FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2019/2020

Trabajo de Fin de Máster

Título: Predicción y análisis de los resultados de las elecciones de representantes al senado en el año 2018 en los quince departamentos más afectados por la violencia en Colombia

***Alumno*: Angélica Guerrero López**

***Tutor*: Antonio Sarasa**

Septiembre de 2020

UNIVERSIDAD COMPLUTENSE
MADRID

A mi familia por ser mi guía, mi motor e inspiración
Al amor que se transforma, por siempre acompañar
Gracias.

Contenido

Contenido de gráficas

Contenido de tablas

1. Introducción

Colombia se encuentra ubicado en Sudamérica, se divide en 32 departamentos (33 si tomamos la capital como departamento independiente) y 1120 municipios, con una población aproximadamente de 50 millones de habitantes. Su geografía montañosa ha hecho que la distribución poblacional no sea homogénea en todo el territorio; la mayor parte se encuentra agrupada en el centro y norte del país. Al tener costas en el Océano Pacífico y Atlántico, y un 40% del territorio de selva amazónica, cuenta con una riqueza natural que lo posiciona como la segunda nación más biodiversa del mundo (Ministerio de Ciencias de Colombia, 2016).

Curiosamente, Colombia ha sido reconocido más por sus problemas de violencia que por su riqueza natural. El conflicto armado interno nace hacia 1960 como respuesta de ciertas poblaciones campesinas frente a la desigualdad social y económica, y, por tanto, se organizaron en grupos guerrilleros con tendencia ideológica de izquierda (FARC-EP, ELN, EPL, entre otras guerrillas menores) y se alzaron en armas en contra del Estado. Años más tarde, hacia la década de los noventa e inicios del nuevo milenio, surgieron grupos de extrema derecha denominados paramilitares (AUC, Bacrim, AGC) que ahondaron las heridas y marcas de la guerra en el territorio colombiano. Esta guerra ha creado fenómenos sociales como el desplazamiento interno (El Espectador, 2019) el narcotráfico (Público, 2018) y violencia (Unidad de atención y reparación integral a las víctimas, 2020) que han incrementado la desigualdad y la pobreza en muchas regiones del país.

Durante años, tanto las guerrillas como los paramilitares se apropiaron de diversos territorios del país, evitando que el Estado tuviera el control de dichas zonas, y determinando el futuro de los habitantes según sus normas y sus intereses a través de las presiones electorales que ejercían para votar por cierta tendencia ideológica de candidatos políticos o por ciertos partidos políticos. En los municipios que hacen parte de sus territorios constantemente se veían expuestos a combates entre los grupos armados, estaban obligados a tener cultivos ilícitos y en muchos casos, ser expulsados de sus casas y tierras al demostrar su inconformidad (Centro Nacional de Memoria Histórica, 2018).

En el año 2016 se firmó el Acuerdo de Paz (Jurisdicción Especial para la Paz, 2016) entre el Estado colombiano y una de las principales guerrillas del país, y una de las más antiguas de Latinoamérica, a saber, las Fuerzas Armadas Revolucionarias de Colombia- Ejército del Pueblo (FARC-EP). En el segundo punto del Acuerdo se aprobó que la guerrilla pasara a ser parte del Senado de la Republica y desde allí, siendo oposición, pudiera aportar con leyes y plenarias a mejorar el país.

El Congreso colombiano actualmente cuenta con 108 Senadores, y en el marco del Acuerdo se pactó que se asignarían algunas curules especiales para el nuevo partido político FARC. “Los ciudadanos podrán elegir el candidato de su preferencia en cualquier lugar del país para que sea su representante en la cámara alta. Los candidatos estarán inscritos en una

lista avalada por un partido o grupo significativo de ciudadanos mediante firmas. Los electores deberán marcar el logo del partido y el número correspondiente a su candidato" (Semana, 2018). Allí, los representantes tienen el poder de reformar la constitución política mediante actos legislativos, elaborar, interpretar, reformar y derogar las leyes y códigos en todos los ramos de la legislación, entre otras funciones (Senado, 2020). De esta forma, el ejercicio democrático al que invitaba el Acuerdo abría la posibilidad a que los ciudadanos pudieran, de manera libre y sin coacción, votar legítimamente por un partido de izquierda con el pasado histórico de violencia como lo es FARC, pero también poder votar libremente por otras opciones políticas con opuestas tendencias ideológicas. De esta forma las opciones electorales se ampliaban y las influencias por actores armados disminuiría.

En el marco de las herramientas de análisis de datos que se han fomentado en la Maestría de Minería de Datos e Inteligencia de Negocios, se quiere analizar la intención de voto de cada municipio al Senado en el año 2014 y predecir el de las elecciones del año 2018. Cabe aclarar que tomar como referencia solo dos resultados electorales y unas cuantas variables no es suficiente para afirmar que el Acuerdo de Paz es el único responsable de los cambios ni del comportamiento que se identifique. Al intentar predecir los resultados del año 2018 se pretende analizar el comportamiento de la población que reside en municipios que históricamente fueron los más afectados por la violencia y de esta forma ver si su tendencia electoral guarda relación con las variables que se emplearon para el estudio.

Es por esto que, una vez se obtenga el mejor modelo, se comparará con los resultados reales del 2018, logrando evaluar el ejercicio académico y dejar insumos que permitan hacer un análisis más profundo del efecto del acuerdo de paz en las elecciones de Colombia.

2. Objetivos

Mediante el estudio se quiere predecir los resultados de las elecciones de representantes al Senado del año 2018 en 15 departamentos identificados como unos de los más afectados por la violencia en Colombia. Esto, teniendo como base las elecciones previas al Acuerdo de Paz, realizadas en el año 2014. De esta forma, se pretende analizar su comportamiento desde una perspectiva socioeconómica y teniendo en cuenta algunos factores emergentes generados por el Acuerdo de Paz o en dicho periodo de su aprobación entre los dos años de elecciones legislativas.

- Determinar si en estas regiones hay una preferencia marcada por partidos de derecha.

- Identificar variables o factores socioeconómicos y de violencia para describir el contexto desde el cual se identifica fluidez electoral en el periodo estudiado.

- Observar si los resultados de las elecciones del 2014 cambian de forma drástica en el año 2018 en alguna región históricamente afectada por el conflicto armado.

3. Marco teórico

3.1. Estado del arte

El estudio de las dinámicas electorales hace parte del campo de estudio específico de la Ciencia Política, y desde allí los análisis cualitativos y cuantitativos han ofrecido interpretaciones para revisar distintos fenómenos y contextos que cambian el panorama de votantes en cualquier contexto democrático. Desde esta perspectiva, los análisis de estudiosos como Mario Latorre Rueda (2009) o Francisco Leal Buitrago, que desde los años 70 han plasmado diferentes realidades sociales y políticas que han vivido el país. Sin embargo, el caso excepcional de un proceso de paz que se sostiene a través de un Acuerdo de Paz como el firmado en 2016, cambia la manera de leer el electorado en Colombia. Si bien se tuvo procesos de paz previos como los experimentados con el M-19, el EPL o el Quintin Lame, o incluso los acontecidos con las FARC-EP bajo el gobierno de Andrés Pastrana, desde los inicios del conflicto armado librado por guerrillas en los años 60, no habíamos contado con un proceso de la trascendencia ni la relativa estabilidad que se logró con el Acuerdo de Paz bajo el gobierno de Juan Manuel Santos (Rodríguez Pinzón, 2014).

Los análisis que de este proceso se han hecho han reseñado principalmente los problemas de abstención que son sistemáticos en la historia electoral colombiana, y que en el periodo comprendido entre las elecciones legislativas de 2014 y las de 2018 se esperaba que la participación fuera mayor (Basset & Guavita, 2019). Tal como se muestra en *Radiografía del Desencanto,* los autores logran mapear de forma sociológica el panorama de distribución electoral influenciado por la violencia, los distintos grupos armados que inciden en estas dinámicas. Sin embargo, el estudio muestra las distribuciones de *participación electoral* en el país, lo cual aporta al debate respecto a la alta y sistemática abstención del sistema electoral colombiano. En este estudio, se hace claro que los grupos armados tienen una incidencia importante en la coacción del electorado, pero su foco de atención organizativa no está sobre las elecciones sino sobre el narcotráfico (Basset & Guavita, 2019, pág. 237). Bajo esta lupa, los cambios entre una elección legislativa o presidencial a la siguiente, se convierte en una forma para evaluar el impacto del Acuerdo de Paz en el entendido del retiro de guerrilleros que coaccionaban a la población previamente para votar por algún partido o candidato en especial.

En este sentido, los análisis que han evidenciado trabajos como los de Milanese y Serrano (2019) muestran un panorama de *fluidez electoral* desde la cual se analizan los cambios de preferencias ideológicas o partidistas en un periodo de elecciones, y en específico, en los últimos comicios celebrados en 2018. En dicho trabajo se analizan específicamente las 16 Circunscripciones Transitorias Especiales para la Paz, es decir, 16 agrupaciones de municipios de distintos departamentos que tenían como intención ser una medida de reparación histórica hacia las víctimas que sufrieron la guerra con más fuerza. De esta forma, lo que evidencian Milanese y Serrano a partir del modelo de inferencia ecológica, es que las Circunscripciones no están ajenas a las dinámicas de fluidez altas, es decir, en donde más de un 70% de los electores fluctúan de una opción política a otra que habían tomado en el pasado. De hecho, los autores mencionan que: “estos comportamientos no pueden ser interpretados como perfectamente consistentes, pero sí articulados sobre un eje que estructura una suerte de núcleo duro de votantes caracterizado por su arraigo y que trasciende las nociones convencionales de lo que se entiende por izquierda y derecha” (Milanese & Serrano, 2019, pág. 22). Sin embargo, este es el trabajo que más se acerca a

la lectura frente a la *fluidez electoral,* aunque con un espectro limitado de análisis según el nicho espacial que fue elegido, y que se destaca por haber sido un modelo electoral fallido al no garantizar ni otorgar

Desde esta perspectiva, la invitación metodológica que proponen Barrero y Baquero (2019) para el análisis electoral, y sobre todo en el caso colombiano, es cuidar el directo relacionamiento entre las variables numéricas y las interpretaciones sociológicas a las que tenga lugar quien escribe la investigación, y por tanto, la fiabilidad de los resultados siempre tenderá a tener un sesgo de interpretación. Para esto, lo que proponen los autores es afinar las herramientas de análisis para que desde allí se puedan crear categorías, modelos y nuevas aproximaciones a la complejidad del sistema electoral en Colombia. De esta forma, la medición de volatilidad electoral, o en palabras de Milanese y Serrano, la fluidez electoral, tiene como base de datos fundamental los resultados oficiales de las votaciones, lo cual no permite entender otras dinámicas sociales, económicas o contextuales que puedan ocurrir más allá de los resultados de una jornada electoral.

En este sentido, usar las herramientas de *big data* y *machine learning* permiten integrar múltiples variables a un modelo de predicción que ofrezca líneas de interpretación objetivas y complejas para este tipo de fenómenos políticos. Si bien el campo de la Ciencia Política ha construido su propio anaquel de métodos y procedimientos de análisis, lo que se pretende con el presente trabajo es dar un panorama investigativo que aporte al conocimiento actual que hay respecto al panorama político posacuerdo, complementando con datos comparativos nacionales y focalizados en otras unidades de análisis que no son las Circunscripciones Especiales para la Paz. Sin embargo, cabe resaltar que uno de los trabajos que más aporta al conocimiento de las dinámicas electorales del periodo entre 2014 y 2018 es el texto *Circunscripciones Especiales: la paz en la apatía electoral* de Fernando Giraldo García y Hernán Renán Soto Caballero (2019), en donde resaltan de forma muy focalizada y sólida la baja participación en el Plebiscito para la refrendación de los Acuerdos, y que fue un gazapo para la legitimidad política del Acuerdo y del proceso de paz en sí mismo. Tal como lo describen los autores:

> Uno de los resultados más interesantes es que no en todos los municipios que conforman las CTEP ganó el “Sí”; de hecho, en 35 de los 167 municipios ganó el “No” […], lo que refleja que las condiciones adversas de la pobreza y el conflicto armado no garantizan necesariamente un apoyo automático e incondicional a la paz; prueba de ello es que sólo en siete de las CTEP los municipios votaron totalmente a favor del “Sí”; en el extremo opuesto, en la CTEP 15 todos los municipios votaron a favor del “No”. De este modo, el apoyo a la paz en estas circunscripciones no estará alejado de la polarización que se manifiesta en el resto del país, y se aprecia que no necesariamente la categoría de “víctimas” es sinónimo de apoyo a la paz (2019).

De esta forma, el presente estudio se alinea con las investigaciones que actualmente permiten entender el panorama de cambios electorales que se vivieron en Colombia a raíz de la firma del Acuerdo de Paz como un hito histórico, no solo para la democracia sino para la historia misma de la nación como correlato de violencia y control estatal, que casi después de 50 años de conflicto, se pudo por fin vislumbrar otras nuevas posibilidades electorales. ¿Fueron en exceso volátiles las elecciones entre 2014 y 2018, y en qué medida se podrían usar herramientas de big data para predecir los resultados electorales y algunos

de sus cambios? A continuación, se explican los métodos usados para abordar la fluidez electoral que predice en el periodo Posacuerdo.

3.2. Metodología empleada

3.2.1. Análisis Multivariante

Son técnicas estadísticas que analizan resultados múltiples de individuos; será multivariado si las variables son aleatorias y existe relación entre ellas, pues de esta forma no tendrían un efecto de manera individual.

El objetivo principal de estos métodos es resumir grandes cantidades de datos por medio de un número reducido de parámetros. Es decir, se trata de simplificar y poner orden en la ingente información que se procesa.

Algunas de las técnicas como el análisis de componentes y el análisis factorial están dirigidos hacia las variables, es decir tratan de detectar las relaciones que existen entre ellas (Valecia Delfa & Vicente Hernanz, 2006).

3.2.2. Análisis factorial

A partir de la matriz de covarianzas o de la matriz de correlaciones de un conjunto de variables, es posible detectar relaciones existentes entre dichas variables. El propósito esencial del Análisis factorial es describir, si es posible, estas relaciones en términos de unas pocas variables subyacentes e inobservables variables denominadas Factores.

El primer paso para el análisis fue calcular la matriz de correlaciones R entre todas las variables que entran en el análisis para comprobar si sus características eran adecuadas para realizar este tipo de análisis.

En este caso se tomó el índice KMO:

$$\begin{cases} \text{si } KMO \leq 0.5 \text{ valor inaceptable, se desaconseja A.F.} \\ \text{si } 0.5 \leq KMO \leq 0.6 \text{ valor demasiado bajo} \\ \text{si } 0.6 \leq KMO \leq 0.8 \text{ valor mediocre} \\ \text{si } KMO \succ 0.8 \text{ valor excelente.} \end{cases}$$

La medida de adecuación de la muestra MSA (evaluado para cada variable) donde valores bajos de este índice desaconsejaron el uso del análisis en la misma medida que el índice KMO (se pudo eliminar la variable del estudio y continuar con el resto).

Una vez seleccionadas las variables, se halló la descomposición en función de sus autovalores de la matriz de correlaciones. Según dichos resultados se eligieron los factores a retener y se procedió a identificar las cargas de cada variable en los factores.

Para equilibrar la variabilidad es importante rotar los factores, pues puede llevar a una mejor interpretación de los factores, además con la evaluación de la matriz de cargas se pueden obtener agrupaciones de observaciones por grupos o cluster (Valecia Delfa & Vicente Hernanz, 2006).

3.2.3. Metodología SEMMA

Esta metodología fue desarrollada principalmente por la empresa de software SAS y CRISP-DM desarrollada, entre otros, por la empresa IBM, consta de cinco pasos:

- Sample (muestrear): Si la base de datos es demasiado grande, será necesario tomar una muestra lo suficientemente grande para contener toda la información y lo suficientemente pequeña para poder ser procesada.

- Explore (explorar): Es necesario explorar los datos para detectar relaciones, anomalías y tendencias.

- Modify (modificar): Es recomendable modificar los datos creando, seleccionando y transformando variables para facilitar la modelización.

- Model (modelizar): Hallar el modelo que nos permita predecir la variable objetivo

- Assess (evaluar): Comprobar la calidad de las predicciones y comparar los modelos obtenidos.

Es importante aclarar que no siempre intervienen todas las fases del proceso, además, las fases pueden repetirse y el orden de estas pueden modificarse (Calviño Martínez, 2019).

3.2.4. Regresión Logística

Es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor.

Es importante tener en cuenta que, aunque la regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final se hace en función de las probabilidades predichas.

$$Función\ sigmoide = \ \sigma\ (x) = \frac{1}{1 + \ e^{-x}}$$

Para valores de x muy grandes positivos, el valor de e^{-x} es aproximadamente 0 o lo que el valor de la función sigmoide es 1. Para valores de x muy grandes negativos, el valor e^{-x} tiende a infinito por lo que el valor de la función sigmoide es 0.

Sustituyendo la x de la ecuación 1 por la función lineal $(\beta_0 + \beta_1 X)$ se obtiene que:

$$P(Y = k\ |X = x) = \frac{1}{1 + \ e^{-(\beta_0 + \beta_1 X)}}$$

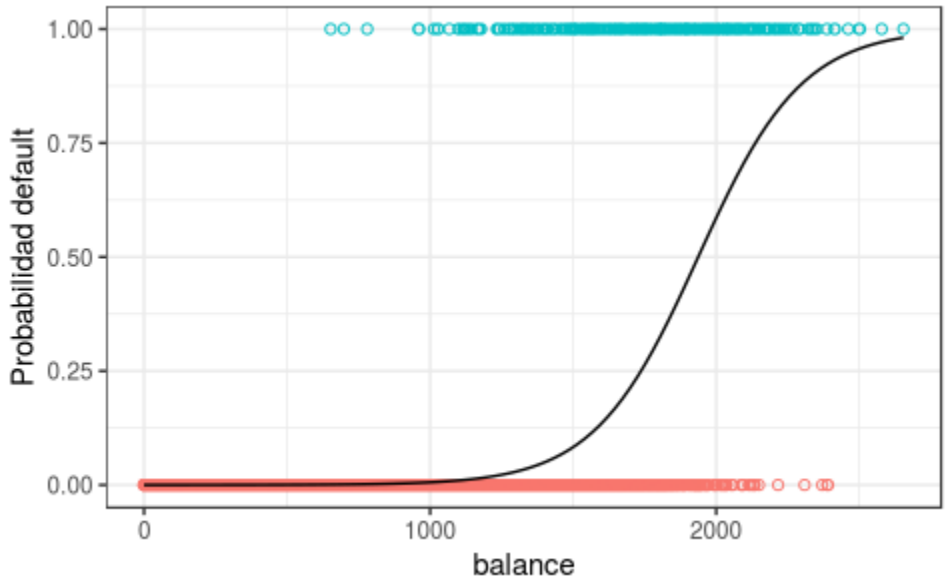
$$\frac{e^{\beta_0 + \beta_1 X}}{1 + \ e^{\beta_0 + \beta_1 X}}$$

Donde $P(Y = k\,|X = x)$ puede interpretarse como: la probabilidad de que la variable cualitativa Y adquiera el valor k (el nivel de referencia, codificado como 1), dado que el predictor X tiene el valor X.

Esta función puede ajustarse de forma sencilla con métodos de regresión lineal si se emplea su versión logarítmica, obteniendo lo que se conoce como LOG of ODDs (Rodrigo, 2020).

$$\ln\left(\frac{p(Y = k|X =)}{1 - p(Y = k|X = x)}\right) = \beta_0 + \beta_1 X$$

Gráfica 1. Regresión logística



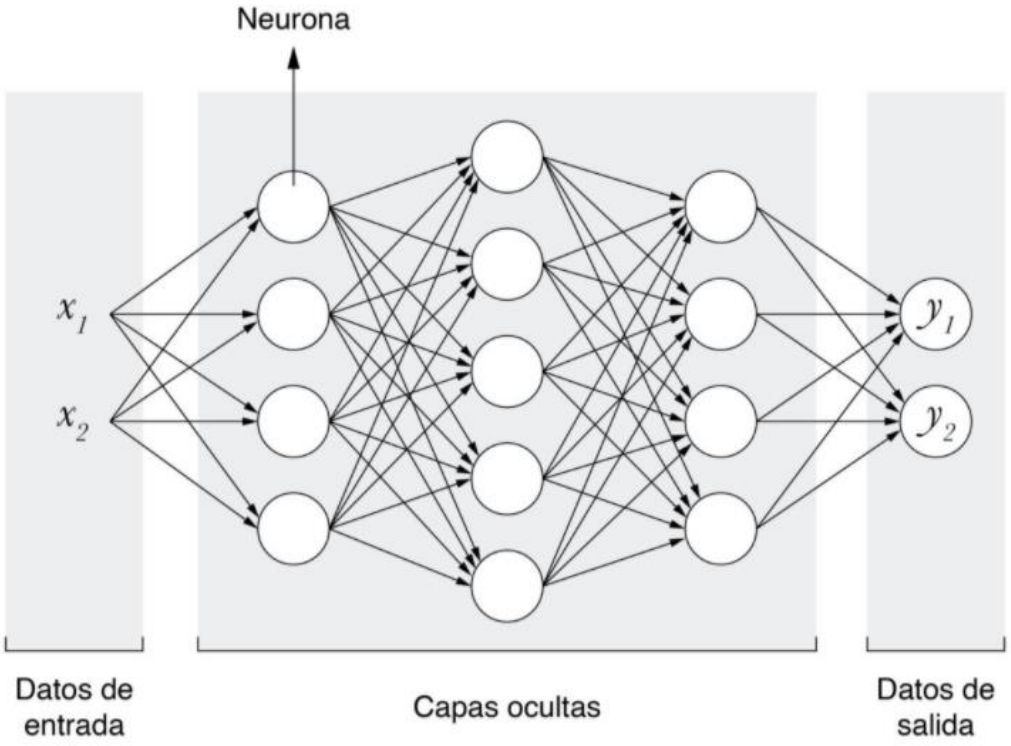
3.2.5. Redes Neuronales

Este algoritmo tal como su nombre lo indica, fue inspirado en cómo se cree que funcionan las neuronas, es por esto que su proceso está compuesto por entidades llamadas neuronas. Dicha entidad se alimenta de datos de otra neurona, generando luego de su debido proceso datos de salida que, a su vez, será enviado a tantas neuronas como capas compongan el modelo hasta llegar a la última neurona de la red, tal como se puede apreciar en la gráfica 2.

Cada conexión entre neuronas tiene diferente peso, este por lo general se representa con la letra ω

Cada vez que llegan datos a una neurona, deben pasar por la función sigmoide:

Gráfica 2. Esquema de redes neuronales



El objetivo fue construir una función que relacione los nodos de entrada con los de salida. En dicha función intervinieron las variables independientes, multiplicándose entre sí por los parámetros y sumándolos; al terminar este proceso, se les aplicó una función no lineal con el fin de complicar o hacer más flexible la función.

3.2.6. Árboles de clasificación

Los árboles son algoritmos que poseen una gran ventaja descriptiva, detectan de forma rápida las interacciones, sus resultados se pueden explicar fácilmente y captan relaciones no lineales.

Los árboles son bastante inestables, pues por algún cambio que se genere en el modelo variará por completo su resultado, también, su eficacia predictiva es baja, pues siempre se genera el mismo valor de predicción: moda para variable objetivo de clase.

3.2.7. Bagging

Diminutivo de *bootstrap aggregation* hace referencia al uso del muestreo repetido (*bootstraping)* con el fin de reducir la varianza de algunos métodos de aprendizaje estadístico, entre ellos los árboles de predicción.

