	rf.mar.4.3500	Modelo 243
			4400	rf.mar.4.4400	Modelo 244
		5	250	rf.mar.5.250	Modelo 245
			500	rf.mar.5.500	Modelo 246
			1000	rf.mar.5.1000	Modelo 247
			1500	rf.mar.5.1500	Modelo 248
			2500	rf.mar.5.2500	Modelo 249
			3500	rf.mar.5.3500	Modelo 250

			4400	rf.mar.5.4400	Modelo 251
		6	250	rf.mar.6.250	Modelo 252
			500	rf.mar.6.500	Modelo 253
			1000	rf.mar.6.1000	Modelo 254
			1500	rf.mar.6.1500	Modelo 255
			2500	rf.mar.6.2500	Modelo 256
			3500	rf.mar.6.3500	Modelo 257
			4400	rf.mar.6.4400	Modelo 258
		7	250	rf.mar.7.250	Modelo 259
			500	rf.mar.7.500	Modelo 260
			1000	rf.mar.7.1000	Modelo 261
			1500	rf.mar.7.1500	Modelo 262
			2500	rf.mar.7.2500	Modelo 263
			3500	rf.mar.7.3500	Modelo 264
			4400	rf.mar.7.4400	Modelo 265
		8	250	rf.mar.8.250	Modelo 266
			500	rf.mar.8.500	Modelo 267
			1000	rf.mar.8.1000	Modelo 268
			1500	rf.mar.8.1500	Modelo 269
			2500	rf.mar.8.2500	Modelo 270
			3500	rf.mar.8.3500	Modelo 271
			4400	rf.mar.8.4400	Modelo 272
		9	250	rf.mar.9.250	Modelo 273
			500	rf.mar.9.500	Modelo 274
			1000	rf.mar.9.1000	Modelo 275
			1500	rf.mar.9.1500	Modelo 276
			2500	rf.mar.9.2500	Modelo 277
			3500	rf.mar.9.3500	Modelo 278
			4400	rf.mar.9.4400	Modelo 279
		10	250	rf.mar.10.250	Modelo 280
			500	rf.mar.10.500	Modelo 281
			1000	rf.mar.10.1000	Modelo 282
			1500	rf.mar.10.1500	Modelo 283
			2500	rf.mar.10.2500	Modelo 284
			3500	rf.mar.10.3500	Modelo 285
			4400	rf.mar.10.4400	Modelo 286
		11	250	rf.mar.11.250	Modelo 287
			500	rf.mar.11.500	Modelo 288
			1000	rf.mar.11.1000	Modelo 289
			1500	rf.mar.11.1500	Modelo 290
			2500	rf.mar.11.2500	Modelo 291
			3500	rf.mar.11.3500	Modelo 292
			4400	rf.mar.11.4400	Modelo 293
		12	250	rf.mar.12.250	Modelo 294
			500	rf.mar.12.500	Modelo 295
			1000	rf.mar.12.1000	Modelo 296
			1500	rf.mar.12.1500	Modelo 297
			2500	rf.mar.12.2500	Modelo 298
			3500	rf.mar.12.3500	Modelo 299
			4400	rf.mar.12.4400	Modelo 300
		13	250	rf.mar.13.250	Modelo 301
			500	rf.mar.13.500	Modelo 302
			1000	rf.mar.13.1000	Modelo 303
			1500	rf.mar.13.1500	Modelo 304
			2500	rf.mar.13.2500	Modelo 305
			3500	rf.mar.13.3500	Modelo 306
			4400	rf.mar.13.4400	Modelo 307
		14	250	rf.mar.14.250	Modelo 308
			500	rf.mar.14.500	Modelo 309
			1000	rf.mar.14.1000	Modelo 310
			1500	rf.mar.14.1500	Modelo 311
			2500	rf.mar.14.2500	Modelo 312
			3500	rf.mar.14.3500	Modelo 313
			4400	rf.mar.14.4400	Modelo 314
		15 (*bagging*)	250	rf.mar.15.250	Modelo 315
			500	rf.mar.15.500	Modelo 316
			1000	rf.mar.15.1000	Modelo 317
			1500	rf.mar.15.1500	Modelo 318

			2500	rf.mar.15.2500	Modelo 319
			3500	rf.mar.15.3500	Modelo 320
			4400	rf.mar.15.4400	Modelo 321
Selección *Miner*	100	5	250	rf.min.16.250	Modelo 322
			500	rf.min.16.500	Modelo 323
			1000	rf.min.16.1000	Modelo 324
			1500	rf.min.16.1500	Modelo 325
			2500	rf.min.16.2500	Modelo 326
			3500	rf.min.16.3500	Modelo 327
			4400	rf.min.16.4400	Modelo 328
		10	250	rf.min.17.250	Modelo 329
			500	rf.min.17.500	Modelo 330
			1000	rf.min.17.1000	Modelo 331
			1500	rf.min.17.1500	Modelo 332
			2500	rf.min.17.2500	Modelo 333
			3500	rf.min.17.3500	Modelo 334
			4400	rf.min.17.4400	Modelo 335
		20	250	rf.min.20.250	Modelo 336
			500	rf.min.20.500	Modelo 337
			1000	rf.min.20.1000	Modelo 338
			1500	rf.min.20.1500	Modelo 339
			2500	rf.min.20.2500	Modelo 340
			3500	rf.min.20.3500	Modelo 341
			4400	rf.min.20.4400	Modelo 342
		30	250	rf.min.30.250	Modelo 343
			500	rf.min.30.500	Modelo 344
			1000	rf.min.30.1000	Modelo 345
			1500	rf.min.30.1500	Modelo 346
			2500	rf.min.30.2500	Modelo 347
			3500	rf.min.30.3500	Modelo 348
			4400	rf.min.30.4400	Modelo 349
		40	250	rf.min.40.250	Modelo 350
			500	rf.min.40.500	Modelo 351
			1000	rf.min.40.1000	Modelo 352
			1500	rf.min.40.1500	Modelo 353
			2500	rf.min.40.2500	Modelo 354
			3500	rf.min.40.3500	Modelo 355
			4400	rf.min.40.4400	Modelo 356
		45	250	rf.min.45.250	Modelo 357
			500	rf.min.45.500	Modelo 358
			1000	rf.min.45.1000	Modelo 359
			1500	rf.min.45.1500	Modelo 360
			2500	rf.min.45.2500	Modelo 361
			3500	rf.min.45.3500	Modelo 362
			4400	rf.min.45.4400	Modelo 363
		50	250	rf.min.50.250	Modelo 364
			500	rf.min.50.500	Modelo 365
			1000	rf.min.50.1000	Modelo 366
			1500	rf.min.50.1500	Modelo 367
			2500	rf.min.50.2500	Modelo 368
			3500	rf.min.50.3500	Modelo 369
			4400	rf.min.50.4400	Modelo 370
		55	250	rf.min.55.250	Modelo 371
			500	rf.min.55.500	Modelo 372
			1000	rf.min.55.1000	Modelo 373
			1500	rf.min.55.1500	Modelo 374
			2500	rf.min.55.2500	Modelo 375
			3500	rf.min.55.3500	Modelo 376
			4400	rf.min.55.4400	Modelo 377
		58 (*bagging*)	250	rf.min.58.250	Modelo 378
			500	rf.min.58.500	Modelo 379
			1000	rf.min.58.1000	Modelo 380
			1500	rf.min.58.1500	Modelo 381
			2500	rf.min.58.2500	Modelo 382
			3500	rf.min.58.3500	Modelo 383
			4400	rf.min.58.4400	Modelo 384
	50	5	4400	rf.min.5.4400.50	Modelo 385
	20			rf.min.5.4400.20	Modelo 386

Modelos Gradient Boosting

Tabla 29. Modelos de gradient boosting

Algoritmo	Selección de variables	Tamaño de hoja	Número de árboles	*Shrinkage*	Porcentaje de variables	Nombre de modelo	Número de modelo
GBM	Agresiva *randomselect*	20	100	0.25	-	gbm.agr.20.100.025	Modelo 387
			500	0.01		gbm.agr.20.500.001	Modelo 388
			1000	0.03		gbm.agr.20.1000.003	Modelo 389
			1500	0.01		gbm.agr.20.1500.001	Modelo 390
		50	100	0.20		gbm.agr.50.100.020	Modelo 391
			500	0.05		gbm.agr.50.500.005	Modelo 392
			1000	0.03		gbm.agr.50.1000.003	Modelo 393
			1500	0.05		gbm.agr.50.1500.005	Modelo 394
		100	100	0.20		gbm.agr.100.100.025	Modelo 395
			500	0.03		gbm.agr.100.500.001	Modelo 396
			1000	0.03		gbm.agr.100.1000.003	Modelo 397
			1500	0.01		gbm.agr.100.1500.001	Modelo 398
	Agresiva	20	100	0.25		gbm.mar.20.100.025	Modelo 399
			500	0.03		gbm.mar.20.500.003	Modelo 400
			1000	0.05		gbm.mar.20.1000.005	Modelo 401
			2000	0.03		gbm.mar.20.2000.003	Modelo 402
		50	100	0.20		gbm.mar.50.100.020	Modelo 403
			500	0.03		gbm.mar.50.500.003	Modelo 404
			1000	0.05		gbm.mar.50.1000.005	Modelo 405
			2000	0.03		gbm.mar.50.2000.003	Modelo 406
		100	100	0.25		gbm.mar.100.100.025	Modelo 407
			500	0.10		gbm.mar.100.500.01	Modelo 408
			1000	0.05		gbm.mar.100.1000.005	Modelo 409
			2000	0.03		gbm.mar.100.2000.003	Modelo 410
	Miner	20	100	0.25		gbm.min.20.100.025	Modelo 411
			500	0.10		gbm.min.20.500.01	Modelo 412
			1000	0.01		gbm.min.20.1000.001	Modelo 413
			2000	0.01		gbm.min.20.2000.001	Modelo 414
			5000	0.001		gbm.min.20.5000.0001	Modelo 415
		50	100	0.05		gbm.min.50.100.005	Modelo 416
			500	0.10		gbm.min.50.500.01	Modelo 417
			1000	0.05		gbm.min.50.1000.005	Modelo 418
			2000	0.01		gbm.min.50.2000.001	Modelo 419
			5000	0.001		gbm.min.50.5000.0001	Modelo 420
		100	100	0.25		gbm.min.100.100.025	Modelo 421
			500	0.05		gbm.min.100.500.005	Modelo 422
			1000	0.03		gbm.min.100.1000.003	Modelo 423
			2000	0.01		gbm.min.100.2000.001	Modelo 424
			5000	0.01		gbm.min.100.5000.0001	Modelo 425
			1500	0.01		gbm.min.100.1500.001	Modelo 426
			2500	0.01		gbm.min.100.2500.001	Modelo 427
			1500	0.02		gbm.min.100.1500.002	Modelo 428
			1500	0.03		gbm.min.100.1500.003	Modelo 429
XGBoost	*Miner*	20	100	0.1	100%	xgb.min.20.100.01	Modelo 430
			500	0.03		xgb.min.20.500.003	Modelo 431
			1000	0.01		xgb.min.20.1000.001	Modelo 432
			2000	0.01		xgb.min.20.2000.001	Modelo 433
		50	100	0.2		xgb.min.50.100.020	Modelo 434
			500	0.05		xgb.min.50.500.005	Modelo 435
			1000	0.03		xgb.min.50.1000.003	Modelo 436
			2000	0.01		xgb.min.50.2000.001	Modelo 437
		20	500	0.03	5%	xgb.min.20.500.003.5	Modelo 438
					12.5%	xgb.min.20.500.003.125	Modelo 439
					25%	xgb.min.20.500.003.25	Modelo 440
					50%	xgb.min.20.500.003.50	Modelo 441
					75%	xgb.min.20.500.003.75	Modelo 442

			2000	0.01	5%	xgb.min.50.500.001.5	Modelo 443
					12.5%	xgb.min.50.500.001.125	Modelo 444
					25%	xgb.min.50.500.001.25	Modelo 445
					50%	xgb.min.50.500.001.50	Modelo 446
					75%	xgb.min.50.500.001.75	Modelo 447

Modelos Support Vector Machines

Tabla 30. Modelos de SVM

Tipo	Selección de variables	*Sigma*	C	Nombre de modelo	Número de modelo
Lineal	Agresiva *randomselect*	-	1	SVMlineal.agr	Modelo 447
	Agresiva			SVMlineal.mar	Modelo 448
	Miner			SVMlineal.min	Modelo 449
Radial/RBF	Agresiva *randomselect*	0.05	1	SVMRBFagr.005.1	Modelo 450
		0.01	1	SVMRBFagr.001.1	Modelo 451
		0.005	2	SVMRBFagr.0005.2	Modelo 452
		0.0001	10000	SVMRBFagr.0001.10000	Modelo 453
	Agresiva	0.05	0.5	SVMRBFmar.005.05	Modelo 454
		0.01	2	SVMRBFmar.001.2	Modelo 455
		0.005	10	SVMRBFmar.0005.10	Modelo 456
		0.0001	10000	SVMRBFmar.0001.10000	Modelo 457
	Miner	0.05	1	SVMRBFmin.005.1	Modelo 458
		0.01	1	SVMRBFmin.001.1	Modelo 459
		0.005	2	SVMRBFmin.0005.2	Modelo 460
		0.0001	10000	SVMRBFmin.0001.10000	Modelo 461

