FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2020/2021

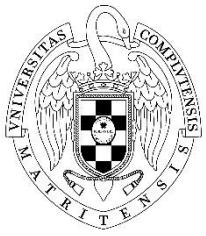
Trabajo de Fin de Máster

TÍTULO: *Aplicación de técnicas de minería de datos para la predicción del riesgo de violencia de género en España*

***Alumna*:** Alexandra Marianova Iordanova

***Tutora*:** Aida Calviño Martínez

Septiembre de 2021



UNIVERSIDAD COMPLUTENSE
MADRID

A las mujeres de mi vida.

Resumen

La violencia de género es un problema grave que las sociedades y gobiernos deben abordar empleando todos los recursos disponibles a su alcance –lo que exige un conocimiento profundo de la magnitud del mismo–, pues afecta a mujeres de todos los países del mundo.

Actualmente, muchas de las víctimas de violencia machista no se atreven a denunciar su situación, impidiendo que pueda investigarse el delito. Es por ello que el objetivo del presente trabajo ha sido el de hallar un modelo de predicción que identifique el posible riesgo de violencia hacia la mujer antes de que se produzca un daño mayor y, en muchos casos, irreversible.

Palabras clave: violencia de género, miedo, mujer, entrevistada, pareja, expareja, aprendizaje automático.

Abstract

Gender-based violence is a serious problem that must be addressed by both societies and governments using all the available resources –which requires a deep knowledge of the magnitude of it–, as it affects women from all around the world.

Currently, many of the victims of macho violence don't dare to denounce their situation, –which prevents the crime from being investigated–. The purpose of this study has been to find a prediction model that identifies possible risks of violence against women, before further and, in many cases, irreversible damage occurs.

Keywords: gender-based violence, fear, woman, interviewee, partner, ex-partner, machine learning.

Índice

Índice de tablas

Índice de ilustraciones

1. Introducción:

Cerca de 641 millones de mujeres sufren violencia de género en el mundo. Sin embargo, si se tiene en consideración el alto grado de estigmatización y el hecho de que muchos actos violentos no se denuncien, estas cifras son, probablemente, mucho mayores (Organización Mundial de la Salud, 2021).

En España, según la Estadística de Víctimas Mortales por Violencia de Género, fueron asesinadas 45 mujeres a manos de sus parejas o exparejas dejando a un total de 26 menores huérfanos el pasado año 2020 (Delegación del Gobierno contra la Violencia de Género, 2021). Pero ¿cuántas mujeres podrían acabar igual? En la última Memoria de la Fiscalía General del Estado correspondiente al año 2019 se contabilizaron, solo en ese año, 168.057 denuncias por violencia de género –lo que supone un crecimiento del 0,67% respecto a las denuncias que se produjeron en 2018–, resaltando que el porcentaje de las falsas rondaba el 0,004. Igualmente, aumentó el número de órdenes de protección y otras medidas cautelares solicitadas, el porcentaje de las concesiones y el número y porcentaje de sentencias condenatorias. A su vez, cabe destacar que las llamadas al 016 aumentaron un 47,3% durante la primera quincena de abril del 2020 con respecto a los datos del mismo periodo del 2019 (La Moncloa, 2020). En cambio, pese a esto, las denuncias en los Juzgados de Violencia sobre la Mujer durante el año 2020, según los datos del Consejo General del Poder Judicial, disminuyeron un 10,31%. No obstante, el gran error está en creer que una disminución de las denuncias implica una reducción de la violencia de género, pues no son un indicador del todo fiable.

Este tipo de maltrato es una lacra social que precisa de su erradicación. Las agresiones físicas no surgen de la nada y es que mucho antes del primer bofetón, empujón, golpe o paliza, se dan los abusos, intimidaciones o amenazas. La peor violencia no es otra que aquella que no se ve y pese a ser un problema de salud mundial de proporciones epidémicas (Organización Mundial de la Salud, 2013), no existe una vacuna para su erradicación. Es por esta razón por la que el propósito del presente trabajo es el de obtener un modelo de predicción que permita identificar el posible riesgo de violencia de género para intentar conseguir así que el menor número de mujeres sean agredidas por sus parejas o exparejas gracias a la temprana actuación por parte de los organismos e instituciones públicas.

Para la correcta ejecución del objetivo principal ha sido necesario el cumplimiento de los siguientes objetivos secundarios:

- Selección de variables útiles para el proyecto a partir de la "*Macroencuesta de Violencia contra la Mujer*" realizada en 2019 por la Delegación del Gobierno contra la Violencia de Género.
- Correcto análisis y tratamiento de los datos con la finalidad de no disponer ni de datos ausentes ni de datos atípicos, junto a unos valores mínimos y máximos razonables.
- Creación de múltiples modelos de predicción mediante el empleo de diversas técnicas de machine learning.
- Comparación y evaluación de los resultados alcanzados.
- Obtención del mejor modelo y análisis del mismo.

2. Estado del arte:

2.1. Concepto y tipos de violencia de género:

La violencia de género fue definida por primera vez por la Organización de las Naciones Unidas (1993) en su "*Declaración sobre la eliminación de la violencia contra la mujer*" como:

> todo acto de violencia basado en la pertenencia al sexo femenino que tenga o pueda tener como resultado un daño o sufrimiento físico, sexual o psicológico para la mujer, así como las amenazas de tales actos, la coacción o la privación arbitraria de la libertad, tanto si se producen en la vida pública como en la vida privada.

Por su parte, la Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género la define en su artículo 1 como aquella que:

> como manifestación de la discriminación, la situación de desigualdad y las relaciones de poder de los hombres sobre las mujeres, se ejerce sobre éstas por parte de quienes sean o hayan sido sus cónyuges o de quienes estén o hayan estado ligados a ellas por relaciones similares de afectividad, aun sin convivencia y, comprende todo acto de violencia física y psicológica, incluidas las agresiones a la libertad sexual, las amenazas, las coacciones o la privación arbitraria de libertad.

Cabe destacar que los términos de violencia doméstica o violencia intrafamiliar no son equivalentes a los de violencia de género, violencia machista o violencia hacia la mujer. El Grupo de Salud Mental del Programa de Actividades de Prevención y Promoción de la Salud de la Sociedad Española de Medicina de Familia y Comunitaria (2003) define a ambos dos primeros como "los malos tratos o agresiones físicas, psicológicas, sexuales o de otra índole, infligidas por personas del medio familiar y dirigida generalmente a los miembros más vulnerables de la misma: niños, mujeres y ancianos". Por lo que, no deben emplearse indistintamente, dado que puede llevar a una comprensión equívoca del problema.

Tras aclarar qué es la violencia de genero conviene a su vez conocer las formas en las que puede manifestarse. La Ley 13/2007, de 26 de noviembre, de medidas de prevención y protección integral contra la violencia de género expone varias tipologías:

a) Violencia física, que incluye cualquier acto no accidental que implique el uso deliberado de la fuerza del hombre contra el cuerpo de la mujer, así como los ejercidos en su entorno familiar o personal como forma de agresión a esta con resultado o riesgo de producir lesión física o daño.
b) Violencia psicológica, que incluye conductas verbales o no verbales, que produzcan en la mujer desvalorización o sufrimiento, a través de amenazas, humillaciones o vejaciones, exigencia de obediencia o sumisión, coerción, control, insultos, aislamiento, culpabilización o limitaciones de su ámbito de libertad, así como las ejercidas en su entorno familiar, laboral o personal como forma de agresión a la mujer.
c) Violencia sexual, que incluye cualquier acto de naturaleza sexual no consentido por la mujer, abarcando la imposición del mismo mediante

fuerza, intimidación o sumisión química, así como el abuso sexual, con independencia de la relación que el agresor guarde con la víctima.

d) Violencia económica, que incluye la privación intencionada y no justificada legalmente de recursos, incluidos los patrimoniales, para el bienestar físico o psicológico de la víctima, de sus hijos o hijas o de las personas de ella dependientes, o la discriminación en la disposición de los recursos que le correspondan legalmente o el imposibilitar el acceso de la mujer al mercado laboral con el fin de generar dependencia económica.

2.2. Inteligencia artificial contra la violencia de género:

El Sistema de Seguimiento Integral en los casos de Violencia de Género (Sistema VioGén), tal y como señala González et al. (2018), surge de la necesidad de desarrollar tareas de valoración y gestión del riesgo para registrar y proteger a las víctimas en función del mismo, así como comunicar a las autoridades judiciales de sus estimaciones. Se trata de una aplicación web a la que acceden usuarios de las Fuerzas y Cuerpos de Seguridad (Policía Nacional y Guardia Civil obligatoriamente, y las Policías Autonómicas y Locales que voluntariamente se adhieran), Instituciones Penitenciarias, Juzgados, Institutos de Medicina Legal y Ciencias Forenses, Oficinas de Asistencia a las Víctimas, Fiscalías, Delegaciones y Subdelegaciones del Gobierno y, Servicios Sociales y Organismos de Igualdad de las diferentes Comunidades Autónomas, quienes han conseguido ya realizar más de tres millones de valoraciones del riesgo.

Las múltiples causas y la baja frecuencia de conductas violentas graves hacen que su predicción sea una tarea difícil, pero, aun así, la Organización Mundial de la Salud (2014) afirma que "se puede predecir y prevenir, y la responsabilidad de abordarla recae sin duda alguna en los gobiernos nacionales". La predicción de la violencia se traduce en estimar la probabilidad de que se produzca una conducta violenta (Echeburúa et al., 2010).

En los últimos años, han surgido diversos instrumentos en el ámbito internacional con el fin de ser empleados en la predicción del riesgo de violencia contra la pareja, tales como el Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER), la Danger Assessment Tool (DA), los Threat Assessment Systems (DV-MOSAIC), la Ontario Domestic Assault Risk Assessment (ODARA), la Spousal Abuse Risk Assessment (SARA), el Domestic Violence Screening Instrument (DVSI y DVSI-R), el Kingston Screening Instrument for Domestic Violence (KSID) y The Spouse Violence Risk Assessment Inventory (SVRA-I). Dado que no están validados en el ámbito nacional, a excepción de la SARA, han ido diseñándose para la población española la Escala de Predicción del Riesgo de Violencia Grave contra la pareja - Revisada (EPV-R), el Protocolo de Valoración del Riesgo de Violencia contra la mujer por parte de su pareja o expareja (RVD-BCN) y el Protocolo de Valoración Policial del Riesgo de Reincidencia en Violencia de Género (VPRVPER).

La EPV-R, diseñada por Echeburúa et al. (2010), está compuesta por 20 ítems ponderados según su capacidad discriminativa para predecir la violencia grave hacia la pareja. Para su construcción y valoración se seleccionaron 450 expedientes de

agresores de pareja denunciados a la Ertzaintza en 2008, lográndose clasificar en tres niveles de riesgo: alto, moderado y bajo.

Por contra, la herramienta RVD-BCN, que nace en 2001 gracias al impulso del Ayuntamiento de Barcelona y el Consorcio Sanitario de Barcelona, tiene como finalidad valorar el riesgo de que a corto plazo se produzcan actos violentos graves por parte de la pareja o expareja, estimándose nuevamente en alto, moderado y bajo y consiguiéndose a través de la formulación de 16 factores (Álvarez et al., 2011).

Finalmente, el VPRVPER, que es el implementado en el Sistema VioGén desde 2007 y que, además, está constituido por los protocolos VPR (Valoración Policial del Riesgo) y VPER (Valoración Policial de la Evolución del Riesgo) –he de ahí su nombre–, clasifica los casos con cinco niveles de riesgo de reincidencia: no apreciado, bajo, medio, alto o extremo. Durante el seguimiento de la situación de la denunciante, para mantener actualizada la valoración del riesgo y actuar en consecuencia, las unidades policiales encargadas de su protección deben notificar siempre que precise de la existencia o inexistencia de un nuevo incidente. Tras lo anterior, se da traslado al Órgano Judicial y al Ministerio Fiscal tanto de la estimación inicial como de las últimas estimaciones que supongan una modificación a mayor o menor gravedad de la última valoración del riesgo manifestada, junto con un informe sobre los principales indicadores del riesgo apreciados (González et al., 2018).

Un estudio longitudinal prospectivo elaborado por López-Ossorio et al. (2017) trató de evaluar la eficacia predictiva de la VPR. Para ello, se llevó a cabo un seguimiento de 3 y 6 meses de 407 mujeres que denunciaron ser víctimas de violencia de género. Los resultados que se obtuvieron por medio de la regresión logística ofrecieron un AUC=0,71 (medida de rendimiento ampliamente utilizada para los modelos predictivos) para intervalos de tiempo en riesgo de 3 meses, mientras que de 0,58 para intervalos de 6 meses. A su vez, la sensibilidad fue del 85% y la especificidad del 53,7%. Por lo que, atendiendo a los datos que se extrajeron en este estudio se pudo concluir que el formulario VPR mostraba una buena capacidad predictiva, siendo altamente útil para tomar decisiones.

Relacionado con lo anterior, el 4 de septiembre de 2021, el periódico El País publicó un artículo titulado "*Matemáticas e inteligencia artificial contra el maltrato machista*", en el que se daba a conocer que un equipo de investigadores estaba desarrollando un nuevo sistema que mejoraba la predicción del riesgo de reincidencia. De hecho, uno de los investigadores del proyecto es José Ángel González-Prieto, profesor ayudante doctor en la Facultad de Ciencias Matemáticas de la Universidad Complutense de Madrid, quien mencionó que la clave de este sistema estaba en el empleo del machine learning.

Se estima que esta herramienta complementaria al Sistema VioGén ayude a mejorar la predicción de entre el 10 y 15%. Es decir, de los más de 600.000 casos que tiene registrados dicho sistema desde su arranque, entre 60.000 y 90.000 mujeres podrían tener una estimación más ajustada a sus necesidades de protección frente a la reincidencia. Pero, a pesar de los buenos resultados previstos, el sistema aun presenta ciertas limitaciones. Una de ellas, por ejemplo, es que el programa solo puede aplicarse cuando una víctima de violencia de género denuncia el maltrato ante los cuerpos de seguridad del Estado, por lo que no está preparado para integrarse en el sistema de asistencia de los servicios sociales.

Tal y como mencionó Pinedo (2021) en su publicación, "el mecanismo que vertebra este nuevo proyecto no es innovador, aunque sí su aplicación al ámbito de la prevención de la violencia de género", pues como punto de partida se ha tomado el funcionamiento de los sistemas de recomendación de plataformas como Netflix, dado que lo que se busca es aprender de los datos, como hacen muchas de las compañías que tienen perfilados a sus consumidores.

Dos de los términos que recién se acaban de mencionar son: machine learning y sistemas de recomendación. De acuerdo a Shalev-Shwartz y Ben-David (2014) el término de machine larning hace referencia a la detección automatizada de patrones significativos en los datos, siendo una de las ramas de la inteligencia artificial. Mientras que, los sistemas de recomendación son algoritmos que buscan predecir sugerencias que puedan ser de interés para el usuario. Son muchas las plataformas que emplean estos sistemas, como es el caso de Spotify recomendando canciones similares al género que suele uno escuchar, YouTube sugiriendo vídeos parecidos a los que recientemente se visualizaron o Amazon ofreciendo productos relacionados con el que se está interesado, haciendo incrementar así el valor de la venta.

Por último y no por ello menos importante, la entidad DigitalFems ha impulsado la plataforma "*Datos contra el ruido*" (https://datoscontraelruido.org/) con el propósito de visibilizar la realidad de la violencia de género, dado que en numerosas ocasiones se ve alterada por mentiras y opiniones subjetivas. También, a disposición de cualquiera, puede encontrarse GenderDataLab, un repositorio con archivos de información que contienen datos desagregados por género y/o con perspectiva de género y en formato abierto. Para conseguir tal fin, cuentan con un equipo especialista en el tratamiento de datos, OpenData y visualización de datos.

3. Metodología y técnicas empleadas:

3.1. Metodología SEMMA:

El esquema que se siguió para poder alcanzar los objetivos establecidos fue el basado en la metodología SEMMA –desarrollada por la empresa de software SAS–. El acrónimo anterior conforma las distintas fases del proceso (Azevedo y Santos, 2008), siendo estas:

- Sample (muestrear): extracción de una muestra lo suficientemente grande como para que contenga información significativa a la vez que lo suficientemente pequeña como para que sea fácilmente procesada.
- Explore (explorar): examinación de los datos con el fin de detectar relaciones, tendencias y anomalías que hagan comprender el conjunto de información.
- Modify (modificar): creación, selección y transformación de las variables para una óptima modelización.
- Model (modelizar): empleo de herramientas analíticas para buscar la combinación de datos que prediga de una forma fiable la variable de interés.
- Assess (evaluar): valoración de la calidad de las predicciones y comparación de diversos modelos logrados.

3.2. Selección de variables:

Para determinar las variables que iban a formar parte de los diversos modelos a elaborar se hizo uso de los programas SAS Enterprise Miner Workstation 14.1 y SAS 9.4, empleándose en cada uno de ellos procesos distintos de selección. Esto fue así porque en el caso de que se dispusiesen de variables input con relaciones no lineales, podía no ajustar bien la selección que se obtuvo mediante el modelo de regresión logística y el método stepwise con SAS 9.4. Para tratar de evitarlo, por medio del nodo de Selección de variables de SAS Enterprise Miner Workstation 14.1 se llevó a cabo una preselección más genérica y menos estricta, ya que, las técnicas basadas en árboles disponen de sus propios métodos de selección de variables.

3.2.1. Selección de variables para técnicas basadas en árboles:

Para las técnicas basadas en árboles, la selección de las variables más importantes, tal y como se acaba de introducir, se consiguió a través del nodo de Selección de variables de SAS Enterprise Miner Workstation 14.1, quien escogió todas las que consideró relevantes en función de su R^2, rechazando, por ende, todas las que no alcanzaron un valor mínimo de este estadístico. De esta forma, se aseguró no estar rechazando variables con cierto poder predictivo.

Dicho programa, para poder calcular el R^2 cuando la variable objetivo es binaria, como en nuestro caso, asigna el valor 1 a la clase mayoritaria y 0 a la clase minoritaria, pudiendo trabajar con ella como si de una variable continua se tratase. Este estadístico se define como:

$$R^2 = 1 - \frac{ASE_{modelo}}{ASE_{no\ modelo}} = 1 - \frac{SSE}{SST} \qquad [1]$$

donde:

$$ASE_{modelo} = SSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad [2]$$

y:

$$ASE_{no\ modelo} = SST = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad [3]$$

siendo ASE error cuadrático medio, SSE suma de cuadrados del error y SST suma de cuadrados total.

El valor del R^2 oscila entre 0 y 1, significando un valor de 0 independencia entre las variables y un valor de 1 dependencia total.

3.2.2. Selección de variables para técnicas no basadas en árboles:

Los modelos de regresión logística tienen su origen en los años 60, pero no es hasta principios de la década de los 80 cuando su uso se vuelve universal gracias a las facilidades informáticas con que se empieza a contar entonces. Dichos modelos tienen el objetivo de predecir la probabilidad de ocurrencia de un evento de la variable dependiente a partir de los datos de otras variables independientes conocidas, ya

sean las mismas cuantitativas o cualitativas (Fiuza y Rodríguez, 2000). De manera que, como lo que se obtienen son probabilidades, el resultado siempre será positivo y, además, variará entre 0 y 1.

La probabilidad de ocurrencia del evento se representa como:

$$P(Y = 1 \mid x_1, x_2, \dots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}} \quad [4]$$

mientras que, la probabilidad de no ocurrencia del evento como:

$$P(Y = 0 \mid x_1, x_2, \dots, x_m) = \frac{e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}} \quad [5]$$

donde $P(Y = 1 \mid x_1, x_2, \dots, x_m)$ es la probabilidad de que Y tome el valor 1 (o el valor 0), $x_1, x_2, \dots, x_m$ es el conjunto de m variables que forman parte del modelo, β_0 es la constante del modelo o término independiente y β_m los coeficientes de las distintas variables, soliéndose emplear el método de máxima verosimilitud para sus estimaciones (Ferre, 2019).

Por otra parte, para interpretar de forma correcta los parámetros de un modelo de regresión logística resulta necesario definir los conceptos de "*odds*" y "*odds-ratio*". Un "*odds*" no es más que el cociente entre la probabilidad de que ocurra un determinado suceso y la probabilidad de que no ocurra éste, expresándose como:

$$odds(Y = 1) = \frac{P(Y = 1)}{1 - P(Y = 1)} \quad [6]$$

mientras que, el "*odds-ratio*" se puede definir como "el cociente entre los odds de un suceso bajo una determinada condición y el odds de ese mismo suceso bajo otra condición, lo que permitirá evaluar el efecto de dichas condiciones sobre las probabilidades del suceso" (Calviño, 2020). Así pues, en este caso se expresaría de la forma:

$$OR = \frac{odds(evento \mid x = 1)}{odds(evento \mid x = 0)} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} \quad [7]$$

Finalmente, señalar que los modelos de regresión además de funcionar como modelos predictivos también lo hacen como métodos de selección de variables automáticos. Los procedimientos más habituales son:

- Método backward o hacia atrás: se comienza considerando todas las variables disponibles del modelo y se van eliminando una a una según su capacidad explicativa, hasta conseguir que todas ellas sean significativas.
- Método forward o hacia delante: se comienza partiendo de un modelo que no contiene ninguna variable y se van incorporando una a una aquellas que mayor mejora produzcan, hasta el punto de que la inclusión de una nueva variable en el modelo ya no aporte información.
- Método stepwise o paso a paso: es uno de los métodos más empleados y combina ambos dos anteriores. Es similar al método forward, pero a diferencia de éste, en el de stepwise es posible que la entrada de una nueva variable haga que otras que ya estaban dentro del modelo sean eliminadas, de acuerdo al método backward, pues pueden terminar resultando redundantes (*Selección de variables explicativas en la regresión*, 2007).

Por tanto, debido a que el método stepwise es el que mayoritariamente se usa fue el que se empleó en la selección de variables para técnicas no basadas en árboles.

3.3. Redes neuronales:

El modelo computacional de las redes neuronales está inspirado en el funcionamiento de las neuronas. La similitud entre una neurona biológica y una neurona artificial según Matich (2001) está en que ambas tienen entradas (dendritas), utilizan pesos (sinapsis) y generan salidas (axón). El esquema de una red neuronal sería tal que:

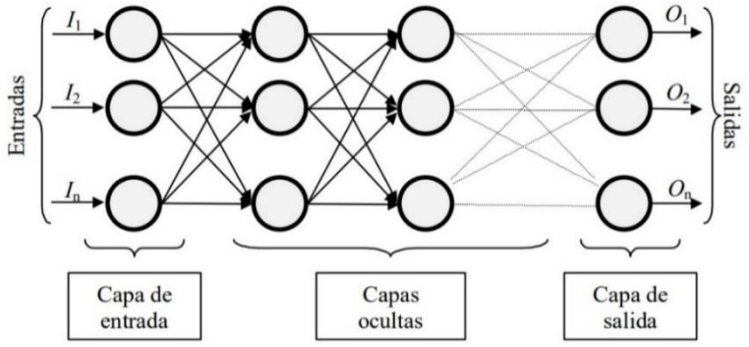


Ilustración 1. Estructura de una red neuronal.

Fuente: (Matich, 2001).

Tal y como puede verse en la Ilustración 1, en una red neuronal pueden distinguirse tres tipos de capas: de entrada, ocultas y de salida. La capa de entrada es aquella que recibe los datos, las ocultas son las intermedias, las encargadas del procesamiento de la información y la de salida es la que proporciona la respuesta de la red. A su vez, las redes también están compuestas por funciones de activaciones y pesos. Una función de activación "representa simultáneamente la salida de la neurona y su estado de activación" (Larranaga et al., 2019) mientras que los pesos miden la fuerza de una conexión de entrada.

Por último, señalar que el tuneo de todos los modelos se llevó a cabo desde RStudio, empleándose para la modelización el paquete "*caret*". En el caso de las redes neuronales los hiperparámetros que pueden modificarse son:

- "*size*": número de nodos en las capas ocultas.
- "*decay*" o "*learning rate*": determina, en un rango de 0 a 1, cuánto se actualizan los pesos en cada iteración. De forma más sencilla, el learning rate o tasa de aprendizaje se puede definir como la rapidez con que nuestra red reemplaza los conceptos que ha aprendido hasta el momento por otros nuevos.
- "*itera*": número de iteraciones máximas de la red.

Aunque con otros programas también puede estudiarse la función de activación y el algoritmo de optimización (teniendo como objetivo estimar los parámetros con el propósito de minimizar la función de error o función objetivo (Portela, 2021), siendo la expresión [2]).

3.4. Modelos basados en árboles:

Los árboles de decisión constituyen uno de los métodos más empleados del aprendizaje inductivo supervisado. Se caracterizan fundamentalmente por la sencillez de los modelos obtenidos y tienen como objetivo dividir y clasificar la información en grupos similares. Un árbol gráficamente se ilustra de la siguiente forma:

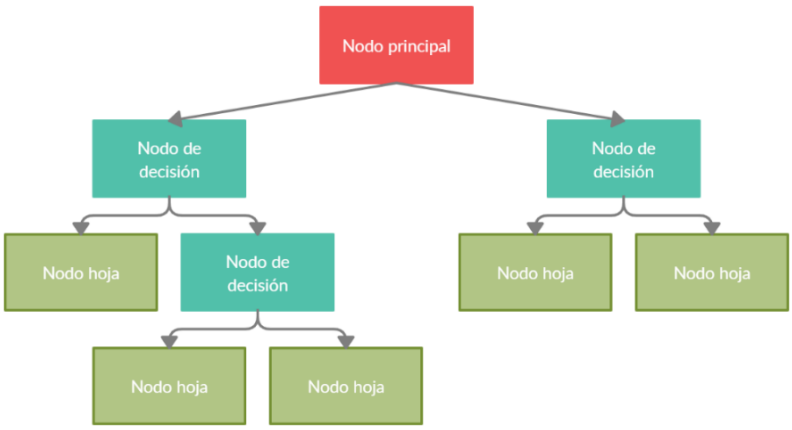


Ilustración 2. Estructura de un árbol de decisión.

El nodo principal es aquel que representa todos los datos de los que se disponen y que posteriormente se dividirá, el nodo de decisión es el encargado de fraccionar los datos en dos grupos, dependiendo, como su nombre ya indica, de una decisión y, finalmente, en el nodo hoja es donde recae la decisión final (Rodríguez, 2018).

Para terminar, los criterios de selección que se suelen utilizar para el punto de corte y la variable a utilizar por el nodo, tal y como recoge la profesora Calviño (2020) en su material docente, son:

- Índice de Gini: mide la varianza entre las clases, es decir, cómo de homogéneas son las observaciones de un nodo.
- Entropía: funciona de una manera muy similar a Gini, pues evalúa como de homogéneo es un nodo.
- Método de la X^2 (ProbChisq): permite evaluar la relación existente entre la variable dependiente y cualquier nueva variable que indique la hoja a la que correspondan las observaciones.

En particular, los modelos basados en árboles que se emplearon y que, por ende, van a explicarse a continuación son: bagging, random forest, gradient boosting y extreme gradient boosting.

3.4.1. Bagging:

El algoritmo de bagging (bootstrap aggregating) se utiliza cuando se tiene como objetivo reducir la varianza de un árbol de decisión. Aquí, la idea que se tiene es la de crear varios subconjuntos de datos a partir de una muestra de entrenamiento elegida al azar con reemplazo. Ahora, cada colección de datos de subconjuntos es

utilizada para entrenar sus árboles de decisión. Como resultado, se termina obteniendo un conjunto de diferentes modelos y se utiliza el promedio de todas las predicciones de los distintos árboles, debido a que se consigue así, una robustez mayor que la de un solo árbol de decisión (Nagpal, 2017).

De igual forma a como se ha expuesto en redes neuronales, en bagging los hiperparámetros que pueden variarse con la librería "*caret*" son:

- "*mtry*": número de variables a sortear en cada nodo, teniendo que ser todas las variables, no produciéndose por ello mismo sorteo.
- "*ntree*": número de árboles a promediar.
- "*nodesize*": tamaño mínimo de nodos finales.
- "*sampsize*": tamaño de la muestra a sortear.
- "*replace*": con reemplazamiento ("*replace=TRUE*") o sin reemplazamiento ("*replace=FALSE*").

Además, con el paquete "*randomForest*" puede plotearse el error OOB (out of bag) a medida que avanzan las iteraciones, siendo ésta una forma de validar el modelo (Bhatia, 2019).

3.4.2. Random forest:

El método de random forest consiste en una gran cantidad de árboles de decisión individuales que operan como un conjunto (Yiu, 2019). Este algoritmo funciona de forma similar al de bagging, salvo por la excepción de que el número de variables a sortear en cada nodo no son todas las variables.

Así pues, dado que los hiperparámetros a modificar son los mismos que en bootstrap aggregating, los modelos que se crearon contuvieron las mismas combinaciones de hiperparámetros aplicadas en dicho algoritmo, estudiándose y variándose tan solo el valor de "*mtry*".

3.4.3. Gradient boosting:

La técnica de gradient boosting se basa en la idea de que los nuevos predictores aprenden de los errores cometidos por los predictores anteriores, necesitándose menos tiempo/iteraciones para aproximarse a las predicciones reales. Pero, han de elegirse cuidadosamente los criterios de parada o de lo contrario, podría producirse un sobreajuste en los datos de entrenamiento (Grover, 2017).

Los principales hiperparámetros que controla este algoritmo son:

- "*shrinkage*": el parámetro de regularización o la tasa de aprendizaje.
- "*n.minobsinnode*": el tamaño máximo de nodos finales, permitiendo medir la complejidad.
- "*n.trees*": el número de árboles que se ejecutan en cada iteración.
- "*interaction.depth*": la profundidad de la iteración, soliéndose especificar de 2 para árboles binarios, dado que funciona bien la mayoría de las veces.

Gradient boosting dispone de todas las ventajas de los árboles, siendo muy agresivo (más que random forest) a la hora de disminuir el error. Además, puede ser sensible a datos atípicos, pero bien tuneado es complicado que sobreajuste, debiendo a esto su popularidad (Portela, 2021).

3.4.4. Extreme gradient boosting:

Bhattacharyya (2020) señala que XGBoost o extreme gradient boosting es el algoritmo más utilizado en competiciones para el aprendizaje automático, ganando popularidad a través de soluciones ganadoras en datos estructuradores y tabulares. Es muy similar al algoritmo anterior solo que éste es una implementación avanzada de incremento de gradiente junto con algunos factores de regularización.

En este caso, los hiperparámetros básicos que pueden modificarse en RStudio son:

- "*nrounds*": el número de iteraciones.
- "*max_depth*": la profundidad máxima de los árboles.
- "*eta*": la tasa de aprendizaje (lo que en gradient boosting llamábamos "*shrinkage*").
- "*gamma*": el coste de regularización (que se suele dejar por defecto en 0).
- "*colsample_bytree*": el porcentaje de sorteo de variables antes de cada árbol (que también se suele dejar por defecto, solo que en este caso en 1).
- "*min_child_weight*": las observaciones mínimas en el nodo final (siendo muy similar al parámetro "*n.minobsinnode*" del algoritmo anterior).
- "*subsample*": el porcentaje de observaciones de variables antes de cada árbol (que también se deja por defecto en 1).

3.5. Support vector machines:

Portela (2021) define que las máquinas de vectores de soporte o máquinas de vector soporte (del inglés, support vector machines, SVM), introducidas por Vapnik y sus compañeros de trabajo, son una familia de algoritmos que tienen como objetivo crear una separación lineal de clases a través de métodos algebraicos mediante la búsqueda de un hiperplano de separación. Se basan en tres ideas relevantes:

- SVM lineal:

    1. Maximal margin classifier o hard margin: el hiperplano que actúa como frontera de decisión se selecciona de tal forma que maximice la distancia desde el hiperplano hasta el punto más próximo del conjunto de entrenamiento, siempre que pueda ser lograda en el conjunto de entrenamiento una perfecta separabilidad entre las clases (Ruiz, 2014).
    2. Soft margin classifier: dado que no es común que se dé una separación perfecta entre las clases a consecuencia del ruido que suelen presentar los datos, el algoritmo del maximal margin no puede ser aplicado en la gran mayoría de los mismos. Este problema motivó al desarrollo de una versión más robusta del algoritmo tolerando ruido y valores atípicos en los datos sin alterar drásticamente el resultado. De esta forma, pueden permitirse

algunas observaciones mal clasificadas teniendo que introducirse un costo adicional para hacerlo (Mammone et al., 2009).

– SVM no lineal:

3. Kernel: hay ocasiones en las que no es posible encontrar un hiperplano que permita separar dos clases de forma lineal, empleándose para solucionar este problema el truco del kernel. Este truco permite "trabajar en un espacio de dimensión superior donde sí tenga sentido la separación lineal" (Portela, 2021). Luego, la idea es que nuestros datos, los cuales no son linealmente separables en nuestro espacio dimensional, puedan serlo en una dimensión mayor.