Promediando un conjunto de observaciones se reduce la varianza, es por esto que una forma de reducir la varianza y aumentar la precisión de un método predictivo es obtener múltiples muestras de la población, ajustar un modelo distinto con cada una de ellas, y hacer la media de las predicciones resultantes (Rodrigo, 2020).

Bagging tiene diferentes ventajas tales como incrementar en gran medida la precisión de las predicciones, no tiene dependencia de los datos y evita la inestabilidad conocida de los árboles. Pero, construye los árboles siempre con las mismas variables, aumentando la probabilidad de que los modelos que construya sean similares entre sí.

3.2.8. Random Forest

Este método es una modificación del proceso *bagging* que consigue mejores resultados gracias a que decorrelaciona los árboles generados en el proceso, pues si la correlación es alta, la reducción de varianza que se puede lograr es pequeña.

Es por esto que dicho algoritmo selecciona aleatoriamente de *m* predictores antes de evaluar cada división. De esta forma, un promedio de *(p-m) / p* divisiones no contempla el predictor influyente, permitiendo que otros predictores puedan ser seleccionados. De esta forma decorrelaciona los árboles, logrando así una reducción de la varianza.

Los métodos de *random forest* y *bagging* siguen siendo el mismo algoritmo, la diferencia en el resultado dependerá del valor *m* escogido. Si *m=p* los resultados de ambos métodos son equivalentes (Rodrigo, 2020).

3.2.9. Gradient boosting

Dada una función de coste el algoritmo trata de encontrar el modelo que minimiza la función de coste, suele iniciarse con la mejor aproximación de la variable respuesta, se calculan los residuos y con ellos se ajusta un nuevo *weak learner* que intente minimizar la función de coste. Este proceso se repite *M* veces, de forma que cada nuevo modelo minimiza los residuos del anterior.

Dado que el objetivo de Gradient Boosting es ir minimizando los residuos iteración a iteración, es susceptible de *overfitting*. Una forma de evitar este problema es empleando un valor de regularización, también conocido como *learning rate* (λ) que limite la influencia de cada modelo en el conjunto del ensamble. Como consecuencia de esta regularización, se necesitan más modelos para formar el ensamble, consiguiendo así mejores resultados (Rodrigo, 2020).

3.2.10. Xgboost

Este algoritmo es una combinación perfecta de técnicas de optimización de software y hardware para producir resultados superiores utilizando menos recursos informáticos en el menor tiempo posible.

Optimiza los diferentes parámetros:

Paralelización: XGBoost aborda el proceso de construcción de árboles secuenciales utilizando una implementación paralelizada. Esto es posible debido a la naturaleza intercambiable de los bucles que se utilizan para construir alumnos básicos. Este

conmutador mejora el rendimiento algorítmico al compensar cualquier sobrecarga de paralelización en el cálculo.

Poda de árboles: El criterio de detención para la división de árboles dentro del marco de GBM es de naturaleza codiciosa, usa el parámetro “max_depth” podando árboles hacia atrás. Este enfoque de profundidad mejora significativamente el rendimiento computacional.

Mejoras algorítmicas:

Regularización: Penaliza a los modelos más complejos mediante la regularización LASSO y Ridge para evitar sobreajuste.

Conciencia de dispersión: Admite naturalmente características dispersas para entradas al aprender automáticamente el mejor valor perdido dependiendo de la perdida de entramiento y maneja diferentes tipos de patrones de dispersión en los datos de manera más eficiente.

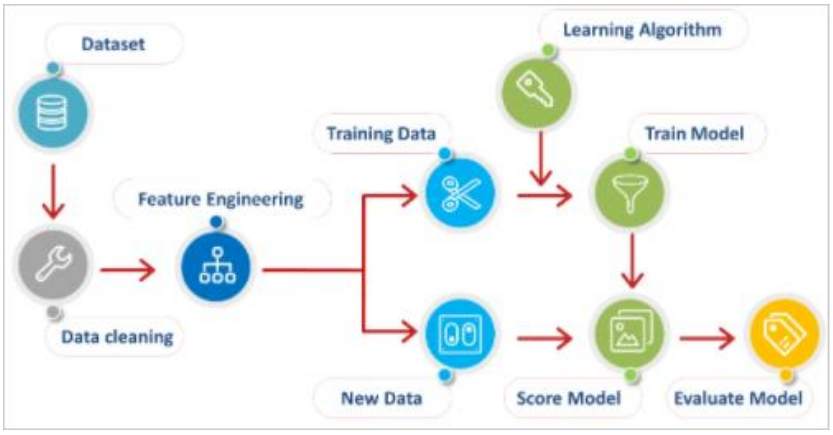
Validación cruzada: El algoritmo viene con un método de validación cruzada incorporado en cada iteración, lo que elimina la necesidad de programar explícitamente esta búsqueda y especificar el número exacto de iteraciones de refuerzo necesarias en una sola ejecución. (Morde, 2020)

3.2.11. Validación del modelo

La forma más común para medir esta capacidad es guardando observaciones para ser usados posteriormente como **validación** de la máquina. Es decir, el conjunto de ejemplos (de los que conocemos el resultado que debe dar), DD, se particiona en dos subconjuntos, Dtrain y Dval de forma que al algoritmo de entrenamiento solo se le enseñan los ejemplos de Dtrain y, una vez realizado el entrenamiento completo, se mide cómo de buenos son los resultados sobre los datos de Dval, que el algoritmo nunca ha visto y de los que conocemos el resultado que debería dar. El error cometido se mide teniendo en cuenta el resultado que la máquina devuelve sobre ellos y el dato, conocido, que debería haber devuelto. (Caparrini, 2020)

En la gráfica 3 se puede apreciar de forma sencilla el proceso.

Gráfica 3. Esquema generación y validación de modelo



4. Desarrollo

4.1. Fuente de datos

Tomando como referencia los 15 departamentos descritos en el apartado 1, se ha creado una base de datos compuesta por 593 observaciones que representan cada uno de los municipios, y 19 variables que permiten conocer la situación de violencia, socioeconómica y resultados electorales de cada observación en el año 2014.

La variable objetivo representa el partido ganador del senador elegido por cada municipio. Esta información se ha transformado en una variable binaria donde el evento de interés agrupa los partidos de derecha (centro derecha y ultraderecha). A continuación, se relacionan las variables obtenidas con su fuente y descripción:

Tabla 1. Descripción de variables

VARIABLES	DESCRIPCIÓN	TIPO	FUENTE
Departamento	ID Departamento	Continua	https://www.registraduria.gov.co/-Datos-abiertos-.html
Código	ID Municipio	Continua	
Censo electoral	Población apta para votar	Continua	
Participación	Total votos	Continua	
Abstención	Población que no votó	Continua	
Ganador circunscripción nacional	Partido político ganador / Objetivo	Binaria	
Minas antipersonas Heridos	No. personas heridas por minas antipersonas	Continua	https://www.datos.gov.co/Inclusi-n-Social-y-Reconciliaci-n/Situaci-n-V-ctimas-Minas-Antipersonal-en-Colombia/yhxn-eqqw
Minas antipersonas Muertos	No. personas muertas por minas antipersonas	Continua	
Área Coca	No. de áreas de coca identificadas	Continua	https://www.datos.gov.co/Justicia-y-Derecho/Densidad-de-Cultivos-de-Coca-2014/s9uu-92nh/data
Promedio área	Promedio de extensión de las áreas de coca identificadas	Continua	
Área máx	Área máx identificada	Continua	
Homicidios Cantidad	No. homicidios por violencia	Continua	https://www.datos.gov.co/Seguridad-y-Defensa/Delito-Homicidio/fbrt-d6qx
Cobertura en educación bruta	No. de personas estudiando	Continua	https://terridata.dnp.gov.co/index-app.html#/descargas
Población en condición de pobreza	No. de personas en condición de pobreza	Continua	
Cobertura de salud	No. de personas afiliadas a algún régimen de salud	Continua	
Afiliados régimen subsidiado	No. de personas afiliadas a reg. subsidiado de salud	Continua	
Cobertura de acueducto	No. de personas con acceso a agua potable	Continua	https://terridata.dnp.gov.co/index-app.html#/descargas
Cobertura de alcantarillado	No. de personas con acceso a tratamiento de aguas residuales	Continua	

Cobertura de energía	No. de personas con acceso a energía	Continua	
Cobertura de aseo	No. de personas con acceso a tratamiento de basuras	Continua	
Prom_edad_mujeres_víctimas	Prom edad de mujeres víctimas del conflicto	Continua	https://www.datos.gov.co/Trabajo/Informaci-n-de-los-buscadores-de-empleo-victimas-d/a4v7-78a9[1]
Total mujeres victimas	No. total de mujeres víctimas del conflicto	Continua	
Prom_edad_hombres_víctimas	Prom edad de hombres víctimas del conflicto	Continua	
Total Hombres victimas	No. total de hombres víctimas del conflicto	Continua	
Desplazados expulsados	No. total de personas desplazadas por la violencia	Continua	https://cifras.unidadvictimas.gov.co/Home/Desplazamiento
Desplazados recibidos	No. total de personas recibidas de otros municipios por desplazamiento forzado por la violencia	Continua	

Dichos datos están disponibles tanto para el año 2014 como para el 2018, desagregada por departamento y municipio.

Las bases de datos fueron modificadas mediante la herramienta *R studio* creando una única base compuesta por los 1122 municipios.

Una vez construida la base se ha trabajado en la herramienta de *SAS Miner* para realizar el proceso de exploración y depuración de la base. Por último, para ser explorada y poder entender de forma clara la información, se ha montado en la herramienta *PowerBi,* donde se ha cruzado con información geográfica de cada departamento y de esta forma ver de forma gráfica, dinámica y agregada el comportamiento de cada variable ya sea por municipio o departamento. Se explicará paso a paso el proceso desarrollado.

4.2. Análisis exploratorio

Dada la naturaleza de las variables seleccionadas, para hacer un análisis más específico se dividen en tres grandes grupos: Violencia, cobertura de servicios, y resultados de votaciones.

Para iniciar, se ha tomado la base de los 33 departamentos (incluyendo a la capital como un departamento independiente) y 1122 municipios para explorar las variables, tal como veremos en la gráfica 4 en el grupo de variables relacionadas con violencia y grupos armados, encontramos que el número de víctimas reportadas en el año 2014 en situación de búsqueda de empleo afectó un 48% más a mujeres que a hombres, en promedio de 33 y 35 años respectivamente.

[1] Dichos datos corresponden a personas que se encuentran en búsqueda de empleo, mayores de edad y victimas de situaciones de violencia propiciada en el año en mención.

Las minas antipersonas tuvieron un porcentaje menor de fallecidos frente a los heridos, siendo 42 las víctimas fatales durante ese periodo de tiempo.
En el caso de los desplazados, vemos como la cifra de los expulsados en cada municipio es nueve veces mayor que la cifra de los recibidos por los demás municipios, lo que indica que el desplazamiento externo en este año fue mucho mayor que el interno.

En cuanto a la cobertura de servicios públicos en los hogares, a nivel nacional vemos que los servicios con menos cobertura son los de alcantarillado, aseo y acueducto, y en promedio, estos tres servicios presentan baja cobertura, pues en promedio 40% de la población aún tiene dificultad para acceder a agua potable

La mayor proporción de pobreza identificada entre el total de los municipios se encuentra en un 60% del total de dicha población.

Para las elecciones del senado, encontramos 8 opciones de respuesta, donde seis son partidos políticos y dos de empate y voto en blanco. Estos partidos políticos se han clasificado de acuerdo con su ideología política, donde el uno corresponde a los partidos de centro derecha y ultraderecha, mientras que el 0 corresponde a las demás opciones.

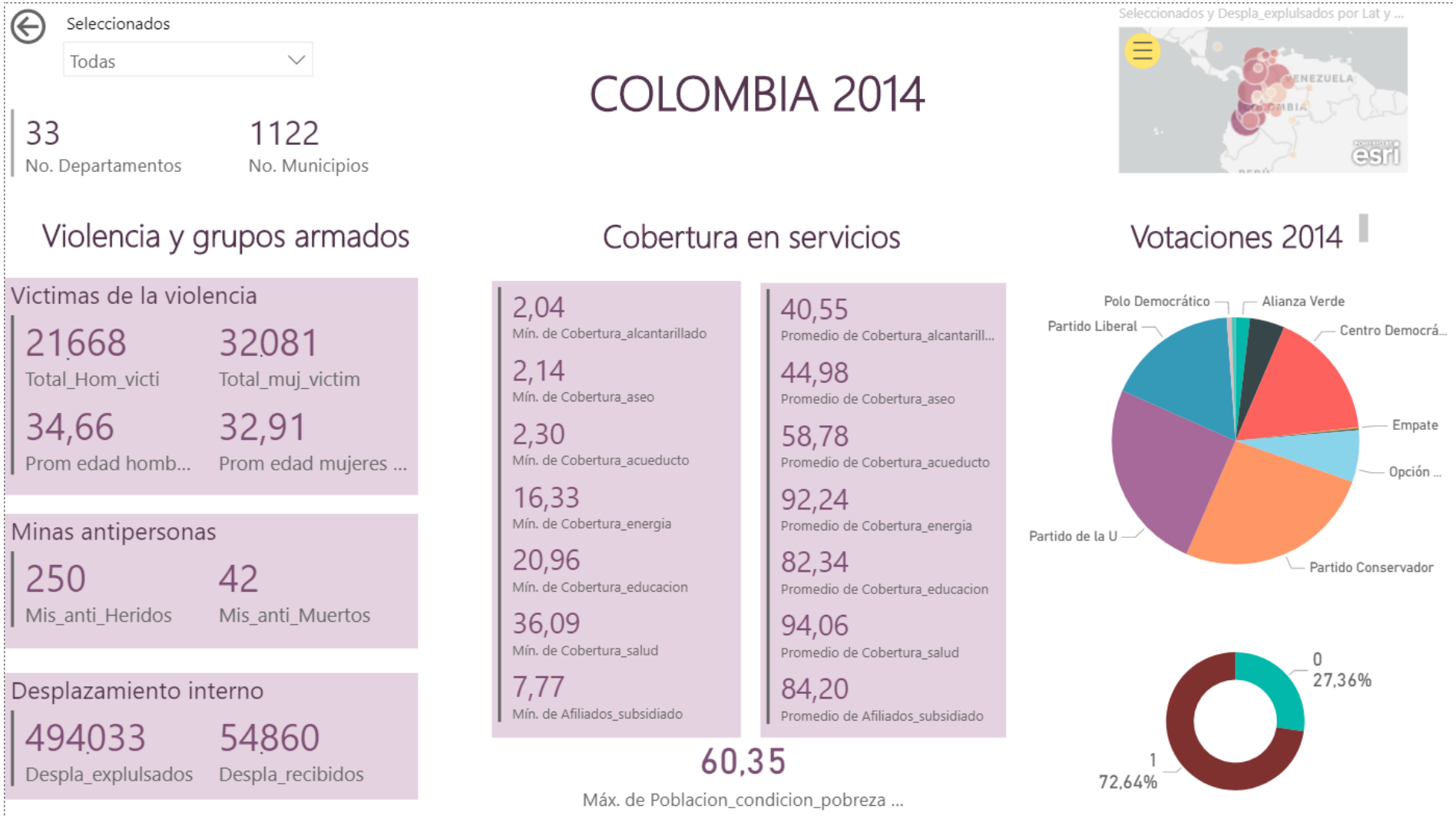
Esta información se puede ver gráficamente en la parte inferior, donde vemos que el 72,6% de los resultados corresponden a partidos de derecha. Dicha variable será nuestra variable objetivo, siendo el uno el evento de interés a predecir.

Teniendo en cuenta el ZOMAC creado por el Ministerio de Hacienda y el DNP (Presidencia de Colombia, 2017) se seleccionaron 15 de los 33 departamentos del país, que tienen más municipios afectados históricamente por la violencia en Colombia: Antioquia, Arauca, Bolívar, Caquetá, Cauca, Casanare, Córdoba, Guaviare, Chocó, Meta, Nariño, Norte de Santander, Santander, Valle del Cauca y Putumayo. No se seleccionaron solo los municipios afectados expresamente en la lista, ya que se quiere ver el comportamiento de la región. La base que se empleará para el análisis cuenta con 593 observaciones correspondientes a 593 municipios de los departamentos mencionados previamente.

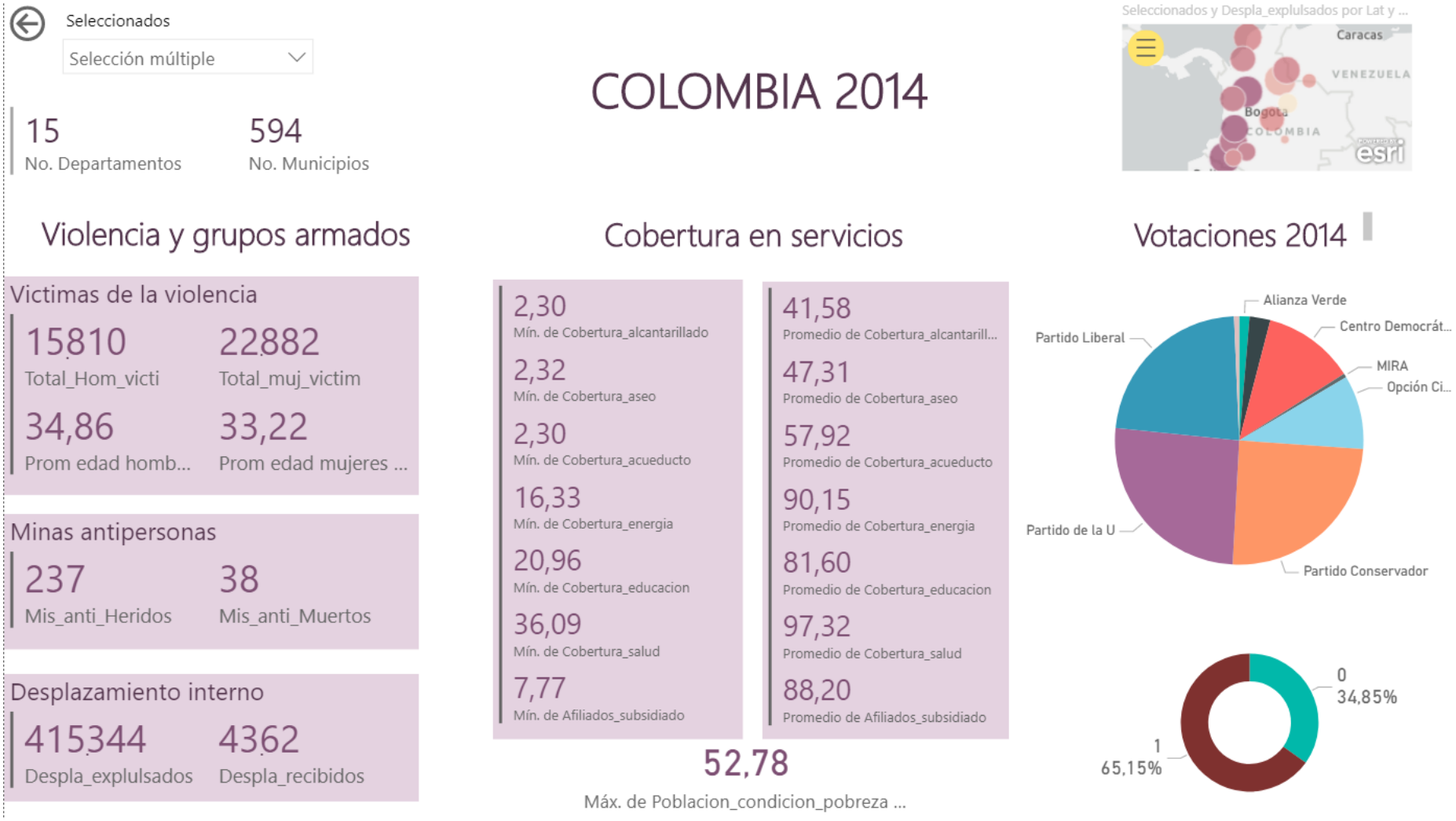
Dicho esto, en la gráfica 5. Vemos las variables solo para los municipios mencionados en el apartado anterior, donde podemos ver como las cifras concuerdan con su afectación por violencia este año, pues del total de víctimas del conflicto la muestra a analizar está representada en un 71% con respecto a las mujeres y en el caso de los hombres con un 73%. En el caso de las minas antipersona en la muestra encontramos casi el 100% de los casos reportados y en las cifras de desplazamiento vemos que el 84% de los expulsados pertenecen a estos departamentos, mientras que de los recibidos solo representa el 7%, indicando que efectivamente abandonan estos lugares buscando ciudades más seguras y diferentes a las ya mencionadas.

En cuento a la cobertura de servicios por hogares, vemos que los valores mínimos se mantienen, pero en promedio la mayoría aumenta un punto porcentual, solo vemos una reducción en los de menor cobertura que son aseo y alcantarillado. Si bien el porcentaje máximo de pobreza es menor al ya visto a nivel nacional, sigue representando más de la mitad de dicho municipio. Lo que nos indica que la muestra representa no solo la más afectada por la violencia, sino que también la de más precariedad en cuanto a atención del estado.

Gráfica 4. Datos agregados Colombia



Gráfica 5. Datos agregados departamentos seleccionados

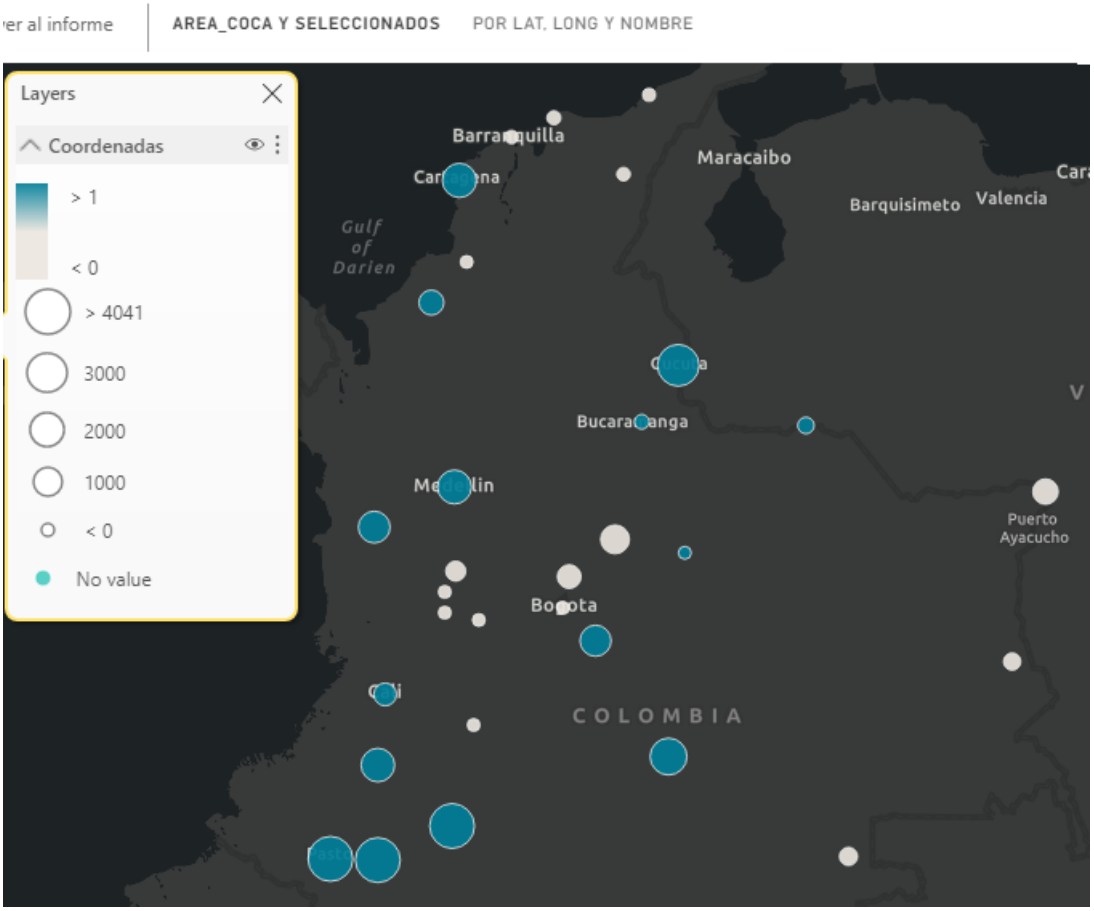


Para finalizar, vemos que a nivel electoral sigue siendo mayor la representación de la derecha con un 65,15% indicando un buen balance.