C. Código

Código SAS®

```
%macro
randomselectlog(data=,listclass=,vardepen=,modelo=,sinicio=,sfinal=,fr
acciontrain=,directorio=);
options nocenter linesize=256;
proc printto print="&directorio\kk.txt";run;
data;file "&directorio\cosa2.txt" ;run;
%do semilla=&sinicio %to &sfinal;
proc surveyselect data=&data rate=&fracciontrain out=sal1234
seed=&semilla;run;

%if &listclass ne %then %do;
ods output type3=parametros;
proc logistic data=sal1234;
      class &listclass;
    model &vardepen= &modelo/ selection=stepwise;
run;
data parametros;length effect $20. modelo $ 20000;retain modelo "
";set parametros end=fin;effect=cat(' ',effect);
if _n_ ne 1 then modelo=catt(modelo,' ',effect);if fin then
do;variable=modelo;output;end;
run;
%end;
%else %do;
ods output  Logistic.ParameterEstimates=parametros;
proc logistic data=sal1234;
        model &vardepen= &modelo/ selection=stepwise;
run;
%end;
ods graphics off;
ods html close;
data;file "&directorio\cosa2.txt" mod;set parametros;
```

```
%if &listclass ne %then %do; put variable @@;%end;
%else %do; if _n_ ne 1 then put variable @@;%end;
run;
%end;
proc printto ;run;
data todos;
infile "&directorio\cosa2.txt";
length efecto $ 400;
input efecto @@;
if efecto ne 'Intercept' then output;
run;
proc freq data=todos;tables efecto /out=sal;run;
proc sort data=sal;by descending count;
proc print data=sal;run;

data todos;
infile "&directorio\cosa2.txt";
length efecto $ 200;
input efecto $ &&;
run;
proc freq data=todos;tables efecto /out=sal;run;
proc sort data=sal;by descending count;
proc print data=sal;run;
data;set sal;put efecto;run;
%mend;


%macro
cruzadalogisticabis(archivo=,vardepen=,conti=,categor=,ngrupos=,sinici
o=,sfinal=,objetivos=,corte=0.5,porcaptura=0,directorio=c:\Windows\Tem
p);
data final;run;
/* contar objetivos */
data _null_;length clase $ 300;
clase="&objetivos";
  nobje= 1;
  do while (scanq(clase, nobje) ^= '');
    nobje+1;
  end;
  nobje+(-1);
  call symput('nobje',left(nobje));
run;
proc printto print="&directorio\outp.txt"
log="&directorio\log.txt";run;/*SE PUEDE QUITAR EL PROC PRINTTO, POR
SI ACASO HAY PROBLEMAS*/
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
      data dos;set &archivo;u=ranuni(&semilla);
      proc sort data=dos;by u;run;
      data dos ;
      retain grupo 1;
      set dos nobs=nume;
      if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
      run;
      data fantasma;run;
      %do exclu=1 %to &ngrupos;
            data tres;set dos;if grupo ne &exclu then vardep=&vardepen;
            ods output ROCAssociation=roca;
            proc logistic data=tres  ROCOPTIONS(NODETAILS)
PLOTS=NONE;/*<<<<<******SE PUEDE QUITAR EL NOPRINT */
```

```
            %if (&categor ne) %then %do;class &categor;model
vardep=&conti;%end;
            %else %do;model vardep=&conti;%end;
            output out=sal p=predi;roc;run;
            data sal2;set sal;pro=1-predi;if pro>&corte then pre11=1;
else pre11=0;
            if grupo=&exclu then output;run;
            proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;
            data estadisticos (drop=count percent pre11 &vardepen);
            retain vp vn fp fn suma 0;
            if _n_=1 then set roca;
            set sal3 nobs=nume;
            suma=suma+count;
            if pre11=0 and &vardepen=0 then vn=count;
            if pre11=0 and &vardepen=1 then fn=count;
            if pre11=1 and &vardepen=0 then fp=count;
            if pre11=1 and &vardepen=1 then vp=count;
            if _n_=nume then do;
            porcenVN=vn/suma;
            porcenFN=FN/suma;
            porcenVP=VP/suma;
            porcenFP=FP/suma;
            sensi=vp/(vp+fn);
            especif=vn/(vn+fp);
            tasafallos=1-(vp+vn)/suma;
            tasaciertos=1-tasafallos;
            precision=vp/(vp+fp);
            F_M=2*Sensi*Precision/(Sensi+Precision);
            Mcc=VP*VN-
FP*FN;be=(VP+FP)*(VP+FN)*(VN+FP)*(VN+FN);be=sqrt(be);
            MCC=MCC/be;
            Youden=especif+sensi-1;
            AUC=Area;
            output;
            end;
            run;

            %if &porcaptura ne 0 %then %do;
            proc sort data=sal2;by descending pre11;
            data sal4;retain suma1 0;set sal2 nobs=nume;
                  if &vardepen=1 then suma1=suma1+1;
                  if _n_=int(&porcaptura*nume) then
do;ncapturados=suma1;capturados=suma1/_n_;ntot=_n_;output;
                  stop;end;
            run;
            data estadisticos;set estadisticos;if _n_=1 then set
sal4;run;
            %end;

data estadisticos;set estadisticos;
keep AUC F_M Mcc Youden ncapturados ntot capturados especif fn fp
porcenFN porcenFP porcenVN porcenVP precision
sensi tasaciertos tasafallos vn vp;
run;
      data fantasma;set fantasma estadisticos;run;
      %end;
      proc means data=fantasma sum noprint;var &objetivos;
      output out=sumaresi sum=suma mean=media1-media&nobje;
      run;
      data sumaresi;set sumaresi;semilla=&semilla;
```

```
      data final (keep=suma media1-media&nobje semilla);set final
sumaresi;if suma=. then delete;run;
      /* renombrar objetivos para entender mejor */
%end;
      data _null_;
      file "&directorio\kk.txt";
      put 'data final;set final;array media{'"&nobje" '};';
      %do i=1 %to &nobje;
      %let vari=%qscan(&objetivos,&i);
      put "&vari" '=media{'"&i"'};';
      %end;
      put 'drop suma media1-media'"&nobje"';output;run;';
      run;

      %include "&directorio\kk.txt";
      proc printto;run;
      proc print data=final;run;
%mend;


%macro
redneuronalbinaria(archivo=,listclass=,listconti=,vardep=,porcen=,semi
lla=,ocultos=,meto=levmar,acti=);

PROC DMDB DATA=&archivo dmdbcat=catauno;
target &vardep;
var &listconti ;
class &vardep &listclass;
run;

data ooo;set &archivo;run;
data datos;set ooo nobs=nume;tr=int(&porcen*nume);call
symput('tr',left(tr));u=ranuni(&semilla);run;
proc sort data=datos;by u;run;
data datos valida;set datos;if _n_>tr then output valida;else output
datos;run;

proc neural data=datos dmdbcat=catauno validata=valida graph;
input &listconti / id=i;
input &listclass / level=nominal;
target &vardep / level=nominal id=o error=ENT;
hidden &ocultos / id=h act=&acti;
nloptions maxiter=10000;
netoptions randist=normal ranscale=0.1 random=15115;
prelim 0;
train maxiter=10000 outest=mlpest estiter=1 technique=&meto;
score data=datos out=mlpout outfit=mlpfit;
score data=valida out=mlpout2 outfit=mlpfit2 role=valid;
run;

data mlpest2 ;
k=3;
retain iterepocas 0;
set mlpest nobs=nume;
call symput('numeroit',left(nume));
eval=_VOBJERR_;
x3=lag3(eval);
x6=lag6(eval);
if _n_>6 and eval>x3 and eval>x6 then iterepocas=_n_;
run;
```

```
data;
set mlpest2 nobs=nume;
if iterepocas ne 0 then do;
call symput('earlystop',left(iterepocas));
stop;
end;
else if _n_=nume and iterepocas=0 then do;iterepocas=&numeroit;
call symput('earlystop',left(iterepocas));
stop;
end;
run;

data fin;j=&earlystop;set mlpest point=j;output;stop;run;

data mlpest;set mlpest nobs=nume; if _n_=&earlystop then do;
cosa1=put(_OBJERR_,20.6) ;
cosa2=put(_VOBJERR_,20.6) ;
end;
else do;cosa1=' ';cosa2=' ';end;
run;

title1
h=2 box=1 j=c c=red 'TRAIN' c=blue '  VALIDA'
h=1.5 j=c c=black "EARLY STOPPING=&earlystop " "semilla=&semilla"
h=1 j=c c=green "NODOS OCULTOS: &ocultos  " " METODO: &meto "
"ACTIVACIÓN: &acti"
h=1 j=c c=black "EL ERROR ES EL VALOR DE LA ENTROPÍA";
;

symbol1 c=red v=circle i=join pointlabel=("#cosa1" h=1 c=red
position=bottom  j=c);
symbol2 c=blue v=circle i=join pointlabel=("#cosa2" h=1 c=blue
position=top j=c);

axis1 label=none;
proc gplot data=mlpest;plot _OBJERR_ *_iter_=1 _VOBJERR_*_iter_=2
/overlay href=&earlystop vaxis=axis1 haxis=axis1 ;run;

proc print data=fin;
var  _iter_ _OBJERR_  _AVERR_  _VNOBJ_   _VOBJ_  _VOBJERR_  _VAVERR_
;run;

%mend;


%macro cruzadabinarianeuralearly(archivo=,vardepen=,
conti=,categor=,ngrupos=,sinicio=,sfinal=,
nodos=,meto=levmar,objetivo=tasafallos,directorio=c:,early=);
data final;run;
proc printto print="directorio\basura.txt"; run;

/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
      data dos;set &archivo;u=ranuni(&semilla);
      proc sort data=dos;by u;run;
      data dos (drop=nume);
      retain grupo 1;
      set dos nobs=nume;
      if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
      run;
      data fantasma;run;
```

```
      %do exclu=1 %to &ngrupos;

      data trestr tresval;
            set dos;if grupo ne &exclu then output trestr;else output
tresval;
            PROC DMDB DATA=trestr dmdbcat=catatres;
            target &vardepen;
            var &conti;
            class &vardepen;
            %if &categor ne %then %do;class &categor &vardepen;%end;
            run;
            proc neural data=trestr dmdbcat=catatres random=789 ;
            input &conti;
            %if &categor ne %then %do;input &categor
/level=nominal;%end;
            target &vardepen /level=nominal;
            hidden &nodos /act=tanh;/*<<<<<******PARA DATOS LINEALES
ACT=LIN (funci?nde activaci?n lineal)
            NORMALMENTE PARA DATOS NO LINEALES MEJOR ACT=TANH */
            /* A PARTIR DE AQUI SON ESPECIFICACIONES DE LA RED, SE
PUEDEN CAMBIAR O A?ADIR COMO PAR?METROS */

            /*nloptions maxiter=500*/;
            netoptions randist=normal ranscale=0.15 random=15459;
            train maxiter=&early outest=mlpest technique=&meto;
            score data=tresval role=valid out=sal ;
            run;
            data sal2;set sal;pro=1-%str(p_&vardepen)0;if pro>0.5 then
pre11=1; else pre11=0;run;
                        proc freq data=sal2;tables
pre11*&vardepen/out=sal3;run;

      data estadisticos (drop=count percent pre11 &vardepen);
            retain vp vn fp fn suma 0;
            set sal3 nobs=nume;
            suma=suma+count;
            if pre11=0 and &vardepen=0 then vn=count;
            if pre11=0 and &vardepen=1 then fn=count;
            if pre11=1 and &vardepen=0 then fp=count;
            if pre11=1 and &vardepen=1 then vp=count;
            if _n_=nume then do;
            /* porcenVN=vn/suma;
            porcenFN=FN/suma;
            porcenVP=VP/suma;
            porcenFP=FP/suma;
            sensi=vp/(vp+fn);
            especif=vn/(vn+fp);
            precision=vp/(vp+fp);
            F_M=2*Sensi*Precision/(Sensi+Precision);
            */
            tasafallos=1-(vp+vn)/suma;
            tasaciertos=1-tasafallos;
            output;
            end;
            run;
      data fantasma;set fantasma estadisticos;run;
      %end;
      proc means data=fantasma sum noprint;var &objetivo;
      output out=sumaresi sum=suma mean=media;
      run;
      data sumaresi;set sumaresi;semilla=&semilla;
```

```
      data final (keep=suma media semilla);set final sumaresi;if
suma=. then delete;run;
%end;
proc printto ;
proc print data=final;run;
%mend;

%macro
cruzadabinarianeural(archivo=,vardepen=,conti=,categor=,ngrupos=,sinic
io=,sfinal=,nodos=,algo=,objetivo=tasafallos,
early=500,acti=tanh,directorio=c:);
title ' ';
data final;run;
proc printto print="&directorio\basura.txt";

/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
      data dos;set &archivo;u=ranuni(&semilla);
      proc sort data=dos;by u;run;
      data dos (drop=nume);
      retain grupo 1;
      set dos nobs=nume;
      if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
      run;
      data fantasma;run;
      %do exclu=1 %to &ngrupos;

      data trestr tresval;
            set dos;if grupo ne &exclu then output trestr;else output
tresval;
            PROC DMDB DATA=trestr dmdbcat=catatres;
            target &vardepen;
            var &conti;
            class &vardepen;
            %if &categor ne %then %do;class &categor &vardepen;%end;
            run;
            proc neural data=trestr dmdbcat=catatres random=789 ;
            input &conti;
            %if &categor ne %then %do;input &categor
/level=nominal;%end;
            target &vardepen /level=nominal;
            hidden &nodos /act=&acti;/*<<<<<******PARA DATOS LINEALES
ACT=LIN (funciónde activación lineal)
            NORMALMENTE PARA DATOS NO LINEALES MEJOR ACT=TANH */
            /* A PARTIR DE AQUI SON ESPECIFICACIONES DE LA RED, SE
PUEDEN CAMBIAR O AÑADIR COMO PARÁMETROS */

            /*nloptions maxiter=500*/;
            netoptions randist=normal ranscale=0.15 random=15459;
            /* Si se desea hacer early stopping se pone prelim 0 y se
marca como comentario
            la línea prelim 15...*/
            /*prelim 0 */
            prelim 15 preiter=10 pretech=&algo;
            train maxiter=&early outest=mlpest technique=&algo;
            score data=tresval role=valid out=sal ;
            run;
            data sal2;set sal;pro=1-%str(p_&vardepen)0;if pro>0.5 then
pre11=1; else pre11=0;run;
                        proc freq data=sal2;tables
pre11*&vardepen/out=sal3;run;
```

```
      data estadisticos (drop=count percent pre11 &vardepen);
            retain vp vn fp fn suma 0;
            set sal3 nobs=nume;
            suma=suma+count;
            if pre11=0 and &vardepen=0 then vn=count;
            if pre11=0 and &vardepen=1 then fn=count;
            if pre11=1 and &vardepen=0 then fp=count;
            if pre11=1 and &vardepen=1 then vp=count;
            if _n_=nume then do;
            porcenVN=vn/suma;
            porcenFN=FN/suma;
            porcenVP=VP/suma;
            porcenFP=FP/suma;
            sensi=vp/(vp+fn);
            especif=vn/(vn+fp);
            tasafallos=1-(vp+vn)/suma;
            tasaciertos=1-tasafallos;
            precision=vp/(vp+fp);
            F_M=2*Sensi*Precision/(Sensi+Precision);
            output;
            end;
            run;
      data fantasma;set fantasma estadisticos;run;
      %end;
      proc means data=fantasma sum noprint;var &objetivo;
      output out=sumaresi sum=suma mean=media;
      run;
      data sumaresi;set sumaresi;semilla=&semilla;
      data final (keep=suma media semilla);set final sumaresi;if
suma=. then delete;run;
%end;
proc printto ;
proc print data=final;run;
%mend;

%macro activalbinariacruza;
%let lista='TANH LOG ARC LIN SIN SOF GAU';
%let nume=6;
%do  i=1 %to &nume;
data _null_;activa=scanq(&lista,&i);call
symput('activa',left(activa));run;
 %cruzadabinarianeural(archivo=uno,vardepen=y,conti=x
z,categor=clase,ngrupos=3,sinicio=12345,sfinal=12347,nodos=10,acti=&ac
tiva);
 data final&i;set final;modelo="&activa";put modelo=;run;
%end;
data union;set %do i=1 %to &nume; final&i %end;
%mend;
```

```
libname tfm "C:\Users\mario\Desktop\TFM"; run;
data tfm;set tfm.redmario2_train; run;
proc print data=tfm; run;
proc dmdb data=tfm; run;
%randomselectlog (data=tfm, listclass=, vardepen=OPOSVACUNA,
modelo=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
```

```
sinicio=12345, sfinal=13444, fracciontrain=0.8,
directorio=C:\Users\mario\Desktop\TFM)
x1=PREF_PRES_No_consta; run;

/* TRABAJO CON REDMARIO REDUCIDO SEGUN RANDOMSELECT*/

/*%macro
cruzadabinarianeural(archivo=,vardepen=,conti=,categor=,ngrupos=,sinic
io=,sfinal=,nodos=,algo=,objetivo=tasafallos,
early=500,acti=tanh,directorio=c:);*/

/*PRIMER PASO: INVESTIGAR UN POCO LAS FUNCIONES DE ACTIVACIÓN/
TRABAJAR CON TODAS ES DEMASIADO ------- ¡¡¡no funciona!!!*/

proc printto log="C:\Users\mario\Desktop\TFM\salida.txt"; run;

%macro activalbinariacruza;
%let lista='TANH LOG ARC SIN SOF GAU';
%let nume=6;
%do  i=1 %to &nume;
data _null_;activa=scanq(&lista,&i);call
symput('activa',left(activa));run;
%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=, ngrupos=5,sinicio=12345,
sfinal=12354,nodos=3,algo=levmar,objetivo=tasafallos,acti=&activa,earl
y=500,directorio=C:\Users\mario\Desktop\TFM);
data final&i;set final;modelo="&activa";put modelo=;run;
%end;
data union;set %do i=1 %to &nume; final&i %end;
%mend;

%activalbinariacruza;run;
data "C:\Users\mario\Desktop\TFM\CARPETA REDES\activa3.sas7bdat";
  set tfm.union;
run;
proc boxplot data=union; plot media*modelo;run;
ods show;
%macro activalbinariacruza;
%let lista='TANH LOG ARC SIN SOF GAU';
%let nume=6;
%do  i=1 %to &nume;
data _null_;activa=scanq(&lista,&i);call
symput('activa',left(activa));run;
%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=, ngrupos=5,sinicio=12345,
sfinal=12354,nodos=9,algo=levmar,objetivo=tasafallos,acti=&activa,earl
y=500,directorio=C:\Users\mario\Desktop\TFM);
data final&i;set final;modelo="&activa";put modelo=;run;
%end;
data union;set %do i=1 %to &nume; final&i %end;
%mend;
```

```
data activa9;set union;
%activalbinariacruza;run;
data "C:\Users\mario\Desktop\TFM\CARPETA REDES\activa9.sas7bdat";
  set tfm.activa9;
run;
proc boxplot data=union; plot media*modelo;run;

%macro activalbinariacruza;
%let lista='TANH LOG ARC SIN SOF GAU';
%let nume=6;
%do  i=1 %to &nume;
data _null_;activa=scanq(&lista,&i);call
symput('activa',left(activa));run;
%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=, ngrupos=5,sinicio=12345,
sfinal=12354,nodos=15,algo=levmar,objetivo=tasafallos,acti=&activa,ear
ly=500,directorio=C:\Users\mario\Desktop\TFM);
data final&i;set final;modelo="&activa";put modelo=;run;
%end;
data union;set %do i=1 %to &nume; final&i %end;
%mend;

%activalbinariacruza;run;
data "C:\Users\mario\Desktop\TFM\activa15.sas7bdat";
  set tfm.union;
run;
proc boxplot data=union; plot media*modelo;run;

%macro activalbinariacruza;
%let lista='TANH LOG ARC SIN SOF GAU';
%let nume=6;
%do  i=1 %to &nume;
data _null_;activa=scanq(&lista,&i);call
symput('activa',left(activa));run;
%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=, ngrupos=5,sinicio=12345,
sfinal=12354,nodos=25,algo=levmar,objetivo=tasafallos,acti=&activa,ear
ly=500,directorio=C:\Users\mario\Desktop\TFM);
data final&i;set final;modelo="&activa";put modelo=;run;
%end;
data union;set %do i=1 %to &nume; final&i %end;
%mend;

%activalbinariacruza;run;
data "C:\Users\mario\Desktop\TFM\activa25.sas7bdat";
  set tfm.union;
run;
proc boxplot data=union; plot media*modelo;run;

libname tfmbis "C:\Users\mario\Desktop\TFM\CARPETA REDES"; run;
data union; set tfmbis.activa3; run;
```

```
proc boxplot data=union; plot media*modelo;run;


/*COMPROBAR EARLY STOPPINGS ---- CON DOS SEMILLAS PARA ASEGURAR*/

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=2,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12344,ocultos=3,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=5,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=7,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=9,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=11,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=15,meto=levmar,acti=tanh)
```

```
%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=19,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_No_cons
ta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=25,meto=levmar,acti=tanh)

/*VALIDACION CRUZADA CON EARLY Y SIN EARLY PARA 2,3,5,7,9,11,15,19,25
NODOS Y FUNCIONES DE ACTIVACIÓN...*/

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final1;set final;modelo="AGR-2-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear2.sas7bdat";
  set final1;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=500,
directorio=C:\Users\mario\Desktop\TFM
);

data final2;set final;modelo="AGR-2-TANH   ";
```

```
data "C:\Users\mario\Desktop\TFM\redagrtanh2.sas7bdat";
  set final2;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=3,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=11,
directorio=C:\Users\mario\Desktop\TFM
);

data final3;set final;modelo="AGR-3-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear3.sas7bdat";
  set final3;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=3,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final4;set final;modelo="AGR-3-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanh3.sas7bdat";
  set final4;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=5,
```

```
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=10,
directorio=C:\Users\mario\Desktop\TFM
);

data final5;set final;modelo="AGR-5-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear5.sas7bdat";
  set final5;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=5,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final6;set final;modelo="AGR-5-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanh5.sas7bdat";
  set final6;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=7,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=9,
directorio=C:\Users\mario\Desktop\TFM
);

data final7;set final;modelo="AGR-7-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear7.sas7bdat";
  set final7;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
```

```
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=7,
sinicio=12345,
sfinal=12394,
nodos=7,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final8;set final;modelo="AGR-7-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanh7.sas7bdat";
  set final8;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=9,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=10,
directorio=C:\Users\mario\Desktop\TFM
);

data final9;set final;modelo="AGR-9-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear9.sas7bdat";
  set final9;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=9,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);
```

```
data final10;set final;modelo="AGR-9-TANH    ";
data "C:\Users\mario\Desktop\TFM\redagrtanh9.sas7bdat";
  set final10;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=11,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=8,
directorio=C:\Users\mario\Desktop\TFM
);

data final11;set final;modelo="AGR-11-EAR-TANH    ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear11.sas7bdat";
  set final11;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=11,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final12;set final;modelo="AGR-11-TANH    ";
data "C:\Users\mario\Desktop\TFM\redagrtanh11.sas7bdat";
  set final12;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
```

```
nodos=15,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=8,
directorio=C:\Users\mario\Desktop\TFM
);

data final13;set final;modelo="AGR-15-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear15.sas7bdat";
  set final13;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=15,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final14;set final;modelo="AGR-15-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanh15.sas7bdat";
  set final14;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=19,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=8,
directorio=C:\Users\mario\Desktop\TFM
);

data final15;set final;modelo="AGR-19-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear19.sas7bdat";
  set final15;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
```

```
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=19,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final16;set final;modelo="AGR-19-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanh19.sas7bdat";
  set final16;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=25,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=8,
directorio=C:\Users\mario\Desktop\TFM
);

data final17;set final;modelo="AGR-25-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanhear25.sas7bdat";
  set final17;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=25,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);
```

```
data final18;set final;modelo="AGR-25-TANH   ";
data "C:\Users\mario\Desktop\TFM\redagrtanh25.sas7bdat";
  set final18;
run;


data final1;set tfm.redagrtanhear2; run;
data final2;set tfm.redagrtanh2; run;
data final3;set tfm.redagrtanhear3; run;
data final4;set tfm.redagrtanh3; run;

data final5;set tfm.redagrtanhear5; run;
data final6;set tfm.redagrtanh5; run;
data final7;set tfm.redagrtanhear7; run;
data final8;set tfm.redagrtanh7; run;


data final9;set tfm.redagrtanhear9; run;
data final10;set tfm.redagrtanh9; run;
data final11;set tfm.redagrtanhear11; run;
data final12;set tfm.redagrtanh11; run;

data final13;set tfm.redagrtanhear15; run;
data final14;set tfm.redagrtanh15; run;
data final15;set tfm.redagrtanhear19; run;
data final16;set tfm.redagrtanh19; run;

data final17;set tfm.redagrtanhear25; run;
data final18;set tfm.redagrtanh25; run;

data union; set final1 final2 final3 final4 final5 final6 final7
final8 final9 final10 final11 final12 final13 final14 final15 final16
final17 final18; run;
proc boxplot data=union; plot media*modelo;run;


/*REDES CON AGRESIVA - PRIMERO VER EL EARLY STOPPING*/

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_Ines_Ar
rimadas PREF_PRES_inigo_Errejon PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12344,ocultos=2,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_Ines_Ar
rimadas PREF_PRES_inigo_Errejon PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=3,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_Ines_Ar
rimadas PREF_PRES_inigo_Errejon PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
```

```
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12344,ocultos=5,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_Ines_Ar
rimadas PREF_PRES_inigo_Errejon PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=11,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=PREF_PRES_Ines_Ar
rimadas PREF_PRES_inigo_Errejon PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12344,ocultos=17,meto=levmar,acti=tanh)

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=17,
directorio=C:\Users\mario\Desktop\TFM
);

data final19;set final;modelo="MAR-2-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmartanhear2.sas7bdat";
  set final19;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
```

```
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final20;set final;modelo="MAR-2-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmartanh2.sas7bdat";
  set final20;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=3,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=11,
directorio=C:\Users\mario\Desktop\TFM
);

data final21;set final;modelo="MAR-3-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmartanhear3.sas7bdat";
  set final21;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=3,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final22;set final;modelo="MAR-3-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmartanh3.sas7bdat";
  set final22;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
```

```
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=5,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=10,
directorio=C:\Users\mario\Desktop\TFM
);

data final23;set final;modelo="MAR-5-EAR-TANH    ";
data "C:\Users\mario\Desktop\TFM\redmartanhear5.sas7bdat";
  set final23;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=5,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final24;set final;modelo="MAR-5-TANH    ";
data "C:\Users\mario\Desktop\TFM\redmartanh5.sas7bdat";
  set final24;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=11,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
```

```
early=9,
directorio=C:\Users\mario\Desktop\TFM
);

data final25;set final;modelo="MAR-11-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmartanhear11.sas7bdat";
  set final25;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=11,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final26;set final;modelo="MAR-11-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmartanh11.sas7bdat";
  set final26;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=17,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=8,
directorio=C:\Users\mario\Desktop\TFM
);

data final27;set final;modelo="MAR-17-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmartanhear17.sas7bdat";
  set final27;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_Ines_Arrimadas PREF_PRES_inigo_Errejon
PREF_PRES_No_consta PREF_PRES_Otro PREF_PRES_Pablo_Casado
```

```
PREF_PRES_Pablo_Iglesias PREF_PRES_Pedro_Sanchez
PREF_PRES_Santiago_Abascal EFEC_COVID_Economia EFEC_COVID_Salud
PREO_COVID_Mucho PREO_COVID_Nada PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=11,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final28;set final;modelo="MAR-17-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmartanh17.sas7bdat";
  set final28;
run;

data tfm; set tfm.redsel1_train; run;
proc print data=tfm; run;

%redneuronalbinaria(archivo=tfm,listclass=,listconti=CIVIS_COVID_DUDA
CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia     EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=2,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=CIVIS_COVID_DUDA
CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia     EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
```

```
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12344,ocultos=3,meto=levmar,acti=tanh)

%redneuronalbinaria(archivo=tfm,listclass=,listconti=CIVIS_COVID_DUDA
CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia     EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
vardep=OPOSVACUNA,
porcen=0.8,semilla=12345,ocultos=5,meto=levmar,acti=tanh)

/*EMPIEZO CRUZADAS SELVEL */

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=CIVIS_COVID_DUDA CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia
      EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
```

```
early=10,
directorio=C:\Users\mario\Desktop\TFM
);

data final29;set final;modelo="MIN-2-EAR-TANH    ";
data "C:\Users\mario\Desktop\TFM\redmintanhear2.sas7bdat";
  set final29;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=CIVIS_COVID_DUDA CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia
      EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final30;set final;modelo="MIN-2-TANH    ";
data "C:\Users\mario\Desktop\TFM\redmintanh2.sas7bdat";
  set final30;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=CIVIS_COVID_DUDA CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia
      EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
```

```
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=4,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=9,
directorio=C:\Users\mario\Desktop\TFM
);

data final31;set final;modelo="MIN-4-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmintanhear4.sas7bdat";
  set final31;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=CIVIS_COVID_DUDA CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia
      EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=4,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final32;set final;modelo="MIN-4-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmintanh4.sas7bdat";
  set final32;
```

```
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=CIVIS_COVID_DUDA CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia
      EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=5,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=9,
directorio=C:\Users\mario\Desktop\TFM
);

data final33;set final;modelo="MIN-5-EAR-TANH   ";
data "C:\Users\mario\Desktop\TFM\redmintanhear5.sas7bdat";
  set final33;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=CIVIS_COVID_DUDA CIVIS_COVID_NO_CIVISMO EFEC_COVID_Economia
      EFEC_COVID_Salud
PARTICIPACIONG_Si G_PREF_PRES_1     G_PREF_PRES_2     G_PREF_PRES_3
      G_PREO_COVID_1    G_PREO_COVID_2    G_INTENCIONG_1
      G_INTENCIONG_2
G_INTENCIONG_3    G_RECUVOTOG_1     G_RECUVOTOG_2     G_RECUVOTOG_3
      G_RECUVOTOG_4     G_PRO_PRI_1 G_PRO_PRI_2 G_PRO_PRI_3 G_SITLAB_1
      G_SITLAB_2
G_SITLAB_3  G_CNO11_1   G_CNO11_2   G_CNO11_3   G_CNO11_4   G_VAL_ECO_1
      G_VAL_ECO_2 G_PRO_SOC_1 G_PRO_SOC_2 G_PRO_SOC_3 G_PRO_SOC_4
      G_ECIVIL_1
G_ECIVIL_2  G_INTENCIONGALTER_1     G_INTENCIONGALTER_2
      G_INTENCIONGALTER_3     G_INTENCIONGALTER_4
      G_NIVELESTENTREV_1      G_NIVELESTENTREV_2
G_NIVELESTENTREV_3      G_NIVELESTENTREV_4      G_NIVELESTENTREV_5
      G_VAL_ECO_PER_1   G_VAL_ECO_PER_2   G_VAL_ECO_PER_3   G_CCAA_1
      G_CCAA_2
G_CCAA_3    G_CCAA_4    G_CCAA_5    G_GOB_ENCAR_1     G_GOB_ENCAR_2
      G_GOB_ENCAR_3     VAL_PS      EDAD  numMissing,
```

```
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=5,
algo=LEVMAR,
objetivo=tasafallos,
acti=tanh,
early=200,
directorio=C:\Users\mario\Desktop\TFM
);

data final34;set final;modelo="MIN-5-TANH    ";
data "C:\Users\mario\Desktop\TFM\redmintanh5.sas7bdat";
  set final34;
run;



data final1;set tfm.redagrtanhear2; run;
data final2;set tfm.redagrtanh2; run;
data final3;set tfm.redagrtanhear3; run;
data final4;set tfm.redagrtanh3; run;

data final5;set tfm.redagrtanhear5; run;
data final6;set tfm.redagrtanh5; run;
data final7;set tfm.redagrtanhear7; run;
data final8;set tfm.redagrtanh7; run;


data final9;set tfm.redagrtanhear9; run;
data final10;set tfm.redagrtanh9; run;
data final11;set tfm.redagrtanhear11; run;
data final12;set tfm.redagrtanh11; run;

data final13;set tfm.redagrtanhear15; run;
data final14;set tfm.redagrtanh15; run;
data final15;set tfm.redagrtanhear19; run;
data final16;set tfm.redagrtanh19; run;

data final17;set tfm.redagrtanhear25; run;
data final18;set tfm.redagrtanh25; run;

data final19;set tfm.redmartanhear2; run;
data final20;set tfm.redmartanh2; run;

data final21;set tfm.redmartanhear3; run;
data final22;set tfm.redmartanh3; run;

data final23;set tfm.redmartanhear5; run;
data final24;set tfm.redmartanh5; run;

data final25;set tfm.redmartanhear11; run;
data final26;set tfm.redmartanh11; run;

data final27;set tfm.redmartanhear17; run;
data final28;set tfm.redmartanh17; run;

data final29;set tfm.redmintanhear2; run;
data final30;set tfm.redmintanh2; run;
```

```
data final31;set tfm.redmintanhear4; run;
data final32;set tfm.redmintanh4; run;
```

```
data union; set final1 final2 final3 final4 final5 final6 final7
final8 final9 final10 final11 final12 final13 final14
final15 final16 final17 final18 final19 final20 final21 final22
final23 final24 final25 final26 final27 final28 final29 final30
final31 final32; run;
proc boxplot data=union; plot media*modelo;run;
```

```
%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=log,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);
```

```
data final33;set final;modelo="AGR-2-EAR-LOG   ";
data "C:\Users\mario\Desktop\TFM\redagrlogear2.sas7bdat";
  set final33;
run;
```

```
%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=arc,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);
```

```
data final34;set final;modelo="AGR-2-EAR-ARC   ";
data "C:\Users\mario\Desktop\TFM\redagrarcear2.sas7bdat";
  set final34;
run;
```

```
%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=sin,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final35;set final;modelo="AGR-2-EAR-SIN   ";
data "C:\Users\mario\Desktop\TFM\redagrsinear2.sas7bdat";
  set final35;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=sof,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final36;set final;modelo="AGR-2-EAR-SOF   ";
data "C:\Users\mario\Desktop\TFM\redagrsofear2.sas7bdat";
  set final36;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=LEVMAR,
objetivo=tasafallos,
acti=gau,
early=15,
```

```
directorio=C:\Users\mario\Desktop\TFM
);

data final37;set final;modelo="AGR-2-EAR-GAU   ";
data "C:\Users\mario\Desktop\TFM\redagrgauear2.sas7bdat";
  set final37;
run;




data prom; set final1 final33 final34 final35 final36 final37; run;
proc boxplot data=prom; plot media*modelo;run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=quanew,
objetivo=tasafallos,
acti=sof,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final38;set final;modelo="AGR-2-EAR-SOF-QUA   ";
data "C:\Users\mario\Desktop\TFM\redagrsofear2QUA.sas7bdat";
  set final38;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=bprop,
objetivo=tasafallos,
acti=sof,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final39;set final;modelo="AGR-2-EAR-SOF-BPROP   ";
data "C:\Users\mario\Desktop\TFM\redagrsofear2BPROP.sas7bdat";
  set final39;
run;
```

```
%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=congra,
objetivo=tasafallos,
acti=sof,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final40;set final;modelo="AGR-2-EAR-SOF-CON   ";
data "C:\Users\mario\Desktop\TFM\redagrsofear2CON.sas7bdat";
  set final40;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=TRUREG,
objetivo=tasafallos,
acti=sof,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final41;set final;modelo="AGR-2-EAR-SOF-TRU   ";
data "C:\Users\mario\Desktop\TFM\redagrsofear2TRU.sas7bdat";
  set final41;
run;


%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=DBLDOG,
objetivo=tasafallos,
acti=sof,
```

```
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final42;set final;modelo="AGR-2-EAR-SOF-DBL   ";
data "C:\Users\mario\Desktop\TFM\redagrsofear2DBL.sas7bdat";
  set final42;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=quanew,
objetivo=tasafallos,
acti=TANH,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final43;set final;modelo="AGR-2-EAR-TANH-QUA   ";
data "C:\Users\mario\Desktop\TFM\redagrTANHear2QUA.sas7bdat";
  set final43;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=bprop,
objetivo=tasafallos,
acti=TANH,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final44;set final;modelo="AGR-2-EAR-TANH-BPROP   ";
data "C:\Users\mario\Desktop\TFM\redagrTANHear2BPROP.sas7bdat";
  set final44;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
```

```
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=congra,
objetivo=tasafallos,
acti=TANH,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final45;set final;modelo="AGR-2-EAR-TANH-CON   ";
data "C:\Users\mario\Desktop\TFM\redagrTANHear2CON.sas7bdat";
  set final45;
run;

%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=TRUREG,
objetivo=tasafallos,
acti=TANH,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final46;set final;modelo="AGR-2-EAR-TANH-TRU   ";
data "C:\Users\mario\Desktop\TFM\redagrTANHear2TRU.sas7bdat";
  set final46;
run;


%cruzadabinarianeural(archivo=tfm,
vardepen= OPOSVACUNA,
conti=PREF_PRES_No_consta PREF_PRES_Otro
PREF_PRES_Pablo_Iglesias PREF_PRES_Santiago_Abascal
EFEC_COVID_Economia PREO_COVID_Mucho PREO_COVID_Nada
PREO_COVID_Poco VAL_PS EDAD,
categor=,
ngrupos=5,
sinicio=12345,
sfinal=12394,
nodos=2,
algo=DBLDOG,
objetivo=tasafallos,
acti=TANH,
early=15,
directorio=C:\Users\mario\Desktop\TFM
);

data final47;set final;modelo="AGR-2-EAR-TANH-DBL   ";
data "C:\Users\mario\Desktop\TFM\redagrTANHear2DBL.sas7bdat";
```

```
  set final47;
run;
data final1;set tfm.redagrtanhear2; run;
data final36;set tfm.redagrsofear2; run;

data prom; set final36 final38 final40 final41 final42; run;
proc boxplot data=prom; plot media*modelo;run;
```