En este presente trabajo nos hemos centrado en la idea del kernel, empleándose los 3 tipos de estos que se muestran en la Ilustración 3 (aunque resulta interesante destacar que también existe el kernel sigmoid).

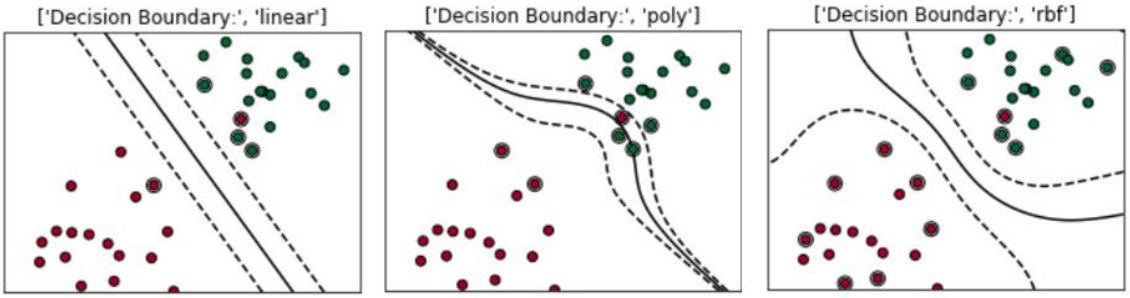


Ilustración 3. Kernel lineal, polinomial y gaussiano.

Fuente: (Chen, 2019).

Para los datos que pueden separarse linealmente, se utiliza el kernel lineal, mientras que para aquellos que no pueden separarse de tal forma se emplea el polinomial y/o gaussiano (también conocido como RBF).

En la creación de los distintos modelos de SVM los hiperparámetros que se pueden variar son:

- "*C*": la constante de regularización (en todos los modelos).
- "*degree*": los grados del polinomio y "*scale*": la escala (en SVM con kernel polinomial).
- "*sigma*": es quien controla el comportamiento del kernel y se utiliza como medida de similitud entre dos puntos (en SVM con kernel gaussiano).

3.6. Modelos de ensamblado:

Los métodos de ensamblado son técnicas que crean múltiples modelos y luego los combinan para producir mejores resultados. Dichos métodos suelen producir soluciones más precisas que un solo modelo (Demir, 2016). Existen varias técnicas básicas de combinado de modelos, como es el caso de bagging, boosting y stacking, entre otras, pero nos centraremos en la de stacking, al ser ésta la que se empleó. A su vez, dentro de esta técnica, que implica ajustar muchos tipos de modelos

diferentes en los mismos y usar otro modelo para aprender a combinar mejor las predicciones, existen tres opciones básicas que de acuerdo con el profesor Portela (2021) son:

1. Averaging (promediado): se calcula el promedio de las predicciones de los diversos modelos a combinar. También, puede utilizarse el promedio ponderado para ello. Remarcar que, en este trabajo se empleó esta opción básica, y, más concretamente, el promedio (y no el ponderado).
2. Voto (para clasificación): se emplea en modelos de clasificación y se predice el resultado con mayoría entre las predicciones.
3. Combinación a partir de otro algoritmo: se introduce en uno de los modelos las predicciones de otros modelos como variables independientes.

Una vez comparados todos los modelos, se combinaron aquellos algoritmos que mejores resultados arrojaron. Por lo que, tan solo se crearon modelos de ensamble con aquellos que obtuvieron tanto una baja tasa de fallos como un elevado AUC (en comparación a los modelos restantes).

3.7. Técnicas y métricas de evaluación de modelos:

3.7.1. Validación cruzada repetida:

Para la evaluación de todos los modelos creados mediante las técnicas de machine learning que recién se acaban de describir, se empleó la validación cruzada repetida, cuyos resultados se recogieron en un boxplot.

La validación cruzada es una estrategia popular para la selección de algoritmos. La idea principal detrás de esta técnica es la de dividir los datos, una o varias veces, tal y como se muestra en la Ilustración 4, donde una parte de los mismos (la muestra de entrenamiento) es empleada para entrenar cada algoritmo, mientras la parte restante (la muestra de validación) es empleada para estimar el riesgo del algoritmo. Por tanto, la validación cruzada selecciona el algoritmo con el menor riesgo estimado (Arlot y Celisse, 2010).

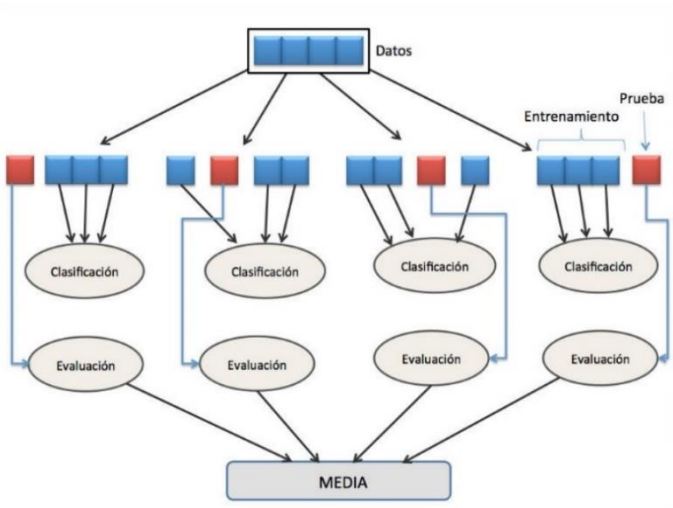


Ilustración 4. Validación cruzada.

Fuente: (Domenech, 2011).

En este trabajo se utilizaron 4 grupos y 5 repeticiones (de ahí que se haya empleado la validación cruzada repetida, pues el proceso descrito con anterioridad se ha repetido el número de veces que se le haya indicado), salvo cuando se compararon los modelos, donde se llevaron a cabo 10 grupos y 20 repeticiones.

3.7.2. Matriz de confusión:

Destacar que sobre el modelo ganador se ha presentado su matriz de confusión, conteniendo la misma el número de observaciones que han sido tanto bien como mal clasificadas. Esta herramienta se estructura de la siguiente forma:

Tabla 1. Matriz de confusión.

	Predicción = 0	Predicción = 1
Realidad = 0	VN	FP
Realidad = 1	FN	VP

Es de ella de donde pueden definirse las medidas de clasificación:

$$Tasa\ de\ aciertos = \frac{VN + VP}{VN + FP + FN + VP} \quad [8]$$

$$Tasa\ de\ fallos = \frac{FP + FN}{VN + FP + FN + VP} \quad [9]$$

$$Sensibilidad = \frac{VP}{FN + VP} \quad [10]$$

$$Especificidad = \frac{VN}{VN + FP} \quad [11]$$

donde VN son los verdaderos negativos (negativos que fueron clasificados correctamente como negativos por el modelo), FP los falsos positivos (negativos que fueron clasificados incorrectamente como positivos), FN los falsos negativos (positivos que fueron clasificados incorrectamente como negativos) y VP los verdaderos positivos (positivos que fueron clasificados correctamente como positivos).

3.7.3. Tasa de fallos y área bajo la curva ROC:

Para finalizar con este punto, indicar que los modelos generados fueron comparados entre sí bajo la tasa de fallos y el área bajo la curva ROC (AUC, del inglés area under the ROC curve).

La primera métrica no es más que el cociente entre las predicciones incorrectas y el total de predicciones, expresándose como se ha hecho en [9].

Y, la segunda métrica puede interpretarse como la probabilidad de que un positivo aleatorio se clasifique antes que un negativo aleatorio (Flach et al., 2011). En nuestro caso esto sería: la probabilidad de que una mujer que alguna vez ha tenido miedo de alguna de su/s pareja/s elegida al azar se clasifique como más propensa a haber tenido miedo que una mujer que nunca lo ha tenido, elegida también al azar.

Un modelo perfecto tendrá un AUC de 1, por lo que, cuanto mayor sea el valor de éste, más poder predictivo tendrá el modelo.

Mencionar que, además, en los boxplots que se obtuvieron una vez realizada la validación cruzada repetida, también se tuvieron en cuenta la varianza (que no es más que el tamaño de la caja) para determinar al modelo ganador y la tasa de aciertos ([8], mostrada anteriormente en el apartado de la matriz de confusión) para establecer los valores de los hiperparámetros de los modelos a crear.

4. Descripción de los datos:

4.1. Origen de los datos:

Los datos sobre los que se trabajó fueron extraídos de la sexta y más reciente "*Macroencuesta de Violencia contra la Mujer*" realizada en 2019 a una muestra representativa de 9.568 mujeres de 16 años o más residentes en España. Esta encuesta es elaborada aproximadamente cada 4 años desde 1999 por la Delegación del Gobierno contra la Violencia de Género, estando, además, incluida en el Plan Estadístico Nacional. Dicho estudio se encuentra dividido en seis grandes módulos, siendo estos:

1. Violencia en la pareja actual.
2. Violencia en las exparejas.
3. Violencia fuera del ámbito de la pareja.
4. Acoso sexual.
5. Acoso repetido.
6. Sociodemográficas.

Indicar que únicamente se seleccionaron aquellas variables que guardaban relación con la propia entrevistada, con la pareja actual y con parejas anteriores, excluyéndose, por tanto, variables ligadas a acoso o violencia por parte de relaciones no afectivas, dado que no se le considera a esto último, violencia de género. Resaltar a su vez que como la base de datos contaba con más de 1.000 variables, se escogieron manualmente aquellas que estaban conexas con los factores de riesgo de violencia contra la pareja que ya han sido identificados en la literatura científica y profesional, más concretamente, en el estudio de Andrés et al. (2008), pudiendo verse los mismos en la Tabla 2. Es importante destacar que los factores en negrita son los predictores más potentes y, por ende, los que más se tuvieron en consideración.

Tabla 2. Factores de riesgo.

	Macro-sistema	Exo-sistema	Micro-sistema	Ontogenético (individual)
Agresor	• Cultura • Valores sociales • Ideología • Creencias sociales	• **Trabajo** • **Nivel educativo** • **Estrés laboral/vital** • Violencia contra familiares (no parejas) • **Ingresos económicos** • Detenciones anteriores • **Edad**	• Víctima infantil de abusos • **Relaciones sexuales forzadas** • Acoso • **Satisfacción pareja** • Separación pareja • Control sobre la pareja • Maltrato animales • **Celos** • **Abuso emocional y/o verbal** • **Historial de agresiones sobre la pareja**	• **Abuso drogas ilegales** • **Odio/hostilidad** • **Actitudes disculpen la violencia contra las mujeres** • **Ideología tradicional en roles sexuales** • **Depresión** • **Abuso de alcohol** • Empatía

	Macro-sistema	Exo-sistema	Micro-sistema	Ontogenético (individual)
Víctima	• Cultura • Valores sociales • Ideología • Creencias sociales	• **Trabajo** • **Nivel educativo** • **Ingresos económicos** • Ayuda social • **Edad**	• Satisfacción pareja • Separación pareja • **Núm./presencia hijos** • **Violencia contra la pareja**	• **Miedo** • Embarazo • Odio/hostilidad • Abuso drogas ilegales • Actitud disculpa la violencia contra las mujeres • **Abuso de alcohol** • **Depresión**

Fuente: (Andrés et al., 2008).

Luego, en base a los factores de riesgo mostrados en la Tabla 2, las variables que se seleccionaron con el propósito de construir el conjunto de datos objeto de estudio son las que se detallan en la Tabla 39 del Anexo I, indicándose en la misma tanto los nombres de éstas como sus respectivas descripciones.

La variable que se determinó como la objetivo se diseñó a partir de dos de las variables seleccionadas: "*miedoparact*" y "*miedoparex*", pudiendo verse la descripción de las mismas en la Tabla 3. De esta forma, se pudo recoger en una nueva, siendo la de "*miedo*" y cuya obtención se detalla en el apartado siguiente, si alguna vez la entrevistada había tenido miedo de alguna de su/s pareja/s. El motivo por el cual se intentó predecir dicha variable fue porque el miedo al agresor, de acuerdo al último estudio elaborado en 2019 por la Delegación del Gobierno para la Violencia de Género y titulado "*Estudio sobre el Tiempo que Tardan las Mujeres Víctimas de Violencia de Género en Verbalizar su Situación*", es el factor que más influye en la decisión de verbalizar y/o denunciar la violencia de género, retrasando o incluso impidiendo tomar esta decisión. Básicamente, el temor a que se produzca más violencia es lo que impide a las mujeres actuar. Por esta razón, esta variable se consideró adecuada para intentar detectar posible riesgo de violencia de género, ya que no todas las mujeres participantes en la macroencuesta la han sufrido.

Tabla 3. Descripción de las variables "*miedoparact*" y "*miedoparex*".

Variable	Descripción
miedoparact	Frecuencia con la que la entrevistada ha tenido miedo de su pareja actual: • 1 (Continuamente) • 2 (Muchas veces) • 3 (Algunas veces) • 4 (Nunca) • 9 (N.C.)
miedoparex	Frecuencia con la que la entrevistada ha tenido miedo de su/s pareja/s anterior/es: • 1 (Continuamente) • 2 (Muchas veces) • 3 (Algunas veces) • 4 (Nunca) • 9 (N.C.)

4.2. Tratamiento previo de los datos:

Dado que no tenía ningún sentido intentar predecir violencia de género sobre aquellas mujeres que nunca habían tenido pareja, se eliminaron todas las observaciones correspondientes a las mismas, siendo en total, 370 mujeres. A su vez, para facilitar la interpretación de los datos, en la Tabla 40 del Anexo I se muestran las primeras modificaciones que se realizaron sobre algunas de las variables seleccionadas.

Para la elaboración de la variable objetivo se creó una variable nueva llamada "*miedo*", donde se tuvo en cuenta tanto si las entrevistadas habían tenido miedo de alguna/s pareja/s pasada/s como de la actual, en el caso de que se dispusiese. Esto, implicó fusionar las respuestas de las variables "*miedoparex*" y "*miedoparact*" (ya presentadas en la Tabla 3) y eliminar ambas dos una vez creada la variable de interés. Necesariamente, como esta variable debía ser binaria tuvo que

transformarse la misma para disponer únicamente de dos posibles valores válidos, 0 y 1, recogiéndose en 0 el nunca haber tenido miedo y en 1 el haberlo tenido alguna vez ((viniendo a ser las respuestas de "*continuamente*" (1), "*muchas veces*" (2) y "*algunas veces*" (3)), justo como se muestra en la Tabla 4. Para ello, a su vez, también fue necesario eliminar las respuestas de todas aquellas entrevistadas que no hubiesen respondido a esta pregunta en concreto o no hubiese sido posible determinar si alguna vez habían sufrido tal temor, siendo tan solo el caso de 33 mujeres.

Tabla 4. Descripción de la variable objetivo.

Variable	Descripción
miedo	Si la entrevistada ha tenido alguna vez miedo de su pareja actual y/o pasada/s: • 0 (No) • 1 (Sí)

Por tanto, finalmente, se acabó trabajando con un total de 9.165 observaciones y 22 variables. En la Tabla 5 se muestra un resumen de las variables que conformaron el conjunto de datos, junto con los posibles valores que tomaron (una vez fueron realizadas las modificaciones pertinentes).

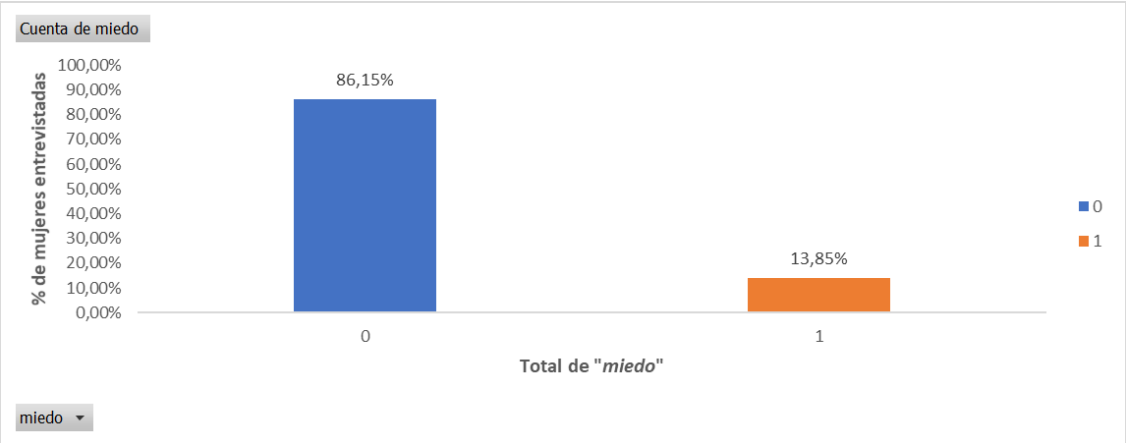
Tabla 5. Descripción de las variables seleccionadas tras ser modificadas.

Variable	Descripción
edad	Edad de la entrevistada.
nacionalidad	Si la entrevistada tiene la nacionalidad española: • 0 (No) • 1 (Sí)
CCAA	Comunidad autónoma a la que pertenece la entrevistada: • 1 (Andalucía) • 2 (Aragón) • 3 (Asturias) • 4 (Baleares) • 5 (Canarias) • 6 (Cantabria) • 7 (Castilla-La Mancha) • 8 (Castilla y León) • 9 (Cataluña) • 10 (Comunidad Valenciana) • 11 (Extremadura) • 12 (Galicia) • 13 (Madrid) • 14 (Murcia) • 15 (Navarra) • 16 (País Vasco) • 17 (La Rioja) • 18 (Ceuta) • 19 (Melilla)
tamuni	Tamaño del municipio en el que vive la entrevistada: • 1 (Menos o igual a 2.000 habitantes) • 2 (2.001 a 10.000 habitantes) • 3 (10.001 a 50.000 habitantes) • 4 (50.001 a 100.000 habitantes) • 5 (100.001 a 400.000 habitantes) • 6 (400.001 a 1.000.000 habitantes) • 7 (Más de 1.000.000 habitantes)
estudios	Estudios de la entrevistada: • 1 (Sin estudios) • 2 (Primaria) • 3 (Secundaria 1ª etapa) • 4 (Secundaria 2ª etapa) • 5 (F.P.) • 6 (Superiores) • 7 (Otros) • 8 (N.S.) • 9 (N.C.)
sitlab	Situación laboral actual de la entrevistada: • 1 (Trabaja) • 2 (Trabaja o colabora de manera habitual en el negocio familiar) • 3 (Jubilada o pensionista (anteriormente ha trabajado)) • 4 (Pensionista (anteriormente no ha trabajado)) • 5 (Parada y ha trabajado antes) • 6 (Parada y busca su primer empleo) • 7 (Estudiante) • 8 (Trabajo doméstico no remunerado) • 9 (Otra situación)
usointernet	Si la entrevistada ha usado alguna vez Internet: • 0 (No) • 1 (Sí) • 9 (N.C.)
hijos	Si la entrevistada tiene hijos/as: • 0 (No) • 1 (Sí) • 9 (N.C.)
visita	Si la entrevistada ha visitado algún psicólogo/a, psicoterapeuta o psiquiatra: • 0 (No) • 1 (Sí) • 9 (N.C.)

Variable	Descripción
ingresos	Fuente principal de ingresos personales de la entrevistada: • 1 (No tiene ingresos (otros) personales) • 2 (Salario del trabajo por cuenta ajena) • 3 (Salario del trabajo por cuenta propia) • 4 (Pensión por jubilación) • 5 (Prestación de desempleo) • 6 (Subsidio (ayudas sociales)) • 7 (Pensión de viudedad) • 8 (Pensión compensatoria) • 9 (Beca) • 10 (Rentas, ahorro (viviendas, tierras, acciones, etc.)) • 11 (Ayuda/asignación de su padre/madre) • 12 (Ayuda/asignación de pareja) • 13 (Ayuda/asignación de la expareja) • 14 (Ayuda/asignación de otra persona) • 15 (Otro no contemplado anteriormente) • 99 (N.C.)
discap	Si la entrevistada tiene un certificado de discapacidad con grado igual o superior al 33%: • 0 (No) • 1 (Sí) • 9 (N.C.)
salud12	Estado de salud en el último año de la entrevistada: • 1 (Muy bueno) • 2 (Bueno) • 3 (Regular) • 4 (Malo) • 5 (Muy malo) • 9 (N.C.)
tenpar	Si la entrevistada tiene pareja actual: • 0 (No) • 1 (Sí)
numpar	Número de parejas que ha tenido la entrevistada a lo largo de su vida, incluida la actual. En el caso de que se indique 99 significa que no se quiere contestar a la pregunta.
sexopar	Sexo de las parejas que ha tenido la entrevistada: • 1 (Solo hombres) • 2 (Solo mujeres) • 3 (Tanto hombres como mujeres) • 9 (N.C.)
miedo	Si la entrevistada ha tenido alguna vez miedo de su pareja actual y/o pasada/s: • 0 (No) • 1 (Sí)
control	Si alguna vez la entrevistada ha sufrido control de actividades por parte de su pareja actual y/o pasada/s: • 0 (No) • 1 (Sí)
hablar	Si la entrevistada tiene una persona cercana con la que pueda hablar con plena confianza sobre los problemas en sus relaciones personales: • 0 (No) • 1 (Sí) • 9 (N.C.)
frecreunion	Frecuencia con la que se reúne la entrevistada con sus parientes, familiares o amistades con quienes no convive: • 1 (Todos o casi todos los días) • 2 (Varias veces a la semana) • 3 (Varias veces al mes) • 4 (Una vez al mes) • 5 (Varias veces al año) • 6 (Una vez al año) • 7 (Menos de una vez al año) • 8 (Nunca) • 9 (N.C.)
acoger	Si la entrevistada tiene amistades o familiares/parientes con quienes podría irse un par de días si no pudiera estar en su casa por alguna razón: • 0 (No) • 1 (Sí) • 9 (N.C.)
suicidio	Si la entrevistada ha pensado alguna vez en terminar con su vida: • 0 (No) • 1 (Sí) • 9 (N.C.)
religion	Cómo se define la entrevistada en materia religiosa: • 1 (Católica) • 2 (Creyente de otra religión) • 3 (No creyente) • 4 (Atea) • 5 (Agnóstica) • 9 (N.C.)

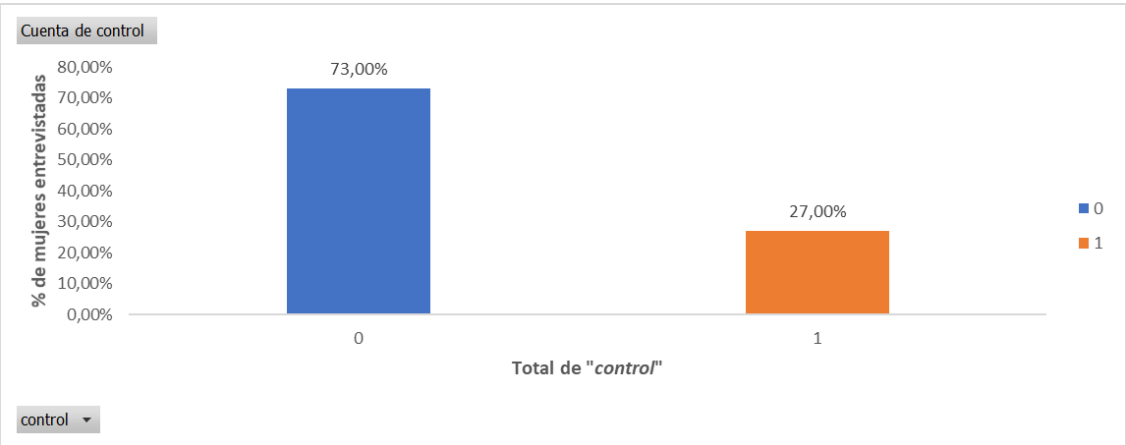
4.3. Análisis descriptivo de las variables:

Previamente a la depuración de los datos fue conveniente llevar a cabo un análisis de las variables que se seleccionaron, pues probablemente, gran parte de ellas fuesen corregidas, transformadas y/o agrupadas con posterioridad. Para lo recién mencionado, se hizo uso del programa Microsoft Excel y, más concretamente, de los gráficos dinámicos.



<u>Ilustración 5</u>. Histograma de la variable "*miedo*".

Analizando en primer lugar los valores que arrojaba la variable objetivo "*miedo*", se observa en la Ilustración 5 que el 86,15% de las mujeres entrevistadas (7.896 de 9.165) nunca había tenido miedo de alguna de su/s pareja/s, mientras que, el 13,85% restante de ellas, sí (1.269 de 9.165). Además, se puede apreciar que existía una desproporción entre las dos posibles clases de la variable objetivo, siendo este hecho muy común en los modelos de clasificación. Pero, el verdadero problema se da cuando los conjuntos de datos exhiben desequilibrios significativos y en su mayoría, extremos, pues generalmente, la clase minoritaria suele ser la de mayor interés, suponiendo un alto coste clasificar erróneamente casos pertenecientes a dicha clase.



<u>Ilustración 6</u>. Histograma de la variable "*control*".

En segundo lugar, si uno se fija en los valores de la variable "*control*" recogidos en la Ilustración 6 se aprecia como el 73% de las mujeres entrevistadas no había sido controlado por parte de alguna de su/s pareja/s, mientras que el 27% faltante sí, pudiendo ser esto un claro indicador de que gran parte de las mujeres no sientan miedo o identifiquen que están sufriendo violencia de género porque normalicen y no detecten ciertas actitudes por parte de su/s pareja y/o expareja/s (ya que, solo el 13,85% dijo haber sentido alguna vez miedo por parte de ésta/s mientras que un 27% de las mismas había sufrido control).

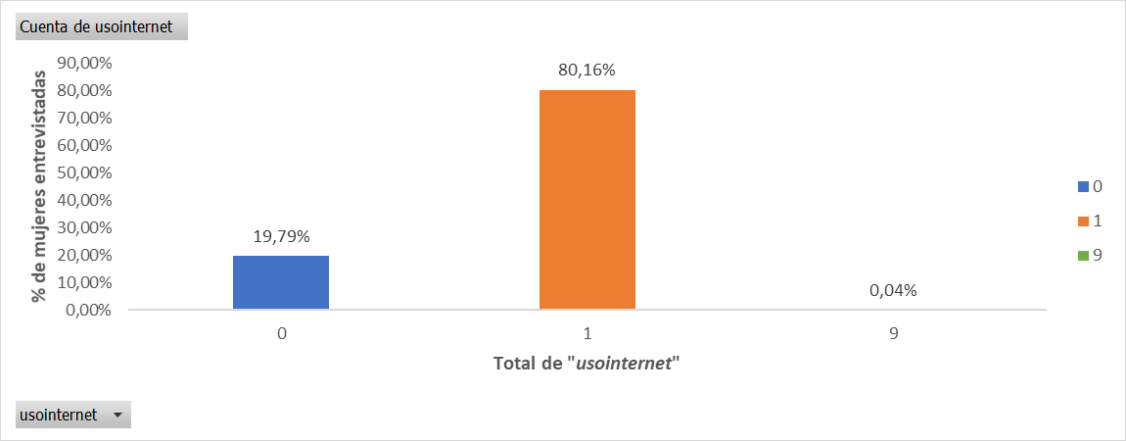


Ilustración 7. Histograma de la variable "*usointernet*".

Seguidamente, si se representaba el total de mujeres que había usado alguna vez Internet, como en la Ilustración 7, sorprendía que casi el 20% de las entrevistadas nunca lo hubiese hecho. Resulta un tema bastante preocupante pues un considerable porcentaje de las mismas no está accediendo a valiosa información donde podría leer noticias sobre violencia de género en las que se sintiese reflejada, relatos de otras víctimas o incluso poder pedir ayuda por esta vía.

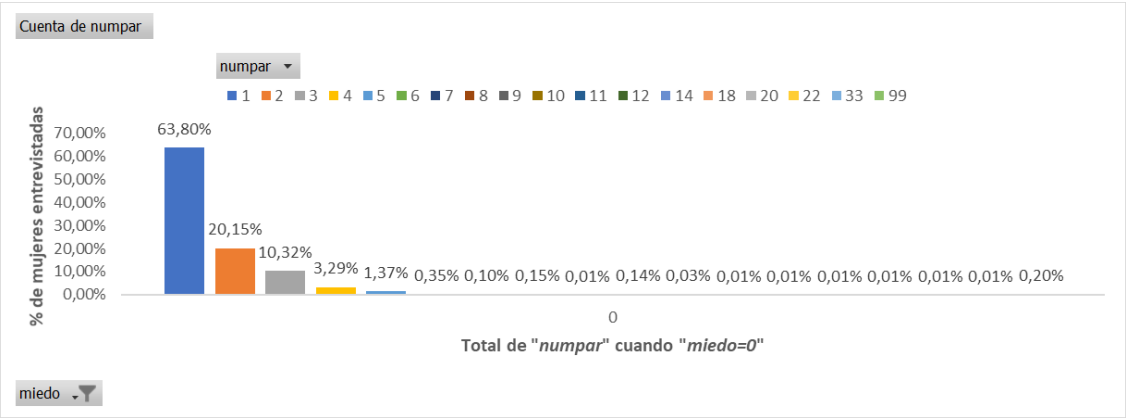


Ilustración 8. Histograma de la variable "*numpar*" cuando "*miedo=0*".

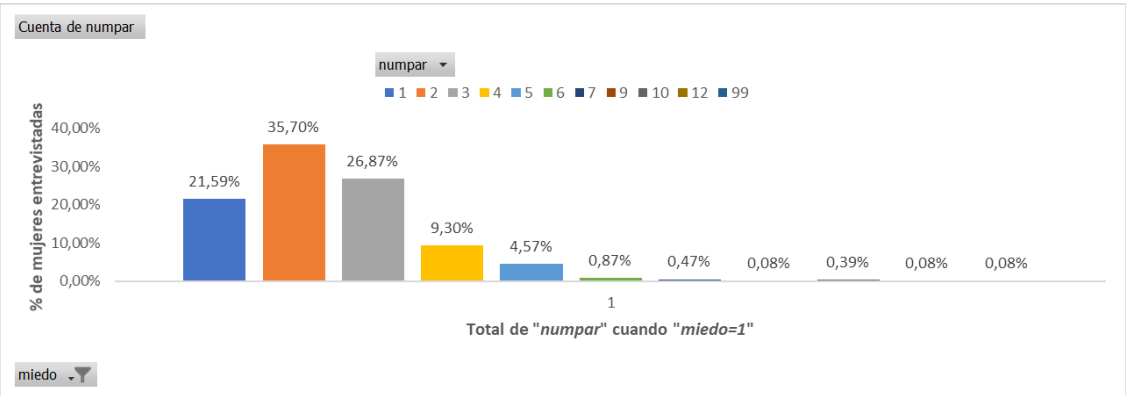
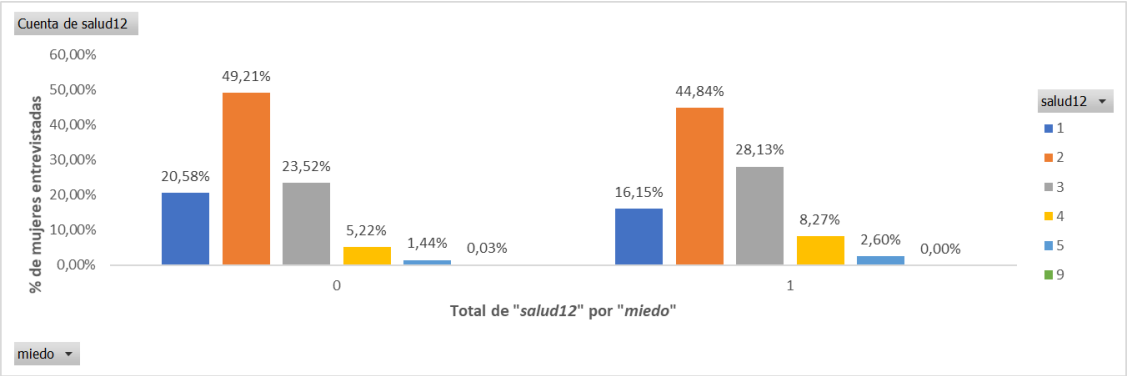


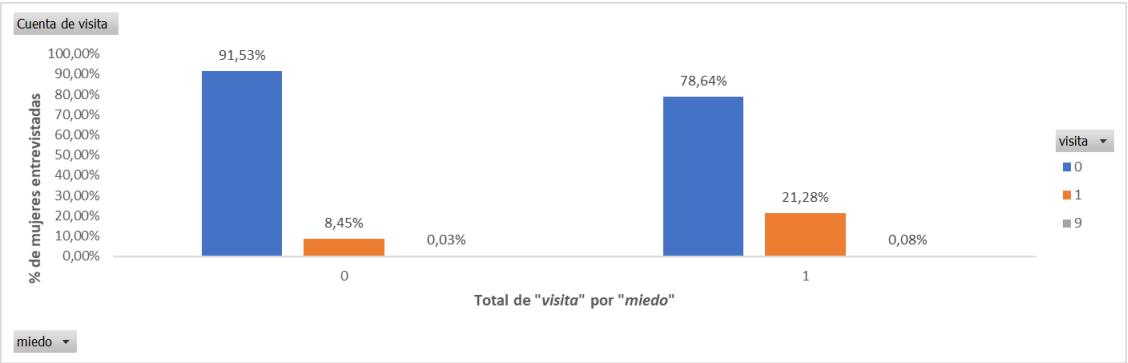
Ilustración 9. Histograma de la variable "*numpar*" cuando "*miedo=1*".