Para visualizar gráficamente como los departamentos seleccionados guardan la mayor información de las variables de violencia frente al total del país, se han seleccionado tres variables para mostrar geográficamente su relación, donde por el color azul identificaremos los departamentos con los que se trabajaron en la base y en blanco los que no.

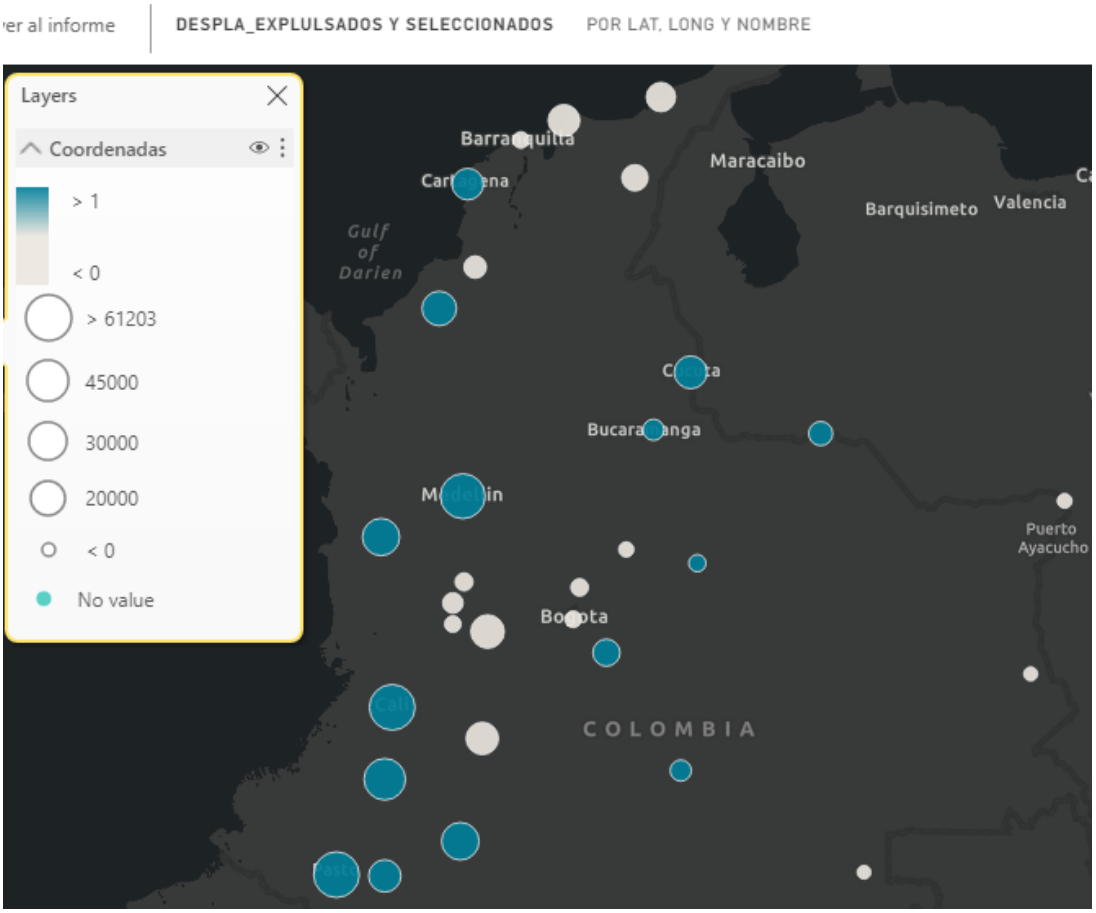
En la gráfica 6 el tamaño de las burbujas representa el número de áreas de coca que se identificaron en el año 2014 en cada uno de los departamentos, mostrando que los azules en su mayoría son de mayor tamaño que los blancos.

Gráfica 6. Mapa de áreas de coca en Colombia por departamento



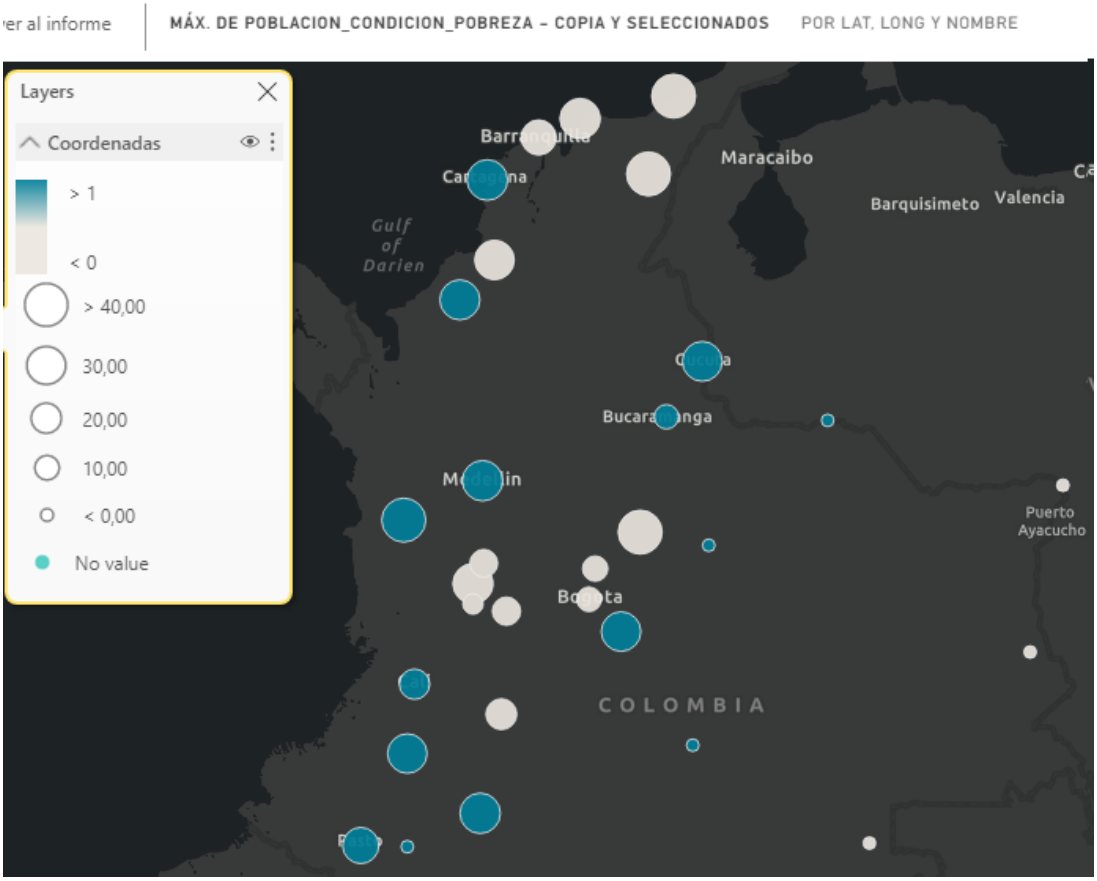
En la gráfica 7 el tamaño de las burbujas representa el número de desplazados expulsados por departamentos, mostrando una vez más que los azules son de mayor tamaño que los blancos, exceptuando en este caso los departamentos Tolima y Huila, siendo estos los círculos que en la zona centro presentan un tamaño considerable.

Gráfica 7. Mapa de desplazados expulsados en Colombia por departamento



Por último, en la gráfica 8. el tamaño de las burbujas se representará la proporción máxima de pobreza identificada por departamento, si bien las burbujas azules mantienen un tamaño importante de la información encontramos en el norte del país que cuatro departamentos más presentan proporciones altas, siendo la Guajira el departamento con un 60% de su población en condición de pobreza.

Gráfica 8. Mapa de proporción de pobreza de la población de Colombia por departamento



4.3. Depuración

Para hacer el análisis preliminar de la base y su información se cargó la base en el software *SAS Miner* con el fin de detectar posibles errores, datos atípicos y ausentes y darles un tratamiento adecuado antes de iniciar con el planteamiento del modelo de redes neuronales.

Tabla 2. Variables depuradas

VARIABLES	DESCRIPCIÓN	TIPO
Departamento	ID Departamento	Nominal
Código	ID Municipio	ID
Censo electoral	Población apta para votar	Continua
Total votos	Total votos	Continua
Abstención	Población que no votó	Continua
Ganador binario	Partido político ganador / Objetivo	Binaria
Mis anti Heridos	No. personas heridas por minas antipersonas	Continua
Mis anti Muertos	No. personas muertas por minas antipersonas	Continua
Área Coca	No. de áreas de coca identificadas	Continua
Promedio área	Promedio de extensión de las áreas de coca identificadas	Continua
Área máx	Área máx identificada	Continua
Homicidios Cantidad	No. homicidios por violencia	Continua
Cobertura en educación	No. de personas estudiando	Continua
Población en cond. de pobreza	No. de personas en condición de pobreza	Continua
Cobertura de salud	No. de personas afiliadas a algún régimen de salud	Continua
Afiliados régimen subsidiado	No. de personas afiliadas a régimen subsidiado de salud	Continua
Cobertura de acueducto	No. de personas con acceso a agua potable	Continua
Cobertura de alcantarillado	No. de personas con acceso a trat. de aguas residuales	Continua
Cobertura de energía	No. de personas con acceso a energía	Continua
Cobertura de aseo	No. de personas con acceso a tratamiento de basuras	Continua
Prom_edad_mujeres_víctimas	Prom edad de mujeres víctimas del conflicto	Continua
Total mujeres victimas	No. total de mujeres víctimas del conflicto	Continua
Prom_edad_hombres_víctimas	Prom edad de hombres víctimas del conflicto	Continua
Total Hombres victimas	No. total de hombres víctimas del conflicto	Continua
Desplazados expulsados	No. total de personas desplazadas por la violencia	Continua
Desplazados recibidos	No. total de personas recibidas de otros municipios por desplazamiento forzado por la violencia	Continua

Luego de ejecutar el nodo DMDB observamos los estadísticos de las variables donde identificamos los siguientes resultados:

Gráfica 9. Resultados nodo DMBB – Variables continuas

Variable	Etiqueta	Ausente ▲	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría
Abstencion	Abstencion	0	594	398	1023998	15177.08	60753.74	12.55956
Afiliados_re...	Afiliados_re...	0	594	773	693172	19087.69	50164.91	9.260464
Area_Coca	Area_Coca	0	594	0	1707	50.27946	177.5137	5.246852
Censo_ele...	Censo_ele...	0	594	805	1545205	27508.94	99406.64	11.59722
Cobertura_...	Cobertura_...	0	594	805	1545205	27119.08	99234.98	11.65986
Despla_ex...	Despla_ex...	0	594	0	102025	699.2323	4532.792	19.70162
Despla_rec...	Despla_rec...	0	594	0	1327	7.343434	63.00454	16.73038
Mis_anti_H...	Mis_anti_H...	0	594	0	33	0.39899	2.130023	9.977392
Mis_anti_M...	Mis_anti_M...	0	594	0	7	0.063973	0.409765	11.08478
Participacio...	Participacio...	0	594	407	586102	12331.85	39339.44	10.32745
Promedio_...	Promedio_...	0	594	0	564134.1	25363.82	62874.28	3.707459
area_max	area_max	0	594	0	5911250	207998.1	627698.9	5.398511
Cobertura_...	Cobertura_...	2	592	623.07	1534389	26433.2	99324.33	11.56244
Cobertura_...	Cobertura_...	4	590	272.48	1407617	24298.41	90523	11.38369
Cobertura_...	Cobertura_...	39	555	129.35	1335688	19884.98	82118.52	12.00136
Poblacion_...	Poblacion_...	45	549	1.1529	295134.2	3170.951	19500.64	11.03814
Cobertura_...	Cobertura_...	58	536	163.1192	7031	3416.94	2109.993	0.380956
Cobertura_...	Cobertura_...	74	520	52.0665	1282198	17936.8	81801.36	11.92077
Homicidios...	Homicidios...	126	468	1	1545	17.12607	80.63991	15.81137
PROM_ED...	PROM_ED...	271	323	20	62	34.86471	7.295629	0.745226
PROM_ED...	PROM_ED...	271	323	18	72	33.2192	7.076543	1.295313
Total_Hom...	Total_Hom...	271	323	1	2128	48.94737	207.8313	6.871185
Total_muj_...	Total_muj_...	271	323	1	4598	70.84211	350.407	9.313543

- En las variables que indican total de víctimas y promedio de edad identificamos cerca del 50% de ausentes, por lo que las cuatro variables resaltadas se rechazaron.
- En las demás variables en las que se identificaron ausentes al ser menos del 50% se decidió imputar dichos valores más adelante.
- En los intervalos de las variables no se identifican valores extraños, por lo que no se corrigieron.

Por otro lado, en las variables de clase no se encontraron valores ausentes. A continuación, se presenta los niveles de cada una:

Gráfica 10. Resultado nodo DMDB – Variables de clase

Variable	Número de niveles
G_Departamento	5
Ganador_binario	2

Rol de los datos=TRAIN

Rol de los datos=TRAIN

Rol de los datos	Nombre de la variable	Rol	Nivel	Número de ocurrencias	Porcentaje
TRAIN	Partido_Ganador	TARGET	1	389	65.5987
TRAIN	Partido_Ganador	TARGET	0	204	34.4013

Se tratará la base como balanceada al ser poca la diferencia entre las clases

4.3.1. Agrupación variable de clase

La variable “Departamento” al tener tantas categorías dificulta su análisis, por esta razón se usó el nodo de selección de variables para agrupar las diferentes categorías de la variable según su relación con la variable objetivo. A continuación, se muestran los resultados:

R-cuadrados para variable objetivo: Partido_ Ganador

Effect	DF	R-Square
Class: Departamento	14	0.329035
Group: Departamento	4	0.325934

Tabla 3. Nodo selección de variables - Agrupación variables de clase

GRUPO	CATEGORÍA	NOMBRES DEPTOS
1	95 – 50 - 23	Córdoba, Guaviare, Meta
2	5 – 13 – 27 - 54 – 85 - 76	Antioquia, Bolívar, Casanare, Chocó, Nte Sant, Valle
3	18 - 52 –81	Arauca, Caquetá, Nariño
4	19 - 86	Cauca, Putumayo
5	68	Santander

Se dejó solo la variable agrupada, pues a pesar de no tener una diferencia significativa en cuanto al R-Square, al estar por grupos nos permitió maximizar su análisis.
Para conocer la relación entre la variable objetivo y cada grupo en la siguiente gráfica se representa la cantidad de municipios que representan los valores 0 y 1 en cada uno de ellos:

Tabla 4. Departamentos agrupados

	VARIABLE OBJETIVO		
G_Departamento	**0**	**1**	**Total**
1	1	61	**62**
2	59	243	**302**
3	33	53	**86**
4	40	15	**55**
5	74	13	**87**
Total general	**207**	**385**	**593**

Los grupos 1 y 2 presentan una tendencia muy fuerte hacia el evento de interés (1), mientras que los grupos 4 y 5 presentan una relación inversa. El grupo 3 a pesar de tener una representación más alta en los valores 1, la diferencia no es tan fuerte, siendo una relación más débil con el evento de interés.

4.3.2. Tratamiento de atípicos y faltantes

Mediante distintas técnicas se buscó identificar datos atípicos en las variables input, para ello teniendo en cuenta su naturaleza y valores tales como su mediana y la distribución de los datos emplearemos el método MAD o Percentiles extremos, pues en este caso ninguna variable cumplía los requisitos de la desviación estándar. Luego de la clasificación obtuvimos los siguientes resultados:

Tabla 5. Tratamiento datos atípicos

VARIABLES	MÉTODO	TOTAL	ATÍPICOS	%	ATÍPICO
Abstencion	MAD	593	41	6,91%	
Afiliados_reg_sub	MAD	593	37	6,24%	
Area_Coca	Percentiles extremos	593	2	0,34%	Si
area_max	Percentiles extremos	593	2	0,34%	Si
Censo_electoral	MAD	593	43	7,25%	
Cobertura_acueducto	MAD	554	44	7,94%	
Cobertura_alcantarillado	MAD	520	52	10,00%	
Cobertura_aseo	MAD	536	0	0,00%	
Cobertura_educacion	MAD	475	44	9,26%	
Cobertura_energia	MAD	589	47	7,98%	
Cobertura_salud	MAD	593	41	6,91%	
Despla_explulsados	MAD	558	69	12,37%	
Despla_recibidos	MAD	558	0	0,00%	
Homicidios_Cantidad	MAD	467	41	8,78%	
Mis_anti_Heridos	Percentiles extremos	593	1	0,17%	Si
Mis_anti_Muertos	Percentiles extremos	593	2	0,34%	Si
Poblacion_en_pobreza	MAD	547	61	11,15%	
Promedio_area	Percentiles extremos	593	2	0,34%	Si
Total_votos	MAD	593	49	8,26%	

Al ser nuestro objetivo identificar datos atípicos (raros) solo se tomaron las variables que no superen el 6% de los datos, dichos valores los transformaremos en missings para luego tratarlos de manera correcta, los demás no fueron transformados.
Ahora usaremos el nodo de código SAS para crear una variable (numMissing) que nos permita contar por observación los valores ausentes, de esta forma sabremos si es necesario eliminar observaciones o no. Si tenemos observaciones con 9 o más valores ausentes eliminaremos la observación.

Gráfica 11. Código SAS - Vacíos por observación

Variable	Máximo
numMissing	9

Eliminamos una observación dado que el valor fue igual que el máximo establecido. Usamos el nodo imputar para sustituir los valores ausentes, para ello en nuestras variables de intervalo se reemplazaron los vacíos con el método de “distribución” que nos permitió mantener su variabilidad. Al tiempo, se creó una variable que nos indicaba cuántos valores se reemplazaron por observación.

4.3.3. Análisis de relación entre las variables input con la objetivo

Antes de transformar las variables y luego de depurar la base, se hizo análisis factorial con el fin de conocer las correlaciones altas y bajas existentes, de esta forma se pretendió continuar el ejercicio exploratorio el cual dio información útil para seleccionar el grupo de variables que se emplearon más adelante en los modelos.

Para este ejercicio se usó el software *SAS.* Al tener algunas variables en escala diferente se usó la base estandarizada.

Gráfica 12. Correlaciones entre variables

Correlaciones

	CENSO_	DESPLA_	ABSTE	DESPLA_	COB_SA	AFI_REG	TOTAL_V	COB_AC	COB_AL	COB_AS	COB_EDU	COB_ENE	HOM_CAN	POB_POB	AREA_CO	M_ANTI_	M_ANTI_	PROME	AREA_MA	GANADOR	G_DEPAR
CENSO_ELEC	1,000	0,226	0,996	0,083	0,999	0,954	0,989	0,780	0,940	0,172	0,994	1,000	0,863	0,874	-0,010	0,002	-0,018	-0,011	-0,007	0,038	-0,061
DESPLA_EXP	0,226	1,000	0,223	-0,011	0,227	0,307	0,228	0,153	0,189	-0,023	0,222	0,225	0,210	0,134	0,239	0,195	0,037	0,245	0,317	-0,051	-0,027
ABSTENCION	0,996	0,223	1,000	0,082	0,995	0,937	0,971	0,773	0,935	0,166	0,983	0,995	0,891	0,869	-0,005	0,005	-0,013	-0,006	-0,001	0,036	-0,054
DESPLA_REC	0,083	-0,011	0,082	1,000	0,081	0,016	0,085	0,077	0,101	0,088	0,066	0,084	-0,001	-0,018	-0,033	-0,014	-0,022	-0,042	-0,038	0,025	-0,020
COB_SALUD	0,999	0,227	0,995	0,081	1,000	0,956	0,989	0,779	0,939	0,170	0,994	0,999	0,865	0,876	-0,009	0,003	-0,017	-0,010	-0,005	0,037	-0,060
AFI_REG_SUB	0,954	0,307	0,937	0,016	0,956	1,000	0,963	0,722	0,856	0,163	0,952	0,953	0,808	0,892	0,027	0,037	-0,006	0,035	0,041	0,036	-0,090
TOTAL_VOT	0,989	0,228	0,971	0,085	0,989	0,963	1,000	0,776	0,930	0,178	0,994	0,989	0,805	0,865	-0,019	-0,003	-0,025	-0,019	-0,015	0,040	-0,072
COB_ACU	0,780	0,153	0,773	0,077	0,779	0,722	0,776	1,000	0,770	0,111	0,784	0,779	0,648	0,658	-0,032	-0,014	-0,029	-0,039	-0,035	0,054	-0,077
COB_ALCAN	0,940	0,189	0,935	0,101	0,939	0,856	0,930	0,770	1,000	0,177	0,941	0,940	0,795	0,786	-0,030	-0,016	-0,024	-0,030	-0,025	0,038	-0,054
COB_ASEO	0,172	-0,023	0,166	0,088	0,170	0,163	0,178	0,111	0,177	1,000	0,170	0,172	0,118	0,149	0,042	0,005	-0,020	0,041	0,030	0,194	-0,232
COB_EDUCA	0,994	0,222	0,983	0,066	0,994	0,952	0,994	0,784	0,941	0,170	1,000	0,994	0,825	0,866	-0,015	-0,001	-0,019	-0,019	-0,014	0,038	-0,061
COB_ENERG	1,000	0,225	0,995	0,084	0,999	0,953	0,989	0,779	0,940	0,172	0,994	1,000	0,862	0,873	-0,019	0,000	-0,020	-0,017	-0,014	0,039	-0,060
HOM_CANTI	0,863	0,210	0,891	-0,001	0,865	0,808	0,805	0,648	0,795	0,118	0,825	0,862	1,000	0,783	0,023	0,043	0,006	0,006	0,018	0,056	-0,035
POB_POBR	0,874	0,134	0,869	-0,018	0,876	0,892	0,865	0,658	0,786	0,149	0,866	0,873	0,783	1,000	-0,017	0,010	-0,009	0,000	-0,012	0,011	-0,056
AREA_COCA	-0,010	0,239	-0,005	-0,033	-0,009	0,027	-0,019	-0,032	-0,030	0,042	-0,015	-0,019	0,023	-0,017	1,000	0,318	0,205	0,629	0,721	-0,108	0,033
M_ANTI_H	0,002	0,195	0,005	-0,014	0,003	0,037	-0,003	-0,014	-0,016	0,005	-0,001	0,000	0,043	0,010	0,318	1,000	0,198	0,221	0,173	-0,070	-0,024
M_ANTI_M	-0,018	0,037	-0,013	-0,022	-0,017	-0,006	-0,025	-0,029	-0,024	-0,020	-0,019	-0,020	0,006	-0,009	0,205	0,198	1,000	0,250	0,266	-0,066	-0,034
PROM_AREA	-0,011	0,245	-0,006	-0,042	-0,010	0,035	-0,019	-0,039	-0,030	0,041	-0,019	-0,017	0,006	0,000	0,629	0,221	0,250	1,000	0,886	-0,128	-0,032
AREA_MAX	-0,007	0,317	-0,001	-0,038	-0,005	0,041	-0,015	-0,035	-0,025	0,030	-0,014	-0,014	0,018	-0,012	0,721	0,173	0,266	0,886	1,000	-0,125	-0,030
GANADOR_B	0,038	-0,051	0,036	0,025	0,037	0,036	0,040	0,054	0,038	0,194	0,038	0,039	0,056	0,011	-0,108	-0,070	-0,066	-0,128	-0,125	1,000	-0,566
G_DEPTO	-0,061	-0,027	-0,054	-0,020	-0,060	-0,090	-0,072	-0,077	-0,054	-0,232	-0,061	-0,060	-0,035	-0,056	0,033	-0,024	-0,034	-0,032	-0,030	-0,566	1,000

Se generó la matriz de correlación usando un mapa de calor con el fin de observar relaciones importantes entre variables. Se resaltaron las variables de “Censo electoral” y “cob_salud” pues en la matriz vemos que guarda una relación muy similar con todas las variables. Al estar tan correladas fueron predecibles entre sí, por lo que se eliminaron una de ellas y conservaron la misma información.