Código R

Tratamiento variables

```
library(sas7bdat)
library(dummies)
library(MASS)
library(reshape)
library(caret)
library (readxl)
library(writexl)
data<-read_excel("TFMESTAND_TRAIN.xlsx")
dput(names(data))
cat<- c("CAPITAL", "TAMUNI", "SEXO", "NACIONALIDAD", "CCAA",
        "CERCANIA", "CNO11", "GOB_ENCAR", "INTENCIONGALTER", "PRACTICARELIG",
        "PREF_PRES", "RECUVOTOG", "AFE_VIDPER", "AFE_VIDSOC", "CIVIS_COVID",
        "CLASESOCIAL", "ECIVIL", "EFEC_COVID", "ESCUELA", "ESTUDIOS",
        "INTENCIONG", "NIVELESTENTREV", "PARTICIPACIONG", "PREO_COVID",
        "PRO_PRI", "PRO_SOC", "RELIGION", "SINT_COVID", "SITLAB", "VAL_ECO",
        "VAL_ECO_PER")
cont<-c("M_ESCIDEOL", "M_VAL_IA", "M_VAL_PC", "M_VAL_PI",
         "M_VAL_PS", "M_VAL_SA", "ESCIDEOL", "VAL_IA", "VAL_PC", "VAL_PI",
         "VAL_PS", "VAL_SA", "EDAD", "aleat1", "aleat2", "numMissing")



data<- dummy.data.frame(data,cat, sep = ".")
data <- data[,-
c(1:4,11,13:15,33,43,53,58,67,74,83,94,98,102,105,111,116,119:121,127,137,148,150,154,165,175,181,182,183,192,197)]

library(writexl)
write_xlsx (data, "ESTAND_DUMM.xlsx")


datasel <- read_excel("ESTAND_SELVAL_TRAIN.xlsx")
dput(names(datasel))
# "VACUNA", "CIVIS_COVID", "EFEC_COVID", "PARTICIPACIONG", "G_PREF_PRES",
# "G_PREO_COVID", "G_INTENCIONG", "G_RECUVOTOG", "G_PRO_PRI", "G_SITLAB",
# "G_CNO11", "G_VAL_ECO", "G_PRO_SOC", "G_ECIVIL", "G_INTENCIONGALTER",
# "G_NIVELESTENTREV", "G_VAL_ECO_PER", "G_CCAA", "G_GOB_ENCAR",
# "VAL_PS", "EDAD", "numMissing"
contsel <- c("VAL_PS", "EDAD", "numMissing")
catsel <- c("CIVIS_COVID", "EFEC_COVID", "PARTICIPACIONG", "G_PREF_PRES",
        "G_PREO_COVID", "G_INTENCIONG", "G_RECUVOTOG", "G_PRO_PRI", "G_SITLAB",
        "G_CNO11", "G_VAL_ECO", "G_PRO_SOC", "G_ECIVIL", "G_INTENCIONGALTER",
        "G_NIVELESTENTREV", "G_VAL_ECO_PER", "G_CCAA", "G_GOB_ENCAR")

datasel<- dummy.data.frame(datasel, catsel, sep =".")
datasel <- datasel[,-c(2,5,8,10,14,17,21,26,30,34,39,42,47,50,55,61,65,71)]
datasel$VACUNA[datasel$VACUNA=="Si"]<-"Yes"
datatodo <- read_excel ("ESTAND_DUMM.xlsx")
datatodo$VACUNA[datatodo$VACUNA=="Si"]<-"Yes"
write_xlsx (datasel, "ESTANDSELVAL_DUMM.xlsx")


datamario <- read.sas7bdat("estandselmary_train.sas7bdat")
dput(names(datamario))

contm <- c("VAL_PS", "EDAD")
catm <- c( "PREF_PRES", "EFEC_COVID", "PREO_COVID")
```

```
datamario <-dummy.data.frame(datamario, catm, sep =".")
datamario <- datamario[, -c(2,11,14)]
datamario$VACUNA[datamario$VACUNA=="Si"]<-"Yes"
summary (datamario)

dput(names(datamario))

#------------------------------RANGO-----------------------

rangtodo <- read.sas7bdat("rang_todo_train.sas7bdat")
dput (names(rangtodo))

catrt<-c("TAMUNI", "SEXO", "CCAA", "VACUNA", "CERCANIA", "CNO11", "GOB_ENCAR",
      "INTENCIONGALTER", "PRACTICARELIG", "PREF_PRES", "RECUVOTOG",
      "AFE_VIDPER", "AFE_VIDSOC", "CIVIS_COVID", "CLASESOCIAL", "ECIVIL",
      "EFEC_COVID", "ESTUDIOS", "INTENCIONG", "NIVELESTENTREV", "PARTICIPACIONG",
      "PREO_COVID", "PRO_PRI", "PRO_SOC", "RELIGION", "SITLAB", "VAL_ECO",
      "VAL_ECO_PER")

rangtodo <-dummy.data.frame(rangtodo, catrt, sep =".")
rangtodo <- rangtodo[,-c(1,8,10,29,39,49,54,63,70,79,90,98,101,107,112,115,121,131,144,148,158,169,175,181,189)]

library(writexl)
write_xlsx (rangtodo, "RANG_TODO.xlsx")

rangtodo <- read_excel("RANG_TODO.xlsx")

rangselval <- read.sas7bdat("rang_selval_train.sas7bdat")
dput(names(rangselval))
crsv <- c("CIVIS_COVID", "EFEC_COVID", "PARTICIPACIONG", "G_PREF_PRES",
"G_PREO_COVID", "G_INTENCIONG", "G_RECUVOTOG", "G_PRO_PRI", "G_SITLAB",
"G_CNO11", "G_VAL_ECO", "G_PRO_SOC", "G_ECIVIL", "G_INTENCIONGALTER",
"G_NIVELESTENTREV", "G_VAL_ECO_PER", "G_CCAA", "G_GOB_ENCAR")
rangselval <-dummy.data.frame(rangselval, crsv, sep=".")
rangselval <- rangselval[, -c(3,5,8,10,14,17,21,26,30,34,39,42,47,50,55,61,65,71)]
rangmario <- read.sas7bdat ("rang_selmar_train.sas7bdat")
dput(names(rangmario))
crsm <- c( "PREF_PRES", "EFEC_COVID", "PREO_COVID")
rangmario <- dummy.data.frame(rangmario, crsm, sep=".")
rangmario <- rangmario [, -c(2,11,14)]

#REDES
redmario<- datamario
redmario$OPOSVACUNA <- ifelse(redmario$VACUNA=="Yes", "No", "Yes")
redmario$VACUNA <- NULL
write_xlsx(redmario, "redmario.xlsx")

redsel<- datasel
redsel$OPOSVACUNA <- ifelse(redsel$VACUNA=="Yes", "No", "Yes")
redsel$VACUNA <- NULL
write_xlsx(redsel, "redsel.xlsx")

redtodo<- datatodo
redtodo$OPOSVACUNA <- ifelse(redtodo$VACUNA=="Yes", "No", "Yes")
redtodo$VACUNA <- NULL

redrandom <- redmario[, -c(1,2,5,7,10)]
```