Posteriormente, teniéndose en consideración el número de parejas que habían tenido las entrevistadas, se aprecia en la Ilustración 8 e Ilustración 9 como aquellas que

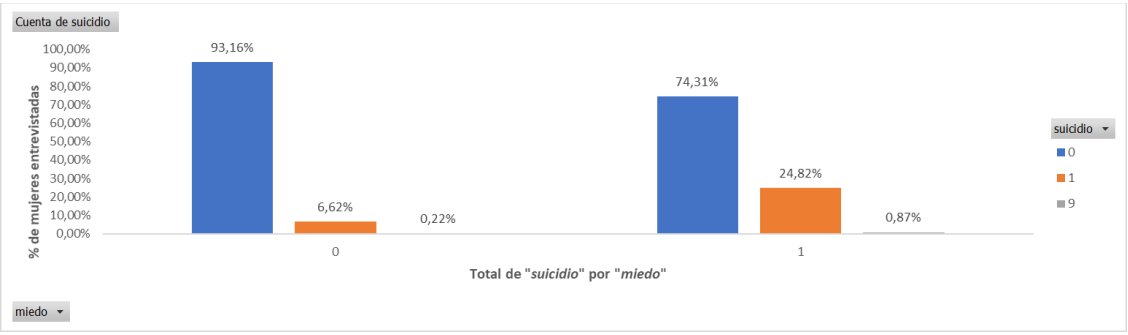
nunca habían tenido miedo de alguna de las mismas habían dispuesto mayoritariamente de una única a lo largo de su vida (63,80%), a diferencia de aquellas que habían sufrido tal temor, dado que habían dispuesto fundamentalmente de dos (35,70%) y de tres (26,87%), respectivamente. Por lo que, la gran mayoría de mujeres que habían vivido dicha angustia habían tenido más de una pareja (aunque cierto es que el 21,59% de ellas había tenido tan solo una única).



<u>Ilustración 10</u>. Histograma de la variable "*salud12*" por "*miedo*".



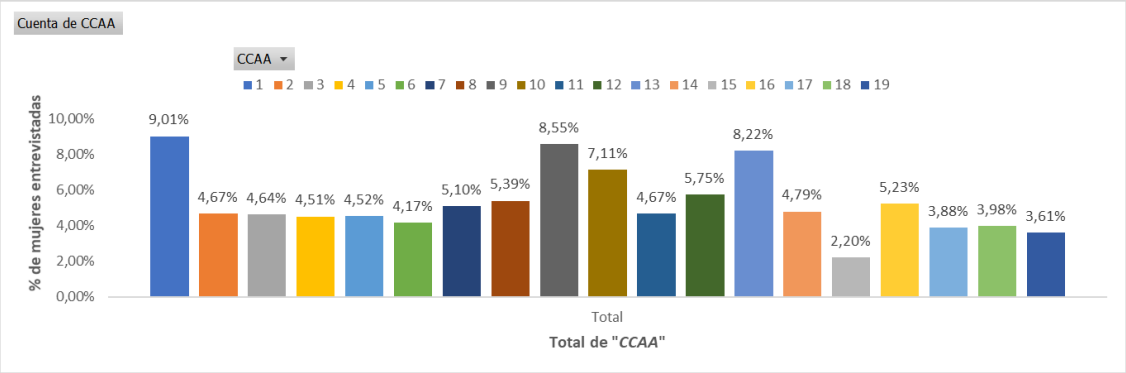
<u>Ilustración 11</u>. Histograma de la variable "*visita*" por "*miedo*".



<u>Ilustración 12</u>. Histograma de la variable "*suicidio*" por "*miedo*".

En penúltimo lugar, en la Ilustración 10, Ilustración 11 e Ilustración 12 se refleja la salud de las entrevistadas. Señalar que alrededor del 60/70% de las mismas había disfrutado de un estado de salud muy bueno o bueno en el último año, a diferencia del 30/40% restante para el que había sido regular, malo o muy malo. Pero, lo verdaderamente preocupante es que un 24,82% de las que en alguna ocasión había tenido miedo de su/s pareja/s había pensado alguna vez en terminar con su vida, una cifra bastante considerable. Además, tan solo el 21,28% de ellas había visitado

algún psicólogo/a, psicoterapeuta o psiquiatra, siendo ésta una realidad enormemente ignorada. Destacar, finalmente, que parecía no haber relación entre la percepción de la salud y el miedo, ya que los gráficos eran muy similares, no siendo esto así para el suicidio y la visita a los profesionales de la salud.



<u>Ilustración 13</u>. Histograma de la variable "*CCAA*".

Para terminar, en la Ilustración 13 se aprecia cuáles son las comunidades autónomas que mayor número de entrevistadas recogieron, siendo en este caso: Andalucía, Cataluña y Madrid, que son a su vez, las comunidades autónomas que mayor población tienen, respectivamente.

4.4. Depuración de los datos:

4.4.1. Asignación de roles y clasificación de las variables:

A continuación, se procedió a depurar los datos que contenía el conjunto de datos creado. Para ello, se hizo uso del programa SAS Enterprise Miner Workstation 14.1 donde, primeramente, fue imprescindible importar el conjunto de datos listo para trabajar sobre él, a la vez que se definieron los roles y niveles de cada una de las variables que contenía el mismo. Esta asignación de funciones y tipologías puede verse en la Tabla 6.

<u>Tabla 6</u>. Rol y nivel de las variables.

Variable	Rol	Nivel
nacionalidad	Input	Binaria
miedo	Objetivo	Binaria
tenpar	Input	Binaria
control	Input	Binaria
numpar	Input	Intervalo
edad	Input	Intervalo
sitlab	Input	Nominal
suicidio	Input	Nominal
salud12	Input	Nominal
sexopar	Input	Nominal
usointernet	Input	Nominal

Variable	Rol	Nivel
visita	Input	Nominal
tamuni	Input	Nominal
acoger	Input	Nominal
hablar	Input	Nominal
hijos	Input	Nominal
estudios	Input	Nominal
frecreunion	Input	Nominal
discap	Input	Nominal
religion	Input	Nominal
ingresos	Input	Nominal
CCAA	Input	Nominal

4.4.2. Análisis descriptivo del conjunto de datos, detección y corrección de errores:

En esta segunda fase lo que se llevó a cabo fue una exploración de los datos, permitiendo esto detectar cualquier problema con los mismos. Los distintos aspectos que se evaluaron fueron:

1. Análisis del número de datos ausentes.
2. Datos faltantes codificados no señalados como tal.
3. Análisis de los límites de las variables de intervalo.
4. Análisis de la frecuencia de las categorías de las variables nominales.
5. Categorías de variables nominales equívocas.

Primeramente, se empleó el nodo DMDB que brinda información inicial sobre la presencia de datos faltantes, la existencia de categorías erróneas y/o de valores fuera de rango. Empezando por las variables de intervalo, como bien puede verse en la Tabla 7, ninguna de ellas presentó valores ausentes, pero el máximo de "*numpar*" lo era, viniendo esto a significar que no se quiso contestar a la pregunta.

Tabla 7. Estadísticos descriptivos de las variables de intervalo.

Variable	Ausente	N	Mínimo	Máximo	Media
edad	0	9.165	16	96	50,48
numpar	0	9.165	1	99	1,93

Y, siguiendo por las variables categóricas, al igual que con las de intervalo, ninguna de ellas presentó datos faltantes, viniendo esto recogido en la Tabla 8. Aunque, por la Tabla 5, se supo que algunas contestaciones dadas por las entrevistadas también lo eran. Por lo que, esto es algo que debía ser corregido con posterioridad en muchas de las variables de clase, más exactamente en: "*acoger*", "*discap*", "*estudios*", "*frecreunion*", "*hablar*", "*hijos*", "*ingresos*", "*religion*", "*salud12*", "*sexopar*", "*suicidio*", "*usointernet*" y "*visita*".

Tabla 8. Estadísticos descriptivos de las variables categóricas.

Variable	Niveles	Ausente
CCAA	19	0
acoger	3	0
control	2	0
discap	3	0
estudios	9	0
frecreunion	9	0
hablar	3	0
hijos	3	0
ingresos	16	0
miedo	2	0

Variable	Niveles	Ausente
nacionalidad	2	0
religion	6	0
salud12	6	0
sexopar	4	0
sitlab	9	0
suicidio	3	0
tamuni	7	0
tenpar	2	0
usointernet	3	0
visita	3	0

Otro de los nodos que genera estadísticos descriptivos, al igual que el nodo anterior, es el de Explorador de estadísticos. Gracias a éste, pudo verse el número de ocurrencias que presentaba cada variable y con ello, el porcentaje que suponía del total. De esta forma, se supo si era necesario reagrupar, siéndolo en el caso de aquellas categorías que presentasen menos del 5% de los datos. En este caso, fue necesario en alguna que otra de las variables, más concretamente en: "*CCAA*", "*estudios*", "*frecreunion*", "*ingresos*", "*salud12*", "*sexopar*" y "*sitlab*", pudiendo verse esto más detallado en la Tabla 41 del Anexo I.

Para realizar las correcciones pertinentes hubo que emplear el nodo de Reemplazo. En primer lugar, para las variables de intervalo lo que se hizo fue sustituir el máximo de la variable "*numpar*" como dato ausente, tal y como se muestra en la Tabla 9. De esta forma, es como se detectaron un total de 17 valores missing.

Tabla 9. Reemplazo variables de intervalo.

Variable	Método de límite	Límite inferior de reemplazo	Límite superior de reemplazo	Método de sustitución
numpar	Especificado por el usuario	.	98	Ausente

Y, en segundo lugar, para las distintas variables de clase que necesitaban ser corregidas, lo que se llevó a cabo es:

- "*CCAA*": se agruparon las categorías por proximidad, es decir, se intentó que las comunidades autónomas quedasen cerca. Las agrupaciones concretas que se hicieron fueron: 3 (Asturias), 6 (Cantabria) y 12 (Galicia); 15 (Navarra), 16 (País Vasco) y 17 (La Rioja); 8 (Castilla y León) y 11 (Extremadura); 5 (Canarias), 18 (Ceuta) y 19 (Melilla); 10 (Comunidad Valenciana) y 14 (Murcia); 2 (Aragón), 4 (Baleares) y 9 (Cataluña) y, 1 (Andalucía), 7 (Castilla-La Mancha) y 13 (Madrid) se dejaron solas. Esto puede quedar mucho más claro con la Ilustración 14.

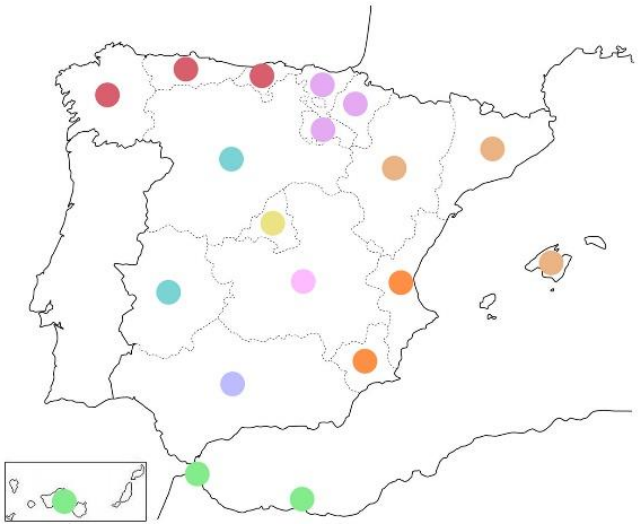


Ilustración 14. Agrupamiento de las categorías de la variable "*CCAA*".

- "*estudios*": la categoría 7 (Otros) fue agrupada con la 1 (Sin estudios).
- "*frecreunion*": se agruparon las categorías 5 (Varias veces al año), 6 (Una vez al año), 7 (Menos de una vez al año) y 8 (Nunca) con la 4 (Una vez al mes), recogiéndose así, las visitas de máximo 1 vez al mes.
- "*ingresos*": se agruparon por un lado las categorías 6 (Subsidio (ayudas sociales)), 9 (Beca), 10 (Rentas, ahorro (viviendas, tierras, acciones, etc.)), 11 (Ayuda/asignación de su padre/madre), 12 (Ayuda/asignación de pareja), 13 (Ayuda/asignación de la expareja), 14 (Ayuda/asignación de otra persona) y 15 (Otro no contemplado anteriormente) con la 5 (Prestación por desempleo) y por el otro, las categorías 7 (Pensión de viudedad) y 8 (Pensión compensatoria) con la 4 (Pensión por jubilación), juntándose así por un lado ayudas y por el otro, pensiones.

- "*salud12*": se juntaron las categorías 1 (Muy bueno) y 2 (Bueno) y, 3 (Regular), 4 (Malo) y 5 (Muy malo), distinguiéndose un estado de salud bueno de otro no tanto.
- "*sexopar*": se agruparon las categorías 2 (Solo mujeres) y 3 (Tanto hombres como mujeres), dejándose la 1 (Solo hombres) sola.
- "*sitlab*": se juntaron las categorías 1 (Trabaja) y 2 (Trabaja o colabora de manera habitual en el negocio familiar), 3 (Jubilada o pensionista (anteriormente ha trabajado)) y 4 (Pensionista (anteriormente no ha trabajado)), 5 (Parada y ha trabajado antes) y 6 (Parada y busca su primer empleo) y, para finalizar, 7 (Estudiante), 8 (Trabajo doméstico no remunerado) y 9 (Otra situación).

Además, todas las categorías que habían sido detectadas como missing en las variables "*estudios*", "*ingresos*", "*hijos*", "*discap*", "*salud12*", "*visita*", "*sexopar*", "*hablar*", *frecreunion*", "*acoger*", "*suicidio*", "*religion*" y "*usointernet*" fueron reemplazadas por "*_MISSING_*".

Ahora, todas las categorías presentaban alrededor del 5% de los datos y los valores faltantes ya habían sido detectados como tal, pudiendo comprobarse este hecho en la Tabla 42 del Anexo I.

4.4.3. Búsqueda y gestión de datos atípicos:

Los métodos de detección de outliers solo se aplican en variables de intervalo a través del nodo Reemplazo. Previamente, es conveniente saber que una distribución es simétrica si el coeficiente de asimetría se encuentra entre -1 y 1, mientras que, en el caso contrario, es asimétrica. En cada caso, se aplica el siguiente método de límite:

- Si la distribución es simétrica se emplea el de la desviación típica.
- Si la distribución es asimétrica con mediana distinta a cero se usa el de la desviación absoluta media.
- Si la distribución es asimétrica con mediana igual a cero se utiliza el de los percentiles extremos.

En la Tabla 10 se recogen la mediana y asimetría de las dos únicas variables de intervalo que se disponían: "*edad*" y "*REP_numpar*".

Tabla 10. Mediana y asimetría de las variables de intervalo.

Variable	Mediana	Asimetría
edad	50	0,10
REP_numpar	1	4,85

La variable "*edad*" presentó una asimetría de 0,10, lo que implicó disponer de una distribución simétrica y por lo que se aplicó el método de la desviación típica, mientras que la variable "*REP_numpar*" presentó una asimetría de 4,85, por lo que la distribución era asimétrica y pese a que la mediana no era 0, como hubo más del 50% de los datos que tomaban el valor 1, la mediana de las diferencias absolutas era 0, teniendo que emplearse por ello mismo el método de los percentiles extremos. Luego, desde el nodo de Reemplazo que se ha mencionado con anterioridad, lo que se hizo es lo que se muestra en la Tabla 11.

Tabla 11. Editor de reemplazo variables de intervalo.

Variable	Método de límite
edad	Desviación estándar
REP_numpar	Percentiles extremos

Para saber para cada una de las variables cuántos atípicos encontró, dentro de los resultados del nodo hubo que fijarse en la ventana de "*Cuentas de reemplazo total*" y más concretamente, en la columna de "*Entrenamiento*". Exactamente, el número de valores atípicos que se hallaron es el que se recoge en la Tabla 12.

Tabla 12. Número de atípicos detectados en las variables de intervalo.

Variable	Entrenamiento
edad	0
REP_numpar	39

Dado que se contaba con 9.165 observaciones, un valor no es considerado atípico cuando supera el 5% de dicha cantidad, es decir, 458. En este caso, como ninguna de las dos variables presentó más de dichas observaciones como atípicos, pues en "*edad*" no se encontró ninguna y en "*REP_numpar*" solo 39, no hubo que corregir posteriormente nada.

4.4.4. Tratamiento de datos faltantes:

Para analizar detalladamente los datos faltantes, puede optarse por una o varias de las siguientes estrategias: eliminación, imputación y recategorización. Empezando por la estrategia de eliminación se tiene que:

- Por variables: es necesario fijarse en los ausentes y ver si son muchos. De ser así, rechazar esas variables. En la Tabla 13 y Tabla 14 puede observarse el número de ausentes tanto por variables de intervalo como por variables de clase que se tuvieron, comprobándose así, que, dado que los valores ausentes no supusieron más del 50% de los datos, no fue necesario rechazar ninguna variable.

Tabla 13. Número de ausentes en las variables de intervalo.

Variable	Ausente
edad	0
REP_REP_numpar	56

Tabla 14. Número de ausentes en las variables de clase.

Variable	Ausente
REP_CCAA	0
REP_acoger	41
REP_discap	8
REP_estudios	15
REP_frecreunion	6
REP_hablar	16
REP_hijos	2
REP_ingresos	46
REP_religion	235
REP_salud12	2

Variable	Ausente
REP_sexopar	4
REP_sitlab	0
REP_suicidio	28
REP_usointernet	4
REP_visita	3
control	0
miedo	0
nacionalidad	0
tamuni	0
tenpar	0

- Por observaciones: se crea una nueva variable llamada "*numMissing*", la cual recoge el número de ausentes por observación. Tal y como se muestra en la

Tabla 15, debido a que más de la mitad de las observaciones estaban completas, no se eliminó ninguna.

Tabla 15. Número de ausentes por observación.

Variable	Máximo	Mediana
numMissing	4	0

Tras esto, nos centramos fundamentalmente en la estrategia de imputación (proceso consistente en sustituir los ausentes por valores válidos), dado que la presencia de ausentes fue baja (inferior al 5%). Para ello, se empleó el nodo de Imputar, asignándole a las variables el método deseado. Tanto para las de intervalo como para las de clase se asignó el de "*Distribución*", consistente en la sustitución por un valor extraído aleatoriamente de la distribución.

Una vez ultimado este proceso, ya no se disponían ni de datos ausentes ni de datos atípicos, presentándose, además, valores mínimos y máximos razonables.

4.4.5. Análisis de la relación de las variables input con la variable objetivo:

Una vez se tienen los datos limpios y no es recomendable hacerlo hasta que estos no lo están, pueden explorarse los mismos desde el punto de vista bivariable. De esta forma, uno puede hacerse una idea de qué variables van a influir más o menos en la predicción de la variable objetivo. Lo que se hace es emplear el nodo de Explorador de estadísticos, quien determina si existe o no alguna relación bivariante entre las variables de entrada y la variable objetivo, junto al de Código SAS, el cual permite crear una variable aleatoria con el fin de saber si las variables tienen menos potencial predictivo que algo que el propio programa acaba de crear aleatoriamente.

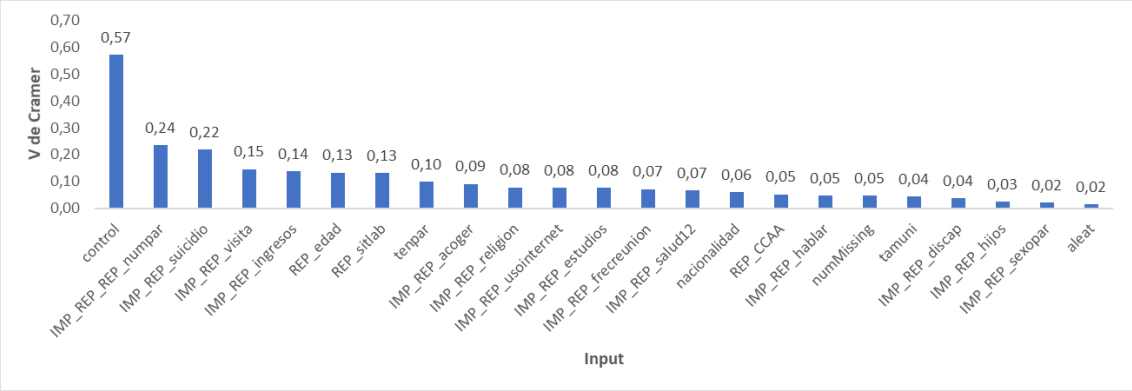
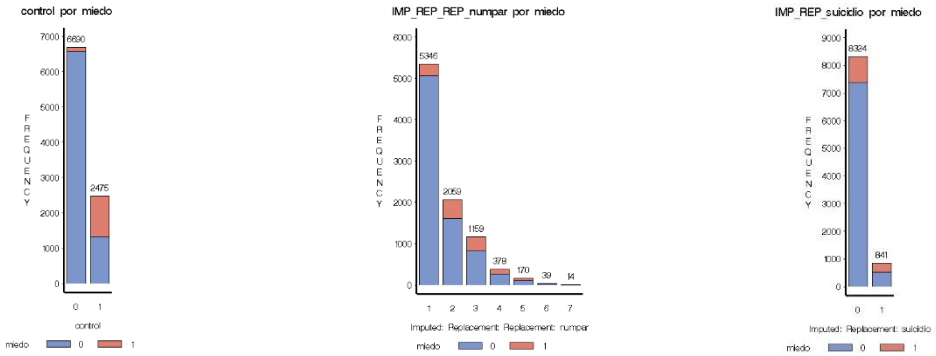


Ilustración 15. V de Cramer.

Puede apreciarse claramente, gracias a la Ilustración 15, que la variable que más relación guardaba con la variable objetivo era la de "*control*", seguida de "*IMP_REP_REP_numpar*" e "*IMP_REP_suicidio*", pues eran aquellas con un V de Cramer más próximo a 1. Remarcar, a su vez, que ninguna variable quedó por debajo de la aleatoria. Luego, todas las variables poseían más potencial predictivo que algo que el mismo programa acababa de crear de forma aleatoria.



<u>Ilustración 16</u>. Gráfico de barras de las variables "*control*", "*IMP_REP_REP_numpar*" e "*IMP_REP_suicidio*" por "*miedo*".

Para analizar la relación de las variables input con la variable objetivo se ha de emplear el nodo Multi gráfico. Gracias a éste, se obtuvieron los gráficos de barras de "*control*", "*IMP_REP_REP_numpar*" e "*IMP_REP_suicidio*" por "*miedo*", tal y como se muestra en la Ilustración 16. Con estas figuras se observó que:

- Las entrevistadas que nunca habían tenido miedo de su pareja actual y/o pasada/s apenas habían sufrido control de actividades por parte de las mismas, a diferencia de aquellas que sí que habían vivido tal temor.
- La gran mayoría de las entrevistadas que nunca habían tenido miedo de su pareja y/o expareja/s habían dispuesto tan solo de una única a lo largo de su vida, mientras que aquellas que sí que habían padecido dicho sentimiento habían dispuesto mayoritariamente de dos y de tres, respectivamente.
- Un considerable número de entrevistadas había pensado alguna vez en terminar con su vida tras haber sufrido miedo de alguna de su/s pareja/s, a diferencia de aquellas que no lo habían sufrido.

4.5. Selección de variables:

4.5.1. Selección de variables con SAS Enterprise Miner Workstation 14.1:

En este primer programa se empleó el nodo de Selección de variables para que el mismo, seleccionase aquellas que considerase relevantes en función de su R^2. Las variables que escogió por tener un R^2 superior a 0,005 fueron:

<u>Tabla 16</u>. Selección de variables $R^2 > 0,005$.

Conjunto	Variables
1	"*IMP_REP_REP_numpar*", "*IMP_REP_acoger*", "*IMP_REP_estudios*", "*IMP_REP_frecreunion*", "*IMP_REP_ingresos*", "*IMP_REP_religion*", "*IMP_REP_suicidio*", "*IMP_REP_usointernet*", "*IMP_REP_visita*", "*REP_edad*", "*REP_sitlab*", "*control*" y "*tenpar*".

4.5.2. Selección de variables con SAS 9.4:

Y, en este segundo programa, lo que se hizo fue emplear la macro "%*randomselectlog*" proporcionada por el profesor Portela (2021), la cual mediante el modelo de regresión logística y el método stepwise permitió identificar las variables más relevantes (una vez fueron importadas y definidas con sus respetivos roles y

niveles). Los modelos que más veces se repitieron tras reproducirse el proceso 101 veces con una partición de 0,8 en entrenamiento y 0,2 en prueba fueron:

Tabla 17. Selección de variables stepwise.

Conjunto	Variables	Frecuencia	% de frecuencia total
2	"*IMP_REP_hijos*", "*IMP_REP_ingresos*", "*IMP_REP_salud12*", "*IMP_REP_suicidio*", "*IMP_REP_visita*", "*control*", "*tenpar*" e "*IMP_REP_REP_numpar*".	9	8,91
3	"*IMP_REP_ingresos*", "*IMP_REP_suicidio*", "*IMP_REP_visita*", "*control*", "*tenpar*" e "*IMP_REP_REP_numpar*".	7	6,93
4	"*IMP_REP_discap*", "*IMP_REP_estudios*", "*IMP_REP_hijos*", "*IMP_REP_ingresos*", "*IMP_REP_suicidio*", "*IMP_REP_visita*", "*control*", "*tenpar*" e "*IMP_REP_REP_numpar*".	5	4,95
5	"*IMP_REP_hijos*", "*IMP_REP_ingresos*", "*IMP_REP_suicidio*", "*IMP_REP_visita*", "*control*", "*tenpar*" e "*IMP_REP_REP_numpar*".	5	4,95

Una vez se obtuvieron los distintos modelos con mayor frecuencia se probó realizar validación cruzada repetida sobre los mismos a través de la macro "*%cruzadalogistica*" (Portela, 2021), considerándose también el conjunto logrado con SAS Enterprise Miner Workstation 14.1. De esta forma, tras 21 repeticiones y una partición de la muestra en 10 grupos, se obtuvo un gráfico de cajas el cual permitió evaluar las condiciones de sesgo y varianza.

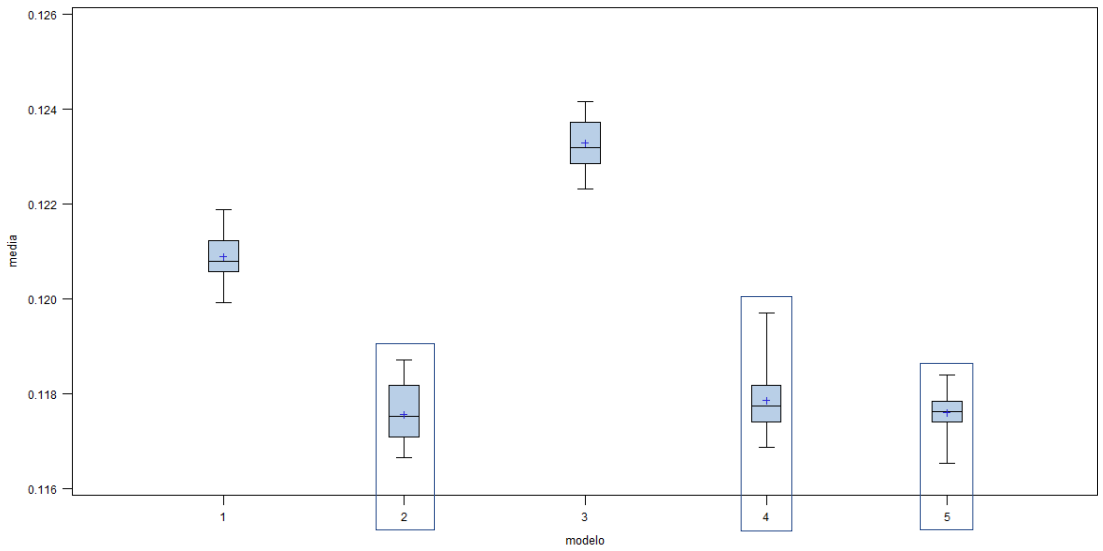
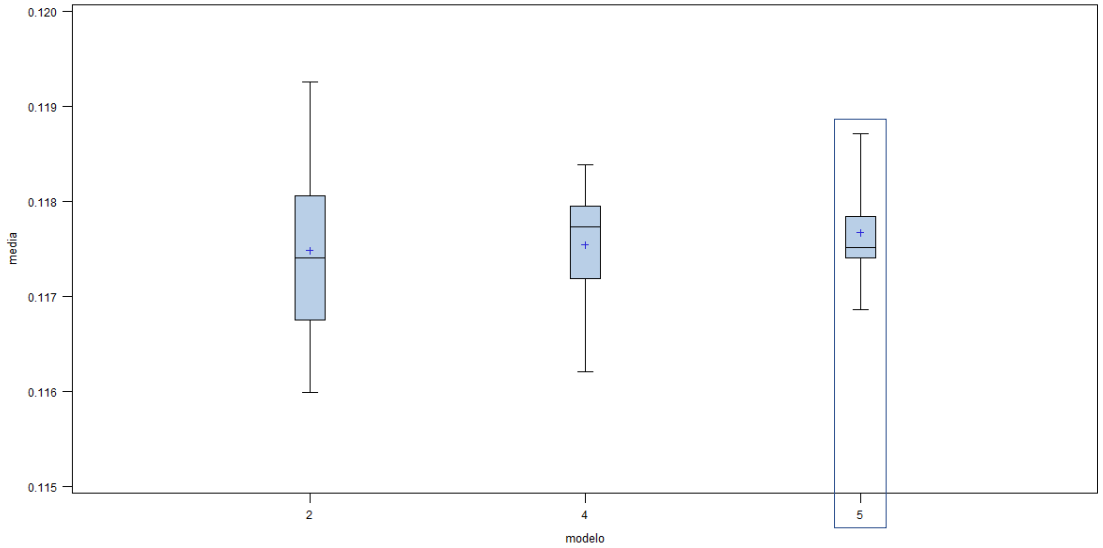


Ilustración 17. Tasa de fallos modelos de regresión logística (stepwise).

Los modelos que llamaron especialmente la atención de la Ilustración 17 por presentar tanto un menor sesgo como una menor varianza fueron el 2, 4 y 5, marcados con un rectángulo. Por ello, se volvió a realizar una pequeña prueba más con los mismos, solo que en este caso, variándose la semilla.

Todos los modelos de la Ilustración 18 presentaban una tasa de fallos similar. Sin embargo, aquel que ofrecía una menor varianza era el modelo 5, siendo de entre los tres, el que menos variables seleccionaba. Luego, las variables que se tomaron con el propósito de comenzar a modelizar con el programa RStudio fueron: "*IMP_REP_hijos*", "*IMP_REP_ingresos*", "*IMP_REP_suicidio*", "*IMP_REP_visita*",

"*control*", "*tenpar*" e "*IMP_REP_REP_numpar*". Pero, dado que esta selección podía no ajustar bien con variables input con relaciones no lineales, se optó por emplear el primer conjunto de datos, es decir, aquel compuesto por todas aquellas variables con un R^2 superior a 0,005 para las técnicas basadas en árboles, debido a que dichas técnicas disponen de sus propios métodos de selección de variables.



<u>Ilustración 18</u>. Tasa de fallos modelos 2, 4 y 5 de regresión logística.

5. Modelización:

Para la construcción de los diversos modelos se emplearon dos conjuntos de variables, tal y como se venía diciendo: para las técnicas basadas en árboles aquel compuesto por todas aquellas variables con un R^2 mayor a 0,005, mientras que, para las técnicas restantes, se utilizó el conjunto 5 de la Tabla 17. El programa que se utilizó para modelizar fue RStudio, teniéndose como objetivo hallar el mejor modelo que permitiera predecir la variable dependiente. Por ello mismo, se tunearon y crearon diversos modelos de redes neuronales, bagging, random forest, gradient boosting, extreme gradient boosting y support vector machines a través del paquete "*caret*" y los códigos proporcionados por el profesor Portela (2021), empleándose sobre estos, validación cruzada repetida con 4 grupos y 5 repeticiones.