En la matriz observamos un grupo verde de variables que representan las altas correlaciones, si bien para este caso solo eliminamos una al presentar una alta correlación, se tuvieron en cuenta las demás para su posterior selección. Se ejecutó PROC FACTOR para realizar el análisis factorial.

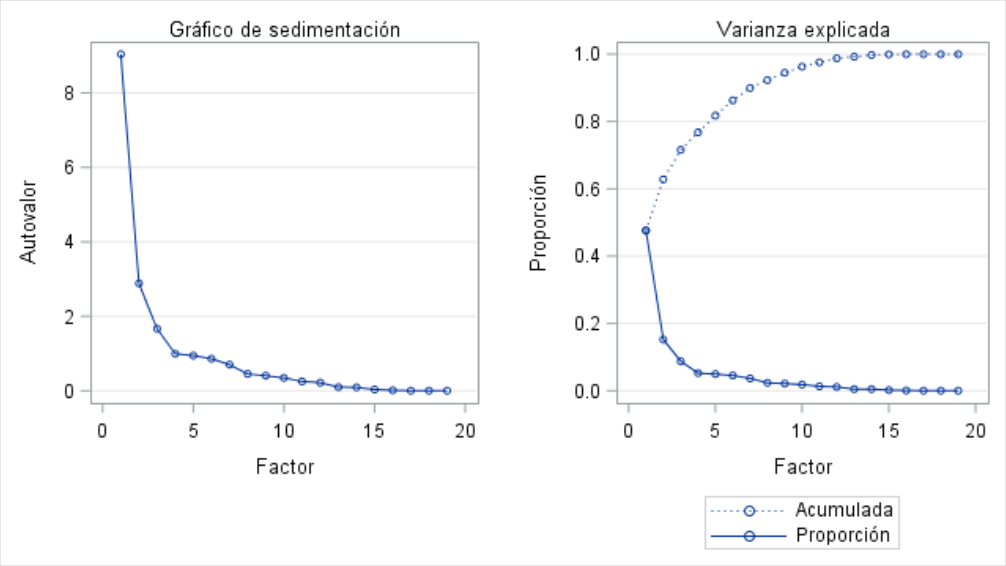
Si el objetivo del estudio fuera la reducción de las variables del conunto de datos, la medida índice de Kaiser sería el indicador ideal para conocer las variables que dada su correlación se ajustan más al estudio:

Gráfica 13. Índice Kaiser

Medida de Kaiser de suficiencia muestral: MSA total= 0.87278378																			
DESPLA_E	ABSTENCI	DESPLA_R	COB_SAL	AFI_REG	TOTAL_	COB_AC	COB_ALC	COB_ASEO	COB_EDUC	COB_ENER	HOM_CAN	POB_POBR	AREA_COC	M_ANTI_H	M_ANTI_M	PROM_A	AREA_MAX	GANADOR_	G_DEPART
0,6734	0,8133	0,2039	0,9067	0,9109	0,8286	0,9936	0,9779	0,8701	0,9411	0,8556	0,9632	0,9562	0,7184	0,5259	0,8009	0,6646	0,6140	0,5376	0,5085

KMO es igual a 0.87, indicando un ajuste alto para el análisis factorial, la variable “Desplazados Recibidos” es la que menor relación tiene entre las demás, por lo que para el estudio la eliminaremos y veremos de que forma se relacionan entre sí.

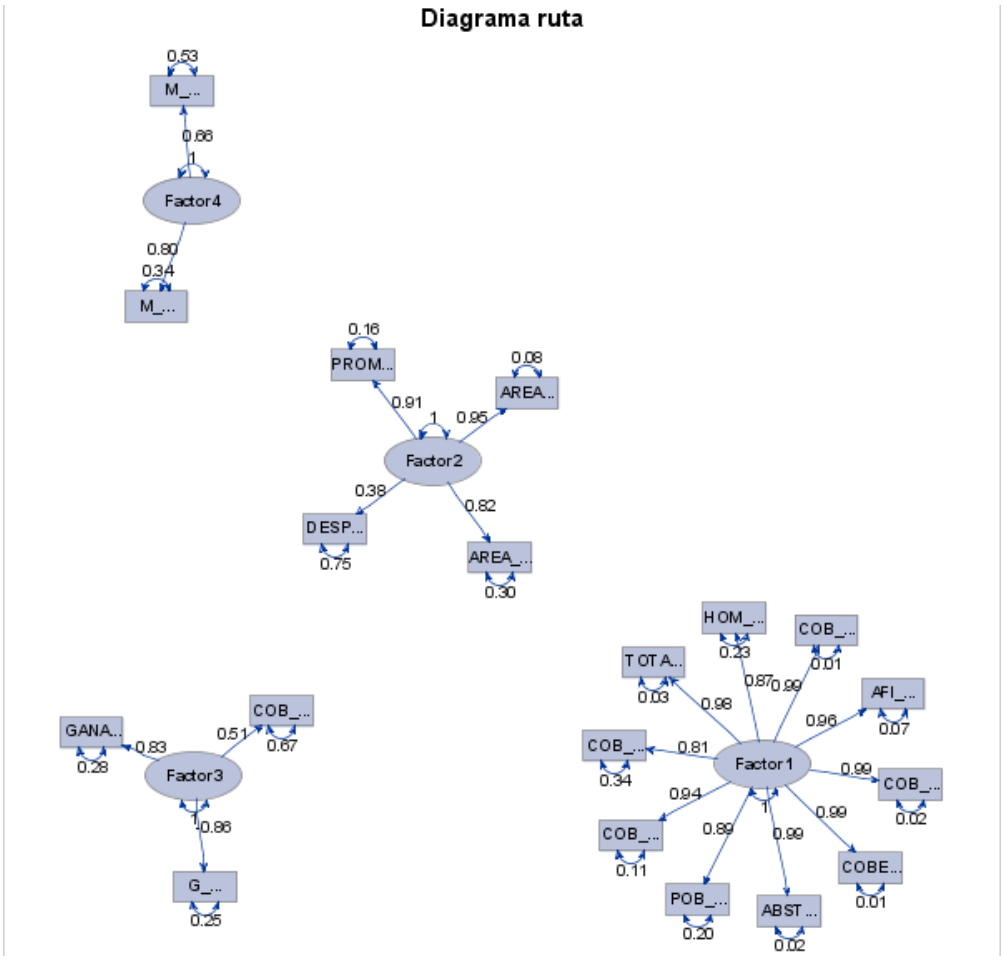
Gráfica 14. Gráfico de sedimentación



Para identificar cuántos factores se retuvieron observamos el gráfico de sedimentación, en donde se identifica el codo o el cambio drástico de tendencia. Se indica el número óptimo de factores para el conjunto de datos, por lo que tomaremos **cuatro**.

Por último, se usó la función ROTATE=VARIMAX para identificar a que factor correspondió cada una de las variables:

Gráfica 15. Diagrama de modelo factorial



Gráfica 16. Plano de modelos factoriales

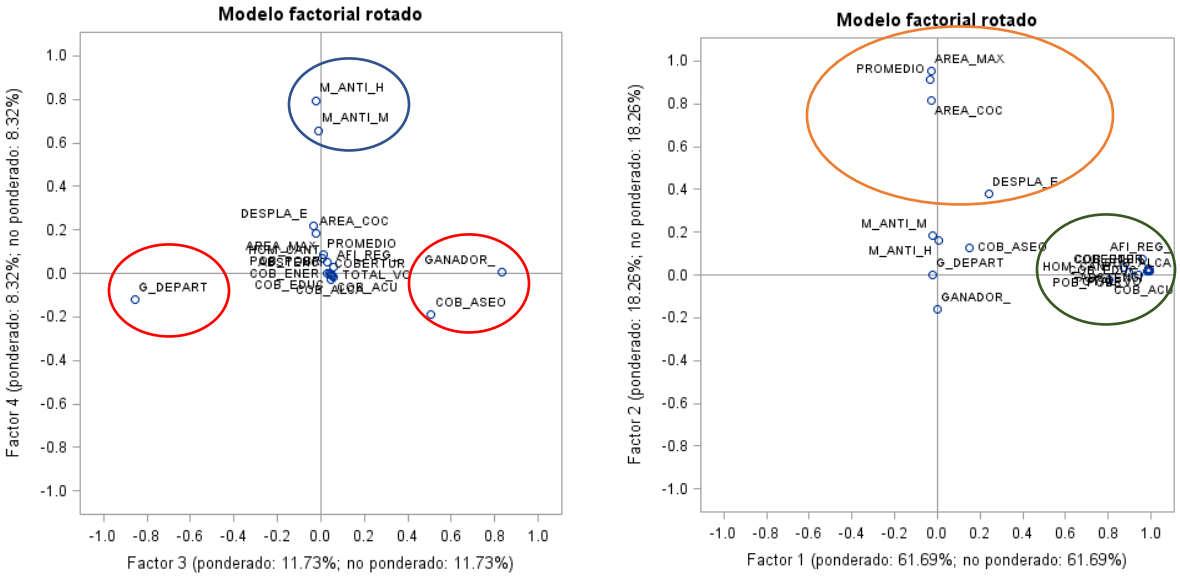


Tabla 6. Modelo factorial

Modelo factorial de rotación				
	Factor1	**Factor2**	**Factor3**	**Factor4**
DESPLA_EXPLULSADOS	0.24024	0.37697	-0.03625	0.22062
ABSTENCION	0.98958	0.02638	0.03991	-0.00094
COBERTURA_SALUD	0.99415	0.02357	0.04448	-0.00476
AFI_REG_SUB	0.95701	0.07549	0.05960	0.03007
TOTAL_VOTOS	0.98299	0.01566	0.05486	-0.01363
COB_ACU	0.80716	-0.02287	0.05740	-0.01445
COB_ALCAN	0.94104	-0.00012	0.04725	-0.02782
COB_ASEO	0.15052	0.12801	0.50854	-0.18743
COB_EDUCACION	0.98830	0.01470	0.04569	-0.00823
COB_ENERGIA	0.99370	0.01486	0.04562	-0.00722
HOM_CANTIDAD	0.87428	0.03499	0.02918	0.05040
POB_POBREZA	0.89419	0.01126	0.02776	-0.00080
AREA_COCA	-0.02903	0.81625	-0.02510	0.18581
M_ANTI_HERIDOS	0.00679	0.16086	-0.02476	0.79552
M_ANTI_MUERTOS	-0.02494	0.18542	-0.00883	0.65631
PROMEDIO_AREA	-0.03435	0.91250	0.01236	0.08487
AREA_MAX	-0.02981	0.95389	0.00801	0.06881
GANADOR_BINARIO	0.00110	-0.16209	0.83258	0.00563
G_DEPARTAMENTO	-0.02296	0.00268	-0.85705	-0.12224

Factor 1:

Variables: abstención, total votos, afiliados régimen subsidiado, cob salud, cob acueducto, cob alcantarillado, cob educación, cob energía, cantidad de homicidios, población en pobreza.

Esta compuesto por las variables que describen la participación electoral de cada municipio y a su vez, por las variables que indican la cobertura de servicios sociales básicos y fenómenos económicos como la pobreza y homicidios.

Factor 2:

Variables: Desplazados expulsados, área coca, promedio área, área máxima.

El factor dos describe las variables directamente relacionadas con la violencia, al ser municipios con mayor presencia de guerrillas, tienen mayor producción de coca y mayor indice de campesinos expulsados de sus hogares, explicando así el porqué la variable desplazados recibidos presentaba una relación baja con el grupo.

Factor 3:

Variables: Cob aseo, ganador, departamento.

"Cobertura de aseo" está muy relacionada con la variable "ganador", ambas guardan una relación inversa con la variable que representa el departamento al que pertenece cada municipio. Al estar en este factor la variable objetivo puede indicar la impotancia de las variables en los modelos de predicción a plantear.

Factor 4:

Variables: Minas antipersonas heridos, minas antipersonas muertos.

Estas dos variables describen otro fenómeno de la guerra, pero está más vinculado con los enfrentamientos entre ejercito y guerrillas, pues las minas antipersonas han sido usadas para debilitar al bando contrario y también para proteger plantaciones de cultivos ilicitos de todo tipo, siendo esta la posible explicación para la poca relación con el factor dos.

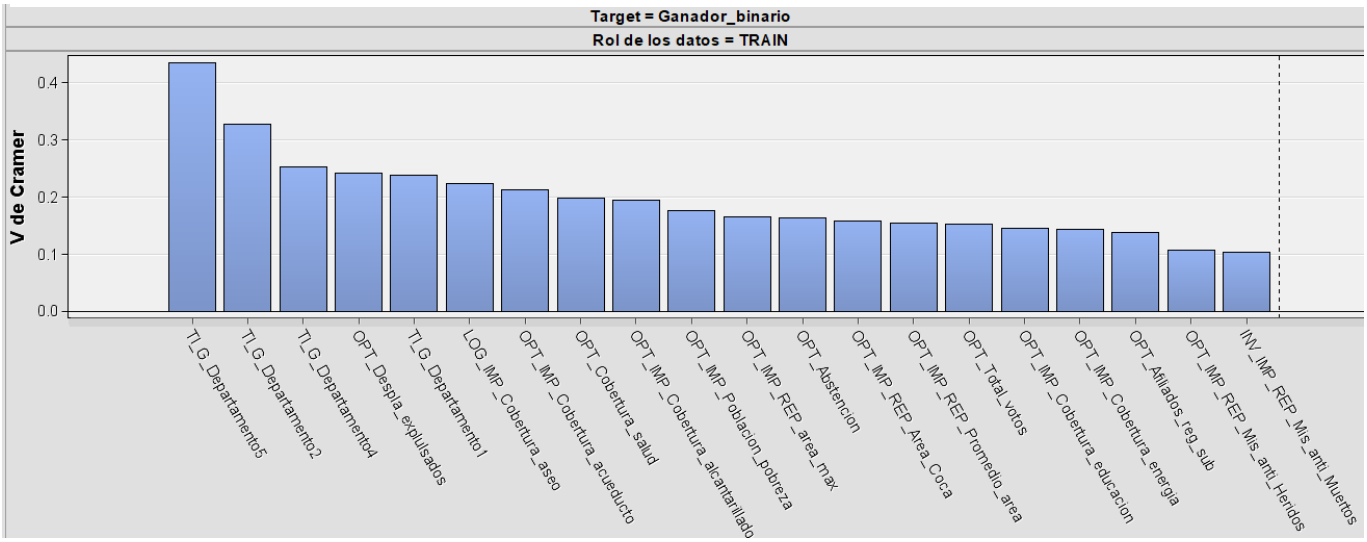
Esta información fue de gran utilidad a la hora de crear los sets de variables, pues al conocer la alta relación entre variables, se intentó seleccionar una de cada factor para optimizar la predicción.

4.3.4. Transformación de variables:

Para aumentar la relación entre la objetivo y las input se usó el nodo de transformación de variables. En dicho nodo se realizaron tres pasos:

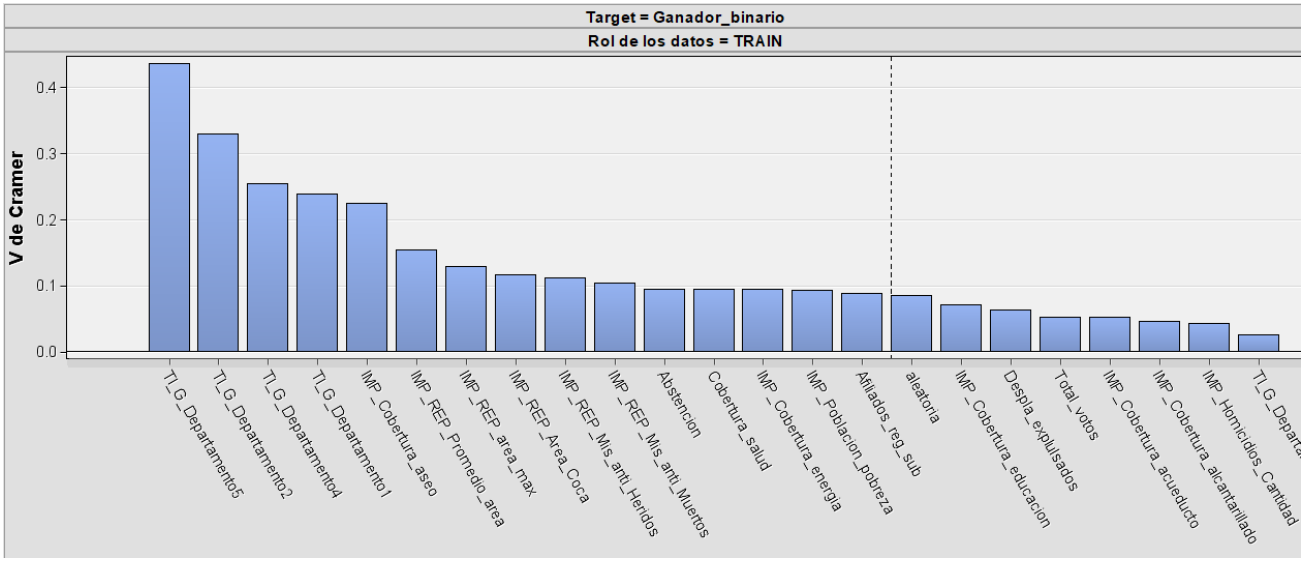
1. Se creó una variable aleatoria que fue una referencia a la hora de identificar la correlación entre cada variable con la objetivo, pues al ser una variable sintética nos permitió suponer que toda aquella que tuviera una menor relación que la aleatoria no nos aportaría información importante en el modelo.
2. Se transformó la variable categórica G_Departamento en 4 variables dummies. A partir de este momento las variables se trataron como continuas.
3. Se transformon las variables input con métodos matemáticos. El software analizó su comportamiento y según esto eligió cual fue el más indicado para aumentar la relación con respecto a la variable objetivo.

Gráfica 17. Variables transformadas - V de Cramer



Al analizar el gráfico de V de Cramer vemos que la variable aleatoria no se encuentra entre las más relacionadas, además, se observa cómo las transformadas por el software fueron las que más presentan relación con la objetivo, por esta razón se ejecutó el nodo solo con las variables originales con el fin de identificar cuáles de ellas presentaron mayor relación

Gráfica 18. Variables originales - V de Cramer



Al analizar el gráfico 18, se observa a la derecha de la aleatoria algunas variables con menor relación. Para obtener mejores efectos en los modelos que se crearon, se eliminaron dichas variables, dejando las transformadas y solo las variables originales que se encuentran a la izquierda de la aleatoria.

4.4. Selección de variables

Según el análisis realizado en el apartado 4.3.4 las variables transformadas guardan mayor relación con la variable objetivo, pero al complejizar el modelo por su interpretabilidad, se han creado dos conjuntos de datos para comparar si los resultados con las transformadas son verdaderamente altos con respecto a las originales. De esta forma se tomará la decisión más adecuada para el modelo:

1. senbis: Incluyó variables transformadas y originales. 32

```
c("P1", "P3", "P4", "P5", "P6", "P7", "P8", "P9", "P10", "P11",
"P12", "P19", "P26", "P30", "P31", "P32", "P33", "gana", "C13",
"C14", "C15", "C16", "C17", "C18", "C20", "C21", "C22", "C23",
"C24", "C25", "C27", "C28", "C29")
```

2. sen1bis: Solo tuvo las originales y variables categóricas transformadas en dummies. 15

```
c("P1", "P3", "P4", "P5", "P6", "P7", "P8", "P9", "P10", "P11",
"P12", "P30", "P31", "P32", "P33", "gana")
```

Luego de realizar la depuración y eliminar las variables que menos aportan información dada su relación con la variable objetivo, tal como se vio en el apartado 4.3.4 de transformación de variables, se crearon grupos de variables con el fin de poder evaluar que combinación fue la más adecuada para la predicción del evento de interés.

Estos grupos de variables solo se usaron en los algoritmos de regresión logística y redes neuronales, pues en ellos se tunearon los parámetros para cada grupo y de esta forma se compararon los resultados. Para crear dichos grupos, se usaron diferentes métodos y softwares para seleccionar las variables que más aportaron información al modelo. A continuación, se describe cada uno:

SAS miner:

Se usaron los nodos de Incremento gradiente y Mínimos cuadrados parciales, se registraron las variables más importantes

SAS Base:

Usando la macro *%randomselectlog* que mediante el método de stepwise permitió identificar las variables más importantes, se han registrado siete grupos de los obtenidos con la función mencionada

R Studio:

Usando la función “Selección” relacionada en anexos, se indetificó la importancia de las variables.

Todas las variables han sido registradas en una tabla organizada según la frecuencia en la que fueron seleccionadas las variables en los diferentes métodos.

Las variables que en casi todos los modelos aparecieron fueron las relacionadas con el departamento del municipio, evidenciando su relación con el análisis previo realizado en el apartado 4.3.3. En el análisis factorial, se observó cómo esta variable tiene una alta probabilidad de predicción de la objetivo, por lo que fue importante que hiciera parte de los grupos de variables.

Si bien las variables transformadas fueron seleccionadas más veces que las originales, se crearon grupos de variables donde solo estuvieron incluidas las no transformadas, esto con el fin de validar la diferencia en la predicción y de hacer más sencilla la interpretación del modelo.