KNN

```
auc<-function(x,y) {
 curvaroc<-roc(response=x,predictor=y)
 auc<-curvaroc$auc
 return(auc)
}
######----------ESTANDARIZADO----------
#####-------TSEL MARIO----------

nums<- list(3,5,7,9,11,15,21,25,31,37,45,55,65,75,85,95,105)
```

```
set.seed(12345)
control<-trainControl(method = "repeatedcv",number=5,repeats=50,
                      savePredictions = "all",classProbs=TRUE)
for (i in nums){
  print(paste0("knnselmar", i))
  knnmod<- train(VACUNA~.,data=datamario,
                 method="knn",
                 trControl=control,tuneGrid=expand.grid(k=c(i)))
  preditest<-knnmod$pred

  preditest$prueba<-strsplit(preditest$Resample,"[.]")
  preditest$Fold <- sapply(preditest$prueba, "[", 1)
  preditest$Rep <- sapply(preditest$prueba, "[", 2)
  preditest$prueba<-NULL

  tabla<-table(preditest$Rep)
  listarep<-c(names(tabla))
  misc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    tasa=1-tasafallos(paso1$pred,paso1$obs)
    misc<-rbind(misc,tasa)
  }
  names(misc)<-"tasa"

  auroc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    auc=suppressMessages(auc(paso1$obs,paso1$Yes))
    auroc<-rbind(auroc,auc)
  }
  names(auroc)<-"auc"
  resumen<-cbind(misc, auroc$auc)
  resumen$modelo<-paste0("knnselmarest", i)
  colnames(resumen) <- c("tasa","auc", "modelo")
  assign(paste0("knnselmarest",i),resumen)
}

#####------- SEL VARIABLES MINER --------

for (i in nums){
  print(paste0("knnselval", i))
  knnmod<- train(VACUNA~.,data=datasel,
                 method="knn",
                 trControl=control,tuneGrid=expand.grid(k=c(i)))
  preditest<-knnmod$pred

  preditest$prueba<-strsplit(preditest$Resample,"[.]")
  preditest$Fold <- sapply(preditest$prueba, "[", 1)
  preditest$Rep <- sapply(preditest$prueba, "[", 2)
  preditest$prueba<-NULL

  tabla<-table(preditest$Rep)
  listarep<-c(names(tabla))
  misc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    tasa=1-tasafallos(paso1$pred,paso1$obs)
    misc<-rbind(misc,tasa)
  }
  names(misc)<-"tasa"

  auroc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    auc=suppressMessages(auc(paso1$obs,paso1$Yes))
    auroc<-rbind(auroc,auc)
  }
  names(auroc)<-"auc"
  resumen<-cbind(misc, auroc$auc)
  resumen$modelo<-paste0("knnselvalest", i)
```

```
 colnames(resumen) <- c("tasa","auc", "modelo")
 assign(paste0("knnselvalest",i),resumen)
}


#####------- TODO ---------

for (i in nums){
 print(paste0("knnseltodo", i))
 knnmod<- train(VACUNA~.,data=datatodo,
           method="knn",
           trControl=control,tuneGrid=expand.grid(k=c(i)))
 preditest<-knnmod$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tabla<-table(preditest$Rep)
 listarep<-c(names(tabla))
 misc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  tasa=1-tasafallos(paso1$pred,paso1$obs)
  misc<-rbind(misc,tasa)
 }
 names(misc)<-"tasa"

 auroc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  auc=suppressMessages(auc(paso1$obs,paso1$Yes))
  auroc<-rbind(auroc,auc)
 }
 names(auroc)<-"auc"
 resumen<-cbind(misc, auroc$auc)
 resumen$modelo<-paste0("knntodoest", i)
 colnames(resumen) <- c("tasa","auc", "modelo")
 assign(paste0("knntodoest",i),resumen)
}


boxplot(data=unionest,auc~modelo,main="AUC")
boxplot(data=unionestselval,auc~modelo,main="AUC")



#----------RANGO--------------

#----------SEL MARIO--------
for (i in nums){
 print(paste0("knnselmar", i))
 knnmod<- train(VACUNA~.,data=datamario,
           method="knn",
           trControl=control,tuneGrid=expand.grid(k=c(i)))
 preditest<-knnmod$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tabla<-table(preditest$Rep)
 listarep<-c(names(tabla))
 misc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  tasa=1-tasafallos(paso1$pred,paso1$obs)
  misc<-rbind(misc,tasa)
```

```
  }
  names(misc)<-"tasa"

  auroc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    auc=suppressMessages(auc(paso1$obs,paso1$Yes))
    auroc<-rbind(auroc,auc)
  }
  names(auroc)<-"auc"
  resumen<-cbind(misc, auroc$auc)
  resumen$modelo<-paste0("knnselmarRAN", i)
  colnames(resumen) <- c("tasa","auc", "modelo")
  assign(paste0("knnselmarRAN",i),resumen)
}


#------------SEL MINER----------

for (i in nums){
  print(paste0("knnselval", i))
  knnmod<- train(VACUNA~.,data=datamario,
                 method="knn",
                 trControl=control,tuneGrid=expand.grid(k=c(i)))
  preditest<-knnmod$pred

  preditest$prueba<-strsplit(preditest$Resample,"[.]")
  preditest$Fold <- sapply(preditest$prueba, "[", 1)
  preditest$Rep <- sapply(preditest$prueba, "[", 2)
  preditest$prueba<-NULL

  tabla<-table(preditest$Rep)
  listarep<-c(names(tabla))
  misc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    tasa=1-tasafallos(paso1$pred,paso1$obs)
    misc<-rbind(misc,tasa)
  }
  names(misc)<-"tasa"

  auroc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    auc=suppressMessages(auc(paso1$obs,paso1$Yes))
    auroc<-rbind(auroc,auc)
  }
  names(auroc)<-"auc"
  resumen<-cbind(misc, auroc$auc)
  resumen$modelo<-paste0("knnselvalRAN", i)
  colnames(resumen) <- c("tasa","auc", "modelo")
  assign(paste0("knnselvalRAN",i),resumen)
}


#--------------TODO------------

for (i in nums){
  print(paste0("knntodo", i))
  knnmod<- train(VACUNA~.,data=datamario,
                 method="knn",
                 trControl=control,tuneGrid=expand.grid(k=c(i)))
  preditest<-knnmod$pred

  preditest$prueba<-strsplit(preditest$Resample,"[.]")
  preditest$Fold <- sapply(preditest$prueba, "[", 1)
  preditest$Rep <- sapply(preditest$prueba, "[", 2)
  preditest$prueba<-NULL

  tabla<-table(preditest$Rep)
  listarep<-c(names(tabla))
```

```
 misc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  tasa=1-tasafallos(paso1$pred,paso1$obs)
  misc<-rbind(misc,tasa)
 }
 names(misc)<-"tasa"

 auroc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  auc=suppressMessages(auc(paso1$obs,paso1$Yes))
  auroc<-rbind(auroc,auc)
 }
 names(auroc)<-"auc"
 resumen<-cbind(misc, auroc$auc)
 resumen$modelo<-paste0("knntodoRAN", i)
 colnames(resumen) <- c("tasa","auc", "modelo")
 assign(paste0("knntodoRAN",i),resumen)
}
```

Regresión logística

```
summary(regorig)
regorig[,as.vector(which(sapply(regorig, class)=="character"))] <- lapply(regorig[,as.vector(which(sapply(regorig, class)=="character"))] , as.factor)

null<-glm(OPOSVACUNA~1,data=regorig,family=binomial)
full<-glm(OPOSVACUNA~.,data=regorig,family=binomial)
RegOrigBackAIC<-step(full, scope=list(lower=null, upper=full), direction="backward")
RegOrigStepAIC<-step(null, scope=list(lower=null, upper=full), direction="both")
RegOrigForwAIC<-step(null, scope=list(lower=null, upper=full), direction="forward")
RegOrigBackBIC<-step(full, scope=list(lower=null, upper=full), direction="backward",k=log(nrow(regorig)))
RegOrigStepBIC<-step(null, scope=list(lower=null, upper=full), direction="both",k=log(nrow(regorig)))
RegOrigForwBIC<-step(null, scope=list(lower=null, upper=full), direction="forward",k=log(nrow(regorig)))

modelosorig <- list (RegOrigBackAIC, RegOrigStepAIC, RegOrigForwAIC, RegOrigBackBIC, RegOrigStepBIC, RegOrigForwBIC)
sapply(modelosorig,function(x) x$rank)
sapply(modelosorig,function(x) roc(regorig$OPOSVACUNA, predict(x,regorig,type = "response"), direction="<")$auc)


#####ORIGINALES E INTERACCIONES##### NO SE EJECUTA
# fullInt<-glm(VACUNA~.^2, data=regorig, family=binomial)
# RegOrigIntBackAIC<-step(fullInt, scope=list(lower=null, upper=fullInt), direction="backward")
# RegOrigIntStepAIC<-step(null, scope=list(lower=null, upper=fullInt), direction="both")
# RegOrigIntForwAIC<-step(null, scope=list(lower=null, upper=fullInt), direction="forward")
# RegOrigIntBackBIC<-step(fullInt, scope=list(lower=null, upper=fullInt), direction="backward",k=log(nrow(regorig)))
# RegOrigIntStepBIC<-step(null, scope=list(lower=null, upper=fullInt), direction="both",k=log(nrow(regorig)))
# RegOrigIntForwBIC<-step(null, scope=list(lower=null, upper=fullInt), direction="forward",k=log(nrow(regorig)))

# modelosorigint <- list(RegOrigIntBackAIC, RegOrigIntStepAIC, RegOrigIntForwAIC, RegOrigIntBackBIC, RegOrigIntStepBIC, RegOrigIntForwBIC)
# sapply(modelosorigint,function(x) x$rank)
# sapply(modelosorigint,function(x) roc(regorig$VACUNA, predict(x,regorig,type = "response"), direction="<")$auc)


#####ORIGINALES Y TRANSFORMADAS#####
nullTrans<-glm(OPOSVACUNA~1,data=regtrans,family=binomial)
fullTrans<-glm(OPOSVACUNA~.,data=regtrans,family=binomial)
#si solo quiero usar las variables originales, uso lo siguiente:
# full<-glm(barrioCaro~.,data=data_train[,1:20],family=binomial)
RegTransBackAIC<-step(fullTrans, scope=list(lower=nullTrans, upper=fullTrans), direction="backward")
RegTransStepAIC<-step(nullTrans, scope=list(lower=nullTrans, upper=fullTrans), direction="both")
RegTransForwAIC<-step(nullTrans, scope=list(lower=nullTrans, upper=fullTrans), direction="forward")
RegTransBackBIC<-step(fullTrans, scope=list(lower=nullTrans, upper=fullTrans), direction="backward",k=log(nrow(regtrans)))
RegTransStepBIC<-step(nullTrans, scope=list(lower=nullTrans, upper=fullTrans), direction="both",k=log(nrow(regtrans)))
RegTransForwBIC<-step(nullTrans, scope=list(lower=nullTrans, upper=fullTrans), direction="forward",k=log(nrow(regtrans)))


modelostrans <- list (RegTransBackAIC, RegTransStepAIC, RegTransForwAIC, RegTransBackBIC, RegTransStepBIC, RegTransForwBIC)
```

```
sapply(modelostrans,function(x) x$rank)
sapply(modelostrans,function(x) roc(regtrans$OPOSVACUNA, predict(x,regtrans,type = "response"), direction="<")$auc)


#####ORIGINALES Y TRANSFORMADAS CON INTERACCIONES#####NO SE EJECUTA
# fullIntTrans<-glm(VACUNA~.^2, data=regtrans, family=binomial)
# RegTransIntBackAIC<-step(fullTrans, scope=list(lower=nullTrans, upper=fullIntTrans), direction="backward")
# RegTransIntStepAIC<-step(nullTrans, scope=list(lower=nullTrans, upper=fullIntTrans), direction="both")
# RegTransIntForwAIC<-step(nullTrans, scope=list(lower=nullTrans, upper=fullIntTrans), direction="forward")
#                 RegTransIntBackBIC<-step(fullTrans,                 scope=list(lower=nullTrans,                 upper=fullIntTrans),
direction="backward",k=log(nrow(regtrans)))
# RegTransIntStepBIC<-step(nullTrans, scope=list(lower=nullTrans, upper=fullIntTrans), direction="both",k=log(nrow(regtrans)))
#                 RegTransIntForwBIC<-step(nullTrans,                 scope=list(lower=nullTrans,                 upper=fullIntTrans),
direction="forward",k=log(nrow(regtrans)))
#
# modelostransint<- list (RegTransIntBackAIC, RegTransIntStepAIC, RegTransIntForwAIC, RegTransIntBackBIC, RegTransIntStepBIC,
RegTransIntForwBIC)
# sapply(modelostransint,function(x) x$rank)
# sapply(modelostransint,function(x) roc(regtrans$VACUNA, predict(x,regtrans,type = "response"), direction="<")$auc)



#####ORIGINALES CON DUMMIES######

nulldum<-glm(OPOSVACUNA~1,data=regorigdum,family=binomial)
fulldum<-glm(OPOSVACUNA~.,data=regorigdum,family=binomial)
#si solo quiero usar las variables originales, uso lo siguiente:
# full<-glm(barrioCaro~.,data=data_train[,1:20],family=binomial)
start <- Sys.time()
RegOrigDumBackAIC<-step(fulldum, scope=list(lower=nulldum, upper=fulldum), direction="backward")
RegOrigDumStepAIC<-step(nulldum, scope=list(lower=nulldum, upper=fulldum), direction="both")
RegOrigDumForwAIC<-step(nulldum, scope=list(lower=nulldum, upper=fulldum), direction="forward")
RegOrigDumBackBIC<-step(fulldum, scope=list(lower=nulldum, upper=fulldum), direction="backward",k=log(nrow(regorigdum)))
RegOrigDumStepBIC<-step(nulldum, scope=list(lower=nulldum, upper=fulldum), direction="both",k=log(nrow(regorigdum)))
RegOrigDumForwBIC<-step(nulldum, scope=list(lower=nulldum, upper=fulldum), direction="forward",k=log(nrow(regorigdum)))

modelosorigdum <- list (RegOrigDumBackAIC, RegOrigDumStepAIC, RegOrigDumForwAIC, RegOrigDumBackBIC,
RegOrigDumStepBIC, RegOrigDumForwBIC)
sapply(modelosorigdum,function(x) x$rank)
sapply(modelosorigdum,function(x) roc(regorigdum$OPOSVACUNA, predict(x,regorigdum,type = "response"), direction="<")$auc)


#####ORIGINALES Y TRANSFORMADAS CON DUMMIES#######

nullTransdumm<-glm(OPOSVACUNA~1,data=regtransdumm,family=binomial)
fullTransdumm<-glm(OPOSVACUNA~.,data=regtransdumm,family=binomial)
#si solo quiero usar las variables originales, uso lo siguiente:
# full<-glm(barrioCaro~.,data=data_train[,1:20],family=binomial)
#RegTransDumBackAIC<-step(fullTransdumm, scope=list(lower=nullTransdumm, upper=fullTransdumm), direction="backward")
RegTransDumStepAIC<-step(nullTransdumm, scope=list(lower=nullTransdumm, upper=fullTransdumm), direction="both")
RegTransDumForwAIC<-step(nullTransdumm, scope=list(lower=nullTransdumm, upper=fullTransdumm), direction="forward")
#RegTransDumBackBIC<-step(fullTransdumm,                 scope=list(lower=nullTransdumm,                 upper=fullTransdumm),
direction="backward",k=log(nrow(regtransdumm)))
RegTransDumStepBIC<-step(nullTransdumm,                 scope=list(lower=nullTransdumm,                 upper=fullTransdumm),
direction="both",k=log(nrow(regtransdumm)))
RegTransDumForwBIC<-step(nullTransdumm,                 scope=list(lower=nullTransdumm,                 upper=fullTransdumm),
direction="forward",k=log(nrow(regtransdumm)))

modelostransdum <- list (RegTransDumStepAIC, RegTransDumForwAIC, RegTransDumStepBIC, RegTransDumForwBIC)
sapply(modelostransdum,function(x) x$rank)
sapply(modelostransdum,function(x)      roc(regtransdumm$OPOSVACUNA,      predict(x,regtransdumm,type      =      "response"),
direction="<")$auc)
end<- Sys.time()
end-start

modelostransvcr <- list(RegTransBackAIC,RegTransForwAIC, RegTransBackBIC, RegTransForwBIC)
formulaModelosTrans<-sapply(modelostransvcr,formula)
nombModelosTrans <- list ("RegTransBackAIC","RegTransForwAIC", "RegTransBackBIC", "RegTransForwBIC")
counter=1
for (i in 1:length(modelostransvcr)){
  set.seed(12345)
```

```
 print(nombModelosTrans[i])
 regmod<- train(as.formula(formulaModelosTrans[[i]]),data=regtrans,
           method="glm",
           trControl=control)
 preditest<-regmod$pred
 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL
 tabla<-table(preditest$Rep)
 listarep<-c(names(tabla))
 misc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  tasa=1-tasafallos(paso1$pred,paso1$obs)
  misc<-rbind(misc,tasa)
 }
 names(misc)<-"tasa"
 auroc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  auc=suppressMessages(auc(paso1$obs,paso1$Yes))
  auroc<-rbind(auroc,auc)
 }
 names(auroc)<-"auc"
 resumen<-cbind(misc, auroc$auc)
 resumen$modelo<-paste0(nombModelosTrans[i])
 colnames(resumen) <- c("tasa","auc", "modelo")
 assign(paste0("error", nombModelosTrans[i]),resumen)
 counter=counter+1
}
errorRegTransBackAIC$rank<-106
errorRegTransForwAIC$rank<-105
errorRegTransBackBIC$rank<-14
errorRegTransForwBIC$rank<-12
unionregtrans<-rbind(errorRegTransBackAIC,errorRegTransForwAIC, errorRegTransBackBIC, errorRegTransForwBIC)
####
# summary(regorigdum)
# regorigdum$VACUNA <- as.character (regorigdum$VACUNA)
# regorigdum$VACUNA[regorigdum$VACUNA== "Si"]<- "Yes"
modelosorigdumvcr <- list(RegOrigDumBackAIC, RegOrigDumForwAIC, RegOrigDumStepAIC, RegOrigDumBackBIC,
                  RegOrigDumForwBIC)
formulaModelosOrigDum<-sapply(modelosorigdumvcr,formula)
nombModelosOrigDum <- list ("RegOrigDumBackAIC", "RegOrigDumForwAIC", "RegOrigDumStepAIC", "RegOrigDumBackBIC",
                   "RegOrigDumForwBIC")
counter=1
for (i in 1:length(modelosorigdumvcr)){
 set.seed(12345)
 print(nombModelosOrigDum[i])
 regmod<- train(as.formula(formulaModelosOrigDum[[i]]),data=regorigdum,
           method="glm",
           trControl=control)
 preditest<-regmod$pred
 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL
 tabla<-table(preditest$Rep)
 listarep<-c(names(tabla))
 misc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  tasa=1-tasafallos(paso1$pred,paso1$obs)
  misc<-rbind(misc,tasa)
 }
 names(misc)<-"tasa"
 auroc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  auc=suppressMessages(auc(paso1$obs,paso1$Yes))
  auroc<-rbind(auroc,auc)
```

```
 }
 names(auroc)<-"auc"
 resumen<-cbind(misc, auroc$auc)
 resumen$modelo<-paste0(nombModelosOrigDum[i])
 colnames(resumen) <- c("tasa","auc", "modelo")
 assign(paste0("error", nombModelosOrigDum[i]),resumen)
 counter=counter+1
}
errorRegOrigDumBackAIC$rank<-78
errorRegOrigDumForwAIC$rank<-59
errorRegOrigDumStepAIC$rank<-60
errorRegOrigDumBackBIC$rank<-22
errorRegOrigDumForwBIC$rank<-19
unionregorigdum<-rbind(errorRegOrigDumBackAIC, errorRegOrigDumForwAIC, errorRegOrigDumStepAIC,
errorRegOrigDumBackBIC,
          errorRegOrigDumForwBIC)
####
# summary(regtransdumm)
# regtransdumm$VACUNA <- as.character (regtransdumm$VACUNA)
# regtransdumm$VACUNA[regtransdumm$VACUNA== "Si"]<- "Yes"
modelostransdummvcr <- list(RegTransDumStepAIC,RegTransDumForwAIC, RegTransDumForwBIC)
formulaModelosTransDum<-sapply(modelostransdummvcr,formula)
nombModelosTransDum <- list ("RegTransDumStepAIC","RegTransDumForwAIC", "RegTransDumForwBIC")
counter=1
for (i in 1:length(modelostransdummvcr)){
 set.seed(12345)
 print(nombModelosTransDum[i])
 regmod<- train(as.formula(formulaModelosTransDum[[i]]),data=regtransdumm,
          method="glm",
          trControl=control)
 preditest<-regmod$pred
 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL
 tabla<-table(preditest$Rep)
 listarep<-c(names(tabla))
 misc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  tasa=1-tasafallos(paso1$pred,paso1$obs)
  misc<-rbind(misc,tasa)
 }
 names(misc)<-"tasa"
 auroc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  auc=suppressMessages(auc(paso1$obs,paso1$Yes))
  auroc<-rbind(auroc,auc)
 }
 names(auroc)<-"auc"
 resumen<-cbind(misc, auroc$auc)
 resumen$modelo<-paste0(nombModelosTransDum[i])
 colnames(resumen) <- c("tasa","auc", "modelo")
 assign(paste0("error", nombModelosTransDum[i]),resumen)
 counter=counter+1
}
errorRegTransDumStepAIC$rank<-60
errorRegTransDumForwAIC$rank<-66
errorRegTransDumForwBIC$rank<-21
unionregtransdum<-rbind(errorRegTransDumStepAIC, errorRegTransDumForwAIC, errorRegTransDumForwBIC)
unionregresion <- rbind(unionregorig, unionregtrans, unionregorigdum, unionregtransdum)



library(dplyr)
library(highcharter)
# Set highcharter options
options(highcharter.theme = hc_theme_smpl(tooltip = list(valueDecimals = 4)))
pb <- rbind(unionregresion[501:600,], unionregresion[701:750,])
#------AUC-----------
```

```
hcboxplot(
  outliers = FALSE,
  x = pb$auc,
  var = pb$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS REGRESIÓN LOGÍSTICA") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")

mean(unionregresion[551:600,]$tasa)
mean(unionregresion[701:750,]$auc)
sd(unionregresion[501:550,]$auc)
sd(unionregresion[551:600,]$auc)
```