Antes de empezar a modelizar fue necesario estandarizar las variables continuas y transformar las categóricas a dummies. Así, por cada categoría que presentaba cada una de las nominales, se creó una nueva variable codificada de manera binaria, esto es, con valores 0 y 1. Tras esto, para la correcta ejecución de los códigos fue también conveniente transformar los valores "*0*" y "*1*" de la variable objetivo a "*No*" y "*Yes*", respectivamente.

5.1. Regresión logística:

Dado que el modelo de regresión logística que se construyó para la selección de variables presentó unos resultados relativamente buenos, como bien puede verse en

la Ilustración 19, resulta de interés mostrar los diferentes coeficientes del mismo, los cuales pueden interpretarse y permiten obtener mayor información de las variables independientes.

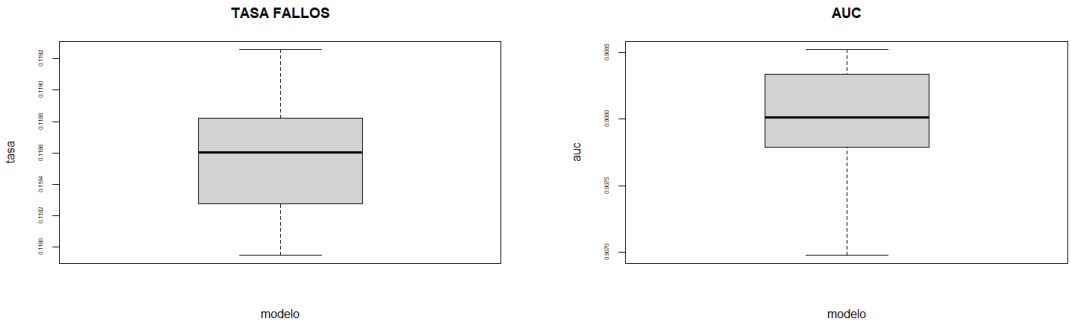


Ilustración 19. Tasa de fallos y AUC regresión logística.

De acuerdo a la Tabla 18 todas las variables que conforman el modelo son significativas, destacando las de "*control.0*", "*IMP_REP_REP_numpar*" e "*IMP_REP_suicidio.1*", siendo justo las que se destacaban en el punto "*4.4.5. Análisis de la relación de las variables input con la variable objetivo*". Por tanto, estas variables son fundamentales a la hora de explicar la probabilidad de haber sentido alguna vez miedo de alguna pareja.

Tabla 18. Odds ratio de las variables del modelo de regresión logística.

Variable	Estimate	Pr(>|z|)	OR
IMP_REP_REP_numpar	0,34293	<2e-16	1,4091
tenpar.0	0,43074	3,81e-07	1,5384
control.0	-3,49506	<2e-16	0,0303
IMP_REP_hijos.0	-0,53259	1,98e-09	0,5871
IMP_REP_ingresos.1	0,33272	0,022128	1,3948
IMP_REP_ingresos.2	0,61964	7,50e-07	1,8583
IMP_REP_ingresos.3	0,44200	0,014947	1,5558
IMP_REP_ingresos.5	0,70468	1,08e-06	2,0232
IMP_REP_suicidio.1	0,79545	4,46e-14	2,2154
IMP_REP_visita.1	0,38337	0,000337	1,4672

Así pues, entre otros, puede interpretarse que:

- Las entrevistadas que nunca han sufrido control de actividades por parte de su/s pareja/s tienen 33 veces menos de posibilidades de tener miedo de alguna de ellas frente a aquellas que sí han sufrido tal control.
- Por cada pareja adicional que han tenido las entrevistadas, las posibilidades de haber tenido miedo de alguna de estas se multiplican por 1,41.
- Si se comparan las entrevistadas que han pensado alguna vez en quitarse voluntariamente la vida con las que nunca han pensado en hacerlo, las probabilidades de que las primeras hayan tenido miedo de alguna pareja son 2,21 veces superiores que las de las segundas.

5.2. Redes neuronales:

Para obtener el número idóneo de nodos que debía tener el modelo en la capa oculta con el propósito de llevar a cabo un estudio inicial sobre el posible tuneo que debía realizarse en redes neuronales, bastó con emplear la ecuación (Portela, 2021):

$$h(k+1)+h+1=obs/p, \qquad [12]$$

donde h es el número de nodos ocultos, k el número de nodos input, obs el número de observaciones y p el número de parámetros.

Luego, teniendo en consideración que se disponían de un total de 9.165 observaciones en el conjunto de datos, de 7 variables (tras haber escogido como ganador el conjunto 5 de la Tabla 17 para técnicas no basadas en árboles) y que lo conveniente es establecer como mínimo 25/30 observaciones por parámetro, escogiéndose en este presente caso, 30, se obtuvo despejando que:

$$h(7+1)+h+1=9.165/30 \rightarrow h=33{,}83$$

Por tanto, el número máximo de nodos a probar debía rondar en torno a los 33/35. Entonces, una vez se halló esto y mediante la librería "*caret*", se creó una rejilla con 5, 10, 15, 20, 25, 30 y 35 nodos y unos learning rates de 0,001, 0,01 y 0,1, donde se obtuvo:

Tabla 19. Resultados rejilla redes neuronales.

Nodos	Learning rate	Tasa de aciertos
5	0,001	0,8795203
5	0,010	0,8809383
5	0,100	0,8818985
10	0,001	0,8742393
10	0,010	0,8804581
10	0,100	0,8805889
15	0,001	0,8737374
15	0,010	0,8787777
15	0,100	0,8792361
20	0,001	0,8715986
20	0,010	0,8776867
20	0,100	0,8791486
25	0,001	0,8715767
25	0,010	0,8773811
25	0,100	0,8785377
30	0,001	0,8700491
30	0,010	0,8775557
30	0,100	0,8789086
35	0,001	0,8732351
35	0,010	0,8770321
35	0,100	0,8778177

De modo que, observando los resultados de la Tabla 19 se pudo apreciar como mayor número de nodos no implicaba mayores tasas de aciertos. De hecho, se conseguía una tasa de aciertos mayor con tan solo 5 nodos. Así que, dado que los mejores resultados se obtuvieron con 5 y 10 nodos y unos learning rates de 0,01 y 0,1, se probó volver a crear una nueva rejilla con 3, 7, 9 y 11 nodos, junto a ambos learning rates mencionados. De esta forma se consiguió:

Tabla 20. Resultados rejilla redes neuronales (2).

Nodos	Learning rate	Tasa de aciertos
3	0,010	0,8829460
3	0,100	0,8829897
7	0,010	0,8818765
7	0,100	0,8803271
9	0,010	0,8808290
9	0,100	0,8803489
11	0,010	0,8811781
11	0,100	0,8799561

En definitiva, tal y como puede verse en la Tabla 20, 11 nodos eran más que suficientes para conseguir una elevada tasa de aciertos. Así pues, las combinaciones de número de nodos y learning rates que mayores tasas de aciertos arrojaron y sobre las cuales se realizó validación cruzada repetida con el objetivo de poder seleccionar el mejor modelo fueron las que se muestran en la Tabla 21 –indicar que las iteraciones de todos los modelos que se construyeron se fijaron en 100–.

Tabla 21. Modelos candidatos redes neuronales.

Modelo	Nodos	Learning rate	Iteraciones
avnnet1	3	0,1000	100
avnnet2	3	0,0100	100
avnnet3	5	0,1000	100
avnnet4	7	0,0100	100
avnnet5	11	0,0100	100

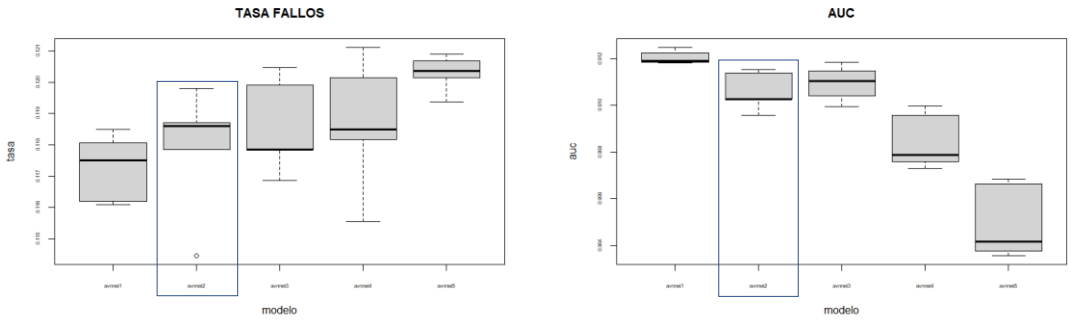


Ilustración 20. Tasa de fallos y AUC redes neuronales.

Pese a que el modelo que presentaba un mayor AUC fue el de "*avnnet1*", se escogió como ganador el de "*avnnet2*", debido a que ofrecía una menor varianza de tasa de fallos. Además, si se tenían en cuenta los valores del eje Y tanto de la tasa de fallos como del AUC, puede verse en la Ilustración 20 como no se daban grandes saltos, por lo que la diferencia entre los mismos era apenas considerable. Por consiguiente, el modelo "*avnnet2*" resultó ser más estable, siendo aquel donde se consideraron tan solo 3 nodos en la capa oculta y un learning rate de 0,01.

5.3. Modelos basados en árboles:

5.3.1. Bagging:

Previamente a la elaboración de los modelos bagging y a través de la librería "*randomForest*", resultó de interés plotear el error OOB a medida que avanzaban las iteraciones con el propósito de conocer a partir de qué cantidad de árboles el error se estabilizaba.

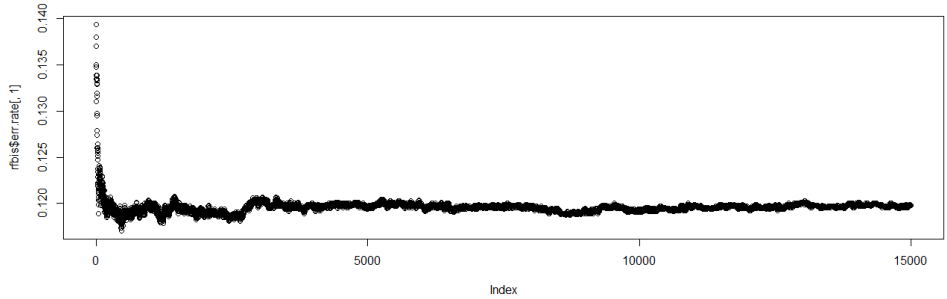


Ilustración 21. Error OBB según avance iteraciones bagging.

La Ilustración 21 muestra cómo a partir de aproximadamente los 3.000/5.000 árboles el error se mantenía fijo. Por lo que, los modelos a comparar contuvieron dicho

número de árboles, pues considerar mayor cantidad de estos no hacía mejorar (aunque tampoco emporar) el modelo.

Tras esto, señalar que el hiperparámetro más importante a establecer en los modelos de bagging es el "*sampsize*", o lo que es lo mismo, el tamaño de la muestra a extraer, calculándose como (Portela, 2021):

$$\left[\left(\frac{1}{k}\right) * (k-1)\right] * n, \qquad [13]$$

donde k es el número de grupos de validación cruzada repetida que se emplean y n el número de observaciones con el que se cuenta. Por tanto, dado que en este presente caso se emplearon 4 grupos y se disponían de 9.165 observaciones, se tuvo que:

$$\left[\left(\frac{1}{4}\right) * (4-1)\right] * 9.165 = 0{,}75 * 9.165 = 6.874$$

El máximo de tamaño de muestra que podía probarse era de 6.874, siendo el que por defecto prueba el paquete en el caso de que no se le especifique lo contrario. De hecho, se probó este valor y el de 5.957 (0,65*9.165).

Finalmente, señalar que se probaron los tamaños mínimos 10, 15 y 20 de nodos finales y que el número de variables a sortear en cada nodo del árbol fueron todas las variables, no produciéndose por ello mismo sorteo.

Luego, los modelos sobre los que se realizó validación cruzada repetida fueron:

Tabla 22. Modelos candidatos bagging.

Modelo	mtry	ntree	nodesize	sampsize
bagging1	27	3.000	10	5.957 (0,65*9.165)
bagging2	27	3.000	15	5.957 (0,65*9.165)
bagging3	27	3.000	20	5.957 (0,65*9.165)
bagging4	27	5.000	10	6.874 (0,75*9.165)
bagging5	27	5.000	15	6.874 (0,75*9.165)
bagging6	27	5.000	20	6.874 (0,75*9.165)

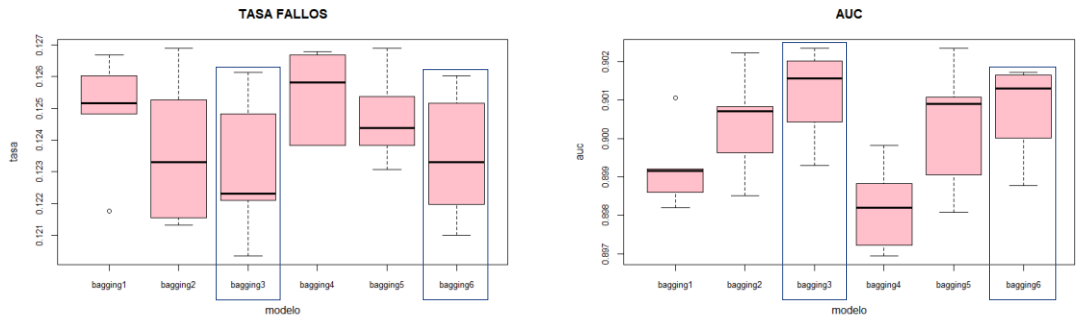


Ilustración 22. Tasa de fallos y AUC bagging.

Entonces, visto que en la Ilustración 22 los dos modelos que presentaban tanto un alto AUC como una baja tasa de fallos eran "*bagging3*" y "*bagging6*", se volvió a realizar validación cruzada repetida sobre los mismos, variándose únicamente la semilla. De esta forma se obtuvo:

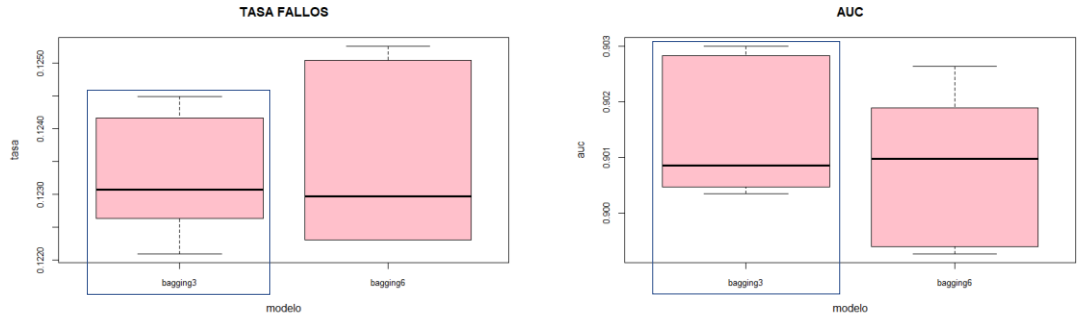


Ilustración 23. Tasa de fallos y AUC mejores modelos bagging.

De modo que, en base a los resultados de la Ilustración 23, el modelo que se escogió como ganador fue el de "*bagging3*", aquel con un "*mtry*" de 27, un "*ntree*" de 3.000, un "*nodesize*" de 20 y un "*sampsize*" de 5.957, dado que presentó tanto un mayor AUC como una menor tasa de fallos.

5.3.2. Random forest:

De forma similar a como se hizo en redes neuronales se creó una rejilla especificando unos valores de "*mtry*" desde 1 hasta 26 con un "ntree" de 300 y un "nodesize" de 10 con la finalidad de estudiar así el número óptimo de variables que debían sortearse en cada nodo.

Tabla 23. Resultados rejilla random forest.

mtry	Tasa de aciertos
1	0,8615385
2	0,8656847
3	0,8739773
4	0,8749591
5	0,8760506
6	0,8744137
7	0,8741956
8	0,8745230
9	0,8733227
10	0,8732132
11	0,8740866
12	0,8753959
13	0,8745226

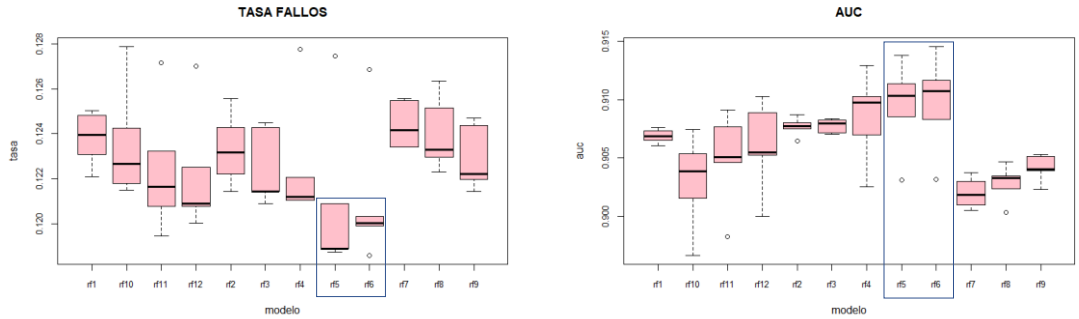
mtry	Tasa de aciertos
14	0,8749594
15	0,8762684
16	0,8747409
17	0,8745230
18	0,8751777
19	0,8753958
20	0,8749596
21	0,8750684
22	0,8732140
23	0,8729956
24	0,8745231
25	0,8736502
26	0,8728864

Gracias a la Tabla 23 se supo que los dos "*mtry*" que ofrecían mayores tasas de aciertos eran 15 y 5, respectivamente. Por lo que, los diversos modelos de random forest a crear contendrían dichos "*mtry*", además de las mismas combinaciones de hiperparámetros que se emplearon en el apartado de bagging. Así pues, los modelos sobre los que se realizó validación cruzada repetida fueron:

Tabla 24. Modelos candidatos random forest.

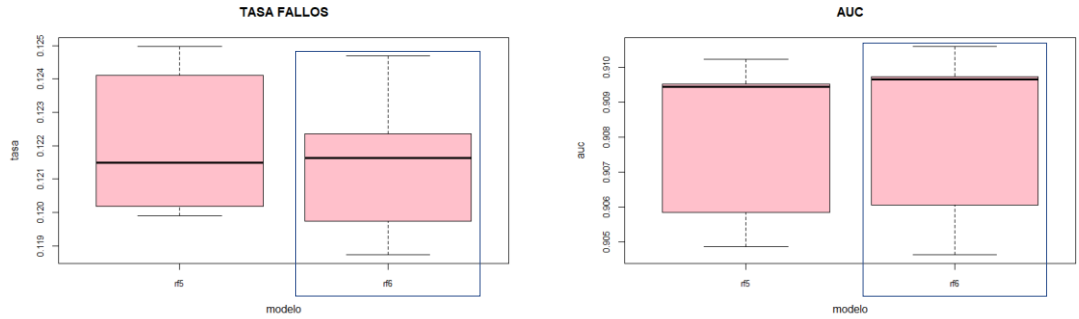
Modelo	mtry	ntree	nodesize	sampsize
rf1	5	3.000	10	5.957 (0,65*9.165)
rf2	5	3.000	15	5.957 (0,65*9.165)
rf3	5	3.000	20	5.957 (0,65*9.165)
rf4	5	5.000	10	6.874 (0,75*9.165)
rf5	5	5.000	15	6.874 (0,75*9.165)
rf6	5	5.000	20	6.874 (0,75*9.165)
rf7	15	3.000	10	5.957 (0,65*9.165)

Modelo	mtry	ntree	nodesize	sampsize
rf8	15	3.000	15	5.957 (0,65*9.165)
rf9	15	3.000	20	5.957 (0,65*9.165)
rf10	15	5.000	10	6.874 (0,75*9.165)
rf11	15	5.000	15	6.874 (0,75*9.165)
rf12	15	5.000	20	6.874 (0,75*9.165)



<u>Ilustración 24</u>. Tasa de fallos y AUC random forest.

Los dos modelos que presentaron tanto menor tasa de fallos como mayor AUC fueron "*rf5*" y "*rf6*", tal y como puede verse en la Ilustración 24. Es por eso que, sobre dichos modelos, se volvió a realizar validación cruzada repetida, lográndose:



<u>Ilustración 25</u>. Tasa de fallos y AUC mejores modelos random forest.

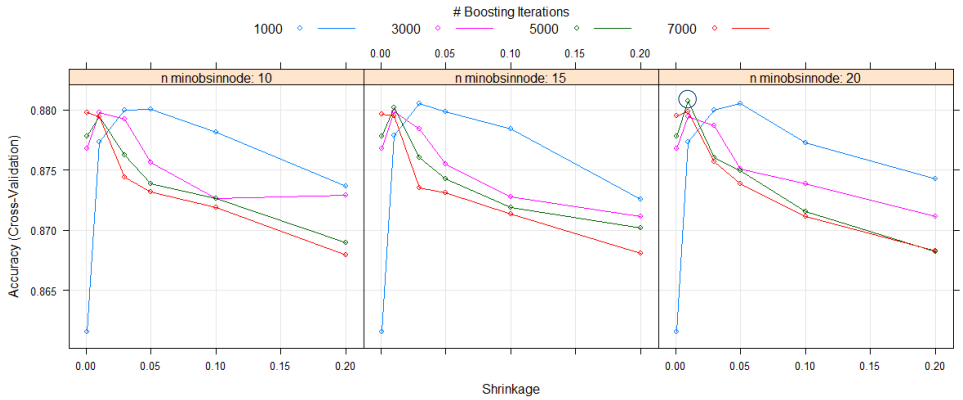
Observando la Ilustración 25 puede apreciarse como ambos modelos presentaron un AUC muy similar. Sin embargo, aquel que presentó una menor tasa de fallos fue el de "*rf6*", escogiéndose por ello mismo éste como ganador. Dicho modelo fue aquel compuesto por un "*mtry*" de 5, un "*ntree*" de 5.000, un "*nodesize*" de 20 y un "*sampsize*" de 6.874.

5.3.3. Gradient boosting:

Para la creación de los diversos modelos de gradient boosting se volvió nuevamente a emplear el paquete "*caret*", cuya rejilla diseñada con el propósito de hallar elevadas tasas de aciertos combinó los siguientes valores de los hiperparámetros:

- "*shrinkage*": 0,001, 0,01, 0,03, 0,05, 0,1 y 0,2
- "*n.minobsinnode*": 10, 15 y 20.
- "*n.trees*": 1.000, 3.000, 5.000 y 7.000.
- "*interaction.depth*": 2.

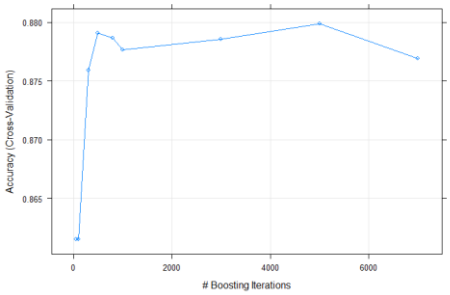
Y, utilizándose validación cruzada repetida sobre la misma se tuvo que:



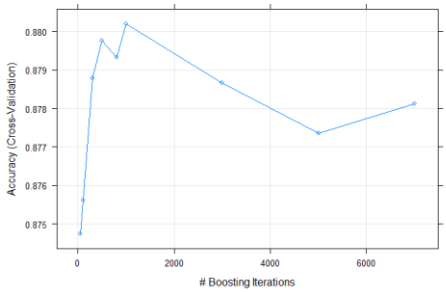
<u>Ilustración 26</u>. Iteraciones gradient boosting.

En la Ilustración 26 puede verse que la mayor tasa de aciertos se consiguió con un "*shrinkage*" de 0,01, un "*n.minobsinnode*" de 20 y un "*n.trees*" de 5.000. Además, puede apreciarse como las tasas de aciertos más altas se consiguieron con unos valores de "*shrinkage*" bajos, entre 0,01 y 0,05, ocurriendo algo semejante con unos valores de "*n.trees*" de 1.000 y 5.000, no siendo necesarios más (ver Tabla 43 y Tabla 44 del Anexo II para comprobarlo). Así pues, se probó fijar un "*shrinkage*" de 0,01, 0,03 y 0,05 y un "*n.minobsinnode*" de 20, en base a los resultados anteriores, con la finalidad de saber cuántas iteraciones debían utilizarse.

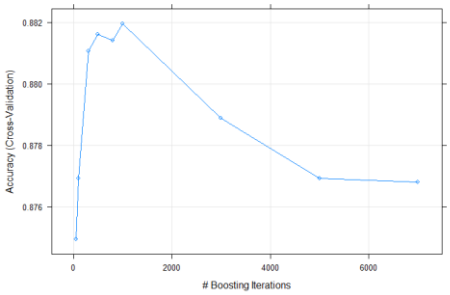
shrinkage=0,01



shrinkage=0,03



shrinkage=0,05



<u>Ilustración 27</u>. Iteraciones gradient boosting con "*shrinkage*"=0,01, 0,03 y 0,05 y "*n.minobsinnode*"=20.

Las iteraciones óptimas a emplear eran las de 1.000, pero, aun así, dado que con un "*shrinkage*" de 0,01 se conseguía una mayor tasa de aciertos con 5.000, como bien puede observarse en la Ilustración 27, se emplaron ambos valores para la creación

de los diversos modelos. De modo que, mediante la validación cruzada repetida se probaron los modelos:

Tabla 25. Modelos candidatos gradient boosting.

Modelo	shrinkage	n.trees	n.minobsinnode	interaction.depth
gb1	0,01	1.000	10	2
gb2	0,01	1.000	15	2
gb3	0,01	1.000	20	2
gb4	0,03	1.000	10	2
gb5	0,03	1.000	15	2
gb6	0,03	1.000	20	2
gb7	0,05	1.000	10	2
gb8	0,05	1.000	15	2
gb9	0,05	1.000	20	2
gb10	0,01	5.000	10	2
gb11	0,01	5.000	15	2
gb12	0,01	5.000	20	2
gb13	0,03	5.000	10	2
gb14	0,03	5.000	15	2
gb15	0,03	5.000	20	2
gb16	0,05	5.000	10	2
gb17	0,05	5.000	15	2
gb18	0,05	5.000	20	2

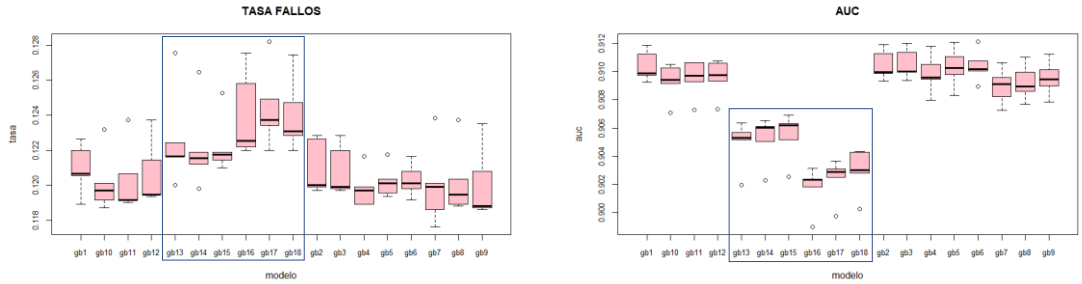


Ilustración 28. Tasa de fallos y AUC gradient boosting.

Puede apreciarse en la Ilustración 28 que hubo 6 modelos con un AUC bastante más bajo que el de los demás. Además, presentaron mayores tasas de fallos, por lo que se procedió a eliminarlos para poder visualizar mejor los valores de los modelos restantes, obteniéndose así:

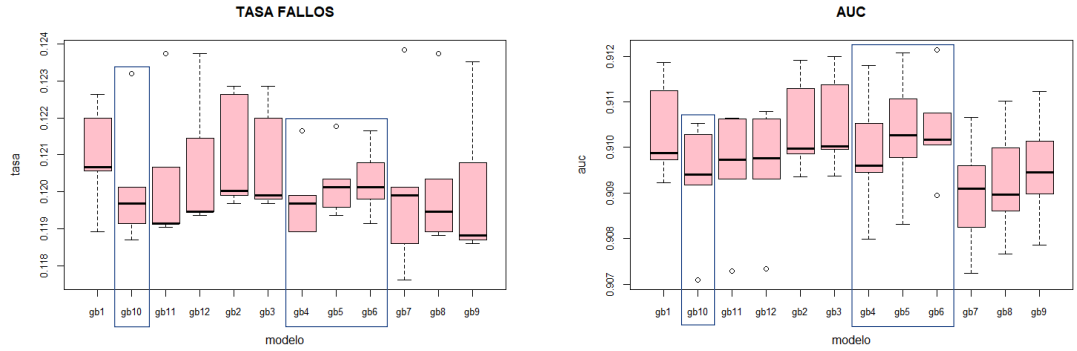
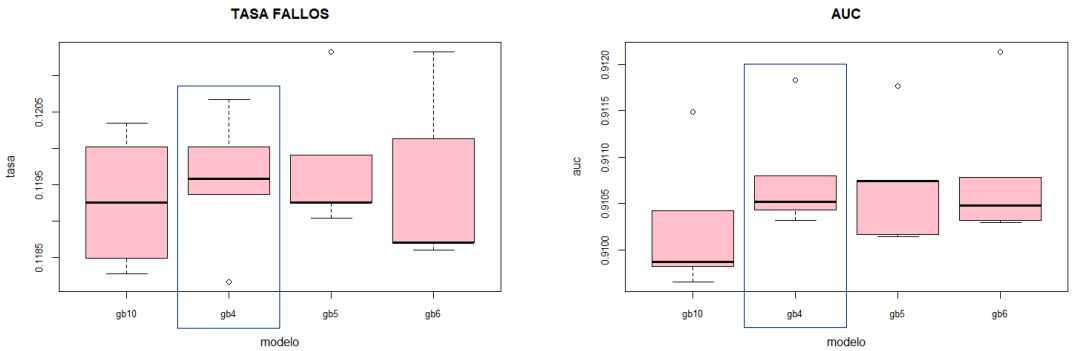


Ilustración 29. Tasa de fallos y AUC gradient boosting (2).

Los 4 modelos que presentaron un alto AUC a la vez que una baja tasa de fallos junto a una baja varianza de ésta en la Ilustración 29 fueron: "*gb4*", "*gb5*", "*gb6*" y "*gb10*".

Luego, es a estos a los que se les volvió a realizar validación cruzada repetida para determinar al ganador.



<u>Ilustración 30</u>. Tasa de fallos y AUC mejores modelos gradient boosting.

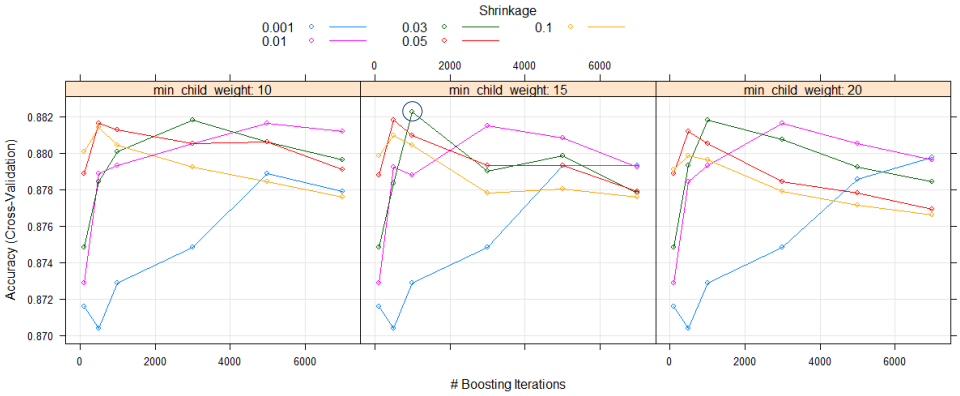
Para finalizar, el modelo que se escogió como ganador de la Ilustración 30 fue "*gb4*", pues de todos los modelos restantes fue el más estable, ya que, presentó tanto una baja varianza de tasa de fallos como de AUC. El modelo elegido estaba compuesto por un "*shrinkage*" de 0,03, un "*n.trees*" de 1.000, un "*n.minobsinnode*" de 10 y una "*interaction.depth*" de 2.

5.3.4. Extreme gradient boosting:

De forma idéntica a como se hizo con el algoritmo anterior, se creó una rejilla con los siguientes valores de los hiperparámetros:

- "*min_child_weight*": 10, 15 y 20.
- "*eta*": 0,001, 0,01, 0,03, 0,05 y 0,1.
- "*nrounds*": 100, 500, 1.000, 3.000, 5.000 y 7.000.
- "*max_depth*": 2.
- "*gamma*": 0.
- "*colsample_bytree*": 1.
- "*subsample*": 1.