A continuación, en la tabla 7 se puede observar las diferentes variables seleccionadas:

Tabla 7. Selección de variables en cada software

Software		SAS BASE					R STUDIO		SAS MINER		
Variables[2]		Rand Sel					R AIC	R BIC	Min cuadrados Parciales	Incregradiante	TOTAL
G_Departamento	P4										0
IMP_REP_Mis_anti_Muertos	P10										0
OPT_Abstencion	C13										0
OPT_Afiliados_reg_sub	C14										0
OPT_Cobertura_salud	C15										0
OPT_IMP_Homicidios_Cantid	C22										0
OPT_IMP_Poblacion_pobreza	C23										0
INV_IMP_REP_Mis_anti_Muer	P26										0
Abstencion	P1						X				1
Cobertura_salud	P3									X	1
IMP_Cobertura_energia	P6									X	1
IMP_Poblacion_pobreza	P7									X	1
IMP_REP_Mis_anti_Heridos	P9								X		1
OPT_IMP_Cobertura_alcantar	C18									X	1
IMP_Cobertura_aseo	P5						X		X		2
IMP_REP_Promedio_area	P11								X	X	2
IMP_REP_area_max	P12								X	X	2
LOG_IMP_Cobertura_aseo	P19								X	X	2
OPT_IMP_Cobertura_educac	C20									X	1
OPT_IMP_Cobertura_energia	C21			X							1
Afiliados_reg_sub	P4						X		X	X	3
TI_G_Departamento2	P31								X	X	2
OPT_Despla_explulsados	C16									X	1
OPT_IMP_REP_Area_Coca	C24					X	X				2
IMP_REP_Area_Coca	P8					X	X		X	X	4
OPT_IMP_Cobertura_acuedu	C17						X		X	X	4
OPT_IMP_REP_area_max	C28	X									2
OPT_Total_votos	C29	X				X	X			X	4
OPT_IMP_REP_Mis_anti_Heri	C25	X		X	X	X	X				6
OPT_IMP_REP_Promedio_area	C27		X	X		X	X	X			6
TI_G_Departamento1	P30	X	X			X	X	X	X		7
TI_G_Departamento4	P32	X	X	X		X	X	X	X		9
TI_G_Departamento5	P33	X	X	X	X	X	X	X	X		10

Una vez se verificó la frecuencia de las variables, se crearon los diferentes grupos que fueron de referencia. Estos se pueden detallar en la tabla 8.

Los grupos 7,8,9,12,14 y 17 solo incluyen variables originales sin transformar, en el grupo 8 se seleccionaron las variables que representan a cada factor, tal como se analizó en el apartado 4.3.3 en el análisis factorial.

[2] Las variables que fueron transformadas en el software *Miner* han sido señaladas con el prefijo C y resaltadas en verde.

Tabla 8. Grupos de variables

Variables		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
G_Departamento	P4																				
IMP_REP_Mis_anti_Muertos	P10							1													
OPT_Abstencion	C13																				
OPT_Afiliados_reg_sub	C14																				
OPT_Cobertura_salud	C15																				
OPT_IMP_Homicidios_Cantid	C22																				
OPT_IMP_Poblacion_pobreza	C23																				
INV_IMP_REP_Mis_anti_Muer	P26																				
Abstencion	P1																				
Cobertura_salud	P3								1												
IMP_Cobertura_energia	P6										1										
IMP_Poblacion_pobreza	P7										1										
IMP_REP_Mis_anti_Heridos	P9								1									1			
OPT_IMP_Cobertura_alcantar	C18																				
IMP_Cobertura_aseo	P5	1								1					1						
IMP_REP_Promedio_area	P11														1						
IMP_REP_area_max	P12							1	1	1	1				1						
LOG_IMP_Cobertura_aseo	P19																			1	
OPT_IMP_Cobertura_educac	C20																			1	
OPT_IMP_Cobertura_energia	C21				1															1	
Afiliados_reg_sub	P4														1						
TI_G_Departamento2	P31														1	1	1				
OPT_Despla_explulsados	C16															1			1		1
OPT_IMP_REP_Area_Coca	C24										1										
IMP_REP_Area_Coca	P8										1		1		1			1			
OPT_IMP_Cobertura_acuedu	C17	1			1						1					1	1		1	1	1
OPT_IMP_REP_area_ma	C28	1	1			1	1									1	1		1		
OPT_Total_votos	C29	1		1	1						1			1							
OPT_IMP_REP_Mis_anti_Heri	C25	1	1			1	1				1			1		1					1
OPT_IMP_REP_Promedio_area	C27			1			1					1		1		1			1		
TI_G_Departamento1	P30	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1
TI_G_Departamento4	P32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TI_G_Departamento5	P33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Para empezar, se ejecutaron los algoritmos en los que se usaron los grupos descritos anteriormente.

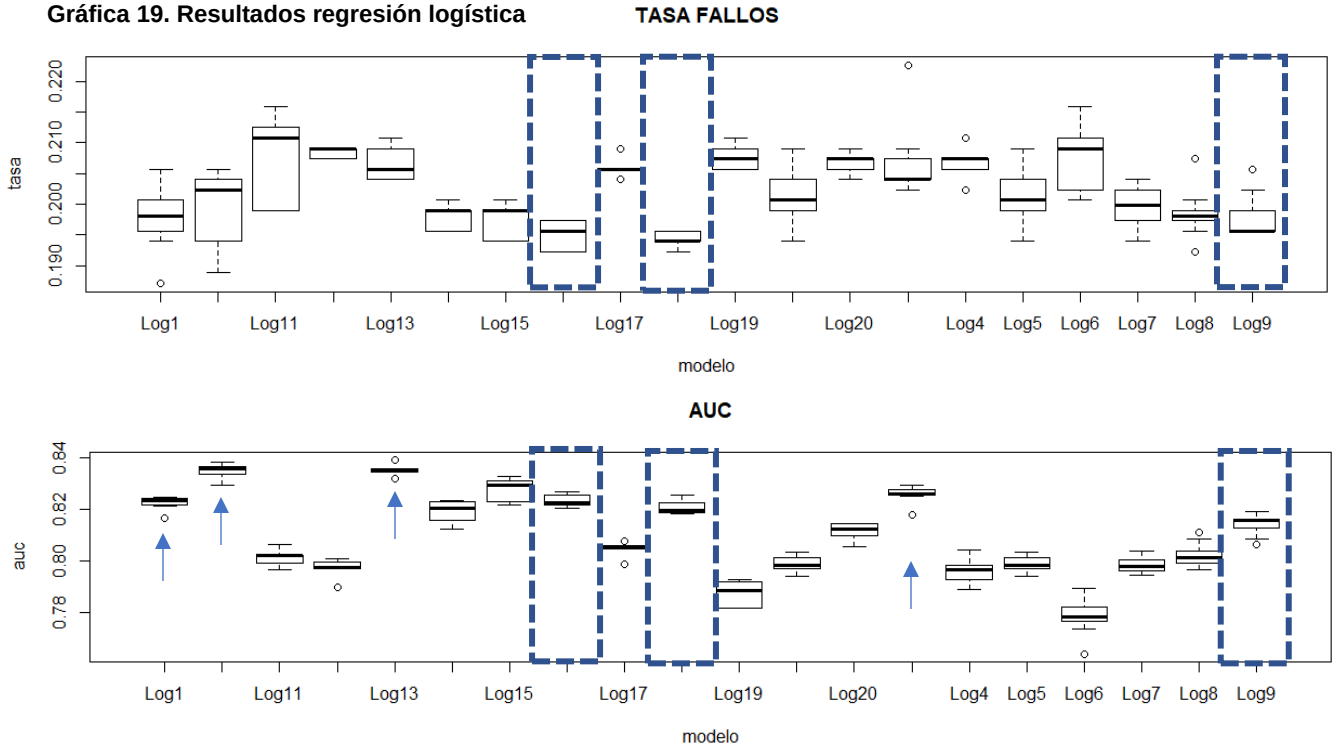
4.5. Modelado

Usando el software *R studio,* se ejecutaron diferentes algoritmos explicados en el apartado 4.4. Selección de variables, con el fin de buscar el que arrojara mejores resultados.

4.5.1. Regresión Logística

Una vez construidos los grupos de variables, se emplea la función *cruzadalogistica* en el software seleccionado, con el fin de ver cuál de ellos obtiene menos error. A continuación, los resultados:

Gráfica 19. Resultados regresión logística



Los modelos indicados con las flechas a pesar de tener un alto *AUC* su tasa de fallo presenta un alto sesgo y alta varianza, por lo que no son la mejor opción. Mientras que los modelos seleccionados en la línea punteada presentan un área bajo la curva mayor que los demás y su tasa de fallos es la más baja en comparación de los otros modelos.

Log14: 9 variables (Cob aseo, promedio área, área máx, afiliados reg sub, área coca y los cuatro grupos de departamentos)

Log18: 7 variables (Desplazados expulsados, cob en acueducto, área máx, promedio de área y tres de los cuatro departamentos) si bien obtiene buenos resultados, tiene más variables que los demás y al tener variables transformadas pierde un poco de interpretabilidad, además, contiene dos variables muy relacionadas.

Log9: 5 Variables originales (Cobertura de aseo, área máx y 3 de los cuatro departamentos, al ser todas relacionadas y sin repetir información es un modelo aceptable.

Se compararon los modelos 18 y 9 al tener menos variables, coherentes con el modelo y buenos resultados.

4.5.2. Redes neuronales

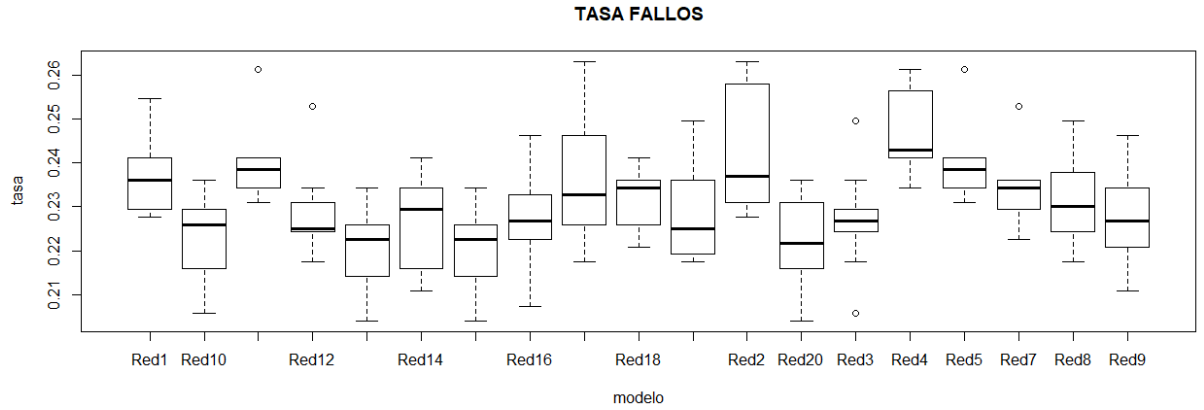
En este algoritmo encontramos múltiples parámetros de los que dependió la óptima ejecución del modelo, en el software R se usó la librería *caret* para tunear dichos parámetros, ya que para cada modelo fueron valores diferentes que según las variables tuvieron un mejor resultado. Los valores obtenidos fueron los siguientes:

Tabla 9. Resultados de parámetros tuneados.

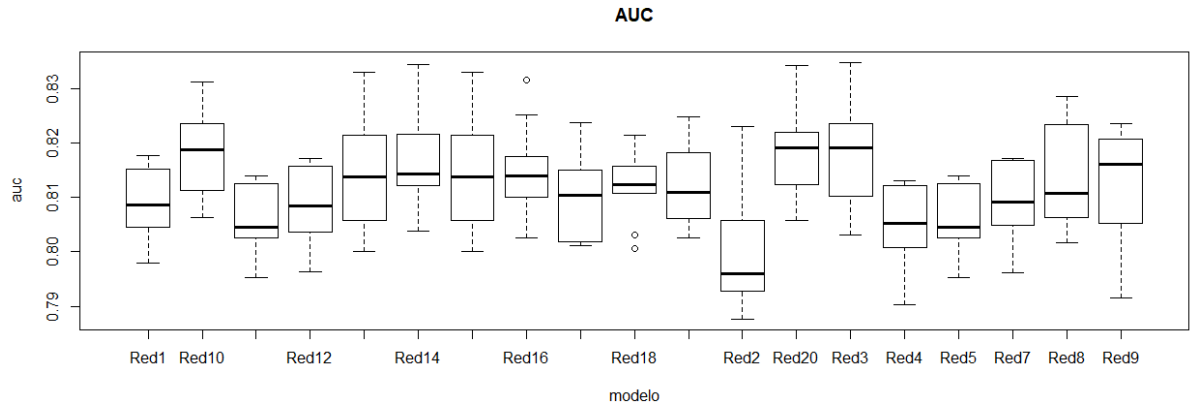
GRUPO DE VARIABLES	BAG	SIZE	DECAY
Grupo 1 P30+P32+P33+C21+C24+C27,	F	6	0.1
Grupo 2 P32+P30+P33+C28+C25,	F	5	0.001
Grupo 3* P30+P32+P33+C25+C27+C29,	F	3	0.1
Grupo 4 P30+P32+P33+C27+C29,	F	4	0.01
Grupo 5 P30+P32+P33+C25+C28,	F	3	0.01
Grupo 6 P30+P31+P32+P33+C17+C28,	F	3	0.1
Grupo 7 P8+P9+P30+P32+P33,	F	8	0.1
Grupo 8 P3+P12+P9+P30+P32+P33,	F	5	0.1
Grupo 9 P5+P12+P30+P32+P33,	F	3	0.01
Grupo 10 P6+P7+P8+P12+P30+P32+P33+C24+C17+C29+C25,	F	3	0.1
Grupo 11 P30+P32+P33+C27,	F	3	0.01
Grupo 12 P8+P30+P32+P33,	F	3	0.01
Grupo 13 P30+P32+P33+C25+C27+C29,	F	3	0.001
Grupo 14 P4+P5+P8+P11+P12+P30+P31+P32+P33,	F	3	0.1
Grupo 15 P30+P32+P33+C16+C17+C28+C16+C25+C27,	F	3	0.001
Grupo 16 P30+P31+P32+P33+C17+C28,	F	3	0.1
Grupo 17 P8+P9+P30+P32+P33,	F	7	0.1
Grupo 18 P30+P32+P33+C16+C17+C28+C27,	F	5	0.1
Grupo 19 P19+P30+P32+P33+C17+C20+C21,	F	4	0.1
Grupo 20 P30+P32+P33+C16+C17+C25,	F	3	0.001

*El grupo 3 es igual que el grupo 13, se ha ejecutado dos veces con parámetros diferentes para evaluar su efectividad.

Gráfica 20. Resultados redes neuronales



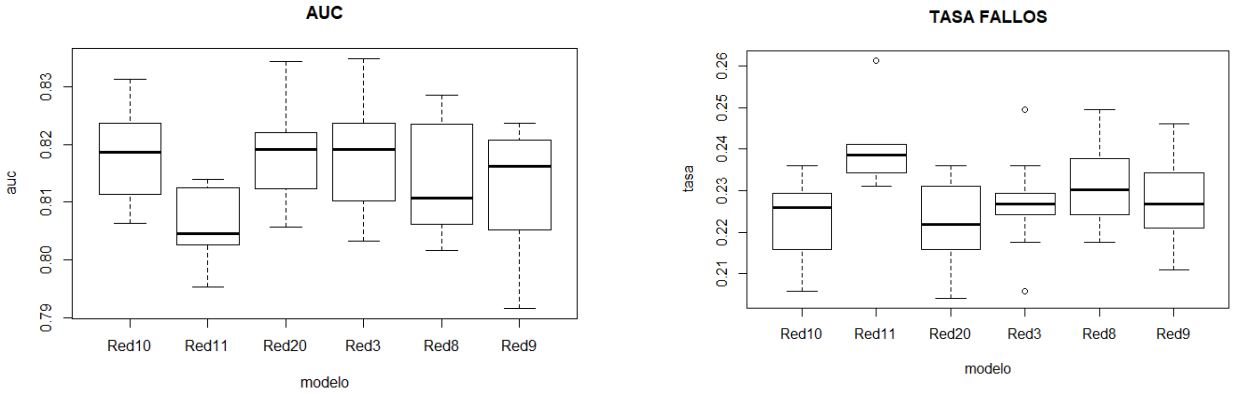
Una vez obtenemos dichos valores, se ejecuta la función *cruzadaavnnetbin* en donde establecemos los valores tanto de los parámetros como los de las variables y graficamos sus resultados:



Ya que los resultados no se diferencian de forma drástica entre modelos, se tomaron los mejores, seleccionando el más adecuado para el estudio.

Los mejores:

Gráfica 21. Resultados mejores modelos redes neuronales



El modelo 10 al tener tantas variables y algunas sin tener mucha relación entre sí no fue tenido en cuenta; al excluirlo, los dos con más área bajo la curva fueron el 20 y el 3. Las diferencias que presentan en sesgo no fueron significativas. En cuanto al número ambos contaron con seis variables, con la única diferencia que en el grupo 3 se encontraron la variable “total votos”, dicho valor, representa una situación momentánea de cada municipio, por lo que puede llegar a variar en el futuro y cambiar el sentido del modelo.

Ambos grupos se compararon con los mejores modelos para evaluar su comportamiento.

4.5.3. Bagging – Random Forest

Otro algoritmo que permitió obtener resultados efectivos fueron los árboles de clasificación. Estos dividen los datos en regiones basadas en intervalos de las variables independientes. Se encontraron regiones que minimizaron la función de error, en este caso fue la tasa de clasificación.

Tal como se describió en el apartado 4.4 Selección de variables, se compararon los resultados de la base con variables transformadas con las originales.

Mediante la librería *caret*, se tunearon tres parámetros importantes de los árboles: cantidad de variables a usar (*mtry*), mínimo tamaño final por nodo (*nodesize*) y el tamaño de la muestra (*sampsize)*. Luego de realizar diferentes combinaciones se creó una tabla donde se pudieron apreciar los mejores resultados obtenidos:

Tabla 10. Resultados parámetros tuneados Random Forest

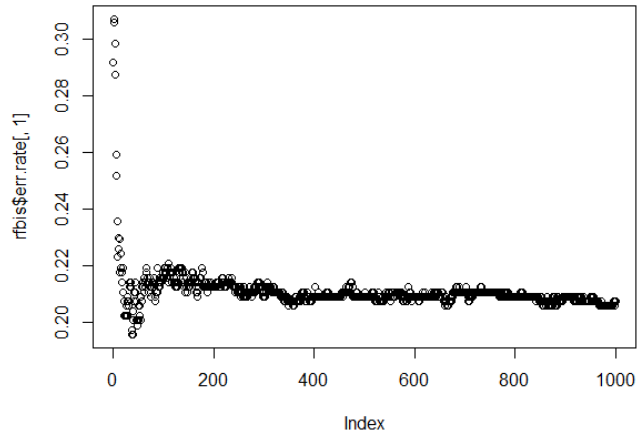
	Base Datos Originales (sen1)		
sampsize	200	300	400
nodosize	15	10	15
mtry	Accuracy	Accuracy	Accuracy
3	0.8044168	0.7993493	0.7958860
4	0.8044168	0.7976601	0.8010563
5	0.8010498	0.7976487	0.8026887
6	0.7993493	0.7959596	0.8026772
7	0.7942930	0.7976601	0.8077679
8	0.7942930	0.8044055	0.8212477
9	0.7926038	0.8010271	0.8060787
10	0.7926038	0.8027050	0.8077679
11	0.7976601	0.7959709	0.8077564
12	0.7875476	0.8077725	0.8060901
13	0.7892368	0.8010158	0.8128585
14	0.7943384	0.7976487	0.8077566
15	0.7858698	0.8027050	0.8060787

	Base con datos transformados (sen)		
sampsize	200	300	400
nodosize	20	10	20
mtry	Accuracy	Accuracy	Accuracy
5	0.7942535	0.7858991	0.8025862
7	0.7992872	0.7808543	0.8075857
9	0.7976093	0.7842326	0.8025522
11	0.8009422	0.7842101	0.8025522
13	0.7958973	0.7959891	0.8025522
15	0.7942308	0.7909669	0.8008630
17	0.7959087	0.7977354	0.8059079
19	0.7925303	0.7960461	0.8008517
21	0.7975865	0.8027917	0.8025523
23	0.7975752	0.7993790	0.8025409
25	0.7925074	0.7993788	0.8008517
30	0.7858189	0.7977124	0.8008860
32	0.7874854	0.8061245	0.8008745

Se han resaltado las Tasa de acierto *Accurancy* más altas, para tener en cuenta dichos valores en el modelo ganador. Cabe aclarar que los resultados en donde *mtry* es igual al total de variables fue el resultado del algoritmo ***Bagging*.**

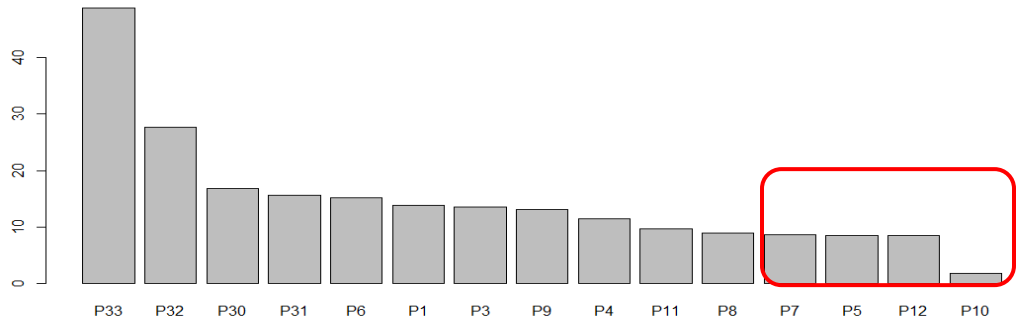
Otro factor importante para la evaluación del modelo fue la cantidad de árboles. En la gráfica 22 se puede visualizar cómo el error se estabiliza luego de los 300 árboles aproximadamente, por lo que en los modelos a comparar se pusieron 500 árboles en todos.

Gráfica 23. Estabilización del error



En cuanto a la importancia de variables, a continuación, se muestran los resultados para cada base de datos. Teniendo en cuenta esta información se eliminaron las que menor información aportaban al modelo.