Redes

```
set.seed(12345)
control<-trainControl(method = "repeatedcv",number=5,repeats=50,
                      savePredictions = "all",classProbs=TRUE)
decays <- list (0.001)
for (i in decays){
  print(paste0("red2sinear", i))
  avnnetgrid <-expand.grid(size=c(2),decay=c(i),bag=FALSE)
  redmod<- train(OPOSVACUNA~., data=redrandom, method="avNNet", linout = FALSE,
                 maxit=100,trControl=control,repeats=5,tuneGrid=avnnetgrid)
  preditest<-redmod$pred

  preditest$prueba<-strsplit(preditest$Resample,"[.]")
  preditest$Fold <- sapply(preditest$prueba, "[", 1)
  preditest$Rep <- sapply(preditest$prueba, "[", 2)
  preditest$prueba<-NULL

  tabla<-table(preditest$Rep)
  listarep<-c(names(tabla))
  misc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    tasa=1-tasafallos(paso1$pred,paso1$obs)
    misc<-rbind(misc,tasa)
  }
  names(misc)<-"tasa"

  auroc<-data.frame()
  for (repi in listarep) {
    paso1<-preditest[which(preditest$Rep==repi),]
    auc=suppressMessages(auc(paso1$obs,paso1$Yes))
    auroc<-rbind(auroc,auc)
  }
  names(auroc)<-"auc"
  resumen<-cbind(misc, auroc$auc)
  resumen$modelo<-paste0("red2sinear", i)
  colnames(resumen) <- c("tasa","auc", "modelo")
  assign(paste0("red2sinear",i),resumen)
}


for (i in decays){
  print(paste0("red3sinear", i))
  avnnetgrid <-expand.grid(size=c(3),decay=c(i),bag=FALSE)
  redmod<- train(OPOSVACUNA~., data=redrandom, method="avNNet", linout = FALSE,
                 maxit=100,trControl=control,repeats=5,tuneGrid=avnnetgrid)
  preditest<-redmod$pred

  preditest$prueba<-strsplit(preditest$Resample,"[.]")
  preditest$Fold <- sapply(preditest$prueba, "[", 1)
  preditest$Rep <- sapply(preditest$prueba, "[", 2)
```

```
 preditest$prueba<-NULL

 tabla<-table(preditest$Rep)
 listarep<-c(names(tabla))
 misc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  tasa=1-tasafallos(paso1$pred,paso1$obs)
  misc<-rbind(misc,tasa)
 }
 names(misc)<-"tasa"

 auroc<-data.frame()
 for (repi in listarep) {
  paso1<-preditest[which(preditest$Rep==repi),]
  auc=suppressMessages(auc(paso1$obs,paso1$Yes))
  auroc<-rbind(auroc,auc)
 }
 names(auroc)<-"auc"
 resumen<-cbind(misc, auroc$auc)
 resumen$modelo<-paste0("red3sinear", i)
 colnames(resumen) <- c("tasa","auc", "modelo")
 assign(paste0("red3sinear",i),resumen)
}

library (readxl)
redessas<-read_excel("red rocs.xlsx")
todasredes <- rbind (redessas, red2sinear0.1, red2sinear0.01, red2sinear0.001, red3sinear0.1, red3sinear0.01, red3sinear0.001)
todasredes$auc<- as.numeric(todasredes$auc)
todasredes$tasa<- as.numeric(todasredes$tasa)
library(dplyr)
library(highcharter)
# Set highcharter options
options(highcharter.theme = hc_theme_smpl(tooltip = list(valueDecimals = 4)))
bp <- todasredes
head(bp,4)
data("bp")
#------AUC-----------
hcboxplot(
 outliers = FALSE,
 x = bp$tasa,
 var = bp$modelo,
 name = "MODELO",
 color = "#92D050"
) %>%
 hc_title(text = "MODELOS RED NEURONAL") %>%
 hc_yAxis(title = list(text = "TASA DE FALLOS")) %>%
 hc_chart(type = "column")
```

Random Forest / Bagging

```
cruzadarfbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,nodesize=20,
  mtry=2,ntree=50,replace=TRUE,sampsize=1)
 {

if (any(listclass==c(""))==FALSE)
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else  {
   databis<-data[,c(vardep,listconti)]
  }

  means <-apply(databis[,listconti],2,mean)
  sds<-sapply(databis[,listconti],sd)
```

```
 datacon<-scale(databis[,listconti], center = means, scale = sds)
 numerocont<-which(colnames(databis)%in%listconti)
 databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

 databis[,vardep]<-as.factor(databis[,vardep])

 formu<-formula(paste("factor(",vardep,")~.",sep=""))

 set.seed(sinicio)
 control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
  savePredictions = "all",classProbs=TRUE)

 rfgrid <-expand.grid(mtry=mtry)

 if  (sampsize==1)
 {
   rf<- train(formu,data=databis,
  method="rf",trControl=control,
  tuneGrid=rfgrid,nodesize=nodesize,replace=replace,ntree=ntree)
 }

else  if  (sampsize!=1)
 {
   rf<- train(formu,data=databis,
  method="rf",trControl=control,
  tuneGrid=rfgrid,nodesize=nodesize,replace=replace,sampsize=sampsize,
  ntree=ntree)
 }


 print(rf$results)

 preditest<-rf$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }
  tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
tasa=1-tasafallos(paso1$pred,paso1$obs)
medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"

 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

  mediasbis<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
auc=suppressMessages(auc(paso1$obs,paso1$Yes))
mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"
 medias$auc<-mediasbis$auc

 return(medias)
```

```
 }

varis<-list(3,4,5,6,7,8,9,10)
muestras<-list(4400, 3500, 2500, 1500, 1000, 500, 250)
for (i in varis){
 for (k in muestras){
  print(paste0(i, ".", k))
  temporal<-cruzadarfbin(data=redrandom, vardep="OPOSVACUNA",
  listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),listclass=c(""),
  grupos=5,sinicio=12345,repe=10,nodesize=100,
  mtry=i,ntree=1000,replace=TRUE, sampsize=k)

  temporal$modelo=paste0("rf.agr.", i, ".", k)
  assign(paste0("rf.agr.", i, ".", k),temporal)
 }
}

varisx<-list(3,4,5,6,7,8,9,10,11,12,13,14,15)
muestras<-list(4400, 3500, 2500, 1500, 1000, 500, 250)
for (i in varisx){
 for (k in muestras){
  print(paste0(i, ".", k))
  temporal<-cruzadarfbin(data=redmario, vardep="OPOSVACUNA",
                listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                        "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),listclass=c(""),
                grupos=5,sinicio=12345,repe=10,nodesize=100,
                mtry=i,ntree=1000,replace=TRUE, sampsize=k)

  temporal$modelo=paste0("rf.mar.", i, ".", k)
  assign(paste0("rf.mar.", i, ".", k),temporal)
 }
}

varisxx<-list(5,10,20,30,40,45,50,55,58)
muestras<-list(4400, 3500, 2500, 1500, 1000, 500, 250)
for (i in varisxx){
 for (k in muestras){
  print(paste0(i, ".", k))
  temporal<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
                grupos=5,sinicio=12345,repe=10,nodesize=100,
                mtry=i,ntree=1000,replace=TRUE, sampsize=k)

  temporal$modelo=paste0("rf.min.", i, ".", k)
  assign(paste0("rf.min.", i, ".", k),temporal)
 }
}

forestran <- rbind(rf.agr.3.250, rf.agr.3.500, rf.agr.3.1000, rf.agr.3.1500,rf.agr.3.2500, rf.agr.3.3500,rf.agr.3.4400,
            rf.agr.4.250, rf.agr.4.500, rf.agr.4.1000, rf.agr.4.1500,rf.agr.4.2500, rf.agr.4.3500,rf.agr.4.4400,
            rf.agr.5.250, rf.agr.5.500, rf.agr.5.1000, rf.agr.5.1500,rf.agr.5.2500, rf.agr.5.3500,rf.agr.5.4400,
            rf.agr.6.250, rf.agr.6.500, rf.agr.6.1000, rf.agr.6.1500,rf.agr.6.2500, rf.agr.6.3500,rf.agr.6.4400,
```

```
            rf.agr.7.250, rf.agr.7.500, rf.agr.7.1000, rf.agr.7.1500,rf.agr.7.2500, rf.agr.7.3500,rf.agr.7.4400,
            rf.agr.8.250, rf.agr.8.500, rf.agr.8.1000, rf.agr.8.1500,rf.agr.8.2500, rf.agr.8.3500,rf.agr.8.4400,
            rf.agr.9.250, rf.agr.9.500, rf.agr.9.1000, rf.agr.9.1500,rf.agr.9.2500, rf.agr.9.3500,rf.agr.9.4400,
            rf.agr.10.250, rf.agr.10.500, rf.agr.10.1000, rf.agr.10.1500,rf.agr.10.2500, rf.agr.10.3500,rf.agr.10.4400)

listforestran<-list(rf.agr.3.250,rf.agr.3.500,rf.agr.3.1000,rf.agr.3.1500,rf.agr.3.2500,rf.agr.3.3500,rf.agr.3.4400,
rf.agr.4.250,rf.agr.4.500,rf.agr.4.1000,rf.agr.4.1500,rf.agr.4.2500,rf.agr.4.3500,rf.agr.4.4400,
rf.agr.5.250,rf.agr.5.500,rf.agr.5.1000,rf.agr.5.1500,rf.agr.5.2500,rf.agr.5.3500,rf.agr.5.4400,
rf.agr.6.250,rf.agr.6.500,rf.agr.6.1000,rf.agr.6.1500,rf.agr.6.2500,rf.agr.6.3500,rf.agr.6.4400,
rf.agr.7.250,rf.agr.7.500,rf.agr.7.1000,rf.agr.7.1500,rf.agr.7.2500,rf.agr.7.3500,rf.agr.7.4400,
rf.agr.8.250,rf.agr.8.500,rf.agr.8.1000,rf.agr.8.1500,rf.agr.8.2500,rf.agr.8.3500,rf.agr.8.4400,
rf.agr.9.250,rf.agr.9.500,rf.agr.9.1000,rf.agr.9.1500,rf.agr.9.2500,rf.agr.9.3500,rf.agr.9.4400,
rf.agr.10.250,rf.agr.10.500,rf.agr.10.1000,rf.agr.10.1500,rf.agr.10.2500,rf.agr.10.3500,rf.agr.10.4400)

nombresforestran<-list("rf.agr.3.250","rf.agr.3.500","rf.agr.3.1000","rf.agr.3.1500","rf.agr.3.2500","rf.agr.3.3500","rf.agr.3.4400","
rf.agr.4.250","rf.agr.4.500","rf.agr.4.1000","rf.agr.4.1500","rf.agr.4.2500","rf.agr.4.3500","rf.agr.4.4400","
rf.agr.5.250","rf.agr.5.500","rf.agr.5.1000","rf.agr.5.1500","rf.agr.5.2500","rf.agr.5.3500","rf.agr.5.4400","
rf.agr.6.250","rf.agr.6.500","rf.agr.6.1000","rf.agr.6.1500","rf.agr.6.2500","rf.agr.6.3500","rf.agr.6.4400","
rf.agr.7.250","rf.agr.7.500","rf.agr.7.1000","rf.agr.7.1500","rf.agr.7.2500","rf.agr.7.3500","rf.agr.7.4400","
rf.agr.8.250","rf.agr.8.500","rf.agr.8.1000","rf.agr.8.1500","rf.agr.8.2500","rf.agr.8.3500","rf.agr.8.4400","
rf.agr.9.250","rf.agr.9.500","rf.agr.9.1000","rf.agr.9.1500","rf.agr.9.2500","rf.agr.9.3500","rf.agr.9.4400","
rf.agr.10.250","rf.agr.10.500","rf.agr.10.1000","rf.agr.10.1500","rf.agr.10.2500","rf.agr.10.3500","rf.agr.10.4400")


forestmar <- rbind(rf.mar.3.250, rf.mar.3.500, rf.mar.3.1000, rf.mar.3.1500,rf.mar.3.2500, rf.mar.3.3500,rf.mar.3.4400,
            rf.mar.4.250, rf.mar.4.500, rf.mar.4.1000, rf.mar.4.1500,rf.mar.4.2500, rf.mar.4.3500,rf.mar.4.4400,
            rf.mar.5.250, rf.mar.5.500, rf.mar.5.1000, rf.mar.5.1500,rf.mar.5.2500, rf.mar.5.3500,rf.mar.5.4400,
            rf.mar.6.250, rf.mar.6.500, rf.mar.6.1000, rf.mar.6.1500,rf.mar.6.2500, rf.mar.6.3500,rf.mar.6.4400,
            rf.mar.7.250, rf.mar.7.500, rf.mar.7.1000, rf.mar.7.1500,rf.mar.7.2500, rf.mar.7.3500,rf.mar.7.4400,
            rf.mar.8.250, rf.mar.8.500, rf.mar.8.1000, rf.mar.8.1500,rf.mar.8.2500, rf.mar.8.3500,rf.mar.8.4400,
            rf.mar.9.250, rf.mar.9.500, rf.mar.9.1000, rf.mar.9.1500,rf.mar.9.2500, rf.mar.9.3500,rf.mar.9.4400,
            rf.mar.10.250, rf.mar.10.500, rf.mar.10.1000, rf.mar.10.1500,rf.mar.10.2500, rf.mar.10.3500,rf.mar.10.4400,
            rf.mar.11.250, rf.mar.11.500, rf.mar.11.1000, rf.mar.11.1500,rf.mar.11.2500, rf.mar.11.3500,rf.mar.11.4400,
            rf.mar.12.250, rf.mar.12.500, rf.mar.12.1000, rf.mar.12.1500,rf.mar.12.2500, rf.mar.12.3500,rf.mar.12.4400,
            rf.mar.13.250, rf.mar.13.500, rf.mar.13.1000, rf.mar.13.1500,rf.mar.13.2500, rf.mar.13.3500,rf.mar.13.4400,
            rf.mar.14.250, rf.mar.14.500, rf.mar.14.1000, rf.mar.14.1500,rf.mar.14.2500, rf.mar.14.3500,rf.mar.14.4400,
            rf.mar.15.250, rf.mar.15.500, rf.mar.15.1000, rf.mar.15.1500,rf.mar.15.2500, rf.mar.15.3500,rf.mar.15.4400)

forestmin<- rbind(rf.min.5.250, rf.min.5.500, rf.min.5.1000, rf.min.5.1500,rf.min.5.2500, rf.min.5.3500,rf.min.5.4400,
            rf.min.10.250, rf.min.10.500, rf.min.10.1000, rf.min.10.1500,rf.min.10.2500, rf.min.10.3500,rf.min.10.4400,
            rf.min.20.250, rf.min.20.500, rf.min.20.1000, rf.min.20.1500,rf.min.20.2500, rf.min.20.3500,rf.min.20.4400,
            rf.min.30.250, rf.min.30.500, rf.min.30.1000, rf.min.30.1500,rf.min.30.2500, rf.min.30.3500,rf.min.30.4400,
            rf.min.40.250, rf.min.40.500, rf.min.40.1000, rf.min.40.1500,rf.min.40.2500, rf.min.40.3500,rf.min.40.4400,
            rf.min.45.250, rf.min.45.500, rf.min.45.1000, rf.min.45.1500,rf.min.45.2500, rf.min.45.3500,rf.min.45.4400,
            rf.min.50.250, rf.min.50.500, rf.min.50.1000, rf.min.50.1500,rf.min.50.2500, rf.min.50.3500,rf.min.50.4400,
            rf.min.55.250, rf.min.55.500, rf.min.55.1000, rf.min.55.1500,rf.min.55.2500, rf.min.55.3500,rf.min.55.4400,
            rf.min.58.250, rf.min.58.500, rf.min.58.1000, rf.min.58.1500,rf.min.58.2500, rf.min.58.3500,rf.min.58.4400)
foresttodos <- rbind (forestran, forestmar, forestmin)

######VCR DE 50 para los mejores######

toprf.5.3500<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
        grupos=5,sinicio=12345,repe=50,nodesize=100,
        mtry=5,ntree=1000,replace=TRUE, sampsize=3500)

toprf.5.4400<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
```

```
                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
        grupos=5,sinicio=12345,repe=50,nodesize=100,
        mtry=5,ntree=1000,replace=TRUE, sampsize=4400)


toprf.10.3500<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
        grupos=5,sinicio=12345,repe=50,nodesize=100,
        mtry=10,ntree=1000,replace=TRUE, sampsize=3500)

toprf.10.4400<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
        grupos=5,sinicio=12345,repe=50,nodesize=100,
        mtry=10,ntree=1000,replace=TRUE, sampsize=4400)



toprf.45.2500<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
```

```
                grupos=5,sinicio=12345,repe=50,nodesize=100,
                mtry=45,ntree=1000,replace=TRUE, sampsize=2500)

toprf.55.1000<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
                grupos=5,sinicio=12345,repe=50,nodesize=100,
                mtry=55,ntree=1000,replace=TRUE, sampsize=1000)


toprf.5.3500$model<-"M327-rf.min.5.3500"
toprf.5.4400$model<-"M328-rf.min.5.4400"

toprf.10.3500$model<-"M334-rf.min.10.3500"
toprf.10.4400$model<-"M335-rf.min.10.4400"

toprf.45.2500$model<-"M361-rf.min.45.2500"

toprf.55.1000$model<-"M373-rf.min.55.1000"


topforest<-rbind(toprf.5.3500, toprf.5.4400, toprf.10.3500, toprf.10.4400, toprf.45.2500, toprf.55.1000)

hcboxplot(
  outliers = FALSE,
  x = topforest[51:100,]$auc,
  var = topforest[51:100,]$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS RF") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")

mean(topforest[51:100,]$tasa)


####VARIOS TAMAÃ□OS DE NODESIZE######

toprf.5.4400.50<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
                grupos=5,sinicio=12345,repe=50,nodesize=50,
                mtry=5,ntree=1000,replace=TRUE, sampsize=4400)

toprf.5.4400.20<-cruzadarfbin(data=redsel, vardep="OPOSVACUNA",
                  listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
```

```
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,nodesize=20,
                    mtry=5,ntree=1000,replace=TRUE, sampsize=4400)

toprf.5.4400.50$model<-"M385-rf.min.5.4400.50"
toprf.5.4400.20$model<-"M386-rf.min.5.4400.20"
forestvarios <- rbind(toprf.5.4400, toprf.5.4400.50, toprf.5.4400.20)
hcboxplot(
  outliers = FALSE,
  x = forestvarios$auc,
  var = forestvarios$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS RF") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")
```