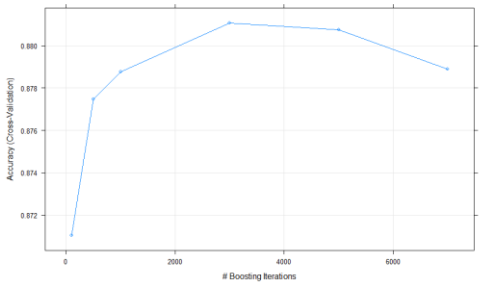
Y, empleándose validación cruzada repetida sobre la misma se tuvo que:



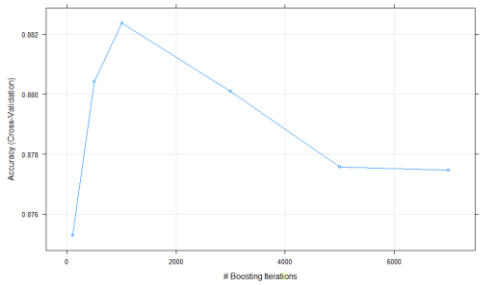
<u>Ilustración 31</u>. Iteraciones extreme gradient boosting.

Se consiguió una mayor tasa de aciertos con un "*nrounds*" de 1.000, un "*max_depth*" de 2, un "*eta*" de 0,03, un "*gamma*" de 0, un "*colsample_bytree*" de 1, un "*min_child_weight*" de 15 y un "*subsample*" de 1, tal y como puede verse en la Ilustración 31. Dado que los 3 "*eta*" que ofrecieron tasas de aciertos más altas fueron las de 0,01, 0,03 y 0,05 se probó establecer los mismos, junto a un "*min_child_weight*" de 10, para conocer cuántas iteraciones debían utilizarse, pues era el valor que más se repitía (fijarse en las Tabla 45 y Tabla 46 del Anexo II).

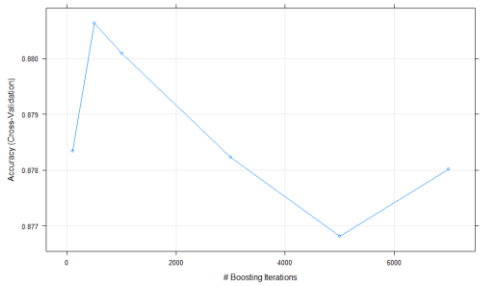
shrinkage=0,01



shrinkage=0,03



shrinkage=0,05

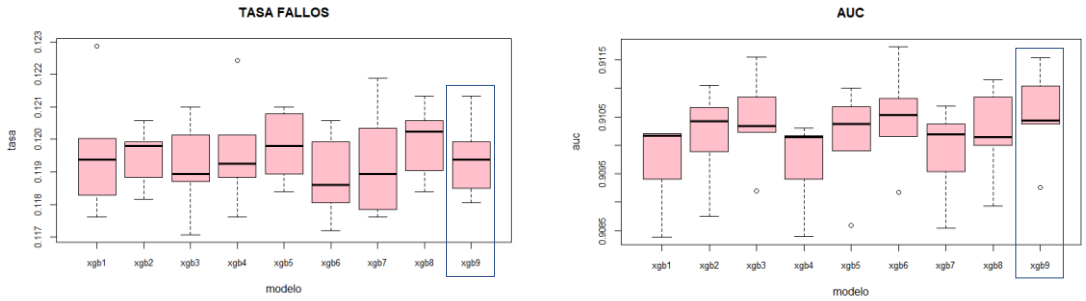


<u>Ilustración 32</u>. Iteraciones extreme gradient boosting con "*eta*"=0,01, 0,03 y 0,05 y "*min_child_weight*"=10.

Dependiendo del "*eta*" empleado se consiguía una tasa de aciertos más alta con unas iteraciones distintas. En el caso de "*eta*"=0,01 era con 3.000, en el de 0,03 con 1.000 y en el de 0,05 con 500, por lo que parecía que a mayor "*eta*", menos iteraciones eran necesarias. Por eso mismo, se crearon modelos con estos 3 valores de "*nrounds*". Así pues, mediante la validación cruzada repetida se crearon los siguientes modelos de extreme gradient boosting:

<u>Tabla 26</u>. Modelos candidatos extreme gradient boosting.

Modelo	eta	nrounds	min_child_weight	max_depth
xgb1	0,01	3.000	10	2
xgb2	0,01	3.000	15	2
xgb3	0,01	3.000	20	2
xgb4	0,03	1.000	10	2
xgb5	0,03	1.000	15	2
xgb6	0,03	1.000	20	2
xgb7	0,05	500	10	2
xgb8	0,05	500	15	2
xgb9	0,05	500	20	2



<u>Ilustración 33</u>. Tasa de fallos y AUC extreme gradient boosting.

Todos los modelos presentaron casi la misma tasa de fallos y AUC, por lo que todos ellos eran muy similares, pero el que se escogió como ganador fue el de "*xgb9*", pues de entre los que presentaron un alto AUC fue el que menor tasa de fallos ofreció, además de una baja varianza. Este modelo fue aquel compuesto por un "*eta*" de 0,05, un "*nrounds*" de 500, un "*min_child_weight*" de 20 y un "*max_depth*" de 2.

5.4. Support vector machines:

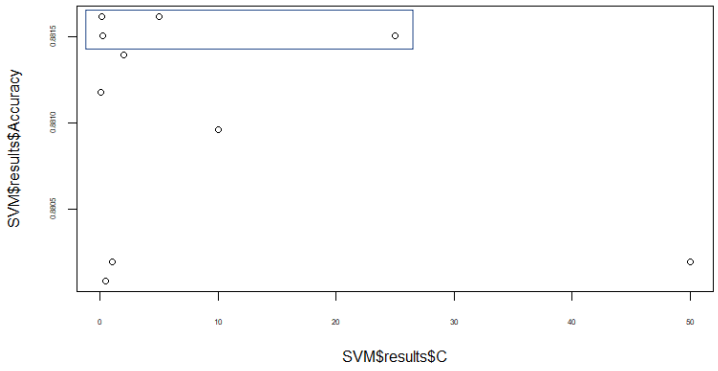
5.4.1. SVM con kernel lineal:

De nuevo, para poder tunear los algoritmos de SVM hizo falta emplear el paquete "*caret*". En el caso de SVM con kernel lineal únicamente pudo tunearse la constante de regularización "*C*", creándose por ello mismo una rejilla con exclusivamente los valores de 0,05, 0,1, 0,2, 0,5, 1, 2, 5, 10, 25 y 50. Las respectivas tasas de aciertos que se obtuvieron fueron:

<u>Tabla 27</u>. Resultados rejilla SVM con kernel lineal.

C	Tasa de aciertos
0,05	0,8811789
0,1	0,8816153
0,2	0,8815062
0,5	0,8800876
1	0,8801968

C	Tasa de aciertos
2	0,8813971
5	0,8816153
10	0,8809606
25	0,8815062
50	0,8801968



<u>Ilustración 34</u>. Resultados rejilla SVM con kernel lineal.

Tal y como afirman la Tabla 27 e Ilustración 34, las mejores tasas de aciertos se consiguieron con un valor de "*C*" de 0,1, 0,2, 5 y 25 y acorde con lo recién mencionado, los modelos que se crearon fueron:

Tabla 28. Modelos candidatos SVM con kernel lineal.

Modelo	C
svml1	0,1
svml2	0,2
svml3	5
svml4	25

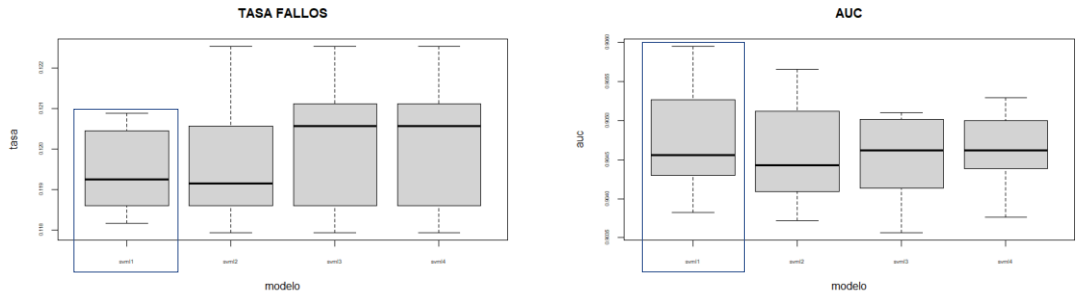


Ilustración 35. Tasa de fallos y AUC SVM con kernel lineal.

Así que, en base a los resultados de la Ilustración 35 el modelo que se determinó como ganador fue el de "*svml1*", debido a que presentó no solo un mayor AUC sino también una menor tasa de fallos. O sea que el valor de "*C*" que brindó mejores resultados en SVM con kernel lineal fue el de 0,1.

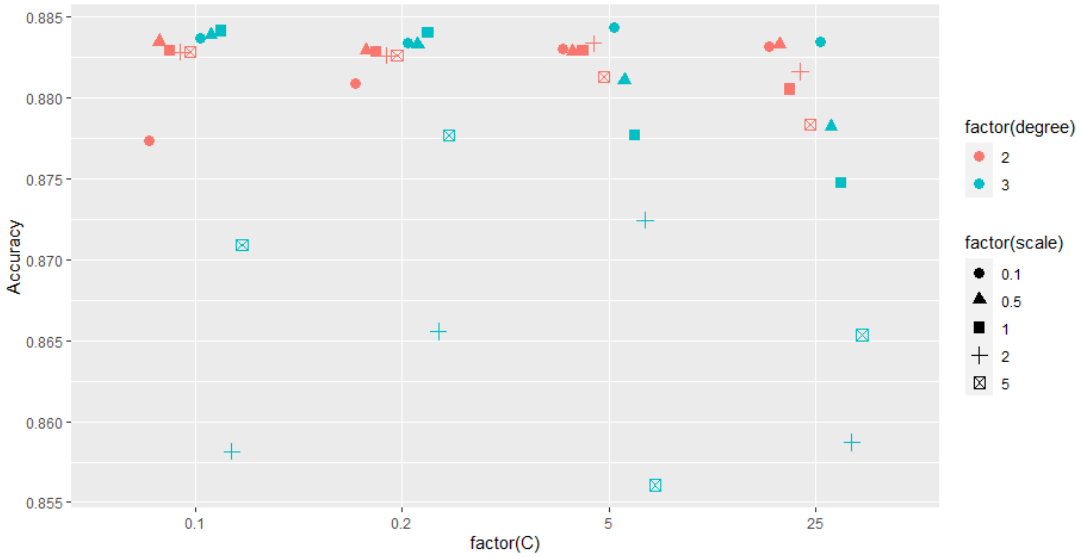
5.4.2. SVM con kernel polinomial:

A diferencia del algoritmo de SVM con kernel lineal, en el de kernel polinomial puedieron tunearse, a parte de la constante "*C*", los grados del polinomio ("*degree*") y la escala ("*scale*"). La rejilla que se diseñó comprendió unos valores de "*C*" iguales a 0,1, 0,2, 5 y 25, ya que fueron los que mejores tasas de acierto arrojaron con el algoritmo anterior de SVM, unos valores de 2 y de 3 para "*degree*" y unos de 0,1, 0,5, 1, 2 y 5 para "*scale*". De esta forma, se tuvo que:

Tabla 29. Resultados rejilla SVM con kernel polinomial.

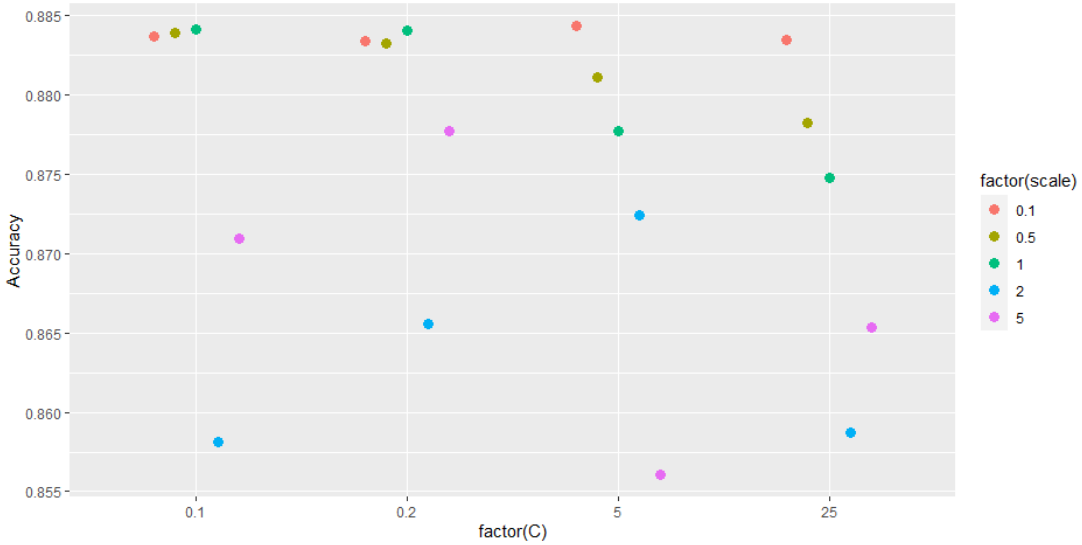
C	degree	scale	Tasa de aciertos
0,1	2	0,1	0,8773599
0,2	2	0,1	0,8808518
5	2	0,1	0,8830341
25	2	0,1	0,8831433
0,1	2	0,5	0,8834706
0,2	2	0,5	0,8829251
5	2	0,5	0,8828159
25	2	0,5	0,8832523
0,1	2	1	0,8829251
0,2	2	1	0,8828159
5	2	1	0,8829251
25	2	1	0,8805246
0,1	2	2	0,8828159
0,2	2	2	0,8825977
5	2	2	0,8833612
25	2	2	0,8816160
0,1	2	5	0,8828159
0,2	2	5	0,8825978
5	2	5	0,8812883
25	2	5	0,8783428

C	degree	scale	Tasa de aciertos
0,1	3	0,1	0,8836889
0,2	3	0,1	0,8833616
5	3	0,1	0,8843432
25	3	0,1	0,8834704
0,1	3	0,5	0,8839069
0,2	3	0,5	0,8832522
5	3	0,5	0,8810697
25	3	0,5	0,8782329
0,1	3	1	0,8841251
0,2	3	1	0,884016
5	3	1	0,8776878
25	3	1	0,874742
0,1	3	2	0,8581575
0,2	3	2	0,8655749
5	3	2	0,87245
25	3	2	0,8587037
0,1	3	5	0,8709228
0,2	3	5	0,8776876
5	3	5	0,8560835
25	3	5	0,8653578



<u>Ilustración 36</u>. Resultados rejilla SVM con kernel polinomial.

La Ilustración 36 facilita ver que un "*degree*" de 3 exhibía mayores tasas de aciertos que uno de 2, es por eso que se redujo el gráfico a dichos grados del polinomio, obteniéndose:



<u>Ilustración 37</u>. Resultados rejilla SVM con kernel polinomial para "*degree*"=3.

De esta forma, se consiguió ver gracias a la Ilustración 37 que para unos valores de "*C*" iguales a 0,1 y 0,2 se obtenían mayores tasas de aciertos con una escala de 1 mientras que, para unos iguales a 5 y 25, era con una escala de 0,1. Por tanto, los modelos que se construyeron fueron:

<u>Tabla 30</u>. Modelos candidatos SVM con kernel polinomial.

Modelo	C	degree	scale
svmp1	0,1	3	1
svmp2	0,2	3	1
svmp3	5	3	0,1
svmp4	25	3	0,1

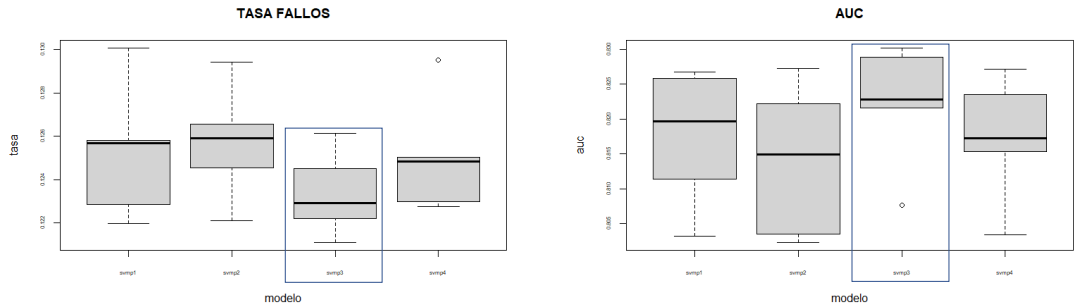


Ilustración 38. Tasa de fallos y AUC SVM con kernel polinomial.

El mejor modelo fue sin lugar a dudas el de "*svmp3*", debido a su baja tasa de fallos y elevado AUC, siendo aquel con un valor de "*C*" de 5, un "*degree*" de 3 y un "*scale*" de 0,1.

5.4.3. SVM con kernel gaussiano:

Finalmente, en el caso de SVM con kernel gaussiano pudo tunearse tanto el valor de "*C*", igual que en los otros dos algoritmos anteriores, como el de "*sigma*", que es quien controla el comportamiento del kernel. Para ello, se volvió a diseñar una nueva rejilla con los mismos valores de "*C*" que se venían empleando y unos valores de "*sigma*" desde 0,001 hasta 0,5. Es de ahí de donde se obtuvieron los siguientes resultados:

Tabla 31. Resultados rejilla SVM con kernel gaussiano.

C	sigma	Tasa de aciertos
0,1	0,001	0,8615385
0,1	0,01	0,8615385
0,1	0,02	0,8660121
0,1	0,05	0,8748504
0,1	0,1	0,8734316
0,1	0,2	0,8661211
0,1	0,5	0,8629569
0,2	0,001	0,8615385
0,2	0,01	0,8681943
0,2	0,02	0,8765960
0,2	0,05	0,8758322
0,2	0,1	0,8788875
0,2	0,2	0,8770325
0,2	0,5	0,8671029

C	sigma	Tasa de aciertos
5	0,001	0,8777962
5	0,01	0,8794328
5	0,02	0,8824877
5	0,05	0,8822698
5	0,1	0,8775782
5	0,2	0,8731041
5	0,5	0,8708126
25	0,001	0,8805235
25	0,01	0,8821604
25	0,02	0,8829245
25	0,05	0,8786695
25	0,1	0,8744136
25	0,2	0,8723401
25	0,5	0,8707034

Por tanto, como resultado a la Tabla 31 e Ilustración 39, se aprecia que para un determinado valor de "*C*" el "*sigma*" que ofrecía mayores tasas de aciertos era distinto. Es por ello que para un "*C*" de 0,1 se asignó un "*sigma*" de 0,05, para uno de 0,2 de 0,1 y para unos de 5 y 25 uno de 0,02, tal y como se recoge en la Tabla 32, siendo los modelos sobre los que se realizó validación cruzada repetida. Las tasas de fallos y AUC que se obtuvieron fueron las que se observan en la Ilustración 40, remarcándose en azul el modelo que mejores valores de estos brindó.

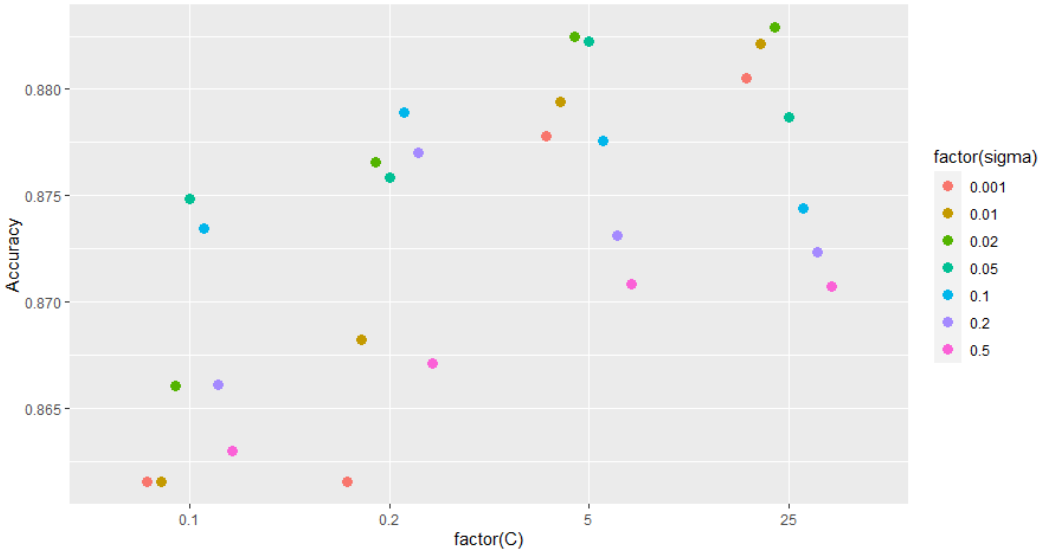


Ilustración 39. Resultados rejilla SVM con kernel gaussiano.

Claramente, el modelo que se escogió como ganador en la Ilustración 40 fue el primero, el de "*svmrbf1*", pues es el que, con diferencia, mayor AUC presentó, además de una tasa de fallos similar a la de los restantes modelos, pero con una baja varianza.

Tabla 32. Modelos candidatos SVM con kernel gaussiano.

Modelo	C	sigma
svmrbf1	0,1	0,05
svmrbf2	0,2	0,1
svmrbf3	5	0,02
svmrbf4	25	0,02

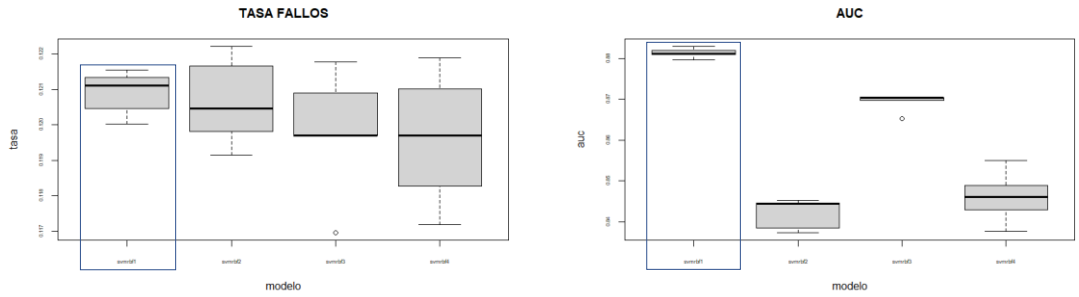


Ilustración 40. Tasa de fallos y AUC SVM con kernel gaussiano.

El modelo ganador fue por tanto aquel que tuvo un valor de la constante "*C*" de 0,1 y un valor de "*sigma*" de 0,05.

6. Resultados:

6.1. Comparación de modelos:

En primer lugar y a modo de resumen, se recogen en la Tabla 33 todas las características de los diversos modelos que resultaron ganadores. Señalar que los modelos de regresión logística que se generaron y cuyas variables resultantes se

emplearon tanto en las técnicas basadas en árboles como en las técnicas no basadas en estos, también se compararon.

Tabla 33. Características de los modelos ganadores.

Técnica	Modelo	Características
Regresión logística	log1	- 13 variables (selección por R^2 > 0,005 con SAS Enterprise Miner Workstation 14.1)
	log2	- 7 variables (selección por método stepwise con SAS 9.4)
Red neuronal	avnnet	- N.º de nodos en la capa oculta: 3 - Learning rate: 0,01 - Iteraciones: 100
Bagging	bg	- N.º de variables: 27 - N.º de árboles que se ejecutan en cada iteración: 3.000 - Tamaño mínimo de las hojas finales de cada árbol: 20 - Tamaño de la muestra que se reemplaza: 5.957 - Con reemplazamiento
Random forest	rf	- N.º de variables: 5 - N.º de árboles que se ejecutan en cada iteración: 5.000 - Tamaño mínimo de las hojas finales de cada árbol: 20 - Tamaño de la muestra que se reemplaza: 6.874 - Con reemplazamiento
Gradient boosting	gbm	- Parámetro de regularización: 0,03 - N.º de árboles que se ejecutan en cada iteración: 1.000 - Tamaño máximo de nodos finales: 10 - Profundidad de la iteración: 2
Extreme gradient boosting	xgbm	- Tasa de aprendizaje: 0,05 - N.º de iteraciones: 500 - N.º de observaciones mínimas en el nodo final: 20 - Profundidad máxima de los árboles: 2 - Coste de regularización: 0 - % de sorteo de variables antes de cada árbol: 1 - % de observaciones de variables antes de cada árbol: 1
Support vector machine	svml	- Constante de regularización "*C*": 0,1
	svmp	- Constante de regularización "*C*": 5 - Grados del polinomio: 3 - Escala: 0,1
	svmrbf	- Constante de regularización "*C*": 0,1 - Valor de "*sigma*": 0,05

Para realizar este estudio comparativo de modelos señalar que volvieron a construirse los mismos y que tras esto, se empleó sobre ellos la técnica de validación cruzada repetida, solo que, en este nuevo caso, con 10 grupos y 20 repeticiones.

En la Ilustración 41 se recogen los resultados que se obtuvieron. Hay 3 modelos que destacaron por su bajo AUC, siendo los de "*bg*", "*svmrbf*" y "*svmp*", aunque sin duda, este último fue el peor de todos ellos.

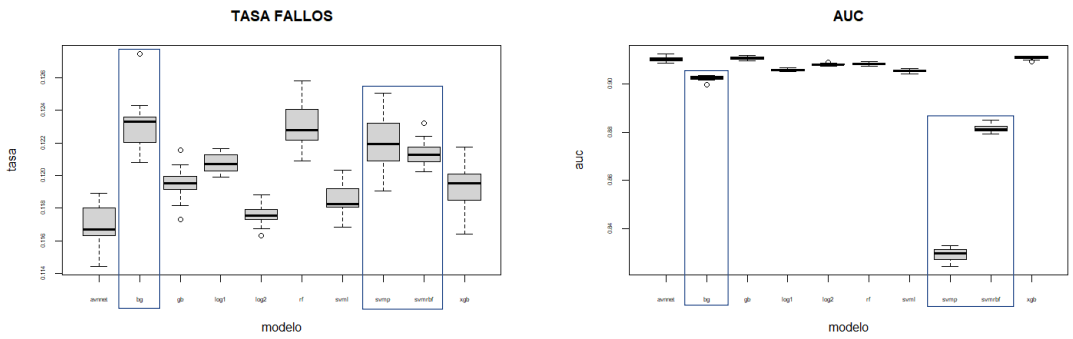
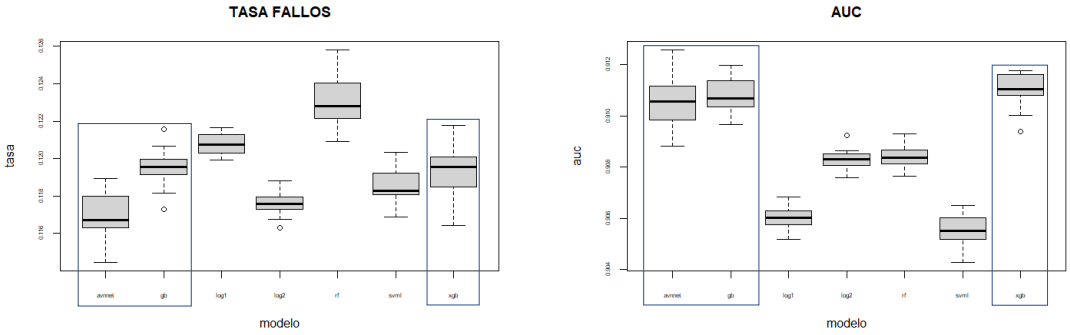


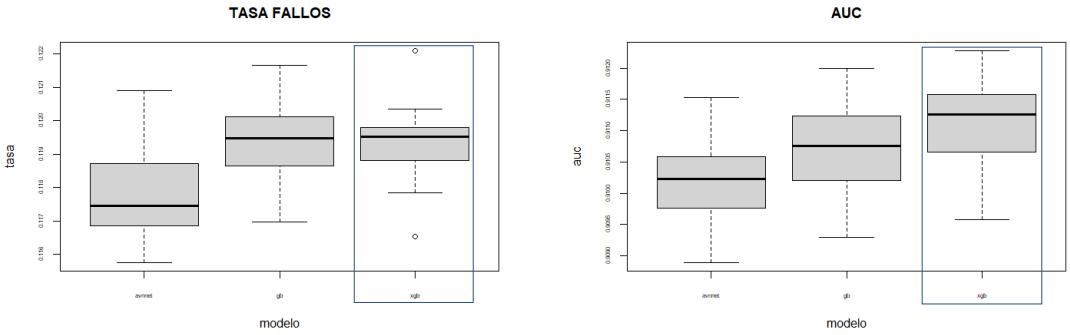
Ilustración 41. Tasa de fallos y AUC modelos ganadores.

Una vez se rehicieron los gráficos tras no considerarse los modelos que recién se acaban de mencionar, se tuvo que:



<u>Ilustración 42</u>. Tasa de fallos y AUC modelos ganadores (2).

Los mejores modelos de la Ilustración 42 y, por tanto, aquellos que se emplearían en el ensamble fueron: "*avnnet*", "*gbm*" y "*xgbm*". Si sobre estos volvía a realizarse validación cruzada repetida variando tan solo la semilla se conseguía:



<u>Ilustración 43</u>. Tasa de fallos y AUC modelos ganadores (3).

El modelo que a priori se acabó escogiendo como ganador fue el de "*xgb*" dado que de los tres fue el que mayor AUC presentó. A su vez, también ofreció una tasa de fallos similar a la de los restantes modelos, junto a una baja varianza de ésta.

6.2. Modelos de ensamble:

En pocas palabras, los modelos de ensamble construyen predicciones a partir de la combinación de varios modelos. Por consiguiente, se realizó un promedio de los modelos que mejores resultados arrojaron bajo los algoritmos de redes neuronales, gradient boosting y extreme gradient boosting, siendo el objetivo final a alcanzar el de intentar reducir todavía más el error de los modelos que se habían obtenido hasta el momento. La combinación de modelos que se probaron es la recogida en la Tabla 34, habiéndose empleado validación cruzada repetida sobre los mismos con un total de 10 grupos y 20 repeticiones.

Tabla 34. Descripción modelos ensamble.

Modelo	Descripción
avnnet	RED
gbm	GB
xgbm	XGB
predi4	RED + GB
predi5	RED + XGB
predi6	GB + XGB
predi7	RED + GB + XGB

Los resultados que se obtuvieron fueron:

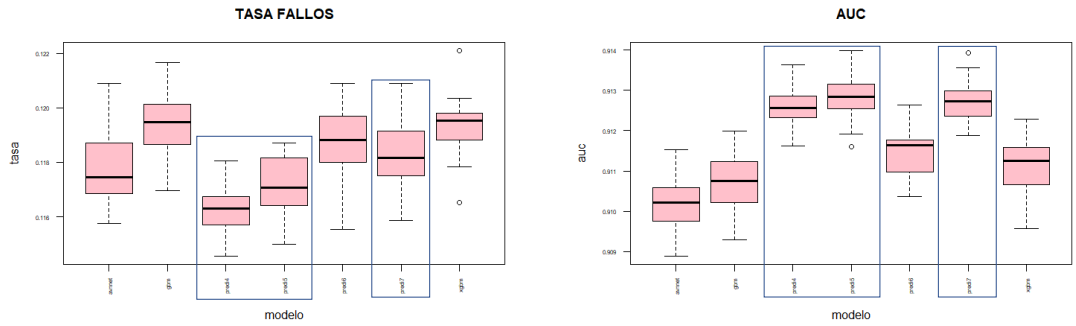


Ilustración 44. Tasa de fallos y AUC modelos ensamble.

Tal y como se observa en la Ilustración 44, los modelos de ensamble consiguieron reducir el error y aumentar el AUC, destacando los de "*predi4*", "*predi5*" y "*predi7*". Aun así, si nos fijábamos en los valores del eje Y, dado que la mejora era muy pequeña, no compensaba el aumento de complejidad que estos modelos generaban. Por eso mismo, se acabó determinando como modelo ganador definitivo al de "*xgbm*", es decir, aquel que tenía una tasa de aprendizaje de 0,05, un número de iteraciones de 500, un número de observaciones mínimas en el nodo final de 20, una profundidad máxima de los árboles de 2, un coste de regularización igual a 0, un porcentaje de sorteo de variables antes de cada árbol de 1 y un porcentaje de observaciones de variables antes de cada árbol de 1.

6.3. Análisis del modelo ganador:

Tras determinarse como ganador definitivo al modelo "*xgbm*" de la Tabla 33 se procedió a analizarlo. Primeramente, se obtuvo un gráfico de la importancia de las variables a través de los gráficos dinámicos de Microsoft Excel con el propósito de conocer así cuáles de ellas eran las que realizaban un mayor aporte al modelo para predecir la variable objetivo.