Gráfica 22. Importancia de variables originales



VARIABLES	NO	YES	MEAN DECREASE ACCURACY
P33	479522453	27370790	49149145
P32	255312690	15659025	25589145
P6	-19410041	18219186	19303107
P30	161881685	17136871	19119737
P31	157488715	11523637	18126530
P3	-19337468	15439848	16680037
P1	-27527104	15688153	15696704
P4	-0.3078986	13322568	14795285
P5	69743180	9404482	12400010
P11	59979116	5600602	10631849
P9	55364359	9178565	10020742
P7	39011832	6702857	9008020
P12	51279863	3745136	8004393
P8	48211751	3454759	7609625
P10	-0.7935012	1742992	1033610

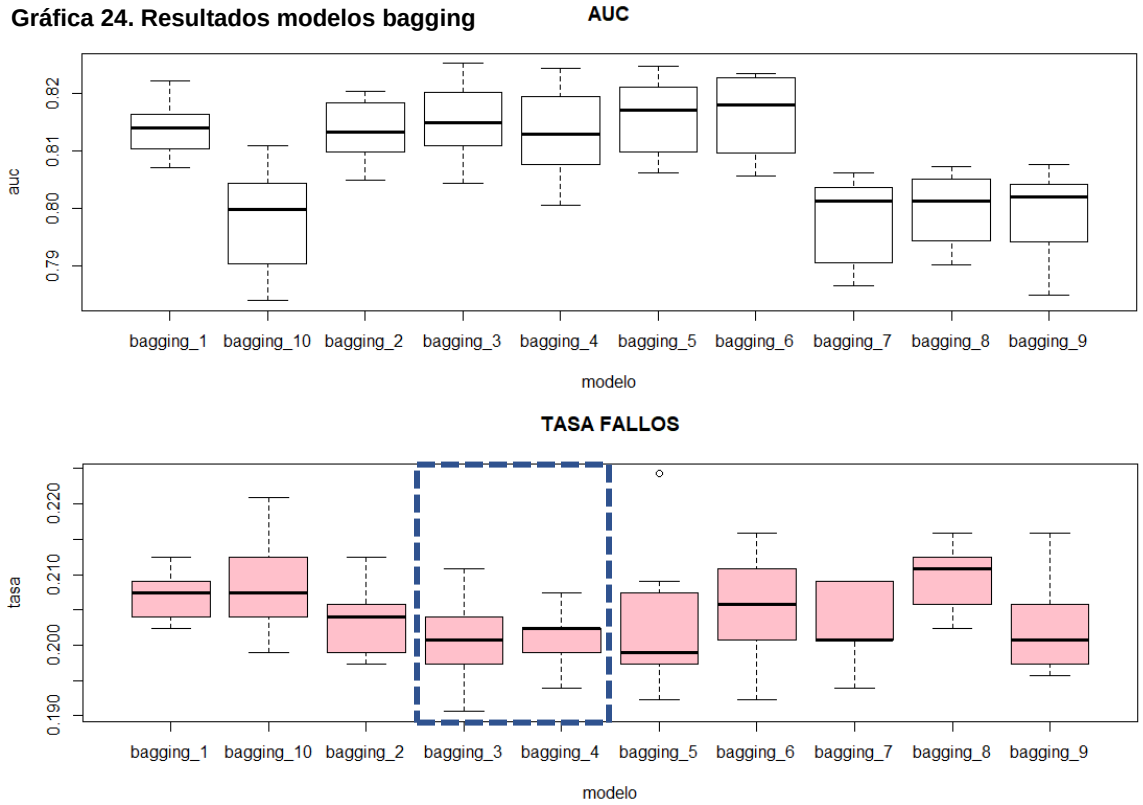
Tabla 11. Importancia de variables transformadas

VARIABLES	NO	YES	MEAN DECREASE ACCURACY
P33	4375271185	321103732	4471630199
P32	2824598439	207273438	2773675985
P31	1421842773	90934276	1428158438
P30	1268635717	131120357	1392712291
P6	-740707373	133684754	1297206865
P1	-400262190	123071049	1266442215
P3	-365128295	116865372	1232937381
P7	0.65087279	113843398	1112500044
P19	-160939437	100046094	1106347060
P5	-0.80890095	95381139	1094312939
C2902:4631-high	-371794767	103151290	1074438070
P4	-339014192	98506050	1033607540
C2502:4.5-high	698874127	100034801	1028808683
P9	426777941	85334406	932133664
P11	399932884	70038091	903875075
P12	119622877	70328964	809508354
P8	345574026	52888544	696126168
C2702:72649.43-high	342210737	54995293	655267423
C2402:104.5-high	314692517	58399791	633360420
C2802:614310-high	338899824	41137930	540990250
C1603:1026.5-high	-0.37138638	50444849	488570955
C1302:1000.5-high, MISSING	0.05795906	48007758	487338139
C1703:16316.756-56275.345	529841965	12137126	460051523
C2002:7956.4004-high	-121469914	39473048	430863026
C2102:12239.638-high	-220784957	39018886	405637409
C1602:6.5-1026.5, MISSING	-123029622	43926848	388793458
C2303:3522.4392-63945.77	228290311	21932541	313396665
C2302:36.42485-3522.4392, MI	-262262757	32920570	281176673
P26	251792214	13069599	238313633
C1502:2132.5-40743.5, MI	129775364	14221970	189837194
C1503:40743.5-86136.5	194023518	11605047	185140469
C1402:2240-high, MISSING	0.90740412	14258301	141073353
C1802:1089.8274-high, MI	167764509	-0.3475422	139805065
P10	103117686	0.3683201	122301181
C1504:86136.5-high	-0.27169016	0.7554653	0.54143422
C1704:56275.345-high	187496830	-0.7022799	0.16348703
C2202:1.5-high, MISSING	137506123	-0.6042369	0.10629176
C1702:1338.3704-16316.756, M	-0.58286156	0.1757928	-0.08433225
C2304:63945.77-high	-0.52671470	-0.1054778	-0.38692219

Tabla 12. Parámetros por modelo bagging

MODELO	BAG 1	BAG 2	BAG 3	BAG 4	BAG 5	BAG 6	BAG 7	BAG 8	BAG 9	BAG 0
Base	Senbis	Senbis	Senbis	Senbis	Senbis	Senbis	Sen1bis	Sen1bis	Sen1bis	Sen1bis
Variables	Todas	Todas	Todas	Todas	Todas	Todas	Orig	Orig	Orig	Orig
NTREE	500	500	500	500	500	500	500	500	500	500
NODESIZE	20	20	10	5	20	10	5	15	10	20
SAMPSIZE	200	200	300	200	400	400	200	200	300	400
MTRY	32	19	32	32	32	32	17	17	17	17

Gráfica 24. Resultados modelos bagging



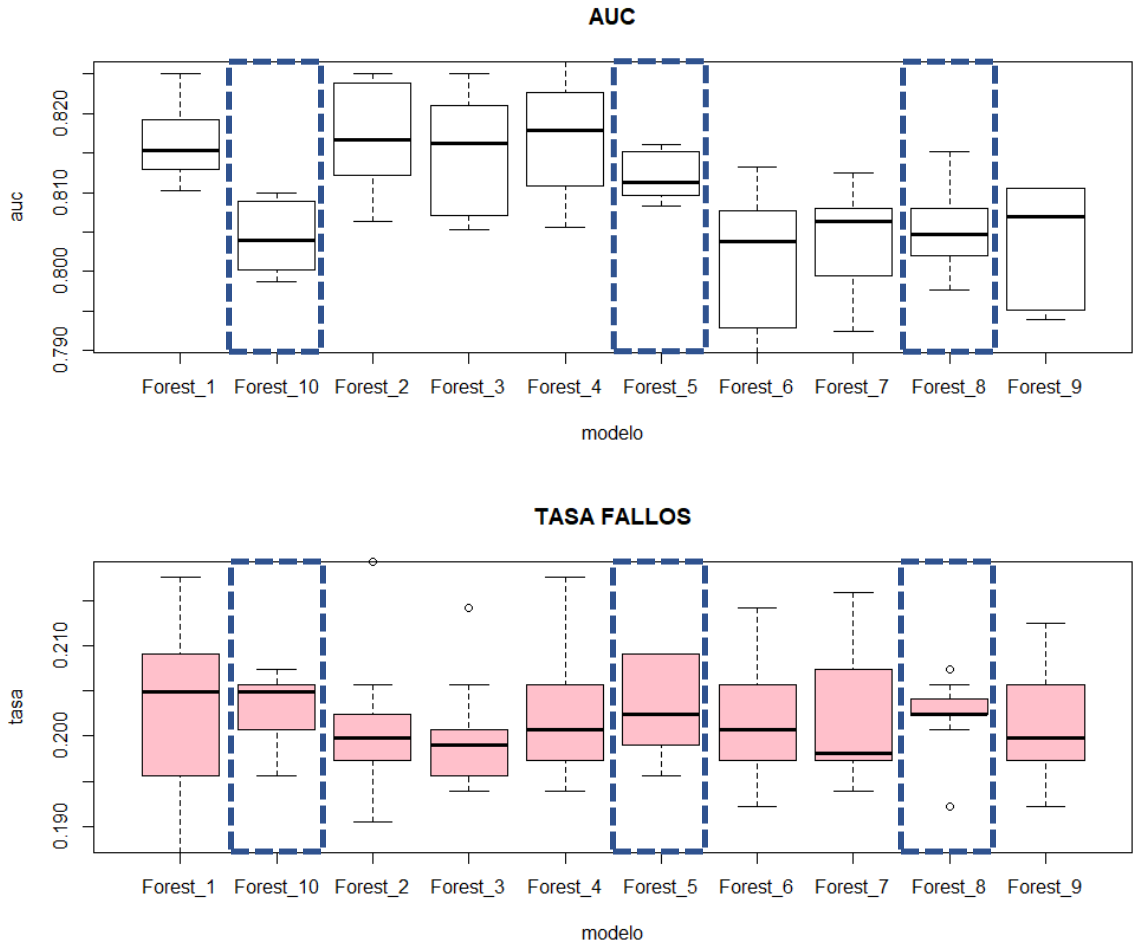
Al graficar los resultados se observó un mayor AUC con los modelos que contienen las variables transformadas, pero al compararlo con la tasa de fallos no se encontró una diferencia significativa. Se seleccionaron los modelos con menor sesgo y menor tasa de fallos y comparando sus parámetros se vio que el modelo 3 tuvo un número mínimo de nodos mayor al modelo 4, siendo el primero mejor al evitar el sobreajuste del modelo, por lo que se tomó el modelo 3 como el mejor y se procedió a graficar los resultados de Random Forest.

Tabla 13. Parámetros modelos Random forest

MODELO	RAN 1	RAN 2	RAN 3	RAN 4	RAN 5	RAN 6	RAN 7	RAN 8	RAN 9	RAN 0
Base	Senbis	Senbis	Senbis	Senbis	Senbis	Senbis	Sen1bis	Sen1bis	Sen1bis	Sen1bis
MTRY	7	21	29	11	11	12	9	4	11	4
NTREE	500	500	500	500	500	500	500	500	500	500
NODESIZE	20	10	10	10	20	10	10	15	15	15
SAMPSIZE	400	300	300	300	200	300	300	200	400	200

En el modelo dos y diez se han eliminado las variables seleccionadas previamente, con el fin de obtener mejores resultados:

Gráfica 25. Resultados modelos random forest



Al comparar los resultados de AUC se evidencia que de nuevo los modelos con variables transformadas obtuvieron resultados mejores manteniendo un sesgo alto. En cuanto a la tasa de fallos, los mismos modelos mantienen sesgo algo; los de menor sesgo fueron los modelos de variables originales.

Ya que el error en todos fue muy similar se han resaltado los modelos que tuvieron menor sesgo, pero al tener en cuenta que en el modelo 10 fueron eliminadas algunas variables por la poca información aportada, aumentó la probabilidad de que en modelo la combinación de variables fuera más acertada; por esta razón se tomó como el mejor.

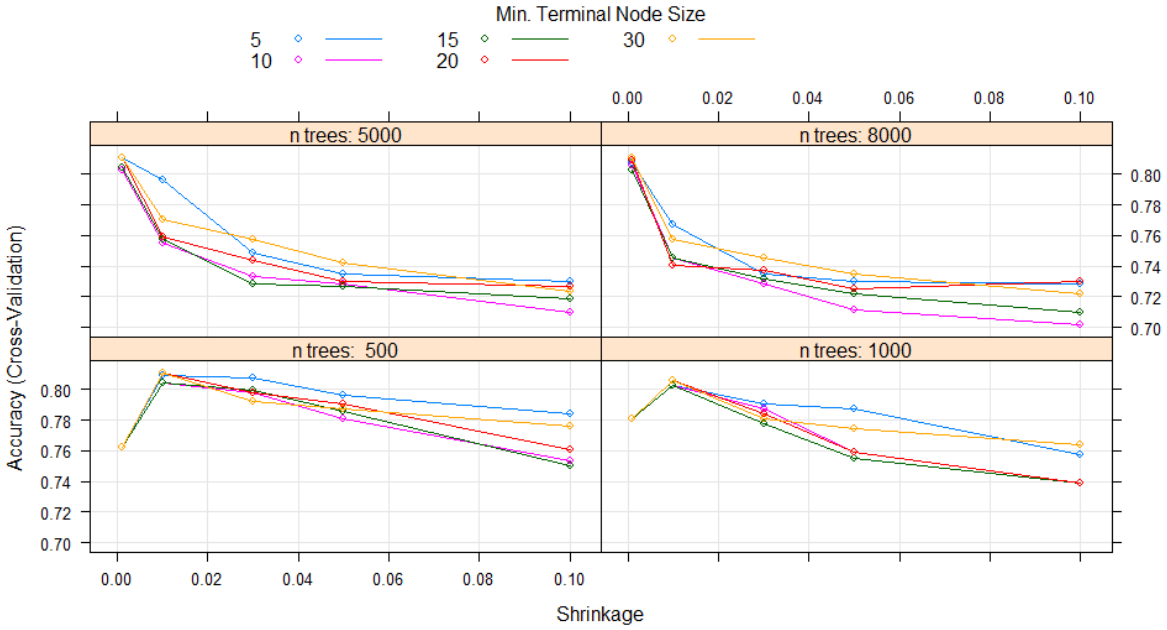
4.5.4. Gradient Boosting

Los parámetros que definen este algoritmo son:

N trees	Iteraciones
Shrinkage	Constante de regularización Entre (0,001 y 0,3)
Interaction.depth	Al ser categórica debe ser 2
Size	Profundidad # hojas al final
Minobsinnode	Observaciones min por hoja

Usando la librería *caret* se tunearon los parámetros anteriormente descritos y se iniciaron con la base de datos **Senbis**. Sus resultados a continuación:

Gráfica 26. Resultados tuneado de parámetros Senbis

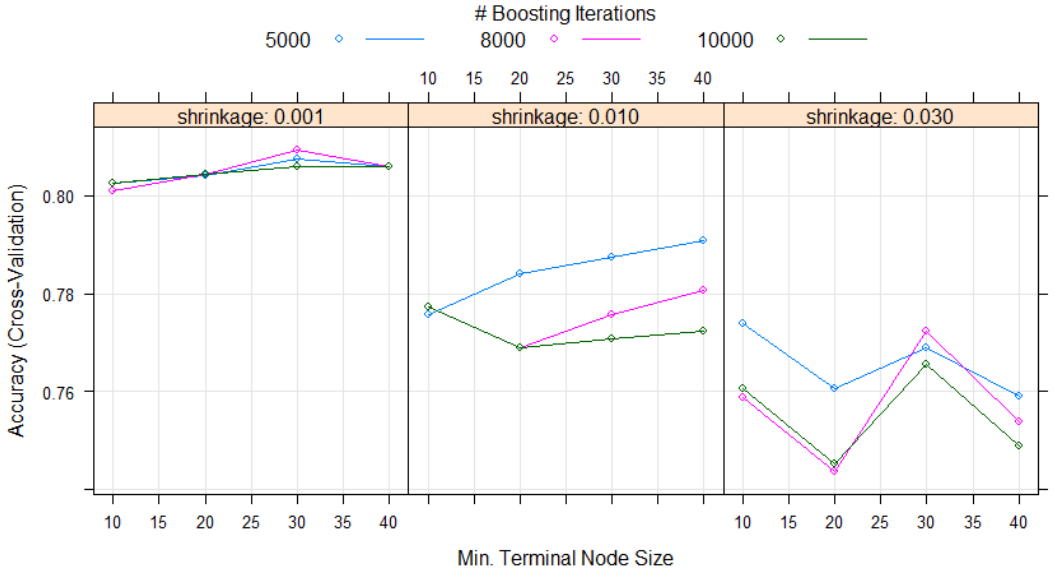


shrinkage	n.minobsinnode	n.trees	Accuracy	Kappa
0.010	10	5000	0.7554242	0.4390063
0.010	10	8000	0.7452431	0.4188296
0.010	15	500	0.8043780	0.5434058
0.010	15	1000	0.8027002	0.5430495
0.010	15	5000	0.7571592	0.4473556
0.010	15	8000	0.7453800	0.4278690
0.010	20	500	0.8111808	0.5601874
0.010	20	1000	0.8061130	0.5517224
0.010	20	5000	0.7588257	0.4503104
0.010	20	8000	0.7403238	0.4131704
0.010	30	500	0.8111808	0.5610525
0.010	30	1000	0.8061244	0.5518269
0.010	30	5000	0.7706390	0.4772185
0.010	30	8000	0.7572050	0.4448099

```
Tuning parameter 'interaction.depth' was held constant at a value of 2
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 500, interaction.depth = 2, shrinkage = 0.01
 and n.minobsinnode = 20.
```

En la gráfica 26 se identificó que a menor *Shrinkage* mayor *Accurancy*, es por esto que se tunea de nuevo dándole prioridad a los valores bajos de *Shrinkage,* y de esta forma identificar los mejores resultados:

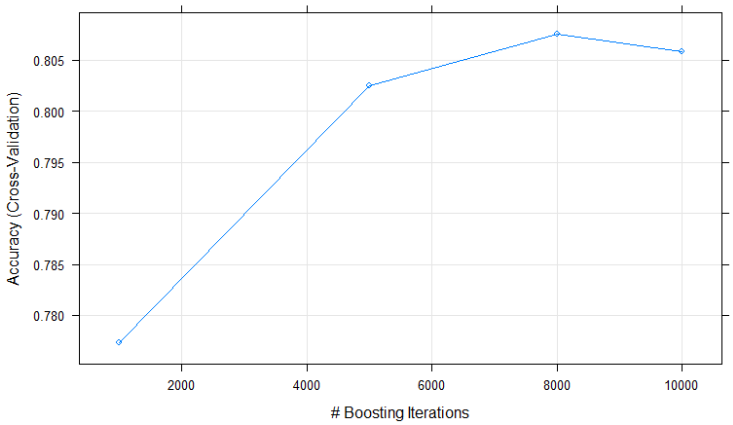
Gráfica 27. Resultados tuneado de parámetros



shrinkage	n.minobsinnode	n.trees	Accuracy
0.001	10	5000	0.8026155
0.001	10	8000	0.8009348
0.001	10	10000	0.8026297
0.001	20	5000	0.8042539
0.001	20	8000	0.8042681
0.001	20	10000	0.8042681
0.001	30	5000	0.8076437
0.001	30	8000	0.8093811
0.001	30	10000	0.8059771
0.001	40	5000	0.8059488
0.001	40	8000	0.8059913
0.001	40	10000	0.8059913
0.010	10	5000	0.7755963

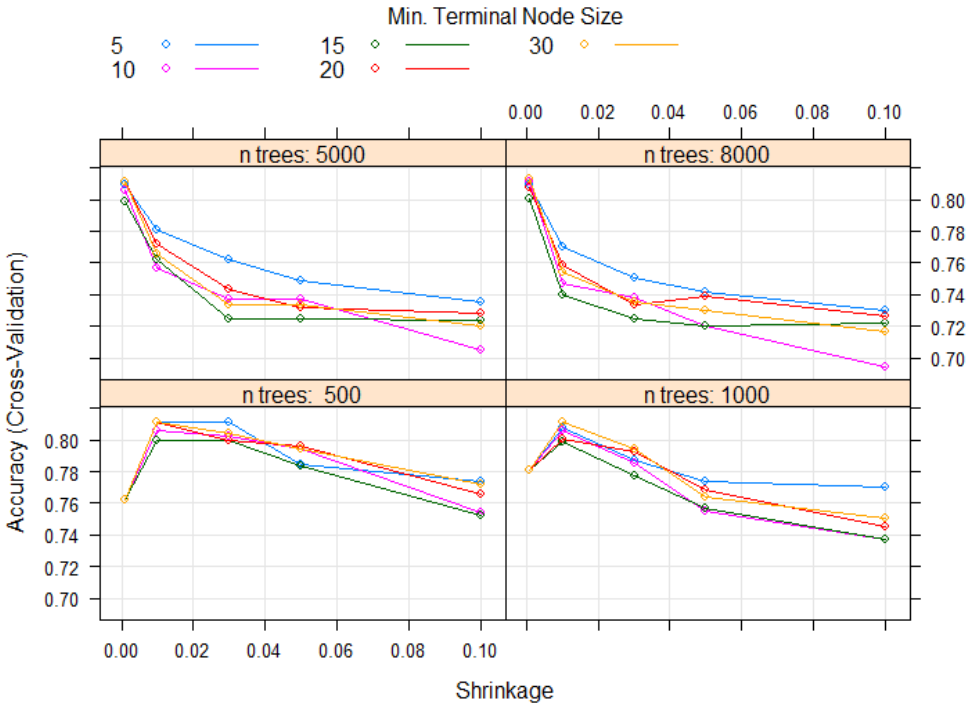
Los resultados son muy similares entre sí, tomaremos la información resaltada anteriormente para hacer el estudio de early stopping y confirmar el *ntree* óptimo: 8000 obtiene el mayor *Accurancy*

Gráfica 28. Early stopping



Para comparar que base de datos nos genera mejores resultados, se tunearan los parámetros previamente descritos en ambas, a continuación, los resultados de **Sen1bis**

Gráfica 29. Resultado tuneado de parámetros Sen1bis

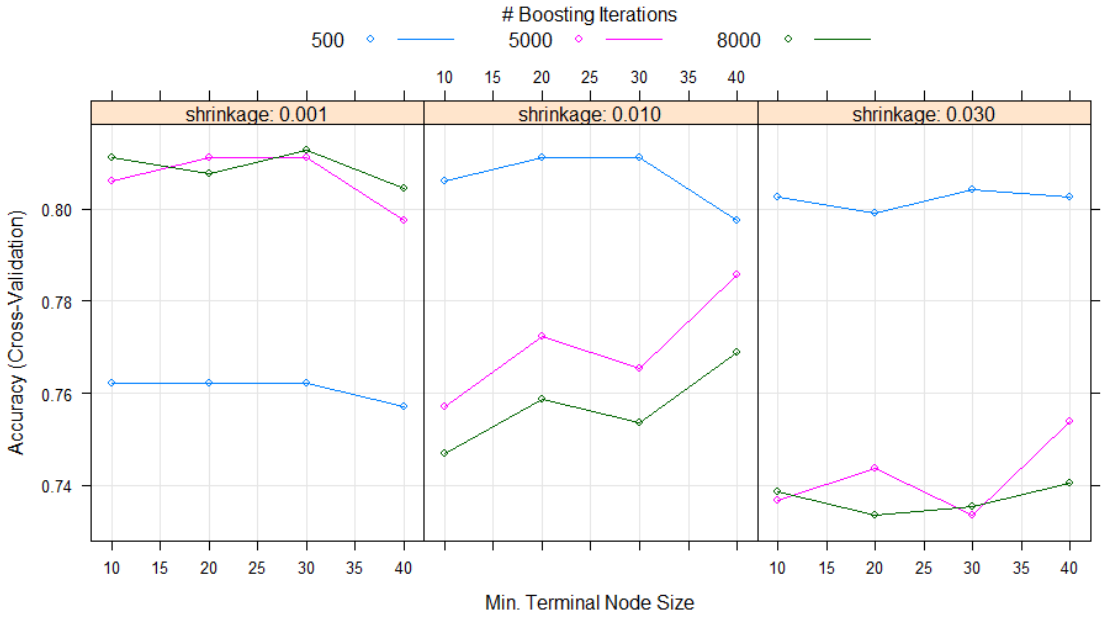


0.001	20	5000	0.8111691	0.5600296
0.001	20	8000	0.8077449	0.5553050
0.001	30	500	0.7622156	0.4000136
0.001	30	1000	0.7808201	0.4633387
0.001	30	5000	0.8111463	0.5609801
0.001	30	8000	0.8128470	0.5663520
0.010	5	500	0.8111578	0.5568785
0.010	5	1000	0.8078136	0.5511502
0.010	20	500	0.8111691	0.5600296
0.010	20	1000	0.8009995	0.5397581
0.010	20	5000	0.7722485	0.4828496
0.010	20	8000	0.7587685	0.4552866
0.010	30	500	0.8111463	0.5609801
0.010	30	1000	0.8111350	0.5633510
0.010	30	5000	0.7654916	0.4706870
0.010	30	8000	0.7537011	0.4423318
0.030	5	500	0.8111805	0.5605593

```
Tuning parameter 'interaction.depth' was held constant at a value of 2
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 8000, interaction.depth = 2, shrinkage
 = 0.001 and n.minobsinnode = 30.
```

Al ver que los resultados más altos de *Accurancy* se obtienen en valores pequeños de *Shrinkage* se ejecutó de nuevo la librería *caret* para verificar cuál fue el número óptimo de nodos:

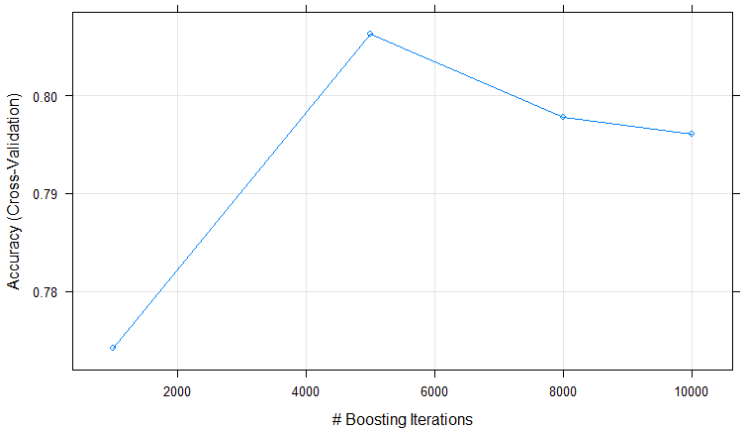
Gráfica 30. Resultados tuneado de parámetros Sen1bis



shrinkage	n.minobsinnode	n.trees	Accuracy	Kappa
0.001	10	500	0.7622156	0.4000136
0.001	10	5000	0.8060671	0.5455135
0.001	10	8000	0.8111463	0.5610032
0.001	20	500	0.7622156	0.4000136
0.001	20	5000	0.8111691	0.5600296
0.001	20	8000	0.8077449	0.5553050
0.001	30	500	0.7622156	0.4000136
0.001	30	5000	0.8111463	0.5609801
0.001	30	8000	0.8128470	0.5663520
0.001	40	500	0.7571820	0.3854273

Al obtener mejores resultados con un *Shrinkage* pequeño, lo ideal es usar un *ntree* alto, por lo que al observar el patrón confirmamos mediante el early stopping el mejor resultado de *ntree*: el óptimo es 5.000, pero mediante validación cruzada confirmamos cuál de estos obtuvo mejores resultados.