Gradient Boosting

```
####PRUEBAS PRELIM#####

set.seed(12345)

gbmgrid<-expand.grid(shrinkage=c(0.2,0.1,0.05,0.03,0.01,0.001, 0.25),
                n.minobsinnode=c(20,50,100),
                n.trees=c(100,500,1000,1500),
                interaction.depth=c(2))

control<-trainControl(method = "cv",number=5,savePredictions = "all",
                 classProbs=TRUE)

gbm5<- train(factor(OPOSVACUNA)~.,data=redrandom,
          method="gbm",trControl=control,tuneGrid=gbmgrid,
          distribution="bernoulli", bag.fraction=1,verbose=FALSE)

gbm5

plot(gbm5)


gbmgrid<-expand.grid(shrinkage=c(0.2,0.25,0.1,0.05,0.03,0.01,0.001),
                n.minobsinnode=c(20),
                n.trees=c(100,500,1000,1500),
                interaction.depth=c(2))

gbm2<- train(factor(OPOSVACUNA)~.,data=redrandom,
          method="gbm",trControl=control,tuneGrid=gbmgrid,
          distribution="bernoulli", bag.fraction=1,verbose=FALSE)

plot(gbm2)

gbmgrid<-expand.grid(shrinkage=c(0.2,0.25,0.1,0.05,0.03,0.01,0.001),
                n.minobsinnode=c(50),
                n.trees=c(100,500,1000,1500),
                interaction.depth=c(2))

gbm3<- train(factor(OPOSVACUNA)~.,data=redrandom,
```

```
        method="gbm",trControl=control,tuneGrid=gbmgrid,
        distribution="bernoulli", bag.fraction=1,verbose=FALSE)


plot(gbm3)


gbmgrid<-expand.grid(shrinkage=c(0.2,0.25,0.1,0.05,0.03,0.01,0.001),
              n.minobsinnode=c(100),
              n.trees=c(100,500,1000,1500),
              interaction.depth=c(2))

gbm4<- train(factor(OPOSVACUNA)~.,data=redrandom,
        method="gbm",trControl=control,tuneGrid=gbmgrid,
        distribution="bernoulli", bag.fraction=1,verbose=FALSE)


plot(gbm4)


gbmgrid<-expand.grid(shrinkage=c(0.2,0.25,0.1,0.05,0.03,0.01,0.001),
              n.minobsinnode=c(20,50,100),
              n.trees=c(100,500,1000,2000,5000),
              interaction.depth=c(2))

gbm4mario<- train(factor(OPOSVACUNA)~.,data=redmario,
        method="gbm",trControl=control,tuneGrid=gbmgrid,
        distribution="bernoulli", bag.fraction=1,verbose=FALSE)


plot(gbm4mario)


gbm4miner<- train(factor(OPOSVACUNA)~.,data=redsel,
            method="gbm",trControl=control,tuneGrid=gbmgrid,
            distribution="bernoulli", bag.fraction=1,verbose=FALSE)
plot(gbm4miner)

cruzadagbmbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,
  n.minobsinnode=20,shrinkage=0.1,n.trees=100,interaction.depth=2)
 {

if (any(listclass==c(""))==FALSE)
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else  {
   databis<-data[,c(vardep,listconti)]
  }


  means <-apply(databis[,listconti],2,mean)
  sds<-sapply(databis[,listconti],sd)

  datacon<-scale(databis[,listconti], center = means, scale = sds)
  numerocont<-which(colnames(databis)%in%listconti)
  databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

  databis[,vardep]<-as.factor(databis[,vardep])

  formu<-formula(paste("factor(",vardep,")~.",sep=""))


  set.seed(sinicio)
  control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
   savePredictions = "all",classProbs=TRUE)
```

```
  gbmgrid <-expand.grid(n.minobsinnode=n.minobsinnode,
   shrinkage=shrinkage,n.trees=n.trees,
   interaction.depth=interaction.depth)

 gbm<- train(formu,data=databis,
  method="gbm",trControl=control,
  tuneGrid=gbmgrid,distribution="bernoulli",verbose=FALSE)

 print(gbm$results)

 preditest<-gbm$pred


 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }


  tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
tasa=1-tasafallos(paso1$pred,paso1$obs)
medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"



 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }


  mediasbis<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
auc=suppressMessages(auc(paso1$obs,paso1$Yes))
mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"


 medias$auc<-mediasbis$auc

 return(medias)

}

gbm.agr.20.100.025 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                        listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                                "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                                "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        n.minobsinnode=20,shrinkage=0.25,n.trees=100,interaction.depth=2)
gbm.agr.20.100.025$model <- "M387-gbm.agr.20.100.025"


gbm.agr.20.500.001 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
```

```
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.01,n.trees=500,interaction.depth=2)
gbm.agr.20.500.001$model <- "M388-gbm.agr.20.500.001"

gbm.agr.20.1000.003 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.03,n.trees=1000,interaction.depth=2)
gbm.agr.20.1000.003$model <- "M389-gbm.agr.20.1000.003"

gbm.agr.20.1500.001 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.01,n.trees=1500,interaction.depth=2)
gbm.agr.20.1500.001$model <- "M390-gbm.agr.20.1500.001"

gbm.agr.50.100.020 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                     listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                             "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                             "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                     listclass=c(""),
                     grupos=5,sinicio=12345,repe=50,
                     n.minobsinnode=50,shrinkage=0.20,n.trees=100,interaction.depth=2)
gbm.agr.50.100.020$model <- "M391-gbm.agr.50.100.020"

gbm.agr.50.500.005 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=50,shrinkage=0.05,n.trees=500,interaction.depth=2)
gbm.agr.50.500.005$model <- "M392-gbm.agr.50.500.005"

gbm.agr.50.1000.003 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=50,shrinkage=0.03,n.trees=1000,interaction.depth=2)
gbm.agr.50.1000.003$model <- "M393-gbm.agr.50.1000.003"

gbm.agr.50.1500.005 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=50,shrinkage=0.05,n.trees=1500,interaction.depth=2)
gbm.agr.50.1500.005$model <- "M394-gbm.agr.50.1500.005"

gbm.agr.100.100.020 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.20,n.trees=100,interaction.depth=2)
gbm.agr.100.100.020$model <- "M395-gbm.agr.100.100.020"
```

```
gbm.agr.100.500.003 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                        listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                                "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                                "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        n.minobsinnode=100,shrinkage=0.03,n.trees=500,interaction.depth=2)
gbm.agr.100.500.003$model <- "M396-gbm.agr.100.500.003"

gbm.agr.100.1000.003 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                         listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                                 "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                                 "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                         listclass=c(""),
                         grupos=5,sinicio=12345,repe=50,
                         n.minobsinnode=100,shrinkage=0.03,n.trees=1000,interaction.depth=2)
gbm.agr.100.1000.003$model <- "M397-gbm.agr.100.1000.003"

gbm.agr.100.1500.001 <- cruzadagbmbin(data=redrandom, vardep="OPOSVACUNA",
                         listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                                 "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                                 "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                         listclass=c(""),
                         grupos=5,sinicio=12345,repe=50,
                         n.minobsinnode=100,shrinkage=0.01,n.trees=1500,interaction.depth=2)
gbm.agr.100.1500.001$model <- "M398-gbm.agr.100.1500.001"

gbmrandom<-rbind(gbm.agr.20.100.025, gbm.agr.20.500.001, gbm.agr.20.1000.003, gbm.agr.20.1500.001,
          gbm.agr.50.100.020, gbm.agr.50.500.005, gbm.agr.50.1000.003, gbm.agr.50.1500.005,
          gbm.agr.100.100.020, gbm.agr.100.500.003, gbm.agr.100.1000.003, gbm.agr.100.1500.001)

#######SELECCIÃ“N AGRESIVA MARIO#########
start<-Sys.time()

gbm.mar.20.100.025 <- cruzadagbmbin(data=redmario,vardep="OPOSVACUNA",
                        listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                                "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                                "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                                "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                                "VAL_PS", "EDAD"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        n.minobsinnode=20,shrinkage=0.25,n.trees=100,interaction.depth=2)
gbm.mar.20.100.025$model <- "M399-gbm.mar.20.100.025"


gbm.mar.20.500.003 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                        listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                                "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                                "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                                "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                                "VAL_PS", "EDAD"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        n.minobsinnode=20,shrinkage=0.03,n.trees=500,interaction.depth=2)
gbm.mar.20.500.003$model <- "M400-gbm.mar.20.500.003"

gbm.mar.20.1000.005 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                         listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                                 "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                                 "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                                 "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                                 "VAL_PS", "EDAD"),
                         listclass=c(""),
                         grupos=5,sinicio=12345,repe=50,
                         n.minobsinnode=20,shrinkage=0.05,n.trees=1000,interaction.depth=2)
gbm.mar.20.1000.005$model <- "M401-gbm.mar.20.1000.005"

gbm.mar.20.2000.003 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                        listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                                "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
```

```
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.03,n.trees=2000,interaction.depth=2)
gbm.mar.20.2000.003$model <- "M402-gbm.mar.20.2000.003"

gbm.mar.50.100.020 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                        "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=50,shrinkage=0.20,n.trees=100,interaction.depth=2)
gbm.mar.50.100.020$model <- "M403-gbm.mar.50.100.020"

gbm.mar.50.500.003 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                        "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=50,shrinkage=0.03,n.trees=500,interaction.depth=2)
gbm.mar.50.500.003$model <- "M404-gbm.mar.50.500.003"

gbm.mar.50.1000.005 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                        "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=50,shrinkage=0.05,n.trees=1000,interaction.depth=2)
gbm.mar.50.1000.005$model <- "M405-gbm.mar.50.1000.005"

gbm.mar.50.2000.003 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                        "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=50,shrinkage=0.03,n.trees=2000,interaction.depth=2)
gbm.mar.50.2000.003$model <- "M406-gbm.mar.50.2000.003"

gbm.mar.100.100.025 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                        "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.25,n.trees=100,interaction.depth=2)
gbm.mar.100.100.025$model <- "M407-gbm.mar.100.100.025"

gbm.mar.100.500.010 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                        "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
```

```
                    n.minobsinnode=100,shrinkage=0.10,n.trees=500,interaction.depth=2)
gbm.mar.100.500.010$model <- "M408-gbm.mar.100.500.010"

gbm.mar.100.1000.005 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                            "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                            "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                            "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.05,n.trees=1000,interaction.depth=2)
gbm.mar.100.1000.005$model <- "M409-gbm.mar.100.1000.005"

gbm.mar.100.2000.003 <- cruzadagbmbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                            "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                            "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                            "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.03,n.trees=2000,interaction.depth=2)
gbm.mar.100.2000.003$model <- "M410-gbm.mar.100.2000.003"

gbmmario<-rbind(gbm.mar.20.100.025, gbm.mar.20.500.003, gbm.mar.20.1000.005, gbm.mar.20.2000.003,
        gbm.mar.50.100.020, gbm.mar.50.500.003, gbm.mar.50.1000.005, gbm.mar.50.2000.003,
        gbm.mar.100.100.025, gbm.mar.100.500.010, gbm.mar.100.1000.005, gbm.mar.100.2000.003)
print("GBM MARIO YA!!!!!!!!!!!!!!!!!!!")

######GBM MINER###########

gbm.min.20.100.025 <- cruzadagbmbin(data=redsel,vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                            "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                            "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                            "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                            "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                            "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                            "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                            "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                            "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                            "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                            "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                            "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                            "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.25,n.trees=100,interaction.depth=2)
gbm.min.20.100.025$model <- "M411-gbm.min.20.100.025"


gbm.min.20.500.010 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                            "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                            "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                            "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                            "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                            "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                            "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                            "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                            "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                            "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                            "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                            "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                            "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.10,n.trees=500,interaction.depth=2)
```

```
gbm.min.20.500.010$model <- "M412-gbm.min.20.500.010"

gbm.min.20.1000.001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                          "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                          "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                          "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                          "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.01,n.trees=1000,interaction.depth=2)
gbm.min.20.1000.001$model <- "M413-gbm.min.20.1000.001"

gbm.min.20.2000.001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                          "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                          "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                          "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                          "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.01,n.trees=2000,interaction.depth=2)
gbm.min.20.2000.001$model <- "M414-gbm.min.20.2000.001"

gbm.min.20.5000.0001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                          "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                          "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                          "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                          "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=20,shrinkage=0.001,n.trees=5000,interaction.depth=2)
gbm.min.20.5000.0001$model <- "M415-gbm.min.20.5000.0001"

gbm.min.50.100.005 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                   listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                         "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                         "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                         "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                         "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                         "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                         "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                         "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
```

```
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                n.minobsinnode=50,shrinkage=0.05,n.trees=100,interaction.depth=2)
gbm.min.50.100.005$model <- "M416-gbm.min.50.100.005"

gbm.min.50.500.010 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                n.minobsinnode=50,shrinkage=0.10,n.trees=500,interaction.depth=2)
gbm.min.50.500.010$model <- "M417-gbm.min.50.500.010"

gbm.min.50.1000.005 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                n.minobsinnode=50,shrinkage=0.05,n.trees=1000,interaction.depth=2)
gbm.min.50.1000.005$model <- "M418-gbm.min.50.1000.005"

gbm.min.50.2000.001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                n.minobsinnode=50,shrinkage=0.01,n.trees=2000,interaction.depth=2)
gbm.min.50.2000.001$model <- "M419-gbm.min.50.2000.001"
```

```
gbm.min.50.5000.0001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                            "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                            "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                            "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                            "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                            "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                            "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                            "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                            "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                            "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                            "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                            "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                            "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=50,shrinkage=0.001,n.trees=5000,interaction.depth=2)
gbm.min.50.5000.0001$model <- "M420-gbm.min.50.5000.0001"

gbm.min.100.100.025 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                            "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                            "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                            "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                            "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                            "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                            "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                            "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                            "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                            "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                            "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                            "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                            "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.25,n.trees=100,interaction.depth=2)
gbm.min.100.100.025$model <- "M421-gbm.min.100.100.025"

gbm.min.100.500.005 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                            "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                            "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                            "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                            "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                            "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                            "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                            "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                            "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                            "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                            "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                            "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                            "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.05,n.trees=500,interaction.depth=2)
gbm.min.100.500.005$model <- "M422-gbm.min.100.500.005"

gbm.min.100.1000.003 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                     listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                            "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                            "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                            "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                            "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                            "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                            "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                            "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                            "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
```

```
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.03,n.trees=1000,interaction.depth=2)
gbm.min.100.1000.003$model <- "M423-gbm.min.100.1000.003"

gbm.min.100.2000.001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.01,n.trees=2000,interaction.depth=2)
gbm.min.100.2000.001$model <- "M424-gbm.min.100.2000.001"

gbm.min.100.5000.001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    n.minobsinnode=100,shrinkage=0.01,n.trees=5000,interaction.depth=2)
gbm.min.100.5000.001$model <- "M425-gbm.min.100.5000.001"

gbmsel<-rbind(gbm.min.20.100.025, gbm.min.20.500.010, gbm.min.20.1000.001, gbm.min.20.2000.001,gbm.min.20.5000.0001,
          gbm.min.50.100.005, gbm.min.50.500.010, gbm.min.50.1000.005, gbm.min.50.2000.001,gbm.min.50.5000.0001,
          gbm.min.100.100.025, gbm.min.100.500.005, gbm.min.100.1000.003, gbm.min.100.2000.001, gbm.min.100.5000.001)
end<-Sys.time()
end-start

gbmtodos <- rbind (gbmrandom, gbmmario, gbmsel)
library(dplyr)
library(highcharter)
# Set highcharter options
options(highcharter.theme = hc_theme_smpl(tooltip = list(valueDecimals = 4)))
hcboxplot(
  outliers = FALSE,
  x = gbmtodos[1201:1950,]$auc,
  var = gbmtodos[1201:1950,]$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS gbm") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")
```

```
#######ULTIMAS PRUEBAS#######
gbm.min.100.1500.001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        n.minobsinnode=100,shrinkage=0.01,n.trees=1500,interaction.depth=2)
gbm.min.100.1500.001$model <- "M426-gbm.min.100.1500.001"

gbm.min.100.2500.001 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        n.minobsinnode=100,shrinkage=0.01,n.trees=2500,interaction.depth=2)
gbm.min.100.2500.001$model <- "M427-gbm.min.100.2500.001"


mejoresgbmmuestra <- rbind (gbm.min.100.1500.001, gbm.min.100.2000.001, gbm.min.100.2500.001)

library(dplyr)
library(highcharter)
# Set highcharter options
options(highcharter.theme = hc_theme_smpl(tooltip = list(valueDecimals = 4)))
hcboxplot(
  outliers = FALSE,
  x = gbmcambioshrinkage$auc,
  var = gbmcambioshrinkage$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS gbm") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")

cruzadaxgbmbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=20,eta=0.1,nrounds=100,max_depth=2,
  gamma=0,colsample_bytree=1,subsample=1,alpha=0,lambda=0)
 {


if (any(listclass==c(""))==FALSE)
  {
```

```
 databis<-data[,c(vardep,listconti,listclass)]
 databis<- dummy.data.frame(databis, listclass, sep = ".")
} else  {
 databis<-data[,c(vardep,listconti)]
}


means <-apply(databis[,listconti],2,mean)
sds<-sapply(databis[,listconti],sd)


datacon<-scale(databis[,listconti], center = means, scale = sds)
numerocont<-which(colnames(databis)%in%listconti)
databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

databis[,vardep]<-as.factor(databis[,vardep])

formu<-formula(paste("factor(",vardep,")~.",sep=""))


set.seed(sinicio)
control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
 savePredictions = "all",classProbs=TRUE)

xgbmgrid <-expand.grid( min_child_weight=min_child_weight,
eta=eta,nrounds=nrounds,max_depth=max_depth,
gamma=gamma,colsample_bytree=colsample_bytree,subsample=subsample)

xgbm<- train(formu,data=databis,
 method="xgbTree",trControl=control,
 tuneGrid=xgbmgrid,verbose=FALSE,
 alpha=alpha,lambda=lambda)

print(xgbm$results)

preditest<-xgbm$pred


preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tasafallos<-function(x,y) {
 confu<-confusionMatrix(x,y)
 tasa<-confu[[3]][1]
 return(tasa)
}


 tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
tasa=1-tasafallos(paso1$pred,paso1$obs)
medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"


 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

  mediasbis<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
auc=suppressMessages(auc(paso1$obs,paso1$Yes))
```

```
mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"

 medias$auc<-mediasbis$auc

 return(medias)

 }
set.seed(12345)
control<-trainControl(method = "cv",number=5,savePredictions = "all",
                classProbs=TRUE)
xgbmgrid<-expand.grid(
 min_child_weight=c(20,50,100),
 eta=c(0.001,0.01,0.03,0.05,0.1,0.2,0.25),
 nrounds=c(100,500,1000,2000,5000),
 max_depth=2,gamma=0,colsample_bytree=1,subsample=1)

xgbm<- train(factor(OPOSVACUNA)~.,data=redsel,
        method="xgbTree",trControl=control,
        tuneGrid=xgbmgrid,verbose=FALSE)

xgbm

plot(xgbm)
#####PRUEBITA GBM CON SHRINKAGE######
gbm.min.100.1500.002 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        n.minobsinnode=100,shrinkage=0.02,n.trees=1500,interaction.depth=2)
gbm.min.100.1500.002$model <- "M428-gbm.min.100.1500.002"

gbm.min.100.1500.003 <- cruzadagbmbin(data=redsel, vardep="OPOSVACUNA",
                        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        n.minobsinnode=100,shrinkage=0.03,n.trees=1500,interaction.depth=2)
gbm.min.100.1500.003$model <- "M429-gbm.min.100.1500.003"

gbmcambioshrinkage <- rbind (gbm.min.100.1500.001,gbm.min.100.1500.002, gbm.min.100.1500.003)
#####XGBOOST SEL MINER######

xgb.20.100.01 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
```

```
                    "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                    "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                    "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                    "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                    "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                    "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                    "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                    "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                    "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                    "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                    "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                    "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                    "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                min_child_weight=20,eta=0.1,nrounds=100,max_depth=2,
                gamma=0,colsample_bytree=1,subsample=1)

xgb.20.100.01$model<-"M430-xgb.20.100.01"

xgb.20.500.003 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                    "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                    "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                    "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                    "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                    "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                    "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                    "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                    "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                    "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                    "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                    "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                    "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                    "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                min_child_weight=20,eta=0.03,nrounds=500,max_depth=2,
                gamma=0,colsample_bytree=1,subsample=1)

xgb.20.500.003$model<-"M431-xgb.20.500.003"

xgb.20.1000.001 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                    "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                    "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                    "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                    "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                    "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                    "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                    "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                    "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                    "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                    "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                    "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                    "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                    "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                min_child_weight=20,eta=0.01,nrounds=1000,max_depth=2,
                gamma=0,colsample_bytree=1,subsample=1)

xgb.20.1000.001$model<-"M432-xgb.20.1000.001"

xgb.20.2000.001 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                    "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                    "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                    "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                    "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                    "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
```

```
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=20,eta=0.01,nrounds=2000,max_depth=2,
                    gamma=0,colsample_bytree=1,subsample=1)

xgb.20.2000.001$model<-"M433-xgb.20.2000.001"


xgb.50.100.02 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=50,eta=0.2,nrounds=100,max_depth=2,
                    gamma=0,colsample_bytree=1,subsample=1)

xgb.50.100.02$model<-"M434-xgb.50.100.02"

xgb.50.500.005 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                     listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                         "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                         "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                         "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                         "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                         "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                         "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                         "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                         "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                         "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                         "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                         "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                         "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                         "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                     listclass=c(""),
                     grupos=5,sinicio=12345,repe=50,
                     min_child_weight=50,eta=0.05,nrounds=500,max_depth=2,
                     gamma=0,colsample_bytree=1,subsample=1)

xgb.50.500.005$model<-"M435-xgb.50.500.005"

xgb.50.1000.003 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                      listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
```

```
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=50,eta=0.03,nrounds=1000,max_depth=2,
                    gamma=0,colsample_bytree=1,subsample=1)

xgb.50.1000.003$model<-"M436-xgb.50.1000.003"

xgb.50.2000.001 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=50,eta=0.01,nrounds=2000,max_depth=2,
                    gamma=0,colsample_bytree=1,subsample=1)

xgb.50.2000.001$model<-"M437-xgb.50.2000.001"


xgboostmejores<-rbind(xgb.20.100.01, xgb.20.500.003, xgb.20.1000.001, xgb.20.2000.001,
            xgb.50.100.02, xgb.50.500.005, xgb.50.1000.003, xgb.50.2000.001)

options(highcharter.theme = hc_theme_smpl(tooltip = list(valueDecimals = 4)))
hcboxplot(
  outliers = FALSE,
  x = xgboostmejores$tasa,
  var = xgboostmejores$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS gbm") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")

####XGBOOST cambiando col_sample

xgb.20.500.003.5 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=20,eta=0.03,nrounds=500,max_depth=2,
                    gamma=0,colsample_bytree=0.05,subsample=1)
```

```
xgb.20.500.003.5$model<-"M438-xgb.20.500.003.5"

xgb.20.500.003.125 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                          "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                          "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                          "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                          "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=20,eta=0.03,nrounds=500,max_depth=2,
                    gamma=0,colsample_bytree=0.125,subsample=1)

xgb.20.500.003.125$model<-"M439-xgb.20.500.003.125"

xgb.20.500.003.25 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                          "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                          "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                          "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                          "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=20,eta=0.03,nrounds=500,max_depth=2,
                    gamma=0,colsample_bytree=0.25,subsample=1)

xgb.20.500.003.25$model<-"M440-xgb.20.500.003.25"

xgb.20.500.003.50 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                          "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                          "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                          "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                          "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=20,eta=0.03,nrounds=500,max_depth=2,
                    gamma=0,colsample_bytree=0.5,subsample=1)

xgb.20.500.003.50$model<-"M441-xgb.20.500.003.50"

xgb.20.500.003.75 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
```

```
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=20,eta=0.03,nrounds=500,max_depth=2,
                    gamma=0,colsample_bytree=0.75,subsample=1)

xgb.20.500.003.75$model<-"M442-xgb.20.500.003.75"

xgb.20.2000.001.5 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    min_child_weight=20,eta=0.01,nrounds=2000,max_depth=2,
                    gamma=0,colsample_bytree=0.05,subsample=1)

xgb.20.2000.001.5$model<-"M443-xgb.20.2000.001.5"

xgb.20.2000.001.125 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                     listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                         "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                         "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                         "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                         "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                         "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                         "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                         "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                         "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                         "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                         "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                         "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                         "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                         "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                     listclass=c(""),
                     grupos=5,sinicio=12345,repe=50,
                     min_child_weight=20,eta=0.01,nrounds=2000,max_depth=2,
                     gamma=0,colsample_bytree=0.125,subsample=1)

xgb.20.2000.001.125$model<-"M444-xgb.20.2000.001.125"

xgb.20.2000.001.25 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
```

```
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                min_child_weight=20,eta=0.01,nrounds=2000,max_depth=2,
                gamma=0,colsample_bytree=0.25,subsample=1)

xgb.20.2000.001.25$model<-"M445-xgb.20.2000.001.25"

xgb.20.2000.001.50 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                min_child_weight=20,eta=0.01,nrounds=2000,max_depth=2,
                gamma=0,colsample_bytree=0.5,subsample=1)

xgb.20.2000.001.50$model<-"M446-xgb.20.2000.001.50"

xgb.20.2000.001.75 <-cruzadaxgbmbin(data=redsel, vardep="OPOSVACUNA",
                listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                        "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                        "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                        "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                        "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                        "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                        "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                        "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                        "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                        "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                        "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                        "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                        "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                min_child_weight=20,eta=0.01,nrounds=2000,max_depth=2,
                gamma=0,colsample_bytree=0.75,subsample=1)

xgb.20.2000.001.75$model<-"M447-xgb.20.2000.001.75"

xgboostxcol<- rbind(xgb.20.500.003,xgb.20.500.003.5, xgb.20.500.003.125, xgb.20.500.003.25,
            xgb.20.500.003.50, xgb.20.500.003.75,xgb.20.2000.001, xgb.20.2000.001.5, xgb.20.2000.001.125, xgb.20.2000.001.25,
            xgb.20.2000.001.50, xgb.20.2000.001.75)

library(dplyr)
library(highcharter)
options(highcharter.theme = hc_theme_smpl(tooltip = list(valueDecimals = 4)))
hcboxplot(
  outliers = FALSE,
  x = xgboostxcol$auc,
  var = xgboostxcol$model,
  name = "MODELO",
  color = "#92D050"
```

```
) %>%
  hc_title(text = "MODELOS xgboost") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")

gradientboost <- rbind (xgb.20.2000.001.125, gbm.min.100.1500.001)
gradientsimples<- rbind (xgb.20.500.003.125, gbm.min.100.500.005)
gradient<- rbind (xgb.20.2000.001.125, gbm.min.100.500.005)
hcboxplot(
  outliers = FALSE,
  x = gradient$auc,
  var = gradient$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS xgboost") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")


mean(gbm.min.100.500.005$auc)
```