Por la Ilustración 45 puede apreciarse que las variables que mayor importancia tenían eran las 5 primeras, siendo estas: "*control*", "*IMP_REP_REP_numpar*", "*REP_edad*", "*IMP_REP_suicidio*" y "*tenpar*", destacando sobre todas ellas la primera.

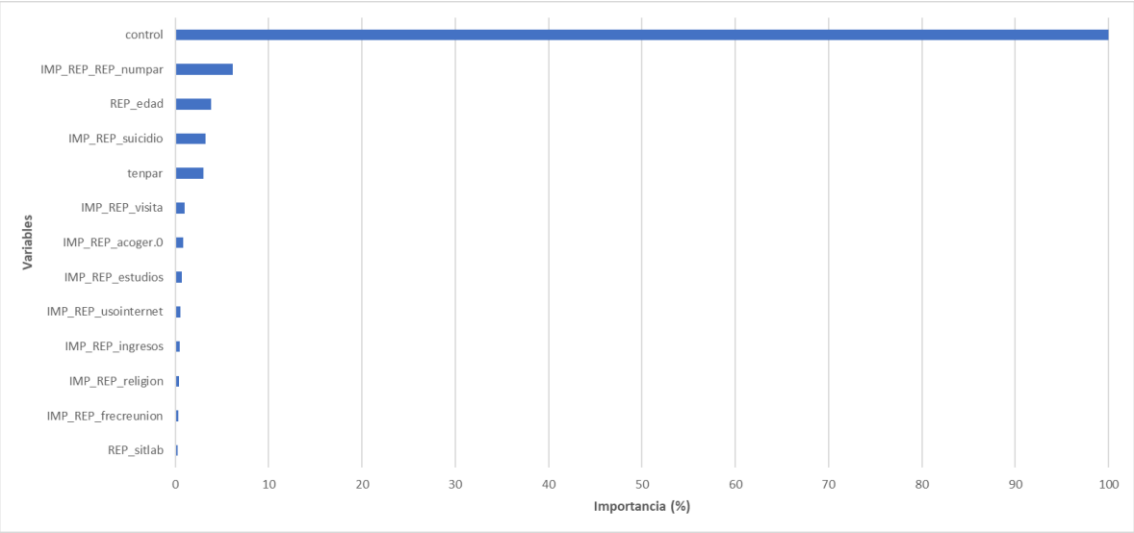


Ilustración 45. Importancia de las variables del modelo ganador en %.

Luego, en base a los resultados obtenidos pudo interpretarse que:

- Sin lugar a dudas, la variable más relevante en el modelo y la que mayor aporte hacía era la de "*control*" pues como ya se venía diciendo, aquellas entrevistadas que nunca habían sentido miedo de alguna de su/s expareja/s y/o pareja apenas habían sufrido control por parte de estas, a diferencia de las que sí que habían sufrido tal temor.
- Además, otra de las variables que se había visto que guardaba relación con la variable objetivo era la de "*IMP_REP_REP_numpar*", donde la gran mayoría de entrevistadas que nunca habían tenido miedo de su pareja actual y/o pasada/s habían dispuesto tan solo de una única a lo largo de su vida, mientras que aquellas que sí que habían padecido dicho sentimiento habían dispuesto mayoritariamente de dos y de tres, respectivamente. Esto es lógico que ocurra porque cuantas más parejas se han tenido, mayor opción existe de haber tenido miedo de alguna de las mismas.
- En cuanto a la variable "*REP_edad*", si se relacionaba con la variable anterior, tenía sentido que aquellas mujeres que disponían de más edad hubiesen tenido más parejas y, por ende, hubiesen sentido alguna vez miedo de las mismas, a diferencia de aquellas con una menor edad.
- En penúltimo lugar, la otra variable que se había considerado relevante con el nodo Multi gráfico fue la de "*IMP_REP_suicidio*", apreciándose que un considerable número de mujeres entrevistada había pensado alguna vez en quitarse voluntariamente la vida tras haber sufrido miedo de alguna de su/s pareja/s, a diferencia de aquellas que no lo habían sufrido.
- Por último, la variable "*tenpar*" también es una de las más relevantes, pudiendo esto indicar que aquellas mujeres que disponían de pareja estuviesen más expuestas al control y por ello mismo, hubiesen sentido miedo de la misma, en comparación con aquellas que no tenían pareja.

Finalizando ya, se comprobó qué medidas de clasificación obtenía el modelo ganador. Para ello, se generó la respectiva matriz de confusión empleando validación cruzada con un punto de corte de 0,5 (una de las estrategias más habituales es la de minimizar la tasa de mal clasificados, lo que se consigue fijando el punto de corte en

dicho valor). De esta forma, si se observan los resultados de la Tabla 35 y Tabla 36, es más que evidente que el modelo detectaba muy bien los casos negativos, pero no tanto los positivos (se le escapaban muchos de ellos, el 52,09% de los mismos (más de la mitad de los positivos)), siendo esto así debido a la muestra desbalanceada con la que se contaba (los datos desbalanceados por lo general perjudican a las clases minoritarias, como ocurre en este caso con los casos positivos). Así pues, se conseguía bien clasificar un 47,91% de los casos positivos y un 94,49% de los casos negativos.

Tabla 35. Matriz de confusión modelo ganador punto de corte=0,5.

	Predicción = 0	Predicción = 1	Total
Realidad = 0	7.461	435	7.896
Realidad = 1	661	608	1.269
Total	8.122	1.043	9.165

Tabla 36. Medidas de clasificación modelo ganador punto de corte=0,5.

Tasa de fallos	0,1196
Tasa de aciertos	0,8804
Sensibilidad	0,4791
Especificidad	0,9449

Dado que la sensibilidad era bastante baja en comparación con la especificidad, resultó conveniente cambiar el punto de corte para conseguir unos mejores resultados de sensibilidad, aunque esto implicase reducir el valor de la especificidad. Para ello, se fijó en 0,1385, determinándose en base a la proporción de eventos (siendo en este caso de tal valor, ya que, 1.269/9.165=0,1385). Con este nuevo punto de corte la matriz de confusión y medidas de clasificación que se obtuvieron fueron:

Tabla 37. Matriz de confusión modelo ganador punto de corte=0,1385.

	Predicción = 0	Predicción = 1	Total
Realidad = 0	6.614	1.282	7.896
Realidad = 1	129	1.140	1.269
Total	6.743	2.422	9.165

Tabla 38. Medidas de clasificación modelo ganador punto de corte=0,1385.

Tasa de fallos	0,1540
Tasa de aciertos	0,8460
Sensibilidad	0,8983
Especificidad	0,8376

En consecuencia, como se venía diciendo, se consiguió aumentar la sensibilidad (en un +87,5%), pero, por ende, la especificidad disminuyó (en un -11,36%). No obstante, como el objetivo del modelo era el de detectar mujeres con miedo, con un punto de corte de 0,1385 se consiguió mejorar el resultado base, ya que, llegó a alcanzarse una sensibilidad del 0,8983 (el "no modelo" supone que ninguna mujer ha tenido miedo, lo que implica una sensibilidad del 0%).

7. Conclusiones y trabajo futuro:

Lo que se pretendía con el presente trabajo era tratar de predecir si alguna de las mujeres entrevistadas había tenido o no miedo de alguna de su/s pareja/s pasada/s y/o de la actual, en el caso de que se dispusiese de ésta.

Para ello, inicialmente, fue necesario seleccionar manual y subjetivamente aquellas variables que guardasen gran relación con los factores de riesgo de violencia machista ya identificados en la literatura científica y profesional, pues la base de datos escogida contaba con más de 1.000 variables.

Tras esto, se procedió a la depuración de los datos y análisis de los mismos donde se pudo apreciar la existencia de una desproporción entre las dos posibles clases de la variable objetivo y que las tres variables que a priori parecía que iban a influir más en la predicción de la misma eran las respuestas de si alguna vez la entrevistada había sufrido control de actividades por parte de alguna de su/s pareja/s, el número de estas que había tenido a lo largo de su vida, incluida la actual y si había pensado en algún momento en terminar voluntariamente con su vida.

Una vez realizado lo anterior, para la construcción de los diversos modelos de machine learning se seleccionaron dos conjuntos de variables distintos: para las técnicas basadas en árboles aquel compuesto por todas aquellas variables con un R^2 superior a 0,005; y para las técnicas no basadas en árboles, el conjunto obtenido empleando regresión logística, el método stepwise y validación cruzada repetida, ya que este segundo conjunto podía no ajustar bien en el caso de que se dispusiesen de variables input con relaciones no lineales.

Finalmente, en base a todos los modelos ganadores generados, se determinó al ganador definitivo tras considerarse su tasa de fallos y AUC, siendo aquel creado con el algoritmo de XGBoost y, más concretamente, el compuesto por una tasa de aprendizaje de 0,05, un número de iteraciones de 500, un número de observaciones mínimas en el nodo final de 20, una profundidad máxima de los árboles de 2, un coste de regularización igual a 0, un porcentaje de sorteo de variables antes de cada árbol de 1 y un porcentaje de observaciones de variables antes de cada árbol de 1.

Así pues, a la luz de los resultados obtenidos se concluye que:

1. Los mejores modelos que se hallaron fueron los de extreme gradient boosting y gradient boosting, mientras que los peores, los de support vector machines, y, más exactamente, aquellos con un kernel polinomial y gaussiano.
2. La combinación de diferentes algoritmos (ensamblado) permitió reducir el error y aumentar el AUC. Aun así, la pequeña mejora obtenida no compensaba con el aumento de complejidad.
3. Las variables más influyentes en el miedo fueron las que recogían si alguna vez la entrevistada había sufrido control de actividades por parte de alguna pareja, el número de éstas que había tenido la misma, la edad, si en alguna ocasión había pensado en suicidarse y si disponía actualmente de pareja o no, destacando sobre todas ellas la primera.
4. Un punto de corte de 0,1385 arrojó mejores resultados que uno de 0,5, ya que se consiguió pasar de una sensibilidad de 0,4791 a una de 0,8983, aunque cierto es que se pasó de una especificidad de 0,9449 a una de 0,8376.

Por lo tanto y a modo de cierre, puede afirmarse que se ha alcanzado el objetivo principal del trabajo, ya que, con el modelo hallado consigue detectarse mujeres con miedo y pese a disponer de una muestra desbalanceada, se alcanzan unos valores de especificidad y sensibilidad elevados.

No obstante, como trabajo futuro se propone:

- No seleccionar las variables que van a componer la base de datos objeto de estudio manual y subjetivamente, pues pueden estar descartándose variables altamente relevantes para la predicción de la variable dependiente.
- Para abordar el problema de los datos desbalanceados sería beneficioso emplear la estrategia básica de métodos de muestreo, y más concretamente, el mecanismo de sobremuestreo aleatorio, pudiendo añadirse datos al conjunto de datos original de bases de datos de otros años. De esta forma, podrían incrementarse los datos de la clase minoritaria.
- Finalmente, sería interesante estudiar futuras macroencuestas que se realicen con la finalidad de analizar y comparar como la COVID-19 ha impactado en la violencia de género y, sobre todo, en nuestra variable objetivo.

Bibliografía:

Álvarez, M., Andrés, A., Augé, M., Choy, A., Fernández, R., Fernández, C., Foulon, H., López, S., Martínez, M. T., Martínez, C., Saiz, M. y Serratusell, L. (2011). RVD-BCN – Protocol de valoración del risc de violencia contra la dona per part de la seva parella o exparella. Circuit Barcelona contra la Violència vers les Dones. https://n9.cl/7k0v9

Andrés, A., López, S. y Álvarez, E. (2008). Valoración del riesgo de violencia contra la pareja por medio de la SARA. *Papeles del psicólogo*, *29*(1), 107-122. https://n9.cl/6oanx

Arlot, S. y Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40-79. https://n9.cl/t92vb

Azevedo, A. y Santos, M. (2008). *KDD, SEMMA and CRISP-DM: A parallel overview*. ResearchGate. 182-185. https://n9.cl/bclft

Bhatia, N. (2019, 26 junio). *What is Out of Bag (OOB) score in Random Forest?* Towards Data Science. https://n9.cl/qjc6b

Bhattacharyya, J. (2020, 2 noviembre). *Understanding XGBoost Algorithm In Detail*. Analytics India Magazine. https://n9.cl/iywm7

Boletín Oficial del Estado. (2004). Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género. https://n9.cl/er50

Boletín Oficial del Estado. (2007). Ley 13/2007, de 26 de noviembre, de medidas de prevención y protección integral contra la violencia de género. https://n9.cl/9vh5l

Calviño, A. (2020). "Técnicas y Metodología de la Minería de Datos (SEMMA)". Universidad Complutense de Madrid – Facultad de Estudios Estadísticos.

Chen, L. (2019, 7 enero). Kernel lineal, polynomial y RBF [Ilustración]. Towards Data Science. https://n9.cl/mz5ql

Consejo General del Poder Judicial. (2021, 15 marzo). *La crisis sanitaria y el confinamiento causaron en 2020 un descenso del 10 por ciento en el número de denuncias y de víctimas de violencia de género*. https://n9.cl/y5evp

Delegación del Gobierno contra la Violencia de Género. (2019). Estudio sobre el Tiempo que Tardan las Mujeres Víctimas de Violencia de Género en Verbalizar su Situación. https://n9.cl/qwye4

Delegación del Gobierno contra la Violencia de Género. (2021). Estadística de Víctimas Mortales por Violencia de Género 2020. https://n9.cl/xr986

Demir, N. (2016, 4 febrero). *Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results*. Toptal Engineering Blog. https://n9.cl/swq6y

Domenech, J. (2011, 8 diciembre). Esquema k-fold cross validation, con k=4 y un solo clasificador [Ilustración]. Wikipedia. https://n9.cl/gqvp4

Echeburúa, E., Amor, P. J., Loinaz, I. y De Corral, P. (2010). Escala de Predicción del Riesgo de Violencia Grave contra la pareja –Revisada– (EPV-R). *Psicothema*, 22(*4*), 1054-1060. https://n9.cl/5tx78

Ferre, M. E. (2019). "FEIR 45: Regresión logística". Universidad de Murcia. https://n9.cl/0g2ev

Fiscalía General del Estado. (2020). Memoria de la Fiscalía General del Estado 2019. https://n9.cl/bz6qi

Fiuza, M. D. y Rodríguez, J. (2000). La regresión logística: una herramienta versátil. *Nefrología*, *20*(6), 477-565. https://n9.cl/f752g

Flach, P., Hernandez-Orallo, J. y Ferri, C. (2011). *A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance*. ICML. 657-664. https://n9.cl/tcjvw

González, J. L., López-Ossorio, J. J. y Muñoz, M. (2018). La valoración policial del riesgo de violencia contra la mujer pareja en España. Ministerio del Interior. https://n9.cl/rv3yb

Grover, P. (2017, 9 diciembre). *Gradient Boosting from scratch*. ML Review. https://n9.cl/ibshs

Grupo de Salud Mental del Programa de Actividades de Prevención y Promoción de la Salud de la Sociedad Española de Medicina de Familia y Comunitaria. (2003). Violencia doméstica. https://n9.cl/6su47

La Moncloa. (2020). *Las llamadas al 016 aumentan un 47,3% en la primera quincena de abril en comparación con el mismo periodo de 2019*. https://n9.cl/grt5m

Larranaga, P., Inza, I., y Moujahid, A. (2019). "Tema 8. Redes Neuronales". Universidad del País Vasco. https://n9.cl/an59m

López-Ossorio, J. J., González, J. L. y Andrés, A. Eficacia predictiva de la valoración policial del riesgo de la violencia de género. *Psychosocial Intervention*, *25*(1), 1-7. https://n9.cl/ga3y

Mammone, A., Turchi, M. y Cristianini, N. (2009). Support vector machines. *Wires: Wiley's Interdisciplinary Reviews in Computational Statistics*, *1*(3), 283-289. https://n9.cl/iq9h0

Matich, D. J. (2001). "Redes Neuronales: Conceptos Básicos y Aplicaciones". Universidad Tecnológica Nacional – Facultad Regional Rosario. https://n9.cl/69n6

Nagpal, A. (2017, 17 octubre). *Decision Tree Ensembles - Bagging and Boosting*. Towards Data Science. https://n9.cl/t5w8

Organización de las Naciones Unidas. (1993). Declaración sobre la eliminación de la violencia contra la mujer. https://n9.cl/fcjx

Organización Mundial de la Salud. (2013). Estimaciones mundiales y regionales de la violencia contra la mujer: prevalencia y efectos de la violencia conyugal y de la violencia sexual no conyugal en la salud. https://n9.cl/n96ix

Organización Mundial de la Salud. (2014). Informe sobre la situación mundial de la prevención de la violencia 2014. https://n9.cl/lwdoe

Organización Mundial de la Salud. (2021, 9 marzo). *La violencia contra la mujer es omnipresente y devastadora: la sufren una de cada tres mujeres*. https://n9.cl/i92f

Pinedo, M. (2021, 2 septiembre). *Matemáticas e inteligencia artificial contra el maltrato machista*. El País. https://n9.cl/bk1e7

Portela, J. (2021). "Técnicas de Machine Learning". Universidad Complutense de Madrid – Facultad de Estudios Estadísticos.

Rodríguez, V. (2018, 17 octubre). *Decision trees / Árboles de decisión para clasificar en Python*. Vicente Rodríguez blog. https://n9.cl/v7ol0

Ruiz, R. (2014). *Análisis de vibraciones mecánicas mediante un clasificador basado en SVM para el mantenimiento predictivo de máquinas: aplicación en una cosechadora agrícola*. (Trabajo de Fin de Máster). Universidad de Valladolid – Escuela Técnica Superior de Ingenieros de Telecomunicación. https://n9.cl/ui3e6

Selección de variables explicativas en la regresión. (2007, 26 octubre). Análisis y comunicación de datos cuantitativos. https://n9.cl/09yo0

Shalev-Shwartz, S. y Ben-David, S. (2014). *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press. ISBN: 978-1-10-705713-5. https://n9.cl/tgzmc

Yiu, T. (2019, 12 junio). *Understanding Random Forest*. Towards Data Science. https://n9.cl/ulf2u

Anexos:

I. Descripción, tratamiento previo y depuración de los datos:

Inicialmente, para la construcción del conjunto de datos objeto de estudio se escogieron 24 variables cuyo nombre y descripción puede verse en la Tabla 39.

Tabla 39. Descripción de las variables seleccionadas manualmente.

Variable	Descripción
edad	Edad de la entrevistada.
nacionalidad	Nacionalidad de la entrevistada: • 1 (La nacionalidad española) • 2 (La nacionalidad española y otra) • 3 (Solo otra nacionalidad)
CCAA	Comunidad autónoma a la que pertenece la entrevistada: • 1 (Andalucía) • 2 (Aragón) • 3 (Asturias) • 4 (Baleares) • 5 (Canarias) • 6 (Cantabria) • 7 (Castilla-La Mancha) • 8 (Castilla y León) • 9 (Cataluña) • 10 (Comunidad Valenciana) • 11 (Extremadura) • 12 (Galicia) • 13 (Madrid) • 14 (Murcia) • 15 (Navarra) • 16 (País Vasco) • 17 (La Rioja) • 18 (Ceuta) • 19 (Melilla)
tamuni	Tamaño del municipio en el que vive la entrevistada: • 1 (Menos o igual a 2.000 habitantes) • 2 (2.001 a 10.000 habitantes) • 3 (10.001 a 50.000 habitantes) • 4 (50.001 a 100.000 habitantes) • 5 (100.001 a 400.000 habitantes) • 6 (400.001 a 1.000.000 habitantes) • 7 (Más de 1.000.000 habitantes)
estudios	Estudios de la entrevistada: • 1 (Sin estudios) • 2 (Primaria) • 3 (Secundaria 1ª etapa) • 4 (Secundaria 2ª etapa) • 5 (F.P.) • 6 (Superiores) • 7 (Otros) • 8 (N.S.) • 9 (N.C.)
sitlab	Situación laboral actual de la entrevistada: • 1 (Trabaja) • 2 (Trabaja o colabora de manera habitual en el negocio familiar) • 3 (Jubilada o pensionista (anteriormente ha trabajado)) • 4 (Pensionista (anteriormente no ha trabajado)) • 5 (Parada y ha trabajado antes) • 6 (Parada y busca su primer empleo) • 7 (Estudiante) • 8 (Trabajo doméstico no remunerado) • 9 (Otra situación)
ingresos	Fuente principal de ingresos personales de la entrevistada: • 1 (No tiene ingresos (otros) personales) • 2 (Salario del trabajo por cuenta ajena) • 3 (Salario del trabajo por cuenta propia) • 4 (Pensión por jubilación) • 5 (Prestación de desempleo) • 6 (Subsidio (ayudas sociales)) • 7 (Pensión de viudedad) • 8 (Pensión compensatoria) • 9 (Beca) • 10 (Rentas, ahorro (viviendas, tierras, acciones, etc.)) • 11 (Ayuda/asignación de su padre/madre) • 12 (Ayuda/asignación de pareja) • 13 (Ayuda/asignación de la expareja) • 14 (Ayuda/asignación de otra persona) • 15 (Otro no contemplado anteriormente) • 99 (N.C.)
hijos	Si la entrevistada tiene hijos/as: • 1 (Sí) • 2 (No) • 9 (N.C.)
discap	Si la entrevistada tiene un certificado de discapacidad con grado igual o superior al 33%: • 1 (Sí) • 2 (No) • 9 (N.C.)

Variable	Descripción
salud12	Estado de salud en el último año de la entrevistada: • 1 (Muy bueno) • 2 (Bueno) • 3 (Regular) • 4 (Malo) • 5 (Muy malo) • 9 (N.C.)
visita	Si la entrevistada ha visitado algún psicólogo/a, psicoterapeuta o psiquiatra: • 1 (Sí) • 2 (No) • 9 (N.C.)
tenpar	Si la entrevistada tiene pareja actual: • 1 (Sí) • 2 (No) • 9 (N.C.)
numpar	Número de parejas que ha tenido la entrevistada a lo largo de su vida, incluida la actual. En el caso de que se indique 99 significa que no se quiere contestar a la pregunta.
sexopar	Sexo de las parejas que ha tenido la entrevistada: • 1 (Solo hombres) • 2 (Solo mujeres) • 3 (Tanto hombres como mujeres) • 9 (N.C.)
miedoparact	Frecuencia con la que la entrevistada ha tenido miedo de su pareja actual: • 1 (Continuamente) • 2 (Muchas veces) • 3 (Algunas veces) • 4 (Nunca) • 9 (N.C.)
miedoparex	Frecuencia con la que la entrevistada ha tenido miedo de su/s pareja/s anterior/es: • 1 (Continuamente) • 2 (Muchas veces) • 3 (Algunas veces) • 4 (Nunca) • 9 (N.C.)
controlparact	Si alguna vez la entrevistada ha sufrido control de actividades por parte de su pareja actual: • 0 (No) • 1 (Sí)
controlparex	Si alguna vez la entrevistada ha sufrido control de actividades por parte de su/s pareja/s anterior/es: • 0 (No) • 1 (Sí)
hablar	Si la entrevistada tiene una persona cercana con la que pueda hablar con plena confianza sobre los problemas en sus relaciones personales: • 1 (Sí) • 2 (No) • 9 (N.C.)
frecreunion	Frecuencia con la que se reúne la entrevistada con sus parientes, familiares o amistades con quienes no convive: • 1 (Todos o casi todos los días) • 2 (Varias veces a la semana) • 3 (Varias veces al mes) • 4 (Una vez al mes) • 5 (Varias veces al año) • 6 (Una vez al año) • 7 (Menos de una vez al año) • 8 (Nunca) • 9 (N.C.)
acoger	Si la entrevistada tiene amistades o familiares/parientes con quienes podría irse un par de días si no pudiera estar en su casa por alguna razón: • 1 (Sí) • 2 (No) • 9 (N.C.)
suicidio	Si la entrevistada ha pensado alguna vez en terminar con su vida: • 1 (Sí) • 2 (No) • 9 (N.C.)
religion	Cómo se define la entrevistada en materia religiosa: • 1 (Católica) • 2 (Creyente de otra religión) • 3 (No creyente) • 4 (Atea) • 5 (Agnóstica) • 9 (N.C.)
usointernet	Si la entrevistada ha usado alguna vez Internet: • 1 (Sí) • 2 (No) • 3 (N.C.)

Tras la creación del conjunto de datos fue conveniente llevar a cabo pequeñas modificaciones sobre ciertas variables para una fácil interpretación de las mismas. Estas transformaciones fueron las que muestran en la Tabla 40.

Tabla 40. Descripción de las variables modificadas.

Variables	Descripción
nacionalidad	Si la entrevistada tiene la nacionalidad española: • 0 (No) • 1 (Sí)
hijos	Si la entrevistada tiene hijos/as: • 0 (No) • 1 (Sí) • 9 (N.C.)
discap	Si la entrevistada tiene un certificado de discapacidad con grado igual o superior al 33%: • 0 (No) • 1 (Sí) • 9 (N.C.)
visita	Si la entrevistada ha visitado algún psicólogo/a, psicoterapeuta o psiquiatra: • 0 (No) • 1 (Sí) • 9 (N.C.)
tenpar	Si la entrevista tiene pareja actual: • 0 (No) • 1 (Sí)
suicidio	Si la entrevistada ha pensado alguna vez en terminar con su vida: • 0 (No) • 1 (Sí) • 9 (N.C.)

Variables	Descripción
acoger	Si la persona tiene amistades o familiares/parientes con quienes podría irse un par de días si no pudiera estar en su casa por alguna razón: • 0 (No) • 1 (Sí) • 9 (N.C.)
hablar	Si la entrevistada tiene una persona cercana con la que pueda hablar con plena confianza sobre los problemas en sus relaciones personales: • 0 (No) • 1 (Sí) • 9 (N.C.)
usointernet	Si la entrevistada ha usado alguna vez Internet: • 0 (No) • 1 (Sí) • 9 (N.C.)
control	Si alguna vez la entrevistada ha sufrido control de actividades por parte de su pareja actual y/o pasada/s: • 0 (No) • 1 (Sí)

La variable "*nacionalidad*" se transformó para que presentase valor 0 si no se tenía la nacionalidad española y 1 en el caso de que sí que se tuviese e incluso se tuviese ésta y otra. Tras esto, se cambiaron los valores de "*hijos*" por 0 en el caso de que no se tuviesen, 1 en el caso de que sí y 9 en el caso de que no se hubiese contestado a la pregunta. Para las variables "*discap*", "*visita*", "*tenpar*", "*hablar*", "*acoger*", "*suicidio*" y "*usointernet*", se hizo exactamente lo mismo. Finalmente, se creó la variable "*control*" que juntó las respuestas de si alguna vez se había sufrido control de actividades por parte de alguna pareja, fuese tanto la actual, si se tenía, como la/s pasada/s ("*controlparact*" y "*controlparex*" (visibles en la Tabla 39 del Anexo I), eliminándose ambas tras la creación de la nueva variable), donde 0 era que no y 1 que sí.

Para saber qué categorías debían de ser reagrupadas en ciertas variables de clase, bastó con ver su número de ocurrencias y el porcentaje que suponía del total, tal y como se recoge en la Tabla 41. Mientras que, para saber qué variables contenían categorías de más, siendo una de ellas la que contenía datos ausentes, bastó con fijarse en la columna del nivel.

Tabla 41. Número de ocurrencias y porcentaje de las variables de clase.

Variable	Nivel	Núm. ocurrencias	%
CCAA	1	826	9,01
CCAA	9	784	8,55
CCAA	13	753	8,22
CCAA	10	652	7,11
CCAA	12	527	5,75
CCAA	8	494	5,39
CCAA	16	479	5,23
CCAA	7	467	5,10
CCAA	14	439	4,79
CCAA	2	428	4,67
CCAA	11	428	4,67
CCAA	3	425	4,64
CCAA	5	414	4,52
CCAA	4	413	4,51
CCAA	6	382	4,17
CCAA	18	365	3,98
CCAA	17	356	3,88
CCAA	19	331	3,61

Variable	Nivel	Núm. ocurrencias	%
ingresos	5	375	4,09
ingresos	6	238	2,60
ingresos	12	167	1,82
ingresos	11	126	1,37
ingresos	15	82	0,89
ingresos	10	62	0,68
ingresos	8	57	0,62
ingresos	99	46	0,50
ingresos	9	30	0,33
ingresos	13	27	0,29
ingresos	14	18	0,20
nacionalidad	1	8417	91,84
nacionalidad	0	748	8,16
religion	1	6086	66,40
religion	3	888	9,69
religion	4	738	8,05
religion	2	722	7,88
religion	5	496	5,41

Variable	Nivel	Núm. ocurrencias	%
CCAA	15	202	2,20
acoger	1	8465	92,36
acoger	0	659	7,19
acoger	9	41	0,45
control	0	6690	73,00
control	1	2475	27,00
discap	0	8604	93,88
discap	1	553	6,03
discap	9	8	0,09
estudios	6	2109	23,01
estudios	3	1895	20,68
estudios	5	1669	18,21
estudios	2	1661	18,12
estudios	4	1187	12,95
estudios	1	628	6,85
estudios	9	11	0,12
estudios	8	4	0,04
estudios	7	1	0,01
frecreunion	1	3459	37,74
frecreunion	2	3156	34,44
frecreunion	3	1538	16,78
frecreunion	5	418	4,56
frecreunion	4	350	3,82
frecreunion	6	104	1,13
frecreunion	7	81	0,88
frecreunion	8	53	0,58
frecreunion	9	6	0,07
hablar	1	8596	93,79
hablar	0	553	6,03
hablar	9	16	0,17
hijos	1	6864	74,89
hijos	0	2299	25,08
hijos	9	2	0,02
ingresos	2	3485	38,03
ingresos	1	2008	21,91
ingresos	4	1267	13,82
ingresos	3	626	6,83
ingresos	7	551	6,01

Variable	Nivel	Núm. ocurrencias	%
religion	9	235	2,56
salud12	2	4455	48,61
salud12	3	2214	24,16
salud12	1	1830	19,97
salud12	4	517	5,64
salud12	5	147	1,60
salud12	9	2	0,02
sexopar	1	9016	98,37
sexopar	3	90	0,98
sexopar	2	55	0,60
sexopar	9	4	0,04
sitlab	1	4080	44,52
sitlab	3	1525	16,64
sitlab	8	1267	13,82
sitlab	5	1246	13,60
sitlab	4	482	5,26
sitlab	7	433	4,72
sitlab	6	57	0,62
sitlab	9	42	0,46
sitlab	2	33	0,36
suicidio	0	8299	90,55
suicidio	1	838	9,14
suicidio	9	28	0,31
tamuni	3	2319	25,30
tamuni	5	1996	21,78
tamuni	4	1627	17,75
tamuni	2	1336	14,58
tamuni	6	747	8,15
tamuni	1	572	6,24
tamuni	7	568	6,20
tenpar	1	6543	71,39
tenpar	0	2622	28,61
usointernet	1	7347	80,16
usointernet	0	1814	19,79
usointernet	9	4	0,04
visita	0	8225	89,74
visita	1	937	10,22
visita	9	3	0,03

Exactamente, hubo que reagrupar las categorías:

- 2 (Aragón), 3 (Asturias), 4 (Baleares), 5 (Canarias), 6 (Cantabria), 11 (Extremadura), 14 (Murcia), 15 (Navarra), 17 (La Rioja), 18 (Ceuta) y 19 (Melilla) de "*CCAA*";
- 7 (Otros) de "*estudios*";

- 4 (Una vez al mes), 5 (Varias veces al año), 6 (Una vez al año), 7 (Menos de una vez al año) y 8 (Nunca) de "*frecreunion*";
- 5 (Prestación de desempleo), 6 (Subsidio (ayudas sociales)), 8 (Pensión compensatoria), 9 (Beca), 10 (Rentas, ahorro (viviendas, tierras, acciones, etc.)), 11 (Ayuda/asignación de su padre/madre), 12 (Ayuda/asignación de pareja), 13 (Ayuda/asignación de la expareja), 14 (Ayuda/asignación de otra persona) y 15 (Otro no contemplado anteriormente) de "*ingresos*";
- 5 (Muy malo) de "salud12";
- 2 (Solo mujeres) y 3 (Tanto hombres como mujeres) de "*sexopar*" y,
- 2 (Trabaja o colabora de manera habitual en el negocio familiar), 6 (Parada y busca su primer empleo), 7 (Estudiante) y 9 (Otra situación) de "*sitlab*".