Gráfica 31. Early stopping

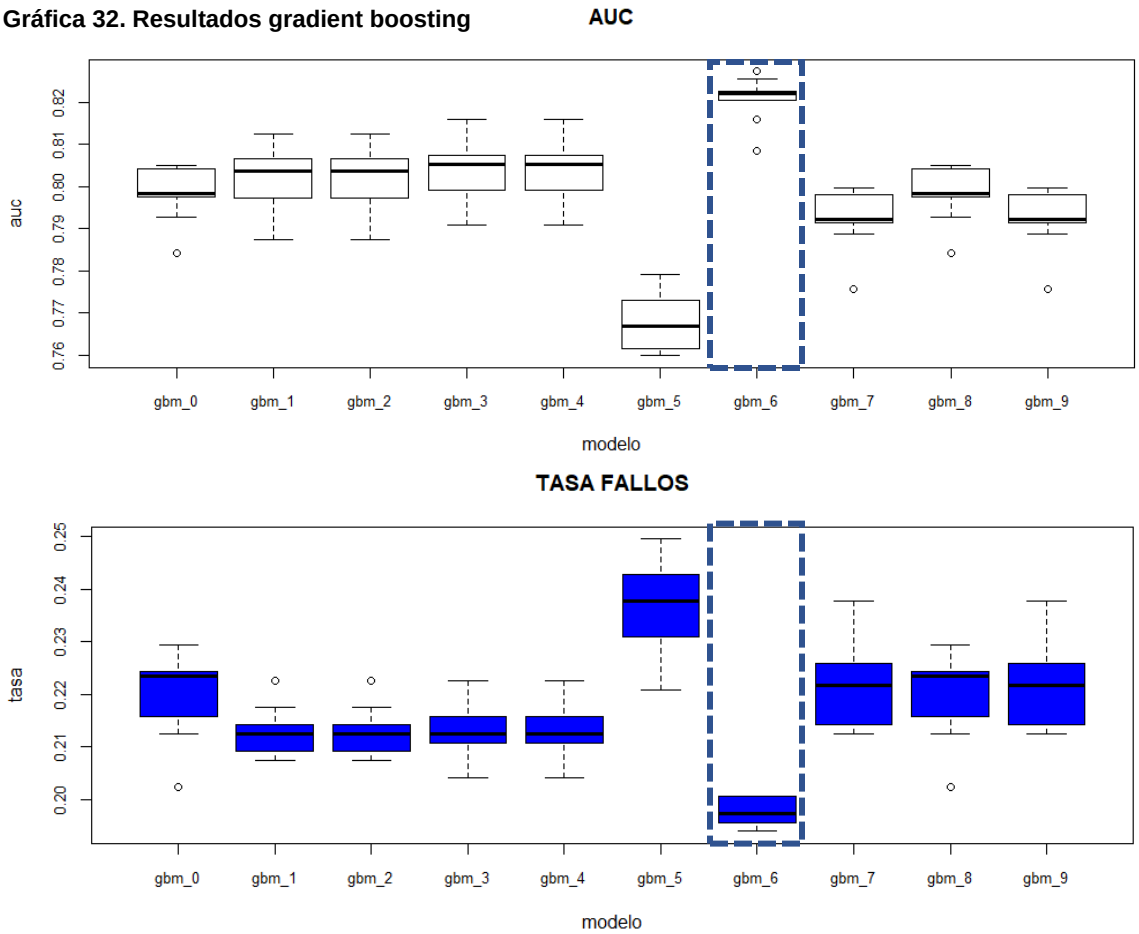


Una vez obtenidos los parámetros óptimos se crearon 10 grupos diferentes con el fin de compararlos en una validación cruzada para así obtener los mejores resultados. Para empezar del grupo 1 al 5 se encuentran los modelos con la base de datos Senbis, mientras que del 6 al 0 se encuentran los modelos generados con la base Sen1bis. De los modelos GBM 2 Y GBM0 se eliminaron las variables seleccionadas en rojo en el cuadro anterior, con el fin de visualizar si al tener en cuenta las que más aportan valor mejoran los resultados. A continuación, los resultados:

Tabla 14. Parámetros por modelo gradient boosting

MODELO	GBM1	GBM2	GBM3	GBM4	GBM5	GBM6	GBM7	GBM8	GBM9	GBM0
Base	Senbis	Senbis	Senbis	Senbis	Senbis	Sen1bis	Sen1bis	Sen1bis	Sen1bis	Sen1bis
NTREE	8000	8000	10000	10000	8000	5000	5000	8000	5000	8000
NODESIZE	30	30	30	30	40	20	30	30	30	30
SHRINKAGE	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001

Gráfica 32. Resultados gradient boosting

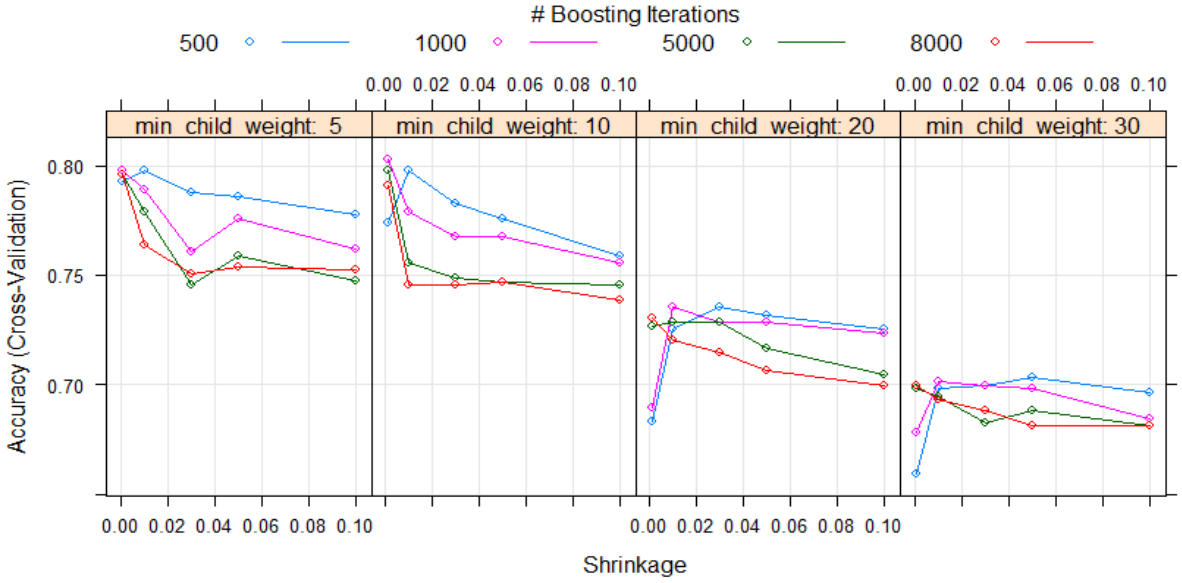


Una vez obtenidos los resultados, se vio que el modelo 6 se diferencia drásticamente de los demás, tal como se vio en el tuneado, en la base **Sen1bis** al usar *ntree*=5000 obtuvimos mejores resultados. Además, se designó un numero de nodos finales igual a 20. En los modelos de la base **Senbis** observamos resultados muy similares y una gran diferencia en el modelo 5 al usar un número de nodos mayor al observado en el tuneado.
Luego del análisis, el mejor fue el modelo 6.

4.5.5. XGBoost

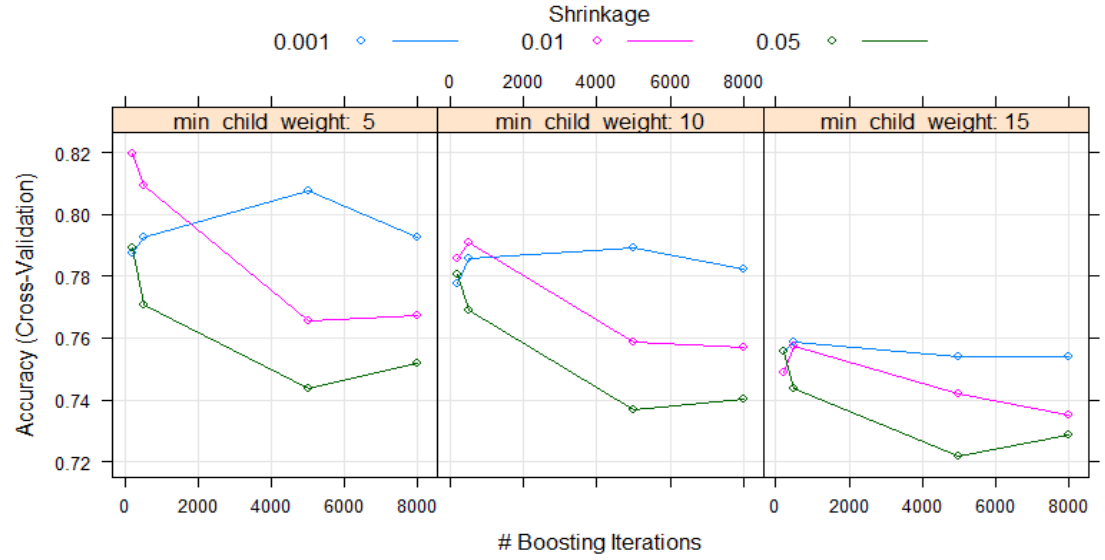
Usando la librería *caret* se buscó la combinación de parámetros que obtuviera una mayor *Accurancy.* Para iniciar, se evaluó la base **Senbis;** los resultados se muestran a continuación:

Gráfica 33. Resultados tuneado parámetros Senbis



Está claro que a mayor *min child weight* obtenemos peores resultados, por eso nos centramos en valores pequeños y manteniendo el intervalo de *Shrinkage* entre 0,001 y 0,05, y obtenemos:

Gráfica 34. Resultados tuneado parámetros Senbis

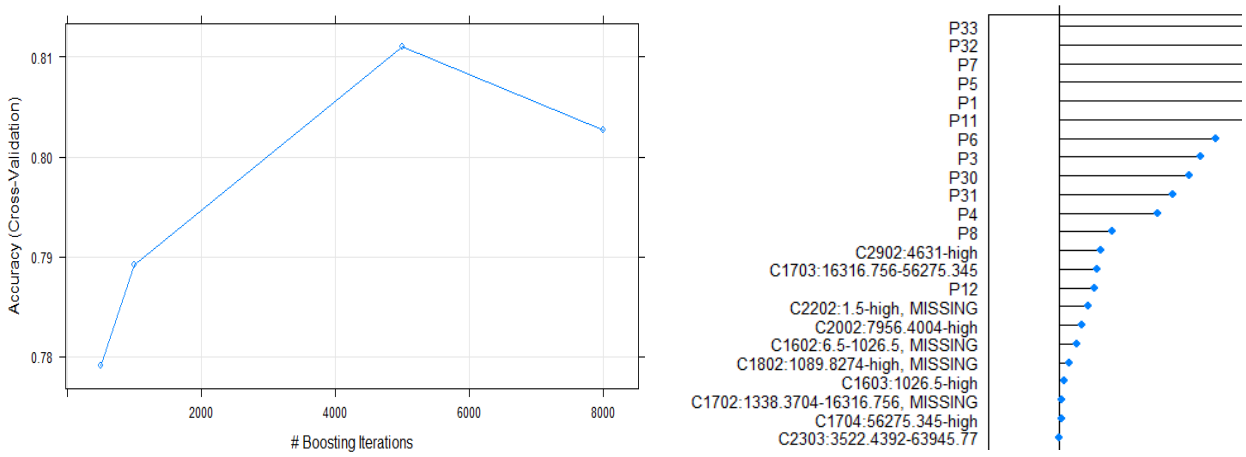


eta	min_child_weight	Nrounds	Accuracy	Kappa
0.001	5	200	0.7875657	0.5188743

0.001	5	500	0.7926449	0.5241011
0.001	5	5000	0.8077679	0.5581052
0.001	5	8000	0.7925991	0.5260241
0.001	10	200	0.7774757	0.4744494
0.001	10	500	0.7858650	0.4906328
0.001	10	5000	0.7891865	0.5056925
0.001	10	8000	0.7824293	0.4897164
0.001	15	200	0.7555270	0.4382080
0.001	15	500	0.7588825	0.4241604
0.001	15	5000	0.7538260	0.4154476
0.001	15	8000	0.7538376	0.4193532
0.010	5	200	0.8196155	0.5849324
0.010	5	500	0.8094458	0.5614482
0.010	5	5000	0.7656283	0.4634947
0.010	5	8000	0.7672948	0.4693821
0.010	10	200	0.7858310	0.4949187

Se evalúa con los mismos intervalos el *early stopping* buscando encontrar la mejor opción:

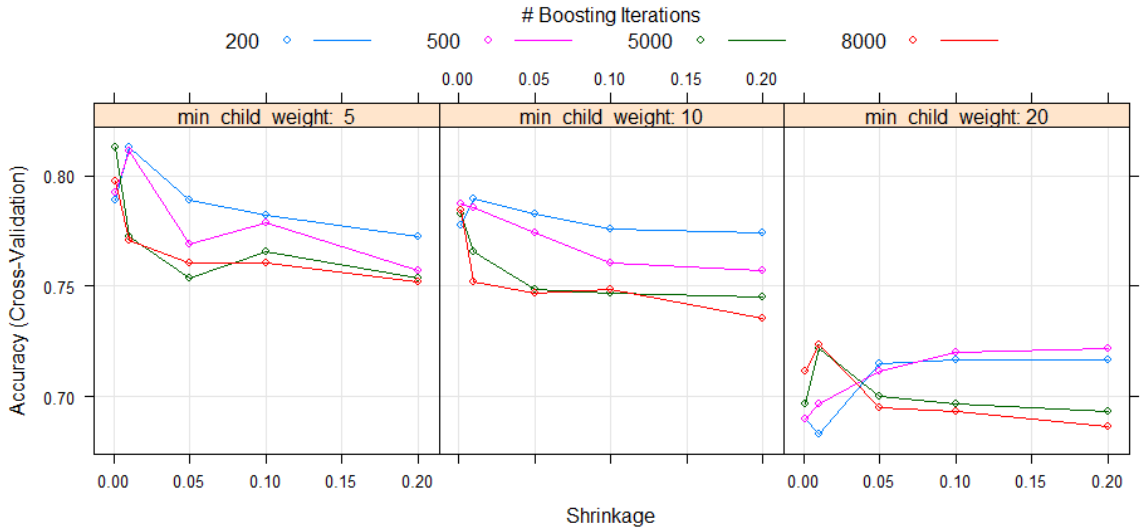
Gráfica 35. Early stopping - Importancia de variables



Una vez se obtienen las variables que más aportan al modelo y número óptimo de iteraciones haremos una validación cruzada para validar los mejores resultados.

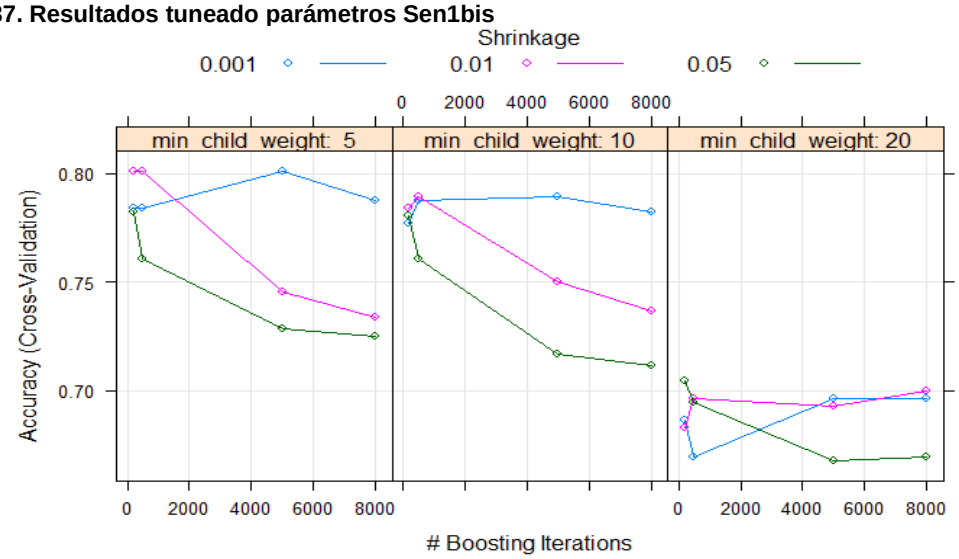
Ahora se muestran las mismas iteraciones con la base **Sen1bis.**

Gráfica 36. Resultados tuneado de parámetros Sen1bis



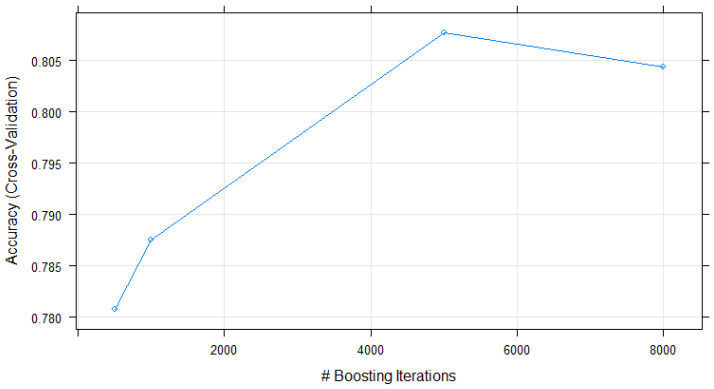
De nuevo vemos que al ser *min child weight* menor se obtienen mejores resultados, de la misma forma se mantuvo el intervalo de *Shrinkage* entre 0,001 y 0,05, y obtuvimos:

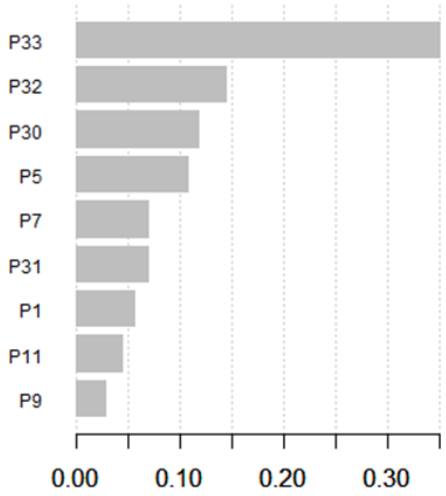
Gráfica 37. Resultados tuneado parámetros Sen1bis



eta	min_child_weight	Nrounds	Accuracy	Kappa
0.001	5	200	0.7841645	0.5044230
0.001	5	500	0.7841645	0.5044230
0.001	5	1000	0.7942772	0.5295838
0.001	5	5000	0.8027462	0.5457734
0.001	10	200	0.7942769	0.5144733
0.001	10	500	0.7959434	0.5149974
0.001	10	1000	0.7976213	0.5257280
0.001	10	5000	0.7740973	0.4699610

Gráfica 38. Early stopping - Importancia de variables



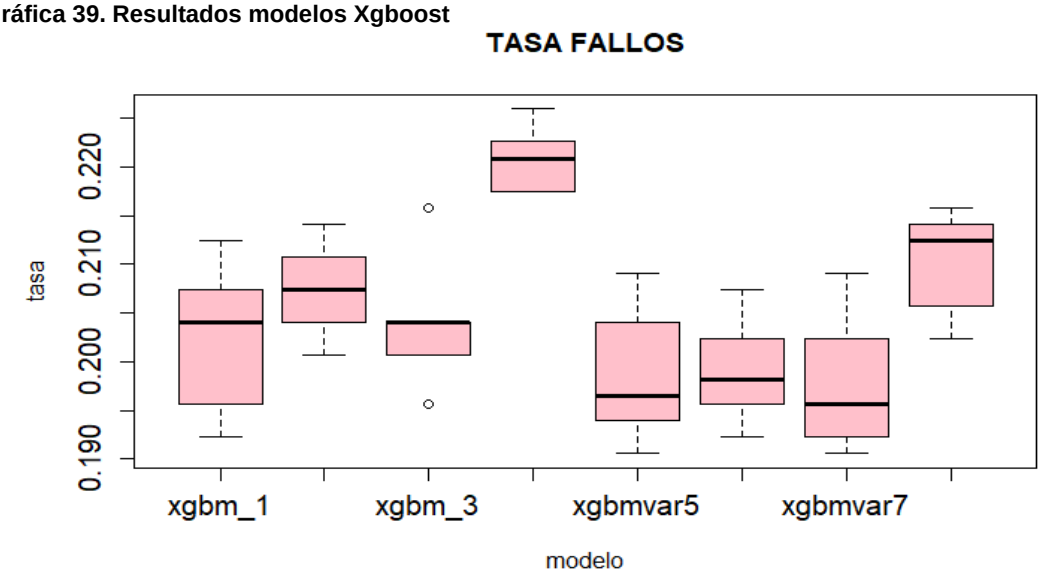


Conociendo los parámetros óptimos para cada una de las bases, se ejecutaron validaciones cruzadas que permitieron elegir lo mejor.