SVM

```
####SEL RANDOMSELECT####
set.seed(12345)
SVMgrid<-expand.grid(C=c(0.0001,0.001,1, 100, 1000, 10000))

control<-trainControl(method = "cv",number=5,savePredictions = "all")

SVMrandomlinear<- train(data=redrandom,factor(OPOSVACUNA)~.,
        method="svmLinear",trControl=control,
        tuneGrid=SVMgrid,verbose=FALSE)


SVMrandomlinear$results
plot(SVMrandomlinear$results$C,SVMrandomlinear$results$Accuracy)


SVMgridradial<-expand.grid(C=c(0.0001,0.001,0.01,0.05,0.1,0.2,0.5,1,2,5,10,25, 100, 1000, 10000, 100000),
            sigma=c(0.0001,0.005,0.01,0.05))

SVMrandomradial<- train(data=redsel,factor(OPOSVACUNA)~.,
        method="svmRadial",trControl=control,
        tuneGrid=SVMgridradial,verbose=FALSE)

SVMrandomradial

dat1<-as.data.frame(SVMrandomradial$results)

ggplot(dat1, aes(x=factor(C), y=Accuracy,
           color=factor(sigma)))+
 geom_point(position=position_dodge(width=0.5),size=3)


#####SEL MARIO#####


control<-trainControl(method = "cv",number=5,savePredictions = "all")

SVMmariolinear<- train(data=redmario,factor(OPOSVACUNA)~.,
                method="svmLinear",trControl=control,
                tuneGrid=SVMgrid,verbose=FALSE)


SVMmariolinear$results
plot(SVMmariolinear$results$C,SVMmariolinear$results$Accuracy)

SVMmarioradial<- train(data=redmario,factor(OPOSVACUNA)~.,
                method="svmRadial",trControl=control,
```

```
                tuneGrid=SVMgridradial,verbose=FALSE)

SVMmarioradial

dat3<-as.data.frame(SVMmarioradial$results)

ggplot(dat3, aes(x=factor(C), y=Accuracy,
            color=factor(sigma)))+
 geom_point(position=position_dodge(width=0.5),size=3)


####SEL miner

set.seed(12345)


control<-trainControl(method = "cv",number=5,savePredictions = "all")

SVMsellinear<- train(data=redsel,factor(OPOSVACUNA)~.,
                method="svmLinear",trControl=control,
                tuneGrid=SVMgrid,verbose=FALSE)


SVMsellinear$results
plot(SVMsellinear$results$C,SVMsellinear$results$Accuracy)

SVMselradial<- train(data=redsel,factor(OPOSVACUNA)~.,
                method="svmRadial",trControl=control,
                tuneGrid=SVMgridradial,verbose=FALSE)


SVMselradial$results
dat5<-as.data.frame(SVMselradial2$results)
library (ggplot2)
ggplot(dat5, aes(x=factor(C), y=Accuracy,
            color=factor(sigma)))+
 geom_point(position=position_dodge(width=0.5),size=3)

load("regresionTFM.RData")


SVMgridradial2<-expand.grid(C=c(0.0001,0.001,0.01,0.05,0.1,0.2,0.5,1,2,5,10,25, 100, 1000, 10000, 100000),
                  sigma=c(0.1))


SVMselradial2<- train(data=redsel,factor(OPOSVACUNA)~.,
               method="svmRadial",trControl=control,
               tuneGrid=SVMgridradial2,verbose=FALSE)


####LINEAL
cruzadaSVMbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,
  C=1,replace=TRUE)
 {

if (any(listclass==c(""))==FALSE)
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else  {
   databis<-data[,c(vardep,listconti)]
  }


  means <-apply(databis[,listconti],2,mean)
  sds<-sapply(databis[,listconti],sd)
```

```
 datacon<-scale(databis[,listconti], center = means, scale = sds)
 numerocont<-which(colnames(databis)%in%listconti)
 databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

 databis[,vardep]<-as.factor(databis[,vardep])

 formu<-formula(paste("factor(",vardep,")~.",sep=""))


 set.seed(sinicio)
 control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
  savePredictions = "all",classProbs=TRUE)


 SVMgrid <-expand.grid(C=C)

 SVM<- train(formu,data=databis,
  method="svmLinear",trControl=control,
  tuneGrid=SVMgrid,replace=replace)

 print(SVM$results)

 preditest<-SVM$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }


  tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
tasa=1-tasafallos(paso1$pred,paso1$obs)
medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"


 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }



  mediasbis<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
auc=suppressMessages(auc(paso1$obs,paso1$Yes))
mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"


 medias$auc<-mediasbis$auc

 return(medias)

}
####SEL RANDOM####
```

```
SVMlinealagr<-cruzadaSVMbin(data=redrandom, vardep="OPOSVACUNA",
                listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                        "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                        "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                listclass=c(""),
                grupos=5,sinicio=12345,repe=50,
                C=1)
SVMlinealagr$model<- "M448-SVMlinealagr"
###SEL MARIO###

SVMlinealmar<-cruzadaSVMbin(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                            "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                            "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                            "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=1)
SVMlinealmar$model<- "M449-SVMlinealmar"
####SEL MINER####
SVMlinealmin<-cruzadaSVMbin(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                            "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                            "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                            "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                            "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                            "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                            "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                            "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                            "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                            "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                            "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                            "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                            "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=1)
SVMlinealmin$model<- "M450-SVMlinealmin"

mejoresSVMlineal <- rbind (SVMlinealagr, SVMlinealmar, SVMlinealmin)

############RBF######

cruzadaSVMbinRBF<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,
  C=1,sigma=1)
 {


if (any(listclass==c(""))==FALSE)
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else  {
   databis<-data[,c(vardep,listconti)]
  }


  means <-apply(databis[,listconti],2,mean)
  sds<-sapply(databis[,listconti],sd)


  datacon<-scale(databis[,listconti], center = means, scale = sds)
  numerocont<-which(colnames(databis)%in%listconti)
  databis<-cbind(datacon,databis[,-numerocont,drop=FALSE ])

  databis[,vardep]<-as.factor(databis[,vardep])
```

```
 formu<-formula(paste("factor(",vardep,")~.",sep=""))


 set.seed(sinicio)
 control<-trainControl(method = "repeatedcv",number=grupos,repeats=repe,
  savePredictions = "all",classProbs=TRUE)


 SVMgrid <-expand.grid(C=C,sigma=sigma)

 SVM<- train(formu,data=databis,
  method="svmRadial",trControl=control,
  tuneGrid=SVMgrid,replace=replace)

 print(SVM$results)

 preditest<-SVM$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold <- sapply(preditest$prueba, "[", 1)
 preditest$Rep <- sapply(preditest$prueba, "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }


  tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
tasa=1-tasafallos(paso1$pred,paso1$obs)
medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"



 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }



  mediasbis<-data.frame()
for (repi in listarep) {
paso1<-preditest[which(preditest$Rep==repi),]
auc=suppressMessages(auc(paso1$obs,paso1$Yes))
mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"



 medias$auc<-mediasbis$auc

 return(medias)

}

##SEL RAN

SVMRBFagr.005.1<-cruzadaSVMbinRBF(data=redrandom, vardep="OPOSVACUNA",
          listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
```

```
        "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
        "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
        listclass=c(""),
        grupos=5,sinicio=12345,repe=50,
        C=1, sigma=0.05)
SVMRBFagr.005.1$model<- "M451-SVMRBFagr.005.1"

SVMRBFagr.001.1<-cruzadaSVMbinRBF(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=1, sigma=0.01)
SVMRBFagr.001.1$model<- "M452-SVMRBFagr.001.1"

SVMRBFagr.0005.2<-cruzadaSVMbinRBF(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=2, sigma=0.005)
SVMRBFagr.0005.2$model<- "M453-SVMRBFagr.0005.2"

SVMRBFagr.00001.10000<-cruzadaSVMbinRBF(data=redrandom, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.No", "PREF_PRES.Otro", "PREF_PRES.PabloIglesias",
                            "PREF_PRES.SantiagoAbascal", "EFEC_COVID.Economia", "PREO_COVID.Mucho",
                            "PREO_COVID.Nada", "PREO_COVID.Poco", "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=10000, sigma=0.0001)
SVMRBFagr.00001.10000$model<- "M454-SVMRBFagr.00001.10000"

#SELMARIO

SVMRBFmar.005.05<-cruzadaSVMbinRBF(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                            "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                            "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                            "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=05, sigma=0.05)
SVMRBFmar.005.05$model<- "M454-SVMRBFmar.005.05"

SVMRBFmar.001.2<-cruzadaSVMbinRBF(data=redmario, vardep="OPOSVACUNA",
                    listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                            "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                            "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                            "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                            "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=2, sigma=0.01)
SVMRBFmar.001.2$model<- "M455-SVMRBFmar.001.2"

SVMRBFmar.0005.10<-cruzadaSVMbinRBF(data=redmario, vardep="OPOSVACUNA",
                     listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
                             "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                             "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                             "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                             "VAL_PS", "EDAD"),
                     listclass=c(""),
                     grupos=5,sinicio=12345,repe=50,
                     C=10, sigma=0.005)
SVMRBFmar.0005.10$model<- "M456-SVMRBFmar.0005.10"

SVMRBFmar.00001.10000<-cruzadaSVMbinRBF(data=redmario, vardep="OPOSVACUNA",
                        listconti=c("PREF_PRES.Ines", "PREF_PRES.inigo", "PREF_PRES.No",
```

```
                        "PREF_PRES.Otro", "PREF_PRES.PabloCa", "PREF_PRES.PabloIg",
                        "PREF_PRES.Pedro", "PREF_PRES.Santiago", "EFEC_COVID.Economia",
                        "EFEC_COVID.Salud", "PREO_COVID.Mucho", "PREO_COVID.Nada", "PREO_COVID.Poco",
                        "VAL_PS", "EDAD"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=10000, sigma=0.0001)
SVMRBFmar.00001.10000$model<- "M457-SVMRBFmar.00001.10000"

####SEL MINER


SVMRBFmin.005.1<-cruzadaSVMbinRBF(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                          "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                          "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                          "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                          "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=1, sigma=0.05)
SVMRBFmin.005.1$model<- "M458-SVMRBFmin.005.1"

SVMRBFmin.001.1<-cruzadaSVMbinRBF(data=redsel, vardep="OPOSVACUNA",
                    listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                    "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                    "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                    "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                    "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                    "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                    "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                    "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                    "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                    "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                    "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                    "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                    "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                    "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                    listclass=c(""),
                    grupos=5,sinicio=12345,repe=50,
                    C=1, sigma=0.01)
SVMRBFmin.001.1$model<- "M459-SVMRBFmin.001.1"

SVMRBFmin.0005.2<-cruzadaSVMbinRBF(data=redsel, vardep="OPOSVACUNA",
                     listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                          "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                          "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                          "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                          "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                          "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                          "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                          "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                          "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                          "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                          "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                          "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                          "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                          "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                     listclass=c(""),
                     grupos=5,sinicio=12345,repe=50,
                     C=2, sigma=0.005)
SVMRBFmin.0005.2$model<- "M460-SVMRBFmin.0005.2"
```

```
SVMRBFmin.00001.10000<-cruzadaSVMbinRBF(data=redsel, vardep="OPOSVACUNA",
                        listconti=c("CIVIS_COVID.DUDA", "CIVIS_COVID.NO CIVISMO", "EFEC_COVID.Economia",
                                "EFEC_COVID.Salud", "PARTICIPACIONG.Si", "G_PREF_PRES.1", "G_PREF_PRES.2",
                                "G_PREF_PRES.3", "G_PREO_COVID.1", "G_PREO_COVID.2", "G_INTENCIONG.1",
                                "G_INTENCIONG.2", "G_INTENCIONG.3", "G_RECUVOTOG.1", "G_RECUVOTOG.2",
                                "G_RECUVOTOG.3", "G_RECUVOTOG.4", "G_PRO_PRI.1", "G_PRO_PRI.2",
                                "G_PRO_PRI.3", "G_SITLAB.1", "G_SITLAB.2", "G_SITLAB.3", "G_CNO11.1",
                                "G_CNO11.2", "G_CNO11.3", "G_CNO11.4", "G_VAL_ECO.1", "G_VAL_ECO.2",
                                "G_PRO_SOC.1", "G_PRO_SOC.2", "G_PRO_SOC.3", "G_PRO_SOC.4", "G_ECIVIL.1",
                                "G_ECIVIL.2", "G_INTENCIONGALTER.1", "G_INTENCIONGALTER.2", "G_INTENCIONGALTER.3",
                                "G_INTENCIONGALTER.4", "G_NIVELESTENTREV.1", "G_NIVELESTENTREV.2",
                                "G_NIVELESTENTREV.3", "G_NIVELESTENTREV.4", "G_NIVELESTENTREV.5",
                                "G_VAL_ECO_PER.1", "G_VAL_ECO_PER.2", "G_VAL_ECO_PER.3", "G_CCAA.1",
                                "G_CCAA.2", "G_CCAA.3", "G_CCAA.4", "G_CCAA.5", "G_GOB_ENCAR.1",
                                "G_GOB_ENCAR.2", "G_GOB_ENCAR.3", "VAL_PS", "EDAD", "numMissing"),
                        listclass=c(""),
                        grupos=5,sinicio=12345,repe=50,
                        C=10000, sigma=0.0001)
SVMRBFmin.00001.10000$model<- "M461-SVMRBFmin.00001.10000"

mejoresSVMRBF<- rbind(SVMRBFagr.005.1, SVMRBFagr.001.1, SVMRBFagr.0005.2, SVMRBFagr.00001.10000,
                SVMRBFmar.005.05, SVMRBFmar.001.2, SVMRBFmar.0005.10, SVMRBFmar.00001.10000,
                SVMRBFmin.005.1, SVMRBFmin.001.1, SVMRBFmin.0005.2, SVMRBFmin.00001.10000)
save.image("regresiontfm.RData")
mejoresSVM <- rbind (mejoresSVMlineal, mejoresSVMRBF)
library(dplyr)
library(highcharter)
options(highcharter.theme = hc_theme_smpl(tooltip = list(valueDecimals = 4)))
hcboxplot(
  outliers = FALSE,
  x = mejoresSVMRBF$auc,
  var = mejoresSVMRBF$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS SVM") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column")


hcboxplot(
  outliers = FALSE,
  x = mejoresSVM$auc,
  var = mejoresSVM$model,
  name = "MODELO",
  color = "#92D050"
) %>%
  hc_title(text = "MODELOS SVM") %>%
  hc_yAxis(title = list(text = "AUC")) %>%
  hc_chart(type = "column"
```