Y, las categorías que hubieron de reemplazarse por "*_MISSING_*" fueron:

- la 9 (N.C.) de "*acoger*",
- la 9 (N.C.) de "*discap*",
- la 8 (N.S.) y 9 (N.C.) de "*estudios*",
- la 9 (N.C.) de "*frecreunion*",
- la 9 (N.C.) de "*hablar*",
- la 9 (N.C.) de "*hijos*",
- la 99 (N.C.) de "*ingresos*",
- la 9 (N.C.) de "*religion*",
- la 9 (N.C.) de "*salud12*",
- la 9 (N.C.) de "*sexopar*",
- la 9 (N.C.) de "*suicidio*",
- la 9 (N.C.) de "*usointernet*" y
- la 9 (N.C.) de "*visita*".

Tras los oportunos reagrupamientos todas las variables disponían de categorías que contaban con más del 5% de los datos y, además, todas aquellas categorías que habían sido detectadas como missing ya se mostraban como tal, pudiendo verse esto en la Tabla 42.

Tabla 42. Número de ocurrencias y porcentaje de las variables de clase tras el oportuno reagrupamiento.

Variable	Nivel	Núm. ocurrencias	%
REP_CCAA	9	1625	17,73
REP_CCAA	12	1334	14,56
REP_CCAA	5	1110	12,11
REP_CCAA	10	1091	11,90
REP_CCAA	15	1037	11,31
REP_CCAA	8	922	10,06
REP_CCAA	1	826	9,01
REP_CCAA	13	753	8,22
REP_CCAA	7	467	5,10
REP_acoger	1	8465	92,36
REP_acoger	0	659	7,19

Variable	Nivel	Núm. ocurrencias	%
REP_religion	1	6086	66,40
REP_religion	3	888	9,69
REP_religion	4	738	8,05
REP_religion	2	722	7,88
REP_religion	5	496	5,41
REP_religion	_MISSING_	235	2,56
REP_salud12	1	6285	68,58
REP_salud12	3	2878	31,40
REP_salud12	_MISSING_	2	0,02
REP_sexopar	1	9016	98,37
REP_sexopar	2	145	1,58

Variable	Nivel	Núm. ocurrencias	%
REP_acoger	_MISSING_	41	0,45
REP_discap	0	8604	93,88
REP_discap	1	553	6,03
REP_discap	_MISSING_	8	0,09
REP_estudios	6	2109	23,01
REP_estudios	3	1895	20,68
REP_estudios	5	1669	18,21
REP_estudios	2	1661	18,12
REP_estudios	4	1187	12,95
REP_estudios	1	629	6,86
REP_estudios	_MISSING_	15	0,16
REP_frecreunion	1	3459	37,74
REP_frecreunion	2	3156	34,44
REP_frecreunion	3	1538	16,78
REP_frecreunion	4	1006	10,98
REP_frecreunion	_MISSING_	6	0,07
REP_hablar	1	8596	93,79
REP_hablar	0	553	6,03
REP_hablar	_MISSING_	16	0,17
REP_hijos	1	6864	74,89
REP_hijos	0	2299	25,08
REP_hijos	_MISSING_	2	0,02
REP_ingresos	2	3485	38,03
REP_ingresos	1	2008	21,91
REP_ingresos	7	1875	20,46
REP_ingresos	5	1125	12,27
REP_ingresos	3	626	6,83
REP_ingresos	_MISSING_	46	0,50

Variable	Nivel	Núm. ocurrencias	%
REP_sexopar	_MISSING_	4	0,04
REP_sitlab	1	4113	44,88
REP_sitlab	3	2007	21,90
REP_sitlab	7	1742	19,01
REP_sitlab	5	1303	14,22
REP_suicidio	0	8299	90,55
REP_suicidio	1	838	9,14
REP_suicidio	_MISSING_	28	0,31
REP_usointernet	1	7347	80,16
REP_usointernet	0	1814	19,79
REP_usointernet	_MISSING_	4	0,04
REP_visita	0	8225	89,74
REP_visita	1	937	10,22
REP_visita	_MISSING_	3	0,03
control	0	6690	73,00
control	1	2475	27,00
nacionalidad	1	8417	91,84
nacionalidad	0	748	8,16
tamuni	3	2319	25,30
tamuni	5	1996	21,78
tamuni	4	1627	17,75
tamuni	2	1336	14,58
tamuni	6	747	8,15
tamuni	1	572	6,24
tamuni	7	568	6,20
tenpar	1	6543	71,39

Señalar que la única categoría que no alcanzó exactamente el 5% de los datos fue la 2 de "*REP_sexopar*" (la cual recogía las respuestas de "*Solo mujeres*" y "*Tanto hombres como mujeres*"). Aun así, como se disponían de muchas observaciones, esto no supuso problema alguno.

Para verificar que los cambios se habían hecho correctamente, bastó con fijarse en si las nuevas variables transformadas tenían delante el prefijo "*REP_NombreVariable*", como bien ocurría.

II. Modelización:

Tras haberse realizado validación cruzada repetida sobre la rejilla diseñada en gradient boosting se tuvo que:

Tabla 43. Resultados rejilla gradient boosting.

shrinkage	n.minobsinnode	n.trees	Tasa de aciertos
0,001	10	1.000	0,8615385

shrinkage	n.minobsinnode	n.trees	Tasa de aciertos
0,001	10	3.000	0,8768145
0,001	10	5.000	0,8777964
0,001	10	7.000	0,8797609
0,001	15	1.000	0,8615385
0,001	15	3.000	0,8768145
0,001	15	5.000	0,8777964
0,001	15	7.000	0,8796518
0,001	20	1.000	0,8615385
0,001	20	3.000	0,8768145
0,001	20	5.000	0,8777964
0,001	20	7.000	0,8795427
0,01	10	1.000	0,8773602
0,01	10	3.000	0,8797606
0,01	10	5.000	0,8794329
0,01	10	7.000	0,8794331
0,01	15	1.000	0,8779057
0,01	15	3.000	0,8798695
0,01	15	5.000	0,8801969
0,01	15	7.000	0,8795424
0,01	20	1.000	0,8773602
0,01	20	3.000	0,8794332
0,01	20	5.000	0,8807425
0,01	20	7.000	0,8798696
0,03	10	1.000	0,8799788
0,03	10	3.000	0,8792148
0,03	10	5.000	0,8762684
0,03	10	7.000	0,8744137
0,03	15	1.000	0,8805243
0,03	15	3.000	0,8784512
0,03	15	5.000	0,8760505
0,03	15	7.000	0,8735408
0,03	20	1.000	0,8799787
0,03	20	3.000	0,8786694
0,03	20	5.000	0,8760505
0,03	20	7.000	0,8757234
0,05	10	1.000	0,8800877
0,05	10	3.000	0,8756138
0,05	10	5.000	0,8738682
0,05	10	7.000	0,8732135
0,05	15	1.000	0,8798695
0,05	15	3.000	0,8755050
0,05	15	5.000	0,8743047
0,05	15	7.000	0,8731045
0,05	20	1.000	0,8805243
0,05	20	3.000	0,8750686
0,05	20	5.000	0,8749598
0,05	20	7.000	0,8738686
0,1	10	1.000	0,8781235
0,1	10	3.000	0,8726679
0,1	10	5.000	0,8726682
0,1	10	7.000	0,8719045
0,1	15	1.000	0,8784510
0,1	15	3.000	0,8727771
0,1	15	5.000	0,8719046
0,1	15	7.000	0,8713592
0,1	20	1.000	0,8772509
0,1	20	3.000	0,8738685
0,1	20	5.000	0,8715773
0,1	20	7.000	0,8711411
0,2	10	1.000	0,8736498
0,2	10	3.000	0,8728868
0,2	10	5.000	0,8689586
0,2	10	7.000	0,8679762
0,2	15	1.000	0,8725587
0,2	15	3.000	0,8711411
0,2	15	5.000	0,8701587

shrinkage	n.minobsinnode	n.trees	Tasa de aciertos
0,2	15	7.000	0,8680854
0,2	20	1.000	0,8743050
0,2	20	3.000	0,8711411
0,2	20	5.000	0,8681951
0,2	20	7.000	0,8683040

Las mayores tasas de aciertos se consiguieron con las combinaciones:

Tabla 44. Mayores tasas de aciertos rejilla gradient boosting.

shrinkage	n.minobsinnode	n.trees	Tasa de aciertos
0,01	20	5000	0,8807425
0,03	15	1000	0,8805243
0,05	20	1000	0,8805243
0,01	15	5000	0,8801969
0,05	10	1000	0,8800877

Por lo que, tal y como puede observarse en la Tabla 44, se obtenían tasas más altas de acierto con unos "*shrinkage*" bajos, entre 0,01 y 0,05, y unos "*n.trees*" entre 1.000 y 5.000.

Mientras que, tras haberse realizado validación cruzada repetida sobre la rejilla diseñada en extreme gradient boosting se tuvo que:

Tabla 45. Resultados rejilla extreme gradient boosting.

eta	min_child_weight	nrounds	Tasa de aciertos
0,001	10	100	0,8715763
0,001	10	500	0,8703759
0,001	10	1.000	0,8728857
0,001	10	3.000	0,8748503
0,001	10	5.000	0,8788875
0,001	10	7.000	0,8779054
0,001	15	100	0,8715763
0,001	15	500	0,8703759
0,001	15	1.000	0,8728857
0,001	15	3.000	0,8748503
0,001	15	5.000	0,8793239
0,001	15	7.000	0,8793237
0,001	20	100	0,8715763
0,001	20	500	0,8703759
0,001	20	1.000	0,8728857
0,001	20	3.000	0,8748503
0,001	20	5.000	0,8785600
0,001	20	7.000	0,8797604
0,01	10	100	0,8728857
0,01	10	500	0,8788875
0,01	10	1.000	0,8793239
0,01	10	3.000	0,8805242
0,01	10	5.000	0,8816153
0,01	10	7.000	0,8811789
0,01	15	100	0,8728857
0,01	15	500	0,8792148
0,01	15	1.000	0,8787780
0,01	15	3.000	0,8815065
0,01	15	5.000	0,8808518
0,01	15	7.000	0,8792151
0,01	20	100	0,8728857
0,01	20	500	0,8784509
0,01	20	1.000	0,8793234
0,01	20	3.000	0,8816155
0,01	20	5.000	0,8805244
0,01	20	7.000	0,8796515
0,03	10	100	0,8748503

eta	min_child_weight	nrounds	Tasa de aciertos
0,03	10	500	0,8784509
0,03	10	1.000	0,8800878
0,03	10	3.000	0,8818337
0,03	10	5.000	0,8806335
0,03	10	7.000	0,8796513
0,03	15	100	0,8748503
0,03	15	500	0,8783418
0,03	15	1.000	0,8822703
0,03	15	3.000	0,8789968
0,03	15	5.000	0,8798697
0,03	15	7.000	0,8777964
0,03	20	100	0,8748503
0,03	20	500	0,8793240
0,03	20	1.000	0,8818337
0,03	20	3.000	0,8807426
0,03	20	5.000	0,8792148
0,03	20	7.000	0,8784511
0,05	10	100	0,8788875
0,05	10	500	0,8816150
0,05	10	1.000	0,8812881
0,05	10	3.000	0,8805242
0,05	10	5.000	0,8806332
0,05	10	7.000	0,8791056
0,05	15	100	0,8787783
0,05	15	500	0,8818338
0,05	15	1.000	0,8809610
0,05	15	3.000	0,8793238
0,05	15	5.000	0,8793238
0,05	15	7.000	0,8779054
0,05	20	100	0,8788873
0,05	20	500	0,8811790
0,05	20	1.000	0,8805245
0,05	20	3.000	0,8784511
0,05	20	5.000	0,8777964
0,05	20	7.000	0,8769233
0,1	10	100	0,8800874
0,1	10	500	0,8813974
0,1	10	1.000	0,8804151
0,1	10	3.000	0,8792149
0,1	10	5.000	0,8784511
0,1	10	7.000	0,8775779
0,1	15	100	0,8798691
0,1	15	500	0,8809610
0,1	15	1.000	0,8804152
0,1	15	3.000	0,8777962
0,1	15	5.000	0,8780147
0,1	15	7.000	0,8775776
0,1	20	100	0,8791052
0,1	20	500	0,8798697
0,1	20	1.000	0,8796516
0,1	20	3.000	0,8779057
0,1	20	5.000	0,8771418
0,1	20	7.000	0,8765963

Las mayores tasas de aciertos se consiguieron con las combinaciones:

Tabla 46. Mayores tasas de aciertos rejilla extreme gradient boosting.

eta	min_child_weight	nrounds	Tasa de aciertos
0,03	15	1.000	0,8822703
0,05	15	500	0,8818338
0,03	10	3.000	0,8818337
0,03	20	1.000	0,8818337
0,01	20	3.000	0,8816155
0,01	10	5.000	0,8816153

eta	min_child_weight	nrounds	Tasa de aciertos
0,05	10	500	0,8816150
0,01	15	3.000	0,8815065
0,1	10	500	0,8813974
0,05	10	1.000	0,8812881
0,05	20	500	0,8811790
0,01	10	7.000	0,8811789

Así pues, gracias a la Tabla 46 se supo que las tasas de aciertos más altas se conseguían con unos "*eta*" entre 0,01 y 0,05, que son los que se especificaron para crear los posibles modelos candidatos a ganadores en extreme gradient boosting.

III. Código SAS 9.4:

III.1. Selección de variables:

```
/*Creación de la librería*/
libname discoc 'C:\Users\alexa\Desktop\MdD\2do
cuatrimestre\TFM\VIOLENCIAMUJER';
data violenciagenero_train;
/*Lee el archivo "violenciagenero_train" de "discoc" y cárgalas en
"violenciagenero_train" (copia temporal)*/
set discoc.violenciagenero_train;run;

/*Información sobre el dataset*/
proc contents data=violenciagenero_train out=salida;run;
proc print data=salida;run;

/*Listado de las variables en la consola*/
data;
set salida;
put name @@;
run;
/*IMP_REP_REP_numpar IMP_REP_acoger IMP_REP_discap IMP_REP_estudios
IMP_REP_frecreunion IMP_REP_hablar IMP_REP_hijos IMP_REP_ingresos
IMP_REP_religion IMP_REP_salud12 IMP_REP_sexopar IMP_REP_suicidio
IMP_REP_usointernet IMP_REP_visita REP_CCAA REP_edad REP_sitlab aleat
control miedo nacionalidad numMissing tamuni tenpar*/

/**********************Definimos el rol y tipo de todas las
variables**********************/
PROC DMDB DATA=violenciagenero_train dmdbcat=cataprueba;
/*Variable objetivo*/
target miedo;
/*Variable de intervalo*/
var aleat IMP_REP_REP_numpar REP_edad;
/*Variable de clase (se incluye la dependiente)*/
class IMP_REP_acoger IMP_REP_discap IMP_REP_estudios
IMP_REP_frecreunion IMP_REP_hablar numMissing IMP_REP_hijos
IMP_REP_ingresos IMP_REP_religion IMP_REP_salud12 IMP_REP_sexopar
IMP_REP_suicidio IMP_REP_usointernet IMP_REP_visita REP_CCAA
REP_sitlab control miedo nacionalidad tamuni tenpar; run;
proc print data=violenciagenero_train;run;

/**********************Selección de variables**********************/
/*La macro "randomselectlog" realiza un método stepwise repetidas
veces con diferentes archivos train.
```

```
La salida incluye una tabla de frecuencias de los modelos que aparecen
seleccionados en los diferentes archivos train.

Los modelos que salen más veces son posibles candidatos a probar con
validación cruzada repetida.

listclass = lista de variables de clase. Atención, en esta lista solo
poner variables que se vayan a usar (bien como efectos principales o
interacciones).
vardepen = variable dependiente
modelo = modelo
sinicio = semilla inicio
sfinal = semilla final
fracciontrain = fracción de datos train
directorio = directorio para archivos basura

El archivo que contiene los efectos se llama: salesfec.
Se incluye en el log el listado para poder copiar y pegar (a veces la
ventana output está limitada).*/

%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;

%randomselectlog(data=violenciagenero_train,listclass=IMP_REP_acoger
IMP_REP_discap IMP_REP_estudios IMP_REP_frecreunion IMP_REP_hablar
IMP_REP_hijos IMP_REP_ingresos IMP_REP_religion IMP_REP_salud12
IMP_REP_sexopar IMP_REP_suicidio IMP_REP_usointernet IMP_REP_visita
REP_CCAA REP_sitlab control nacionalidad tamuni tenpar numMissing,
vardepen=miedo,
modelo=IMP_REP_acoger IMP_REP_discap IMP_REP_estudios
IMP_REP_frecreunion IMP_REP_hablar IMP_REP_hijos IMP_REP_ingresos
IMP_REP_religion IMP_REP_salud12 IMP_REP_sexopar IMP_REP_suicidio
IMP_REP_usointernet IMP_REP_visita REP_CCAA REP_sitlab control
nacionalidad tamuni tenpar aleat IMP_REP_REP_numpar REP_edad
numMissing,
sinicio=12345,sfinal=12445,fracciontrain=0.8,directorio=C:\Users\alexa
\Desktop\MdD\2do cuatrimestre\TFM\VIOLENCIAMUJER);

/**********************Validación cruzada repetida logística para
variables dependientes binarias**********************/
/*

archivo = archivo de datos
vardepen = variable dependiente binaria
categor = lista de variables independientes categóricas
conti = lista de variables independientes continuas Y TODAS LAS
INTERACCIONES
ngrupos = número de grupos validación cruzada repetida
sinicio = semilla inicial para repetición
sfinal = semilla final para repetición
objetivo = tasafallos, sensi, especif, porcenVN, porcenFN, porcenVP,
porcenFP, precision, tasaciertos
El archivo final se llama final.

La variable media es la media del objetivo en todas las pruebas de
validación cruzada repetida (habitualmente tasa de fallos).*/

%macro
cruzadalogistica(archivo=,vardepen=,conti=,categor=,ngrupos=,sinicio=,
sfinal=,objetivo=tasafallos);
title ' ';
data final;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 tres;set dos;if grupo ne &exclu then vardep=&vardepen;
            proc logistic data=tres noprint;/*<<<<<******SE PUEDE
QUITAR EL NOPRINT */
            %if (&categor ne) %then %do;class &categor;model
vardep=&conti &categor ;%end;
            %else %do;model vardep=&conti;%end;
            output out=sal p=predi;run;
            data sal2;set sal;pro=1-predi;if pro>0.5 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;
            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 print data=final;run;
%mend;

%cruzadalogistica
(archivo=violenciagenero_train,vardepen=miedo,
conti=IMP_REP_REP_numpar REP_edad,
categor=IMP_REP_acoger IMP_REP_estudios IMP_REP_frecreunion
IMP_REP_ingresos IMP_REP_religion IMP_REP_suicidio IMP_REP_usointernet
IMP_REP_visita REP_sitlab control tenpar,
ngrupos=10,sinicio=12345,sfinal=12365);
```

```
data final1;set final;modelo=1;
/*Conjunto 1: 13 variables*/

%cruzadalogistica
(archivo=violenciagenero_train,vardepen=miedo,
conti=IMP_REP_REP_numpar,
categor=IMP_REP_hijos IMP_REP_ingresos IMP_REP_salud12
IMP_REP_suicidio IMP_REP_visita control tenpar,
ngrupos=10,sinicio=12345,sfinal=12365);
data final2;set final;modelo=2;
/*Conjunto 2: 8 variables*/

%cruzadalogistica
(archivo=violenciagenero_train,vardepen=miedo,
conti=IMP_REP_REP_numpar,
categor=IMP_REP_ingresos IMP_REP_suicidio IMP_REP_visita control
tenpar,
ngrupos=10,sinicio=12345,sfinal=12365);
data final3;set final;modelo=3;
/*Conjunto 3: 6 variables*/

%cruzadalogistica
(archivo=violenciagenero_train,vardepen=miedo,
conti=IMP_REP_REP_numpar,
categor=IMP_REP_discap IMP_REP_estudios IMP_REP_hijos IMP_REP_ingresos
IMP_REP_suicidio IMP_REP_visita control tenpar,
ngrupos=10,sinicio=12345,sfinal=12365);
data final4;set final;modelo=4;
/*Conjunto 4: 9 variables*/

%cruzadalogistica
(archivo=violenciagenero_train,vardepen=miedo,
conti=IMP_REP_REP_numpar,
categor=IMP_REP_hijos IMP_REP_ingresos IMP_REP_suicidio IMP_REP_visita
control tenpar,
ngrupos=10,sinicio=12345,sfinal=12365);
data final5;set final;modelo=5;
/*Conjunto 5: 7 variables*/

data union;set final1 final2 final3 final4 final5;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;

/*Repetimos validación cruzada repetida variando semilla sobre los
modelos con mayor frecuencia*/
%cruzadalogistica
(archivo=violenciagenero_train,vardepen=miedo,
conti=IMP_REP_REP_numpar,
categor=IMP_REP_hijos IMP_REP_ingresos IMP_REP_salud12
IMP_REP_suicidio IMP_REP_visita control tenpar,
ngrupos=10,sinicio=540,sfinal=560);
data final2;set final;modelo=2;

%cruzadalogistica
(archivo=violenciagenero_train,vardepen=miedo,
conti=IMP_REP_REP_numpar,
categor=IMP_REP_discap IMP_REP_estudios IMP_REP_hijos IMP_REP_ingresos
IMP_REP_suicidio IMP_REP_visita control tenpar,
ngrupos=10,sinicio=540,sfinal=560);
data final4;set final;modelo=4;
```

```
%cruzadalogistica
(archivo=violenciagenero_train,vardepen=miedo,
conti=IMP_REP_REP_numpar,
categor=IMP_REP_hijos IMP_REP_ingresos IMP_REP_suicidio IMP_REP_visita
control tenpar,
ngrupos=10,sinicio=540,sfinal=560);
data final5;set final;modelo=5;

data union;set final2 final4 final5;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;
```

IV. Código RStudio:

IV.1. Librerías:

```
# ********************
# LIBRERÍAS
# ********************
library(sas7bdat)
library(nnet)
library(dummies)
library(MASS)
library(reshape)
library(caret)
library(pROC)
library(randomForest)
library(ggplot2)
library(parallel)
library(doParallel)

cluster <- makeCluster(detectCores() - 1) # number of cores,
convention to leave 1 core for OS
registerDoParallel(cluster) # register the parallel processing

# stopCluster(cluster) # shut down the cluster
# registerDoSEQ(); #  force R to return to single threaded processing
```

IV.2. Generación de los datos:

```
# ********************
# GENERACION DATOS
# ********************
violenciagenero<-read.sas7bdat("C:/Users/alexa/Desktop/MdD/2do
cuatrimestre/TFM/VIOLENCIAMUJER/violenciagenero_train.sas7bdat")
dput(names(violenciagenero))
# c("nacionalidad", "tamuni", "tenpar", "miedo", "control",
# "REP_CCAA", "REP_sitlab", "REP_edad", "numMissing",
# "IMP_REP_acoger", "IMP_REP_discap", "IMP_REP_estudios",
# "IMP_REP_frecreunion", "IMP_REP_hablar", "IMP_REP_hijos",
# "IMP_REP_ingresos", "IMP_REP_religion", "IMP_REP_salud12",
# "IMP_REP_sexopar", "IMP_REP_suicidio", "IMP_REP_usointernet",
# "IMP_REP_visita", "IMP_REP_REP_numpar", "aleat")

# Se eliminan las variables que no han sido seleccionadas ni en el
conjunto 1 ni en el 5.
violenciagenero$nacionalidad<-NULL
violenciagenero$tamuni<-NULL
```

```
violenciagenero$REP_CCAA<-NULL
violenciagenero$numMissing<-NULL
violenciagenero$IMP_REP_discap<-NULL
violenciagenero$IMP_REP_hablar<-NULL
violenciagenero$IMP_REP_salud12<-NULL
violenciagenero$IMP_REP_sexopar<-NULL
violenciagenero$aleat<-NULL
dput(names(violenciagenero))
# c("tenpar", "miedo", "control", "REP_sitlab", "REP_edad",
# "IMP_REP_acoger", "IMP_REP_estudios", "IMP_REP_frecreunion",
# "IMP_REP_hijos", "IMP_REP_ingresos", "IMP_REP_religion",
# "IMP_REP_suicidio", "IMP_REP_usointernet",
# "IMP_REP_visita", "IMP_REP_REP_numpar")

# No pongo la variable objetivo.
continuas<-c("REP_edad", "IMP_REP_REP_numpar")
categoricas<-c("tenpar", "control", "REP_sitlab", "IMP_REP_acoger",
"IMP_REP_estudios", "IMP_REP_frecreunion", "IMP_REP_hijos",
"IMP_REP_ingresos", "IMP_REP_religion", "IMP_REP_suicidio",
"IMP_REP_usointernet", "IMP_REP_visita")

# a) Eliminar las observaciones con missing en alguna variable:
violenciagenero2<-na.omit(violenciagenero,(!is.na(violenciagenero)))

# b) Pasar las categóricas a dummies:
violenciagenero3<- dummy.data.frame(violenciagenero2, categoricas,
sep=".")

# c) Estandarizar las variables continuas:
# Calculo medias y desv. típica de datos y estandarizo (solo las
continuas).
means<-apply(violenciagenero3[,continuas],2,mean)
sds<-sapply(violenciagenero3[,continuas],sd)

# Estandarizo solo las continuas y uno con las categóricas
violenciagenerobis<-scale(violenciagenero3[,continuas], center=means,
scale=sds)
numerocont<-which(colnames(violenciagenero3)%in%continuas)
violenciagenerobis<-cbind(violenciagenerobis,violenciagenero3[,-
numerocont])
# El archivo violenciagenerobis ya está preparado: no hay missings,
las continuas están estandarizadas y las categóricas pasadas a dummies
(salvo la dependiente).
dput(names(violenciagenerobis))
# c("REP_edad", "IMP_REP_REP_numpar", "tenpar.0", "tenpar.1",
# "miedo", "control.0", "control.1", "REP_sitlab.1", "REP_sitlab.3",
# "REP_sitlab.5", "REP_sitlab.7", "IMP_REP_acoger.0",
# "IMP_REP_acoger.1", "IMP_REP_estudios.1", "IMP_REP_estudios.2",
# "IMP_REP_estudios.3", "IMP_REP_estudios.4", "IMP_REP_estudios.5",
# "IMP_REP_estudios.6", "IMP_REP_frecreunion.1",
# "IMP_REP_frecreunion.2","IMP_REP_frecreunion.3",
# "IMP_REP_frecreunion.4", "IMP_REP_hijos.0","IMP_REP_hijos.1",
# "IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
# "IMP_REP_ingresos.5", "IMP_REP_ingresos.7", "IMP_REP_religion.1",
# "IMP_REP_religion.2", "IMP_REP_religion.3", "IMP_REP_religion.4",
# "IMP_REP_religion.5", "IMP_REP_suicidio.0", "IMP_REP_suicidio.1",
# "IMP_REP_usointernet.0", "IMP_REP_usointernet.1",
# "IMP_REP_visita.0", "IMP_REP_visita.1")
```

```
# Transformo los valores "0" y "1" de la variable objetivo a "No" y
"Yes", respectivamente.
violenciagenerobis$miedo<-
ifelse(violenciagenerobis$miedo==1,"Yes","No")
table(violenciagenerobis$miedo)
```

IV.3. Regresión logística:

```
# ********************
# REGRESIÓN LOGÍSTICA
# ********************
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
source("cruzadas avnnet y log binaria.R")

medias1<-cruzadalogistica(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5)

medias1$modelo="Logística1"

union1<-rbind(medias1)

par(cex.axis=0.5)
boxplot(data=union1, tasa~modelo, main="TASA FALLOS")
boxplot(data=union1, auc~modelo, main="AUC")

# ********************
# COEFICIENTES DEL MODELO
# ********************
set.seed(12345)

control<-trainControl(method="none", savePredictions="all",
classProbs=TRUE)

logi<-
train(factor(miedo)~IMP_REP_REP_numpar+tenpar.0+control.0+IMP_REP_hijo
s.0+IMP_REP_ingresos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_i
ngresos.5+IMP_REP_suicidio.1+IMP_REP_visita.1,
data=violenciagenerobis, method="glm", trControl=control)

summary(logi)
```

IV.4. Redes neuronales:

```
# ********************
# REDES NEURONALES
# ********************
# Importante classProbs=TRUE para guardar las probabilidades y definir
la variable de salida con valores alfanuméricos "Yes" y "No".

set.seed(12345)
# Validación cruzada repetida
control<-trainControl(method="repeatedcv", number=4, repeats=5,
savePredictions="all", classProbs=TRUE)
```

```
# ************************************************************
# avNNet: parámetros
# Number of Hidden Units (size, numeric)
# Weight Decay (decay, numeric)
# Bagging (bag, logical)
# ************************************************************
avnnetgrid<-expand.grid(size=c(5,10,15,20,25,30,35),
decay=c(0.001,0.01,0.1), bag=FALSE)

redavnnet<-
train(miedo~IMP_REP_REP_numpar+tenpar.0+control.0+IMP_REP_hijos.0+IMP_
REP_ingresos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.
5+IMP_REP_suicidio.1+IMP_REP_visita.1, data=violenciagenerobis,
method="avNNet", linout=FALSE, maxit=100, trControl=control,
tuneGrid=avnnetgrid, repeats=5)

redavnnet

avnnetgrid2<-expand.grid(size=c(3,7,9,11), decay=c(0.01,0.1),
bag=FALSE)

redavnnet2<-
train(miedo~IMP_REP_REP_numpar+tenpar.0+control.0+IMP_REP_hijos.0+IMP_
REP_ingresos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.
5+IMP_REP_suicidio.1+IMP_REP_visita.1,data=violenciagenerobis,
method="avNNet", linout=FALSE, maxit=100, trControl=control,
tuneGrid=avnnetgrid2, repeats=5)

redavnnet2

# ********************
# MODELOS
# ********************
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
source("cruzadas avnnet y log binaria.R")

medias1<-cruzadaavnnetbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
size=c(3), decay=c(0.1), repeticiones=5, itera=100)

medias1$modelo="avnnet1"

medias2<-cruzadaavnnetbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
size=c(3), decay=c(0.01), repeticiones=5, itera=100)

medias2$modelo="avnnet2"

medias3<-cruzadaavnnetbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
```

```
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
size=c(5), decay=c(0.1), repeticiones=5, itera=100)

medias3$modelo="avnnet3"

medias4<-cruzadaavnnetbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
size=c(7), decay=c(0.01), repeticiones=5, itera=100)

medias4$modelo="avnnet4"

medias5<-cruzadaavnnetbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
size=c(11), decay=c(0.01), repeticiones=5, itera=100)

medias5$modelo="avnnet5"

union1<-rbind(medias1, medias2, medias3, medias4, medias5)

par(cex.axis=0.5)
boxplot(data=union1, tasa~modelo, main="TASA FALLOS")
boxplot(data=union1, auc~modelo, main="AUC")

# ********************
# MODELO GANADOR
# ********************
medias2<-cruzadaavnnetbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
size=c(3), decay=c(0.01), repeticiones=5, itera=100)

medias2$modelo="avnnet2"
```

IV.5. Bagging:

```
# ********************
# BAGGING
# ********************
vardep=c("miedo")

listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5")
```

```
listclass=c("")
paste(listconti,collapse="+")
#"IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudios.2+IMP_REP_estud
ios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_frecreunion.1+IMP_
REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingresos.1+IMP_REP_ing
resos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_REP_religion.1+IMP_R
EP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+IMP_REP_suicidio.0
+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP_REP_REP_numpar+te
npar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitlab.5"

# Para plotear el error OOB a medida que avanzan las iteraciones
# se usa directamente el paquete "randomForest".
set.seed(12345)

rfbis<-
randomForest(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP
_estudios.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+I
MP_REP_frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_R
EP_ingresos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5
+IMP_REP_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_reli
gion.4+IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_e
dad+IMP_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+RE
P_sitlab.5, data=violenciagenerobis, mtry=27, ntree=15000,
sampsize=200, nodesize=15, replace=TRUE)

plot(rfbis$err.rate[,1])

# La función "cruzadarfbin" permite plantear bagging
# (para bagging hay que poner mtry=número de variables independientes)
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
source("cruzada rf binaria.R")

# ********************
# MODELOS
# ********************
medias1<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=10, mtry=27, ntree=3000, replace=TRUE,
sampsize=5957)

medias1$modelo="bagging1"

medias2<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
```

```
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=15, mtry=27, ntree=3000, replace=TRUE,
sampsize=5957)

medias2$modelo="bagging2"

medias3<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=20, mtry=27, ntree=3000, replace=TRUE,
sampsize=5957)

medias3$modelo="bagging3"

medias4<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=10, mtry=27, ntree=5000, replace=TRUE)

medias4$modelo="bagging4"

medias5<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=15, mtry=27, ntree=5000, replace=TRUE)

medias5$modelo="bagging5"

medias6<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
```

```
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=20, mtry=27, ntree=5000, replace=TRUE)

medias6$modelo="bagging6"

union1<-rbind(medias1, medias2, medias3, medias4, medias5, medias6)

par(cex.axis=0.8)
boxplot(data=union1, tasa~modelo, main="TASA FALLOS", col="pink")
boxplot(data=union1, auc~modelo, main="AUC", col="pink")

# ********************
# MEJORES MODELOS
# ********************
medias3<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=8465, repe=5, nodesize=20, mtry=27, ntree=3000, replace=TRUE,
sampsize=5957)

medias3$modelo="bagging3"

medias6<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=8465, repe=5, nodesize=20, mtry=27, ntree=5000, replace=TRUE)

medias6$modelo="bagging6"

union2<-rbind(medias3, medias6)

par(cex.axis=0.8)
boxplot(data=union2, tasa~modelo, main="TASA FALLOS", col="pink")
boxplot(data=union2, auc~modelo, main="AUC", col="pink")

# ********************
# MODELO GANADOR
# ********************
medias3<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
```

```
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=20, mtry=27, ntree=3000, replace=TRUE,
sampsize=5957)

medias3$modelo="bagging3"
```

IV.6. Random forest:

```
# ********************
# RANDOM FOREST
# ********************
# Tuneado de "mtry" con "caret":
set.seed(12345)
rfgrid<-expand.grid(mtry=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,
18,19,20,21,22,23,24,25,26))

control<-trainControl(method="cv", number=4, savePredictions="all",
classProbs=TRUE)

rf<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="rf", trControl=control,
tuneGrid=rfgrid, linout=FALSE, ntree=300, nodesize=10, replace=TRUE,
importance=TRUE)

rf

# ********************
# MODELOS
# ********************
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
source("cruzada rf binaria.R")

medias1<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
```

```
sinicio=1234, repe=5, nodesize=10, mtry=5, ntree=3000, replace=TRUE,
sampsize=5957)

medias1$modelo="rf1"

medias2<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=15, mtry=5, ntree=3000, replace=TRUE,
sampsize=5957)

medias2$modelo="rf2"

medias3<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=20, mtry=5, ntree=3000, replace=TRUE,
sampsize=5957)

medias3$modelo="rf3"

medias4<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=10, mtry=5, ntree=5000, replace=TRUE,
sampsize=6874)

medias4$modelo="rf4"

medias5<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
```

```
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=15, mtry=5, ntree=5000, replace=TRUE,
sampsize=6874)

medias5$modelo="rf5"

medias6<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=20, mtry=5, ntree=5000, replace=TRUE,
sampsize=6874)

medias6$modelo="rf6"

medias7<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=10, mtry=15, ntree=3000, replace=TRUE,
sampsize=5957)

medias7$modelo="rf7"

medias8<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=15, mtry=15, ntree=3000, replace=TRUE,
sampsize=5957)

medias8$modelo="rf8"

medias9<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
```

```
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=20, mtry=15, ntree=3000, replace=TRUE,
sampsize=5957)

medias9$modelo="rf9"

medias10<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=10, mtry=15, ntree=5000, replace=TRUE,
sampsize=6874)

medias10$modelo="rf10"

medias11<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=15, mtry=15, ntree=5000, replace=TRUE,
sampsize=6874)

medias11$modelo="rf11"

medias12<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=20, mtry=15, ntree=5000, replace=TRUE,
sampsize=6874)

medias12$modelo="rf12"
```

```
union1<-rbind(medias1, medias2, medias3, medias4, medias5, medias6,
medias7, medias8, medias9, medias10, medias11, medias12)

par(cex.axis=0.8)
boxplot(data=union1, tasa~modelo, main="TASA FALLOS", col="pink")
boxplot(data=union1, auc~modelo, main="AUC", col="pink")

# ********************
# MEJORES MODELOS
# ********************
medias5<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=9021, repe=5, nodesize=15, mtry=5, ntree=5000, replace=TRUE,
sampsize=6874)

medias5$modelo="rf5"

medias6<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=9021, repe=5, nodesize=20, mtry=5, ntree=5000, replace=TRUE,
sampsize=6874)

medias6$modelo="rf6"

union2<-rbind(medias5, medias6)

par(cex.axis=0.8)
boxplot(data=union2, tasa~modelo, main="TASA FALLOS", col="pink")
boxplot(data=union2, auc~modelo, main="AUC", col="pink")

# ********************
# MODELO GANADOR
# ********************
medias6<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
```

```
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, nodesize=20, mtry=5, ntree=5000, replace=TRUE,
sampsize=6874)

medias6$modelo="rf6"
```

IV.7. Gradient boosting:

```
# ********************
# GRADIENT BOOSTING
# ********************
# El paquete "caret" permite tunear estos parámetros básicos:
#
# shrinkage: parámetro v de regularización, mide la velocidad de
ajuste, a menor v, más lento y necesita más iteraciones, pero es más
fino en el ajuste.
# n.minobsinnode: tamaño máximo de nodos finales (el principal
parámetro que mide la complejidad).
# n.trees: el número de iteraciones (árboles).
# interaction.depth: 2 para árboles binarios.

set.seed(12345)

# El número de iteraciones sí que es algo a tener en consideración en
gradient boosting.
gbmgrid<-expand.grid(shrinkage=c(0.2,0.1,0.05,0.03,0.01,0.001),
n.minobsinnode=c(10,15,20), n.trees=c(1000,3000,5000,7000),
interaction.depth=c(2))

control<-trainControl(method="cv",number=4, savePredictions="all",
classProbs=TRUE)

gbm<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="gbm", trControl=control,
tuneGrid=gbmgrid, distribution="bernoulli", bag.fraction=1,
verbose=FALSE)

gbm

plot(gbm)

# Estudio de early stopping
# Probamos a fijar algunos parámetros para ver cómo evoluciona en
función de las iteraciones.

# shrinkage=c(0.01)
gbmgrid<-expand.grid(shrinkage=c(0.01), n.minobsinnode=c(20),
n.trees=c(50,100,300,500,800,1000,3000,5000,7000),
interaction.depth=c(2))
```

```
control<-trainControl(method="cv", number=4, savePredictions="all",
classProbs=TRUE)
gbm<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="gbm", trControl=control,
tuneGrid=gbmgrid, distribution="bernoulli", bag.fraction=1,
verbose=FALSE)

gbm

plot(gbm)

# shrinkage=c(0.03)
gbmgrid<-expand.grid(shrinkage=c(0.03), n.minobsinnode=c(20),
n.trees=c(50,100,300,500,800,1000,3000,5000,7000),
interaction.depth=c(2))

control<-trainControl(method="cv", number=4, savePredictions="all",
classProbs=TRUE)

gbm2<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="gbm", trControl=control,
tuneGrid=gbmgrid, distribution="bernoulli", bag.fraction=1,
verbose=FALSE)

gbm2

plot(gbm2)

# shrinkage=c(0.05)
gbmgrid<-expand.grid(shrinkage=c(0.05), n.minobsinnode=c(20),
n.trees=c(50,100,300,500,800,1000,3000,5000,7000),
interaction.depth=c(2))

control<-trainControl(method="cv", number=4, savePredictions="all",
classProbs=TRUE)

gbm3<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="gbm", trControl=control,
```

```
tuneGrid=gbmgrid, distribution="bernoulli", bag.fraction=1,
verbose=FALSE)

gbm3

plot(gbm3)

# ********************
# MODELOS
# ********************
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
source("cruzada gbm binaria.R")

medias1<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=10, shrinkage=0.01, n.trees=1000,
interaction.depth=2)

medias1$modelo="gb1"

medias2<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=15, shrinkage=0.01, n.trees=1000,
interaction.depth=2)

medias2$modelo="gb2"

medias3<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=20, shrinkage=0.01, n.trees=1000,
interaction.depth=2)
```

```
medias3$modelo="gb3"

medias4<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=10, shrinkage=0.03, n.trees=1000,
interaction.depth=2)

medias4$modelo="gb4"

medias5<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=15, shrinkage=0.03, n.trees=1000,
interaction.depth=2)

medias5$modelo="gb5"

medias6<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=20, shrinkage=0.03, n.trees=1000,
interaction.depth=2)

medias6$modelo="gb6"

medias7<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
```

```
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""),
grupos=4,sinicio=1234, repe=5, n.minobsinnode=10, shrinkage=0.05,
n.trees=1000, interaction.depth=2)

medias7$modelo="gb7"

medias8<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=15, shrinkage=0.05, n.trees=1000,
interaction.depth=2)

medias8$modelo="gb8"

medias9<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=20, shrinkage=0.05, n.trees=1000,
interaction.depth=2)

medias9$modelo="gb9"

medias10<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=10, shrinkage=0.01, n.trees=5000,
interaction.depth=2)

medias10$modelo="gb10"

medias11<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
```

```
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=15, shrinkage=0.01, n.trees=5000,
interaction.depth=2)

medias11$modelo="gb11"

medias12<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=20, shrinkage=0.01, n.trees=5000,
interaction.depth=2)

medias12$modelo="gb12"

medias13<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=10, shrinkage=0.03, n.trees=5000,
interaction.depth=2)

medias13$modelo="gb13"

medias14<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234,repe=5, n.minobsinnode=15, shrinkage=0.03, n.trees=5000,
interaction.depth=2)

medias14$modelo="gb14"

medias15<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
```

```
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=20, shrinkage=0.03, n.trees=5000,
interaction.depth=2)

medias15$modelo="gb15"

medias16<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=10, shrinkage=0.05, n.trees=5000,
interaction.depth=2)

medias16$modelo="gb16"

medias17<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=15, shrinkage=0.05, n.trees=5000,
interaction.depth=2)

medias17$modelo="gb17"

medias18<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=20, shrinkage=0.05, n.trees=5000,
interaction.depth=2)
```

```
medias18$modelo="gb18"

union1<-rbind(medias1, medias2, medias3, medias4, medias5, medias6,
medias7, medias8, medias9, medias10, medias11, medias12, medias13,
medias14, medias15, medias16, medias17, medias18)

par(cex.axis=0.8)
boxplot(data=union1, tasa~modelo, main="TASA FALLOS", col="pink")
boxplot(data=union1, auc~modelo, main="AUC", col="pink")

union2<-rbind(medias1, medias2, medias3, medias4, medias5, medias6,
medias7, medias8, medias9, medias10, medias11, medias12)

par(cex.axis=0.8)
boxplot(data=union2, tasa~modelo, main="TASA FALLOS", col="pink")
boxplot(data=union2, auc~modelo, main="AUC", col="pink")

# ********************
# MEJORES MODELOS
# ********************
medias4<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=7143, repe=5, n.minobsinnode=10, shrinkage=0.03, n.trees=1000,
interaction.depth=2)

medias4$modelo="gb4"

medias5<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=7143, repe=5, n.minobsinnode=15, shrinkage=0.03, n.trees=1000,
interaction.depth=2)

medias5$modelo="gb5"

medias6<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
```

```
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=7143, repe=5, n.minobsinnode=20, shrinkage=0.03, n.trees=1000,
interaction.depth=2)

medias6$modelo="gb6"

medias10<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=7143, repe=5, n.minobsinnode=10, shrinkage=0.01, n.trees=5000,
interaction.depth=2)

medias10$modelo="gb10"

union3<-rbind(medias4,medias5,medias6,medias10)

par(cex.axis=0.8)
boxplot(data=union3, tasa~modelo, main="TASA FALLOS", col="pink")
boxplot(data=union3, auc~modelo, main="AUC", col="pink")

# ********************
# MODELO GANADOR
# ********************
medias4<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, n.minobsinnode=10, shrinkage=0.03, n.trees=1000,
interaction.depth=2)

medias4$modelo="gb4"
```

IV.8. Extreme gradient boosting:

```
# ********************
# EXTREME GRADIENT BOOSTING
# ********************

# El paquete "caret" permite tunear estos parámetros básicos:
# nrounds (# Boosting Iterations) = número de iteraciones.
# max_depth (Max Tree Depth) = profundida máxima de los árboles.
# eta (Shrinkage) = parámetro v gradient boosting.
```

```
# gamma (Minimum Loss Reduction) = gamma.
   # cte regularización. Dejar a 0 por defecto.
# colsample_bytree (Subsample Ratio of Columns).
   # % sorteo variables antes de cada árbol, al estilo de random
forest pero antes del árbol, no en cada nodo. Dejar a 1 por defecto.
# min_child_weight (Minimum Sum of Instance Weight).
   # Observaciones mínimas en el nodo final. Similar al minobsinnode
del gbm.
# subsample (Subsample Percentage).
   # % sorteo de observaciones antes de cada árbol, al estilo de
random forest. Dejar a 1 por defecto.

set.seed(12345)
xgbmgrid<-expand.grid(min_child_weight=c(10,15,20),
eta=c(0.1,0.05,0.03,0.01,0.001),
nrounds=c(100,500,1000,3000,5000,7000), max_depth=2, gamma=0,
colsample_bytree=1, subsample=1)

control<-trainControl(method="cv",number=4, savePredictions="all",
classProbs=TRUE)

xgbm<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="xgbTree", trControl=control,
tuneGrid=xgbmgrid, verbose=FALSE)

xgbm

plot(xgbm)

# Estudio de early stopping
# Probamos a fijar algunos parámetros para ver como evoluciona en
función de las iteraciones

# eta=0.01
xgbmgrid<-expand.grid(eta=c(0.01), min_child_weight=c(10),
nrounds=c(100,500,1000,3000,5000,7000), max_depth=2, gamma=0,
colsample_bytree=1, subsample=1)
set.seed(12345)
control<-trainControl(method="cv",number=4, savePredictions="all",
classProbs=TRUE)

xgbm<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="xgbTree", trControl=control,
tuneGrid=xgbmgrid, verbose=FALSE)

plot(xgbm)
```

```
# eta=0.03
xgbmgrid<-expand.grid(eta=c(0.03), min_child_weight=c(10),
nrounds=c(100,500,1000,3000,5000,7000), max_depth=2, gamma=0,
colsample_bytree=1, subsample=1)

set.seed(12345)
control<-trainControl(method="cv",number=4, savePredictions="all",
classProbs=TRUE)

xgbm2<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="xgbTree", trControl=control,
tuneGrid=xgbmgrid, verbose=FALSE)

plot(xgbm2)

# eta=0.05
xgbmgrid<-expand.grid(eta=c(0.05), min_child_weight=c(10),
nrounds=c(100,500,1000,3000,5000,7000), max_depth=2, gamma=0,
colsample_bytree=1, subsample=1)

set.seed(12345)
control<-trainControl(method="cv", number=4, savePredictions="all",
classProbs=TRUE)

xgbm3<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="xgbTree", trControl=control,
tuneGrid=xgbmgrid, verbose=FALSE)

plot(xgbm3)

# ********************
# MODELOS
# ********************
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentaci?n
R/Todos los programas y datasets R")
source("cruzada xgboost binaria.R")

medias1<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
```

```
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=10, eta=0.01, nrounds=3000,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias1$modelo="xgb1"

medias2<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234,repe=5, min_child_weight=15, eta=0.01, nrounds=3000,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias2$modelo="xgb2"

medias3<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=20, eta=0.01, nrounds=3000,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias3$modelo="xgb3"

medias4<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=10, eta=0.03, nrounds=1000,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias4$modelo="xgb4"
```

```
medias5<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=15, eta=0.03, nrounds=1000,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias5$modelo="xgb5"

medias6<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=20, eta=0.03, nrounds=1000,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias6$modelo="xgb6"

medias7<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=10, eta=0.05, nrounds=500,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias7$modelo="xgb7"

medias8<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
```

```
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=15, eta=0.05, nrounds=500,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias8$modelo="xgb8"

medias9<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=20, eta=0.05, nrounds=500,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias9$modelo="xgb9"

union1<-rbind(medias1, medias2, medias3, medias4, medias5, medias6,
medias7, medias8, medias9)

par(cex.axis=0.8)
boxplot(data=union1, tasa~modelo, main="TASA FALLOS", col="pink")
boxplot(data=union1, auc~modelo, main="AUC", col="pink")

# ********************
# MODELO GANADOR
# ********************
medias9<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=4,
sinicio=1234, repe=5, min_child_weight=20, eta=0.05, nrounds=500,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias9$modelo="xgb9"
```

IV.9. Support vector machine:

```
# ********************
# SVM KERNEL LINEAL: SOLO PARÁMETRO C
# ********************
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
```

```
source("cruzada SVM binaria lineal.R")

set.seed(12345)
SVMgrid<-expand.grid(C=c(0.05,0.1,0.2,0.5,1,2,5,10,25,50))

control<-trainControl(method = "cv", number=4, savePredictions =
"all")

SVM<-train(data=violenciagenerobis,
factor(miedo)~IMP_REP_REP_numpar+tenpar.0+control.0+IMP_REP_hijos.0+IM
P_REP_ingresos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingreso
s.5+IMP_REP_suicidio.1+IMP_REP_visita.1, method="svmLinear",
trControl=control, tuneGrid=SVMgrid, verbose=FALSE)

SVM$results
plot(SVM$results$C, SVM$results$Accuracy)
# ********************
# MODELOS
# ********************
medias1<-cruzadaSVMbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=0.1)

medias1$modelo="svml1"

medias2<-cruzadaSVMbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=0.2)

medias2$modelo="svml2"

medias3<-cruzadaSVMbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=5)

medias3$modelo="svml3"

medias4<-cruzadaSVMbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=25)

medias4$modelo="svml4"

union1<-rbind(medias1, medias2, medias3, medias4)

par(cex.axis=0.5)
boxplot(data=union1, tasa~modelo, main="TASA FALLOS")
boxplot(data=union1, auc~modelo, main="AUC")
```

```
# ********************
# MODELO GANADOR
# ********************
medias1<-cruzadaSVMbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=0.1)

medias1$modelo="svml1"

# ********************
# SVM KERNEL POLINOMIAL: PARÁMETROS C, DEGREE Y SCALE
# ********************
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
source("cruzada SVM binaria polinomial.R")
SVMgrid<-expand.grid(C=c(0.1,0.2,5,25), degree=c(2,3),
scale=c(0.1,0.5,1,2,5))

control<-trainControl(method = "cv", number=4, savePredictions =
"all")

SVM2<-
train(data=violenciagenerobis,factor(miedo)~IMP_REP_REP_numpar+tenpar.
0+control.0+IMP_REP_hijos.0+IMP_REP_ingresos.1+IMP_REP_ingresos.2+IMP_
REP_ingresos.3+IMP_REP_ingresos.5+IMP_REP_suicidio.1+IMP_REP_visita.1,
method="svmPoly", trControl=control, tuneGrid=SVMgrid, verbose=FALSE)

SVM2

SVM2$results

# ggplot
dat<-as.data.frame(SVM2$results)
ggplot(dat, aes(x=factor(C), y=Accuracy, color=factor(degree),
pch=factor(scale))) + geom_point(position=position_dodge(width=0.5),
size=3)

# ggplot solo degree=3
dat2<-dat[dat$degree==3,]
ggplot(dat2, aes(x=factor(C), y=Accuracy, colour=factor(scale))) +
geom_point(position=position_dodge(width=0.5), size=3)

# ********************
# MODELOS
# ********************
medias5<-cruzadaSVMbinPoly(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=0.1, degree=3, scale=1)

medias5$modelo="svmp1"

medias6<-cruzadaSVMbinPoly(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
```

```
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=0.2, degree=3, scale=1)

medias6$modelo="svmp2"

medias7<-cruzadaSVMbinPoly(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=5, degree=3, scale=0.1)

medias7$modelo="svmp3"

medias8<-cruzadaSVMbinPoly(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=25, degree=3, scale=0.1)

medias8$modelo="svmp4"

union2<-rbind(medias5, medias6, medias7, medias8)

par(cex.axis=0.5)
boxplot(data=union2, tasa~modelo, main="TASA FALLOS")
boxplot(data=union2, auc~modelo, main="AUC")

# ********************
# MODELO GANADOR
# ********************
medias7<-cruzadaSVMbinPoly(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=5, degree=3, scale=0.1)

medias7$modelo="svmp3"

# ********************
# SVM KERNEL RBF/GAUSSIANO: PARÁMETROS C Y SIGMA/GAMMA
# ********************
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
source("cruzada SVM binaria RBF.R")

SVMgrid<-expand.grid(C=c(0.1,0.2,5,25),
sigma=c(0.001,0.01,0.02,0.05,0.1,0.2,0.5))

control<-trainControl(method = "cv", number=4, savePredictions =
"all")

SVM3<- train(data=violenciagenerobis,
factor(miedo)~IMP_REP_REP_numpar+tenpar.0+control.0+IMP_REP_hijos.0+IM
P_REP_ingresos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingreso
s.5+IMP_REP_suicidio.1+IMP_REP_visita.1, method="svmRadial",
trControl=control, tuneGrid=SVMgrid, verbose=FALSE)
```

```
SVM3

dat<-as.data.frame(SVM3$results)

ggplot(dat, aes(x=factor(C), y=Accuracy, color=factor(sigma))) +
geom_point(position=position_dodge(width=0.5), size=3)

# ********************
# MODELOS
# ********************
medias9<-cruzadaSVMbinRBF(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=0.1, sigma=0.05)

medias9$modelo="svmrbf1"

medias10<-cruzadaSVMbinRBF(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=0.2, sigma=0.1)

medias10$modelo="svmrbf2"

medias11<-cruzadaSVMbinRBF(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=5, sigma=0.02)

medias11$modelo="svmrbf3"

medias12<-cruzadaSVMbinRBF(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=25, sigma=0.02)

medias12$modelo="svmrbf4"

union3<-rbind(medias9, medias10, medias11, medias12)

par(cex.axis=0.5)
boxplot(data=union3, tasa~modelo, main="TASA FALLOS")
boxplot(data=union3, auc~modelo, main="AUC")

# ********************
# MODELO GANADOR
# ********************
medias9<-cruzadaSVMbinRBF(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
```

```
"IMP_REP_visita.1"), listclass=c(""), grupos=4, sinicio=1234, repe=5,
C=0.1, sigma=0.05)

medias9$modelo="svmrbf1"
```

IV.10. Comparación de modelos:

```
# ********************
# COMPARACIÓN DE MODELOS
# ********************
# Regresión logística:
medias1<-cruzadalogistica(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=10,
sinicio=1234, repe=20)

medias1$modelo="log1"

medias2<-cruzadalogistica(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=10, sinicio=1234,
repe=20)

medias2$modelo="log2"

# Red neuronal:
medias3<-cruzadaavnnetbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=10, sinicio=1234,
repe=20, size=c(3), decay=c(0.01), repeticiones=5, itera=100)

medias3$modelo="avnnet"

# Bagging:
medias4<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=10,
sinicio=1234, repe=20, nodesize=20, mtry=27, ntree=3000, replace=TRUE,
sampsize=5957)
```

```
medias4$modelo="bg"

# Random forest:
medias5<-cruzadarfbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=10,
sinicio=1234, repe=20, nodesize=20, mtry=5, ntree=5000, replace=TRUE,
sampsize=6874)

medias5$modelo="rf"

# Gradient boosting:
medias6<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=10,
sinicio=1234, repe=20, n.minobsinnode=10, shrinkage=0.03,
n.trees=1000, interaction.depth=2)

medias6$modelo="gb"

# Extreme gradient boosting:
medias7<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=10,
sinicio=1234, repe=20, min_child_weight=20, eta=0.05, nrounds=500,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias7$modelo="xgb"

# SVM con kernel lineal:
medias8<-cruzadaSVMbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
```

```
"IMP_REP_visita.1"), listclass=c(""), grupos=10, sinicio=1234,
repe=20, C=0.1)

medias8$modelo="svml"

# SVM con kernel polinomial:
medias9<-cruzadaSVMbinPoly(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=10, sinicio=1234,
repe=20, C=5, degree=3, scale=0.1)

medias9$modelo="svmp"

# SVM con kernel gaussiano:
medias10<-cruzadaSVMbinRBF(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=10, sinicio=1234,
repe=20, C=0.1, sigma=0.05)

medias10$modelo="svmrbf"

union1<-rbind(medias1, medias2, medias3, medias4, medias5, medias6,
medias7, medias8, medias9, medias10)

par(cex.axis=0.5)
boxplot(data=union1, tasa~modelo, main="TASA FALLOS")
boxplot(data=union1, auc~modelo, main="AUC")

union2<-rbind(medias1, medias2, medias3, medias5, medias6, medias7,
medias8)

par(cex.axis=0.5)
boxplot(data=union2, tasa~modelo, main="TASA FALLOS")
boxplot(data=union2, auc~modelo, main="AUC")

# ********************
# MEJORES MODELOS
# ********************
medias3<-cruzadaavnnetbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1"), listclass=c(""), grupos=10, sinicio=7014,
repe=20, size=c(3), decay=c(0.01), repeticiones=5, itera=100)

medias3$modelo="avnnet"

medias6<-cruzadagbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
```

```
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=10,
sinicio=7014, repe=20, n.minobsinnode=10, shrinkage=0.03,
n.trees=1000, interaction.depth=2)

medias6$modelo="gb"

medias7<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=10,
sinicio=7014, repe=20, min_child_weight=20, eta=0.05, nrounds=500,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias7$modelo="xgb"

union3<-rbind(medias3, medias6, medias7)

par(cex.axis=0.5)
boxplot(data=union3, tasa~modelo, main="TASA FALLOS")
boxplot(data=union3, auc~modelo, main="AUC")

# ********************
# MODELO GANADOR
# ********************
medias7<-cruzadaxgbmbin(data=violenciagenerobis, vardep="miedo",
listconti=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5"), listclass=c(""), grupos=10,
sinicio=1234, repe=20, min_child_weight=20, eta=0.05, nrounds=500,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1, alpha=0,
lambda=0, lambda_bias=0)

medias7$modelo="xgb"
```

IV.11. Ensamblado:

```
# ********************
# MODELOS DE ENSAMBLE
# ********************
# Importante: aquí hay que decidir antes los parámetros a utilizar en
cada algoritmo, no vale el grid.
# Preparación del archivo, de las variables y de la validación
cruzada.
```

```
# Necesario haber cambiado la variable dependiente a "Yes" y "No".

# Leer las cruzadas de ensamblado, son ligeramente diferentes a las
utilizadas anteriormente, aunque se llaman igual.
setwd("C:/Users/alexa/Downloads/Machine Learning tema 5/documentación
R/Todos los programas y datasets R")
source("cruzadas ensamblado binaria fuente.R")

set.seed(12345)
archivo<-violenciagenerobis

vardep=c("miedo")

# Para técnicas no basadas en árboles:
listconti1=c("IMP_REP_REP_numpar", "tenpar.0", "control.0",
"IMP_REP_hijos.0", "IMP_REP_ingresos.1", "IMP_REP_ingresos.2",
"IMP_REP_ingresos.3", "IMP_REP_ingresos.5", "IMP_REP_suicidio.1",
"IMP_REP_visita.1")

# Para técnicas basadas en árboles:
listconti2=c("IMP_REP_acoger.0", "IMP_REP_estudios.1",
"IMP_REP_estudios.2", "IMP_REP_estudios.3", "IMP_REP_estudios.4",
"IMP_REP_estudios.5", "IMP_REP_frecreunion.1",
"IMP_REP_frecreunion.2", "IMP_REP_frecreunion.3",
"IMP_REP_ingresos.1", "IMP_REP_ingresos.2", "IMP_REP_ingresos.3",
"IMP_REP_ingresos.5", "IMP_REP_religion.1", "IMP_REP_religion.2",
"IMP_REP_religion.3", "IMP_REP_religion.4", "IMP_REP_suicidio.0",
"IMP_REP_usointernet.1", "IMP_REP_visita.1", "REP_edad",
"IMP_REP_REP_numpar", "tenpar.0", "control.1", "REP_sitlab.1",
"REP_sitlab.3", "REP_sitlab.5")

listclass=c("")
grupos<-10
sinicio<-7014
repe<-20
# Aplicación cruzadas para ensamblar
medias1<-cruzadaavnnetbin(data=archivo, vardep=vardep,
listconti=listconti1, listclass=listclass, grupos=grupos,
sinicio=sinicio, repe=repe, size=c(3), decay=c(0.01), repeticiones=5,
itera=100)

medias1bis<-as.data.frame(medias1[1])
medias1bis$modelo<-"1"
predi1<-as.data.frame(medias1[2])
predi1$avnnet<-predi1$Yes

medias2<-cruzadagbmbin(data=archivo, vardep=vardep,
listconti=listconti2, listclass=listclass, grupos=grupos,
sinicio=sinicio, repe=repe, n.minobsinnode=10, shrinkage=0.03,
n.trees=1000, interaction.depth=2)

medias2bis<-as.data.frame(medias2[1])
medias2bis$modelo<-"2"
predi2<-as.data.frame(medias2[2])
predi2$gbm<-predi2$Yes

medias3<-cruzadaxgbmbin(data=archivo, vardep=vardep,
listconti=listconti2, listclass=listclass, grupos=grupos,
sinicio=sinicio, repe=repe, min_child_weight=20, eta=0.05,
```

```
nrounds=500, max_depth=2, gamma=0, colsample_bytree=1, subsample=1,
alpha=0, lambda=0, lambda_bias=0)

medias3bis<-as.data.frame(medias3[1])
medias3bis$modelo<-"3"
predi3<-as.data.frame(medias3[2])
predi3$xgbm<-predi3$Yes

union1<-rbind(medias1bis, medias2bis, medias3bis)

par(cex.axis=0.8)
boxplot(data=union1, tasa~modelo, col="pink", main='TASA FALLOS')
boxplot(data=union1, auc~modelo, col="pink", main='AUC')

# Construcción de todos los ensamblados.
# Se utilizarán los archivos surgidos de las funciones llamados
predi1, predi2, etc.
unipredi<-cbind(predi1, predi2, predi3)

# Esto es para eliminar columnas duplicadas
unipredi<- unipredi[, !duplicated(colnames(unipredi))]

# Construcción de ensamblados (cambiar al gusto)
unipredi$predi4<-(unipredi$avnnet+unipredi$gbm)/2
unipredi$predi5<-(unipredi$avnnet+unipredi$xgbm)/2
unipredi$predi6<-(unipredi$gbm+unipredi$xgbm)/2
unipredi$predi7<-(unipredi$avnnet+unipredi$gbm+unipredi$xgbm)/3

# Listado de modelos a considerar (cambiar al gusto)
dput(names(unipredi))

listado<-c("avnnet", "gbm", "xgbm", "predi4", "predi5", "predi6",
"predi7")
# Se define la función de tasa de fallos
tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
}

auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
}

# Se obtiene el número de repeticiones CV y se calculan las medias por
repe en el data frame medias0

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)
  }
}

# Boxplot
par(cex.axis=0.5,las=2)
boxplot(data=medias0, tasa~modelo, col="pink", main="TASA FALLOS")
# Para AUC se utiliza la variable AUC del archivo medias0
boxplot(data=medias0, auc~modelo, col="pink", main="AUC")
```

IV.12. Análisis del modelo ganador:

```
# ********************
# ANÁLISIS DEL MODELO GANADOR
# ********************
# Importancia de las variables del modelo ganador:
set.seed(12345)

xgbmgrid<-expand.grid(min_child_weight=20, eta=0.05, nrounds=500,
max_depth=2, gamma=0, colsample_bytree=1, subsample=1)

control<-trainControl(method="cv", number=4, savePredictions="all",
classProbs = TRUE)

xgbm<-
train(factor(miedo)~IMP_REP_acoger.0+IMP_REP_estudios.1+IMP_REP_estudi
os.2+IMP_REP_estudios.3+IMP_REP_estudios.4+IMP_REP_estudios.5+IMP_REP_
frecreunion.1+IMP_REP_frecreunion.2+IMP_REP_frecreunion.3+IMP_REP_ingr
esos.1+IMP_REP_ingresos.2+IMP_REP_ingresos.3+IMP_REP_ingresos.5+IMP_RE
P_religion.1+IMP_REP_religion.2+IMP_REP_religion.3+IMP_REP_religion.4+
IMP_REP_suicidio.0+IMP_REP_usointernet.1+IMP_REP_visita.1+REP_edad+IMP
_REP_REP_numpar+tenpar.0+control.1+REP_sitlab.1+REP_sitlab.3+REP_sitla
b.5, data=violenciagenerobis, method="xgbTree", trControl=control,
tuneGrid=xgbmgrid, verbose=FALSE)
xgbm

varImp(xgbm)$importance

# Matriz de confusión del modelo ganador con punto de corte 0,5:
sal<-xgbm$pred
confusionMatrix(data=as.factor(ifelse(sal$Yes>=0.5, "Yes", "No")),
reference=sal$obs, positive="Yes")

# Matriz de confusión del modelo ganador con punto de corte 0,1385:
sal<-xgbm$pred

confusionMatrix(data=as.factor(ifelse(sal$Yes>=0.1385, "Yes", "No")),
reference=sal$obs, positive="Yes")
```