Ensamblado

```
.
set.seed(12345)
control<-trainControl(method = "repeatedcv",number=5,repeats=50,
                savePredictions = "all",classProbs=TRUE)

#####KNN######
knnmod<- train(OPOSVACUNA~.,data=redsel,
               trControl=control,method="knn",tuneGrid=expand.grid(k=95))
preditest<-knnmod$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
```

```
medias<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 tasa=1-tasafallos(paso1$pred,paso1$obs)
 medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"

mediasbis<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 auc=suppressMessages(auc(paso1$obs,paso1$Yes))
 mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"

medias$auc<-mediasbis$auc

ensknn<-(list(medias,preditest))
ensknnbis<-as.data.frame(ensknn[1])
ensknnbis$modelo<-"KNN"
predknn<-as.data.frame(ensknn[2])
predknn$knn<-predknn$Yes




#####LOGISTICA####

regresion <- train(RegTransDumStepAIC$formula,data=regtransdumm,
                   trControl=control,method="glm",family = binomial(link="logit"))
preditest<-regresion$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 tasa=1-tasafallos(paso1$pred,paso1$obs)
 medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"

mediasbis<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 auc=suppressMessages(auc(paso1$obs,paso1$Yes))
 mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"

medias$auc<-mediasbis$auc

enslogi<-(list(medias,preditest))
enslogibis<-as.data.frame(enslogi[1])
enslogibis$modelo<-"LogÃ-stica"
predlogi<-as.data.frame(enslogi[2])
predlogi$logi<-predlogi$Yes

#######REDES######
set.seed(12345)
avnnetgrid <-expand.grid(size=c(2),decay=c(0.01),bag=FALSE)
redmod <- train(OPOSVACUNA~., data=redrandom, method="avNNet", linout = FALSE,
                maxit=100,trControl=control,repeats=5,tuneGrid=avnnetgrid)
```

```
preditest<-redmod$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 tasa=1-tasafallos(paso1$pred,paso1$obs)
 medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"

mediasbis<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 auc=suppressMessages(auc(paso1$obs,paso1$Yes))
 mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"

medias$auc<-mediasbis$auc

ensred<-(list(medias,preditest))
ensredbis<-as.data.frame(ensred[1])
ensredbis$modelo<-"Red"
predred<-as.data.frame(ensred[2])
predred$red<-predred$Yes

#####RANDOM FOREST
set.seed(12345)
rfmod <- train(OPOSVACUNA~.,data=redsel,
               method="rf",trControl=control,
               tuneGrid=expand.grid(mtry=5),nodesize=20,replace=TRUE,sampsize=4400,
               ntree=1000)
preditest<-rfmod$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 tasa=1-tasafallos(paso1$pred,paso1$obs)
 medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"

mediasbis<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 auc=suppressMessages(auc(paso1$obs,paso1$Yes))
 mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"

medias$auc<-mediasbis$auc

ensrf<-(list(medias,preditest))
ensrfbis<-as.data.frame(ensrf[1])
ensrfbis$modelo<-"RF"
predrf<-as.data.frame(ensrf[2])
```

```
predrf$rf<-predrf$Yes


######GBM############
set.seed(12345)
gbmmod <- train(factor(OPOSVACUNA)~.,data=redsel,
          method="gbm",trControl=control,tuneGrid=expand.grid(shrinkage=0.01,
                                        n.minobsinnode=100,
                                        n.trees=1500,
                                        interaction.depth=2),
          distribution="bernoulli", bag.fraction=1,verbose=FALSE)
preditest<-gbmmod$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 tasa=1-tasafallos(paso1$pred,paso1$obs)
 medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"

mediasbis<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 auc=suppressMessages(auc(paso1$obs,paso1$Yes))
 mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"

medias$auc<-mediasbis$auc

ensgbm<-(list(medias,preditest))
ensgbmbis<-as.data.frame(ensgbm[1])
ensgbmbis$modelo<-"GBM"
predgbm<-as.data.frame(ensgbm[2])
predgbm$gbm<-predgbm$Yes

####XGBOOST####
set.seed(12345)
xgbmod <- train(factor(OPOSVACUNA)~.,data=redsel,
          method="xgbTree",trControl=control,
          tuneGrid=expand.grid(
            min_child_weight=20,
            eta=0.01,
            nrounds=2000,
            max_depth=2,gamma=0,colsample_bytree=0.125,subsample=1),
          verbose=FALSE)
preditest<-xgbmod$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 tasa=1-tasafallos(paso1$pred,paso1$obs)
 medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"
```

```
mediasbis<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 auc=suppressMessages(auc(paso1$obs,paso1$Yes))
 mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"

medias$auc<-mediasbis$auc

ensxgb<-(list(medias,preditest))
ensxgbbis<-as.data.frame(ensxgb[1])
ensxgbbis$modelo<-"XGBOOST"
predxgb<-as.data.frame(ensxgb[2])
predxgb$xgb<-predxgb$Yes


######SVM####
set.seed(12345)
svmmod<- train(data=redsel,factor(OPOSVACUNA)~.,
           method="svmRadial",trControl=control,
           tuneGrid=expand.grid(C=1,
          sigma=0.05),verbose=FALSE)
preditest<-svmmod$pred

preditest$prueba<-strsplit(preditest$Resample,"[.]")
preditest$Fold <- sapply(preditest$prueba, "[", 1)
preditest$Rep <- sapply(preditest$prueba, "[", 2)
preditest$prueba<-NULL

tabla<-table(preditest$Rep)
listarep<-c(names(tabla))
medias<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 tasa=1-tasafallos(paso1$pred,paso1$obs)
 medias<-rbind(medias,tasa)
}
names(medias)<-"tasa"

mediasbis<-data.frame()
for (repi in listarep) {
 paso1<-preditest[which(preditest$Rep==repi),]
 auc=suppressMessages(auc(paso1$obs,paso1$Yes))
 mediasbis<-rbind(mediasbis,auc)
}
names(mediasbis)<-"auc"

medias$auc<-mediasbis$auc

enssvm<-(list(medias,preditest))
enssvmbis<-as.data.frame(enssvm[1])
enssvmbis$modelo<-"SVM"
predsvm<-as.data.frame(enssvm[2])
predsvm$svm<-predsvm$Yes



#####JUNTAR los DATOS####

unipredi<-cbind(predlogi, predred, predrf, predxgb,predsvm,predknn)
unipredi<- unipredi[, !duplicated(colnames(unipredi))]
dput(names(unipredi))
#COMBINACIONES DE 2

unipredi$logi.red<-(unipredi$logi+unipredi$red)/2
unipredi$logi.rf<-(unipredi$logi+unipredi$rf)/2
unipredi$logi.knn<-(unipredi$logi+unipredi$knn)/2
unipredi$logi.xgb<-(unipredi$logi+unipredi$xgb)/2
unipredi$logi.svm<-(unipredi$logi+unipredi$svm)/2
unipredi$red.rf<-(unipredi$red+unipredi$rf)/2
```

```
unipredi$red.knn<-(unipredi$red+unipredi$knn)/2
unipredi$red.xgb<-(unipredi$red+unipredi$xgb)/2
unipredi$red.svm<-(unipredi$red+unipredi$svm)/2
unipredi$rf.knn<-(unipredi$rf+unipredi$knn)/2
unipredi$rf.xgb<-(unipredi$rf+unipredi$xgb)/2
unipredi$rf.svm<-(unipredi$rf+unipredi$svm)/2
unipredi$knn.xgb<-(unipredi$knn+unipredi$xgb)/2
unipredi$knn.svm<-(unipredi$knn+unipredi$svm)/2
unipredi$xgb.svm<-(unipredi$xgb+unipredi$svm)/2

#COMBINACIONES DE 3
unipredi$logi.red.rf<-(unipredi$logi+unipredi$red+unipredi$rf)/3
unipredi$logi.red.knn<-(unipredi$logi+unipredi$red+unipredi$knn)/3
unipredi$logi.red.xgb<-(unipredi$logi+unipredi$red+unipredi$xgb)/3
unipredi$logi.red.svm<-(unipredi$logi+unipredi$red+unipredi$svm)/3
unipredi$logi.rf.knn<-(unipredi$logi+unipredi$rf+unipredi$knn)/3
unipredi$logi.rf.xgb<-(unipredi$logi+unipredi$rf+unipredi$xgb)/3
unipredi$logi.rf.svm<-(unipredi$logi+unipredi$rf+unipredi$svm)/3
unipredi$logi.knn.xgb<-(unipredi$logi+unipredi$knn+unipredi$xgb)/3
unipredi$logi.knn.svm<-(unipredi$logi+unipredi$knn+unipredi$svm)/3
unipredi$logi.xgb.svm<-(unipredi$logi+unipredi$xgb+unipredi$svm)/3

unipredi$red.rf.knn<-(unipredi$red+unipredi$rf+unipredi$knn)/3
unipredi$red.rf.xgb<-(unipredi$red+unipredi$rf+unipredi$xgb)/3
unipredi$red.rf.svm<-(unipredi$red+unipredi$rf+unipredi$svm)/3
unipredi$red.knn.xgb<-(unipredi$red+unipredi$knn+unipredi$xgb)/3
unipredi$red.knn.svm<-(unipredi$red+unipredi$knn+unipredi$svm)/3
unipredi$red.xgb.svm<-(unipredi$red+unipredi$xgb+unipredi$svm)/3

unipredi$rf.knn.xgb<-(unipredi$rf+unipredi$knn+unipredi$xgb)/3
unipredi$rf.xgb.svm<-(unipredi$rf+unipredi$xgb+unipredi$svm)/3
unipredi$rf.knn.svm<-(unipredi$rf+unipredi$knn+unipredi$svm)/3

unipredi$knn.xgb.svm<-(unipredi$knn+unipredi$xgb+unipredi$svm)/3

#COMBINACIONES DE 4


unipredi$logi.red.rf.knn<-(unipredi$logi+unipredi$red+unipredi$rf+unipredi$knn)/4
unipredi$logi.red.rf.xgb<-(unipredi$logi+unipredi$red+unipredi$rf+unipredi$xgb)/4
unipredi$logi.red.knn.svm<-(unipredi$logi+unipredi$red+unipredi$knn+unipredi$svm)/4
unipredi$logi.red.knn.xgb<-(unipredi$logi+unipredi$red+unipredi$knn+unipredi$xgb)/4
unipredi$logi.red.rf.svm<-(unipredi$logi+unipredi$red+unipredi$rf+unipredi$svm)/4
unipredi$logi.red.svm.xgb<-(unipredi$logi+unipredi$red+unipredi$svm+unipredi$xgb)/4
unipredi$logi.rf.knn.xgb<-(unipredi$logi+unipredi$rf+unipredi$knn+unipredi$xgb)/4
unipredi$logi.rf.knn.svm<-(unipredi$logi+unipredi$rf+unipredi$knn+unipredi$svm)/4
unipredi$logi.rf.xgb.svm<-(unipredi$logi+unipredi$rf+unipredi$xgb+unipredi$svm)/4
unipredi$logi.knn.xgb.svm<-(unipredi$logi+unipredi$knn+unipredi$xgb+unipredi$svm)/4
unipredi$red.rf.xgb.svm<-(unipredi$red+unipredi$rf+unipredi$xgb+unipredi$svm)/4
unipredi$red.rf.knn.svm<-(unipredi$red+unipredi$rf+unipredi$knn+unipredi$svm)/4
unipredi$red.rf.knn.xgb<-(unipredi$red+unipredi$rf+unipredi$knn+unipredi$xgb)/4
unipredi$red.knn.xgb.svm<-(unipredi$red+unipredi$knn+unipredi$xgb+unipredi$svm)/4
unipredi$rf.knn.xgb.svm<-(unipredi$rf+unipredi$knn+unipredi$xgb+unipredi$svm)/4

####COMBINACIONES DE 5
unipredi$logi.red.rf.knn.xgb<-(unipredi$logi+unipredi$rf+unipredi$xgb+unipredi$red+unipredi$knn)/5
unipredi$logi.red.rf.knn.svm<-(unipredi$logi+unipredi$rf+unipredi$knn+unipredi$red+unipredi$svm)/5
unipredi$logi.red.rf.svm.xgb<-(unipredi$logi+unipredi$rf+unipredi$xgb+unipredi$red+unipredi$svm)/5
unipredi$logi.red.svm.knn.xgb<-(unipredi$logi+unipredi$svm+unipredi$xgb+unipredi$red+unipredi$knn)/5
unipredi$logi.svm.rf.knn.xgb<-(unipredi$logi+unipredi$rf+unipredi$xgb+unipredi$svm+unipredi$knn)/5
unipredi$svm.red.rf.knn.xgb<-(unipredi$red+unipredi$rf+unipredi$xgb+unipredi$red+unipredi$knn)/5

unipredi$all<-(unipredi$logi+unipredi$rf+unipredi$xgb+unipredi$red+unipredi$knn+unipredi$svm)/5
dput(names(unipredi))
listado <- c("logi","red",
             "rf",
             "knn", "xgb","svm",
             "logi.red", "logi.rf", "logi.knn", "logi.xgb", "logi.svm", "red.rf",
             "red.knn", "red.xgb", "red.svm", "rf.knn", "rf.xgb", "rf.svm",
             "knn.xgb", "knn.svm", "xgb.svm", "logi.red.rf", "logi.red.knn",
             "logi.red.xgb", "logi.red.svm", "logi.rf.knn", "logi.rf.xgb",
```

```
          "logi.rf.svm", "logi.knn.xgb", "logi.knn.svm", "logi.xgb.svm",
          "red.rf.knn", "red.rf.xgb", "red.rf.svm", "red.knn.xgb", "red.knn.svm",
          "red.xgb.svm", "rf.knn.xgb", "rf.xgb.svm", "rf.knn.svm", "knn.xgb.svm",
          "logi.red.rf.knn", "logi.red.rf.xgb", "logi.red.knn.svm", "logi.red.knn.xgb",
          "logi.red.rf.svm", "logi.red.svm.xgb", "logi.rf.knn.xgb", "logi.rf.knn.svm",
          "logi.rf.xgb.svm", "logi.knn.xgb.svm", "red.rf.xgb.svm", "red.rf.knn.svm",
          "red.rf.knn.xgb", "red.knn.xgb.svm", "rf.knn.xgb.svm", "logi.red.rf.knn.xgb",
          "logi.red.rf.knn.svm", "logi.red.rf.svm.xgb", "logi.red.svm.knn.xgb",
          "logi.svm.rf.knn.xgb", "svm.red.rf.knn.xgb", "all")


repeticiones<-nlevels(factor(unipredi$Rep))
unipredi$Rep<-as.factor(unipredi$Rep)
unipredi$Rep<-as.numeric(unipredi$Rep)


medias0<-data.frame(c())
for (prediccion in listado)
{
 unipredi$proba<-unipredi[,prediccion]
 unipredi[,prediccion]<-ifelse(unipredi[,prediccion]>0.5,"Yes","No")
 for (repe in 1:repeticiones)
 {
  paso <- unipredi[(unipredi$Rep==repe),]
  pre<-factor(paso[,prediccion])
  archi<-paso[,c("proba","obs")]
  archi<-archi[order(archi$proba),]
  obs<-paso[,c("obs")]
  tasa=1-tasafallos(pre,obs)
  t<-as.data.frame(tasa)
  t$modelo<-prediccion
  auc<-suppressMessages(auc(archi$obs,archi$proba))
  t$auc<-auc
  medias0<-rbind(medias0,t)
 }
}
par(cex.axis=0.5,las=2)
boxplot(data=medias0,tasa~modelo,col="pink",main="TASA FALLOS")

library(dplyr)
library(highcharter)
# Set highcharter options
options(highcharter.theme = hc_theme_smpl(tooltip = list(valueDecimals = 4)))
hcboxplot(
 outliers = FALSE,
 x = medias0$auc,
 var = medias0$modelo,
 name = "MODELO",
 color = "#92D050"
) %>%
 hc_title(text = "MODELOS RF") %>%
 hc_yAxis(title = list(text = "AUC")) %>%
 hc_chart(type = "column")

hcboxplot(
 outliers = FALSE,
 x = union1$auc,
 var = union1$modelo,
 name = "MODELO",
 color = "#92D050"
) %>%
 hc_title(text = "MODELOS RF") %>%
 hc_yAxis(title = list(text = "AUC")) %>%
 hc_chart(type = "column")

union1<-rbind(enslogibis,ensredbis,
        ensrfbis,ensknnbis,ensxgbbis,enssvmbis)
par(cex.axis=0.8)
boxplot(data=union1,tasa~modelo,col="pink",main='TASA FALLOS')
boxplot(data=union1,auc~modelo,col="pink",main='AUC')
```

```
observ <- cbind(predlogi$obs, predred$obs, predrf$obs, predgbm$pred, predxgb$obs, predsvm$obs)
aver<-as.data.frame(
  cbind(regtransdumm$OPOSVACUNA, redrandom$OPOSVACUNA, redsel$OPOSVACUNA))
```

Punto de corte

```
probabi<-predict(RegOrigDumForwBIC,testfinal,type = "response")
probabi2<-predict(RegTransDumStepAIC,testfinal,type = "response")
# MEDIDAS CON PUNTO DE CORTE 0.5
confusionMatrix(reference=sal$obs,data=sal$pred, positive="Yes")
testobs<-ifelse(probabi>0.5,"Yes","No")
testobs2<-ifelse(probabi2>0.5,"Yes","No")
corte<-0.2365

sal$predcorte<-ifelse(sal$Yes>corte,"Yes","No")
sal$predcorte<-as.factor(sal$predcorte)
testobs<-as.factor(testobs)
testobs2<-as.factor(testobs2)
vacuna<-as.factor(vacuna)
confusionMatrix(reference=vacuna,data=testobs2, positive="Yes")
confusionMatrix(reference=vacuna, data=testobs, positive="Yes")
roc(vacuna, probabi2, direction="<")

#Bï¿½squeda del mejor punto de corte para el ganador
test_roc<-roc(vacuna, probabi2, direction="<")
plot(test_roc,print.thres="best") #punto de corte maximiza youden
#busco el que iguala sensibilidad y especificidad
test_roc$thresholds[which.min(abs(test_roc$sensitivities-test_roc$specificities))]
#Represento estos, 0.5 y la prop. de eventos
plot(test_roc,print.thres=c(0.192,0.197,0.5))

probabi<-predict(RegTransDumStepAIC,testfinal,type = "response")
length (testfinal$OPOSVACUNA)
vacuna <- testfinal$OPOSVACUNA
length(probabi)
test_roc<-roc(testfinal$OPOSVACUNA, predict(RegTransDumStepAIC,testfinal,type = "response"), direction="<")
#Comparamos las 4 opciones
sensEspCorte(probabi2,testfinal,"OPOSVACUNA",0.2439,"Yes") #max. tasa de acierto
sensEspCorte(modeloStepBIC,data_test,"barrioCaro",0.107,"1") #max. ï¿½ndice de Youden
sensEspCorte(modeloStepBIC,data_test,"barrioCaro",0.016,"1") #igual sens. y esp.
sensEspCorte(modeloStepBIC,data_test,"barrioCaro",0.24,"1") #prop. eventos
#el tercero pierde mucha precisiï¿½n y 0.5 y 0.24 tienen mucha diferencia entre sens y esp
#elijo 0.107

######################PUNTOS CORTE MEJOR MODELO -test#####

probabi2<-predict(RegTransDumStepAIC,testfinal,type = "response")
testobs2<-ifelse(probabi2>0.5,"Yes","No")
testobs2<-as.factor(testobs2)
vacuna<-as.factor(vacuna)
roc(vacuna, probabi2, direction="<")
confusionMatrix(reference=vacuna,data=testobs2, positive="Yes")

probabi2<-predict(RegTransDumStepAIC,testfinal,type = "response")
testobs3<-ifelse(probabi2>0.2353,"Yes","No")
testobs3<-as.factor(testobs3)
vacuna<-as.factor(vacuna)
roc(vacuna, probabi2, direction="<")
confusionMatrix(reference=vacuna,data=testobs3, positive="Yes")

#Bï¿½squeda del mejor punto de corte para el ganador
test_roc<-roc(vacuna, probabi2, direction="<")
plot(test_roc,print.thres="best") #punto de corte maximiza youden
#busco el que iguala sensibilidad y especificidad
test_roc$thresholds[which.min(abs(test_roc$sensitivities-test_roc$specificities))]
#Represento estos, 0.5 y la prop. de eventos
plot(test_roc,print.thres=c(0.192,0.197,0.5))
```

```
testobs4<-ifelse(probabi2>0.192,"Yes","No")
testobs4<-as.factor(testobs4)
vacuna<-as.factor(vacuna)
roc(vacuna, probabi2, direction="<")
confusionMatrix(reference=vacuna,data=testobs4, positive="Yes")

testobs4<-ifelse(probabi2>0.197,"Yes","No")
testobs4<-as.factor(testobs4)
vacuna<-as.factor(vacuna)
roc(vacuna, probabi2, direction="<")
confusionMatrix(reference=vacuna,data=testobs4, positive="Yes")


testobs5<-ifelse(probabi2>0.697,"Yes","No")
testobs5<-as.factor(testobs5)
vacuna<-as.factor(vacuna)
roc(vacuna, probabi2, direction="<")
confusionMatrix(reference=vacuna,data=testobs5, positive="Yes")

###############modelo simple#######

probabi<-predict(RegOrigDumForwBIC,testfinal,type = "response")
roc(vacuna, probabi, direction="<")

testobs6<-ifelse(probabi>0.5,"Yes","No")
testobs6<-as.factor(testobs6)
confusionMatrix(reference=vacuna,data=testobs6, positive="Yes")

testobs7<-ifelse(probabi>0.2335,"Yes","No")
testobs7<-as.factor(testobs7)
confusionMatrix(reference=vacuna,data=testobs7, positive="Yes")



#Búsqueda del mejor punto de corte para el ganador
test_roc<-roc(vacuna, probabi, direction="<")
plot(test_roc,print.thres="best") #punto de corte maximiza youden
#busco el que iguala sensibilidad y especificidad
test_roc$thresholds[which.min(abs(test_roc$sensitivities-test_roc$specificities))]

testobs8<-ifelse(probabi>0.285,"Yes","No")
testobs8<-as.factor(testobs8)
confusionMatrix(reference=vacuna,data=testobs8, positive="Yes")


testobs9<-ifelse(probabi>0.221,"Yes","No")
testobs9<-as.factor(testobs9)
confusionMatrix(reference=vacuna,data=testobs9, positive="Yes")

testfinalver<- cbind(testfinal, probabi)
testfinalver <- testfinalver[, c(8,18,27,39,44,52,73,77,112,141,143,145,146,174,184,148,193,218,220)]
testfinalver<-testfinalver[order(testfinalver$probabi),]


jinetes <- rbind(testfinalver[1,], testfinalver[1380,])
library(reshape2)
yjinetes<-t(jinetes)
```