Tabla 15. Parámetros modelos Xgboost

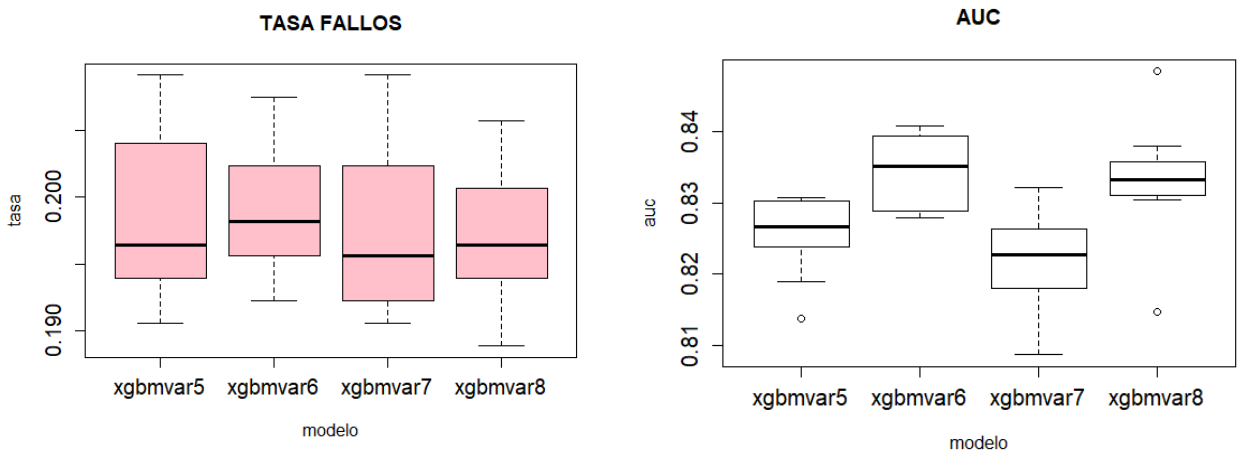
MODELO	XGB1	XGB2	XGB3	XGB4	XGB5	XGB6	XGB7	XGB8
Base	Senbis	Senbis	Sen1bis	Sen1bis	Senbis	Senbis	Sen1bis	Sen1bis
NROUNDS	200	5000	5000	1000	200	5000	200	5000
MIN_CHILD	5	5	5	5	5	5	5	5
SHRINKAGE	0,01	0,001	0,001	0,001	0,001	0,001	0,01	0,001

Gráfica 39. Resultados modelos Xgboost



Una vez se obtienen los resultados, se eligen con los 4 mejores, dos de cada base, evaluando su *Accurancy:*

Gráfica 40. Resultados mejores modelos Xgboost

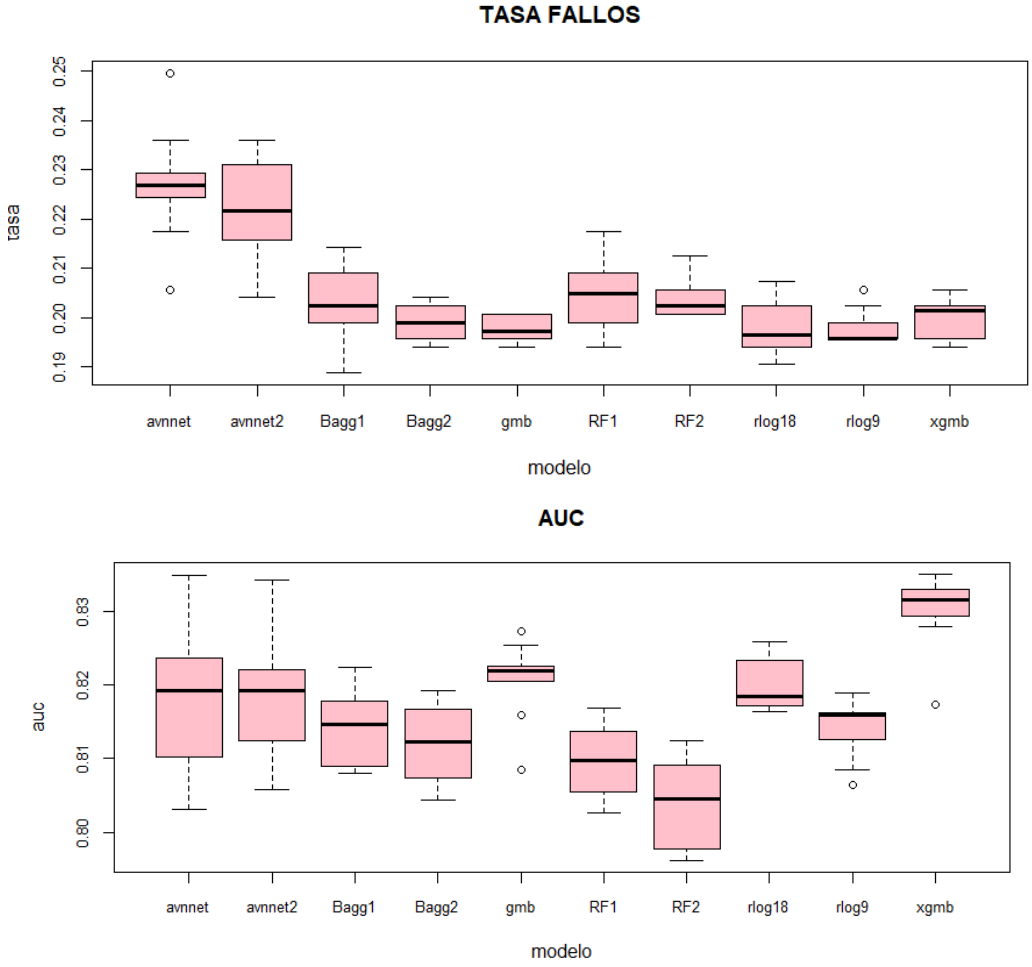


El modelo 6 y 8 presentaron un menor sesgo frente a los otros dos modelos; el grupo 8 contiene solo variables originales. Pese a sus datos atípicos éste fue el modelo para comparar con los demás, ya que permitió un análisis sencillo de la información con resultados óptimos.

4.6. Comparación de modelos

Al ver de forma gráfica los modelos seleccionados encontramos que redes neuronales no es el algoritmo óptimo para los datos, ya que los árboles y la regresión generan resultados muy debajo. Xgboost al ser más potente, minimiza la varianza que se visualiza en los demás algoritmos de árboles y alcanza resultados muy similares a la regresión, obteniendo un sesgo bajo y varianza moderada.

Gráfica 41. Resultado comparación de modelos

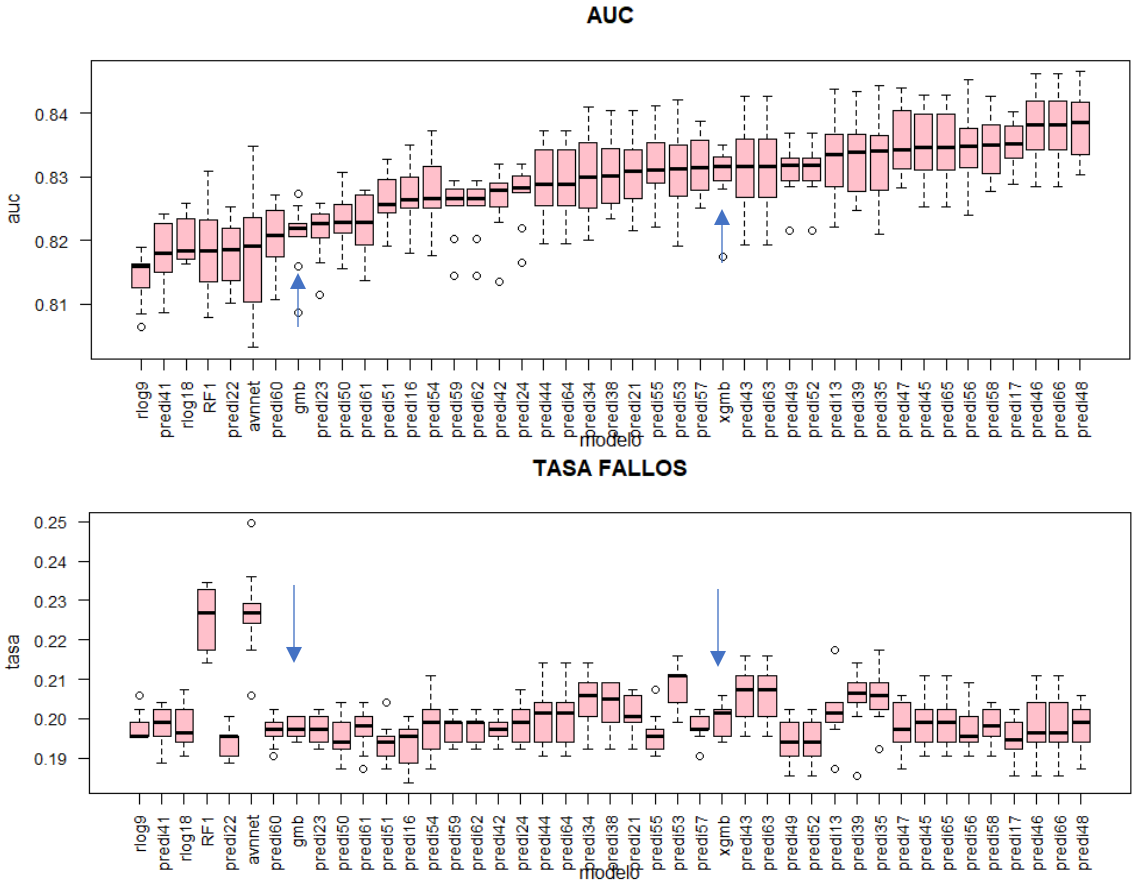


Se empleó un código (Portela, 2020) que permite comparar los modelos y crear combinaciones entre sí; este proceso se llama ensamblado y dependiendo su comportamiento potenciaba la precisión de los algoritmos seleccionados.

4.7. Ensamblado

Una vez ejecutado el código y omitiendo los resultados más bajos, se obtuvieron los siguientes resultados:

Gráfica 42. Resultado modelos ensamblado



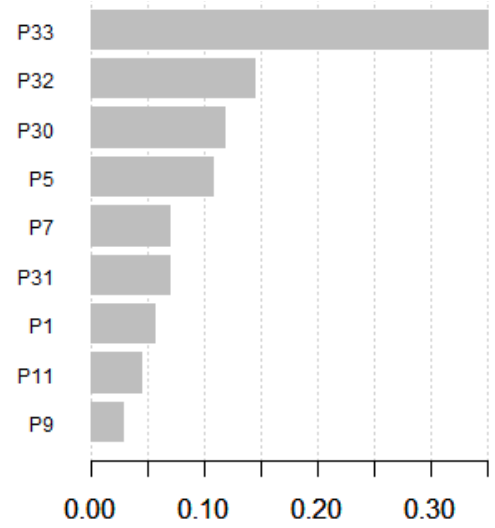
Se han indicado con flechas los modelos originales que obtuvieron resultados competitivos con los modelos ensamblados, mostrando no solo eficacia sino, además, simplicidad tanto al replicar como al analizar sus resultados. Por esta razón, se escogió como modelo ganador a *xgbm*.

4.8. Modelo ganador

Dado el análisis del apartado anterior, el modelo *xgbm* presentó una tasa de fallo baja y un alto AUC superado solo por modelos ensamblados, al ser más sencillo facilita su análisis, por esta razón este ha sido seleccionado el mejor.

Una vez seleccionado, se analizaron las variables que el modelo indicó como importantes con el fin de conocer los factores que más tuvieran relación con el evento de interés seleccionado.

Gráfica 43. Importancia de variables modelo ganador



P33: Grupo de departamentos 5

P32: Grupo de departamentos 4

P30: Grupo de departamentos 1

P5: Cobertura de aseo

P7: Población en cond. pobreza

P31: Grupo de departamentos 2

P1: Abstención

P11: Promedio área coca

P9: Minas antipersonas heridos

Tal como se puede observar en la gráfica 43 todas las variables son originales, la importancia dada por el modelo coincide con el análisis factorial realizado en el apartado 4.3.3, allí se observó que la mayor relación estaba entre las cuatro variables escogidas por el modelo y, además, identifica una variable representativa de los factores previamente detallados.

Entre la relación identificada en el análisis factorial, se identificó una relación negativa entre el departamento y el partido ganador, al dividir dicha variable por grupos, fue posible identificar qué departamentos tienen mayor relación con la variable objetivo. En este caso, la más importante para el modelo está relacionada de forma inversa (grupo 5), en el departamento de Santander los municipios en su mayoría eligieron representantes al senado diferentes a derecha, por lo que al ser la más importante a lo largo de la modelización realizada, indica que para el modelo resulta importante dicha información para la clasificación. El grupo 3 que fue excluido del modelo, históricamente ha sido un fuerte simpatizante de izquierda, lo que coincide con su baja relación con el evento de interés.

En los municipios analizados se observó cómo la pobreza era más alta en proporción al resto del país, una alta incidencia de variables como “población en condición de pobreza” y “abstención” dan una señal de cómo en dichas regiones el abandono del Estado y una mayor presencia de guerrilla ha determinado su comportamiento electoral, lo que en este caso ha aumentado la probabilidad de elegir representantes del senado de derecha.

Por último, al estar las variables de “promedio de área de coca” y “heridos por minas antipersonas” de último lugar en el modelo si bien no son las más importante, mantiene una relación con el evento de interés.

Usando la base de datos que contiene los valores input del periodo siguiente de votaciones correspondiente al 2018, se hizo la predicción de dichas votaciones, probando la eficacia del modelo y la hipótesis planteada al inicio de la investigación.

Mediante la librería *xgboost y* usando la base de datos solo con las variables escogidas como las más importantes para el modelo, se realizó la predicción del evento de interés con el modelo ganador. Se usó como referencia dos resultados importantes, la predicción para la partición *test* y los resultados del *no modelo*[3]*.* Se obtuvieron los siguientes resultados:

TEST

		Predicción +	Predicción -
Obs	+	74	9
	-	16	19

VALIDACIÓN

		Predicción +	Predicción -
Obs	+	337	72
	-	114	70

SIN MODELO

		Predicción +	Predicción -
Obs	+	407	0
	-	186	0

	TEST	VALIDACIÓN	SIN MODELO
Acurracy	78,81%	68,63%	68,63%
Tasa de error	21,19%	31,37%	31,37%
Sensibilidad	89,16%	82,80%	100%
Especificidad	54,29%	37,63%	0,00%
Precisión	82,22%	74,72%	100%
Valor de predicción negativo	67,86%	49,30%	0,00%

Una vez se entrena el modelo, se prueba con la partición test y se obtiene un *Acurrancy* 78,81% este valor coincide con los resultados que se veía en el apartado anterior, donde la tasa de fallos del modelo se encontraba cerca de 0,20, alineado con la precisión (verdaderos positivos acertados) del 82,22%. Al analizar la especificidad del 54,29% se vio que la predicción del “No” fue más difícil para el modelo.

El *no modelo* se tomó como referencia para visualizar los resultados que se tendrían al suponer que los resultados de todas las observaciones fueran “si” o positivas al evento de interés. De esta forma, se sabría con la base de validación si realmente existe una diferencia significativa al usar el modelo.

Lamentablemente, al visualizar los resultados del modelo con la base de validación, se vio que no solo el *Acurrancy* bajó de forma importante, sino que, además, fue exactamente igual que el *no modelo.* Estudiando la especificidad, mostró un bajo porcentaje de acierto en la predicción del “no”, lo que podría ser una señal de la deficiencia del modelo, pues

[3] El no modelo hace referencia a una predicción donde se da por hecho que todas las observaciones tienen un evento de interés verdadero.

aprendió mucho del evento de interés (82,8%), pero al ser tan baja su especificidad pierde precisión *Acurrancy.*

En el modelo las variables más importantes corresponden a los departamentos, es por esto que al analizar sus cambios electorales se pudo entender de qué forma se afectó la predicción.

Tabla 16. Comparación resultados electorales en ambos periodos

	2014		2018		
G_Departamento	**0**	**1**	**0**	**1**	**Var %***
1	1	61	13	49	-20%
2	59	243	71	231	-5%
3	33	53	28	58	9%
4	40	15	23	32	113%
5	74	13	51	36	177%
Total general	**207**	**385**	**186**	**407**	

*Variación porcentual entre periodos electorales

Los grupos de departamentos más importantes para el modelo son el 5, el 4 y el 1. Si analizamos sus resultados, vemos que los dos primeros tienen una relación inversa con el evento de interés, pero, en los resultados de las segundas elecciones, vemos cómo su comportamiento cambia, pues aumenta la preferencia por representantes de partidos de derecha, afectando el patrón identificado inicialmente.

Además, en los grupos donde la preferencia a representantes de derecha es más fuerte, se presentaron variaciones negativas y pequeñas, mientras que en los grupos donde la preferencia por partidos de derecha no era mayoría la variación fue mayor al 100%.

En la tabla 17 realizamos el comparativo de las variables que se ven afectadas por la violencia, en ella se evidencia como la producción de coca aumenta de manera importante en grupos como el 5, 2 y 1, mientras que las minas no presentan cambios importantes. Estas dos variables van muy relacionadas con la presencia de grupos armados por lo que indica que si bien las FARC ya no están, otros grupos al margen de la ley mantienen el negocio.

Tabla 17. Comparación de variables relacionadas con la violencia en ambos periodos

	2014		2018	
	Promedio_area	**Mis_anti_Heridos**	**Promedio_area**	**Mis_anti_Heridos**
1	1.819.247,1	18	2.649.711	15
2	4.984.250,6	75	8.636.080	77
3	4.589.189,6	96	1.728.284	61
4	2.542.107,5	15	2.101.601	3
5	63.158,8	0	3.570.296	0

5. Conclusiones

1. Al entrenar los datos en diferentes modelos y usando diferentes algoritmos, se pudo ver que en el software R, la librería caret permite tunear ampliamente los modelos de árboles, mientras que redes neuronales se queda un poco corto en funciones de activación, impidiendo la optimización de sus resultados.

2. Combinar el análisis multivariante con el aprendizaje automático puede ayudar a entender de forma más adecuada el conjunto de datos, permitiendo una depuración más minuciosa de las variables y un modelo alineado con la realidad de las observaciones.

3. Es importante resaltar qué durante el proceso de modelado, siempre se identificaron las variables de grupos de departamentos como importantes, tal como se evidenció con el modelo ganador: tres de los cinco grupos presentan una tendencia fuerte a elegir representantes al senado de partidos de derecha. En ellos encontramos a Córdoba, Guaviare, Meta, Antioquia, Bolívar, Casanare, Chocó, Norte de Santander, Valle, Arauca, Caquetá y Nariño.

   Una vez se revisan los resultados de la segunda jornada electoral, se encuentran los mismos departamentos fuertemente representados por partidos de derecha, mientras que los otros dos grupos de departamentos si bien aumenta su preferencia, no se encuentra una diferencia significativa entre los periodos.

4. Al analizar los resultados de la base de validación se observa un porcentaje alto de sensibilidad (acierto para el evento de interés), pero bajo para especificidad (acierto para el evento negativo), esto como se mencionó en el apartado anterior, puede indicar que el modelo no obtuvo suficiente información para optimizar la especificidad o, el comportamiento de la población estuvo afectado por factores externos al estudio que cambiaron su preferencia en las elecciones del año 2018.

5. El objetivo inicialmente planteado buscaba conocer la relación entre la elección de representantes al senado con factores económicos y de violencia como lo son la producción de coca y los afectados por minas antipersonas. Si bien han sido seleccionadas por el modelo, no presentan una fuerte relación, además, en los resultados se pudo comprobar cómo luego de la firma del Acuerdo de Paz los cultivos de coca no frenaron su producción; por el contrario, en algunos departamentos se duplicó, evidenciando que dicho negocio ilegal sigue liderado por otros grupos y no dependía exclusivamente de las FARC. Además, las minas antipersonas mantienen su tendencia sin afectarse por la desaparición de dicho grupo armado.

6. Si bien la predicción no es exitosa, el modelo ha permitido contrastar cómo los 15 departamentos seleccionados como más afectados por la violencia históricamente, presentan una tendencia fuerte a seleccionar representantes al senado de derecha, aumentando de un año a otro; además, como se mencionaba en el apartado 3, la presencia de nuevos grupos armados en los territorios donde antes comandaban las FARC intimida a la población de forma indirecta, afectando dinámicas electorales

como la participación (volátil o no) y la abstención electoral. Por último, el control de las nuevas guerrillas hace que factores como los cultivos ilícitos y minas antipersonas se mantengan presentes.

7. En cuanto a líneas de investigación futuras, sería interesante poder incluir más periodos electorales para analizar con mayor profundidad su tendencia, porque si bien en el apartado 3 se habló sobre la volatilidad de la tendencia del voto en los municipios de Colombia, sería interesante identificar si la tendencia por partidos de derecha históricamente ha sido fuerte en dichos departamentos.

   Para el presente trabajo no fue posible conseguir información oficial de presencia de guerrillas, pero teóricamente se sabe que dicho factor es crucial en el comportamiento de la población, por lo que incluir más variables de este tipo sería interesante para otros estudios que puedan emplear modelos similares a los presentados aquí.

Bibliografía

Barrero, F. A., & Baquero, S. Á. (2019). Aportes metodológicos para el análisis electoral y partidista. En F. A. Barrero, *Elecciones presidenciales y de congreso 2018: nuevos acuerdos ante diferentes retos* (págs. 353-377). Bogotá: Fundación Konrad Adenauer.

Basset, Y., & Guavita, L. V. (2019). *Radiografía del desencanto: La participación electoral en Colombia.* Bogotá: Universidad del Rosario.

Calviño Martínez, A. (2019). Técnicas y metodología de la minería de datos SEMMA. Madrid, Comunidad de Madrid.

Caparrini, F. S. (02 de 09 de 2020). *Fernando Sancho Caparrini*. Obtenido de http://www.cs.us.es/~fsancho/?e=231

Centro Nacional de Memoria Histórica. (2018). *Regiones y conflicto armado: Balance de la contribución del CNMH al esclarecimiento histórico.* Bogotá: CNMH. Obtenido de http://www.centrodememoriahistorica.gov.co/micrositios/balances-jep/descargas/balance-regiones.pdf

El Espectador. (19 de 06 de 2019). *El Espectador*. Obtenido de https://www.elespectador.com/colombia2020/pais/colombia-el-pais-con-mas-desplazados-del-mundo-articulo-866644/

Giraldo García, F., & Soto Caballero, H. R. (Abril/Junio de 2019). Circunscripciones Especiales: la paz en la apatía electoral. *Revista mexicana de sociología, 81*(2).

Latorre, M. (2009). *Elecciones y partidos políticos en Colombia.* Bogotá: Ediciones Uniandes.

Milanese, J. P., & Serrano, C. E. (2019). Consistencia espacio-temporal de los apoyos electorales: Un análisis ecológico de la transferencia de votos en las dieciséis fallidas circunscripciones de paz en Colombia. *X Congreso Latinoamericano de Ciencia.* Monterrey. Obtenido de https://alacip.org/cong19/213-milanes-19.pdf

Ministerio de Ciencias de Colombia. (11 de 09 de 2016). *Minciencias*. Obtenido de https://minciencias.gov.co/sala_de_prensa/colombia-el-segundo-pais-mas-biodiverso-del-mundo

Morde, V. (01 de 09 de 2020). *Towards data science*. Obtenido de https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Presidencia de Colombia. (20 de 10 de 2017). *Presidencia de Colombia*. Obtenido de http://es.presidencia.gov.co/normativa/normativa/DECRETO%201650%20DEL%2009%20DE%20OCTUBRE%20DE%202017.pdf

Público. (02 de 12 de 2018). *Público*. Obtenido de https://www.publico.es/economia/narcotrafico-colombia-riqueza-negocio-precariedad-campesino.html

Rodrigo, J. A. (30 de Agosto de 2020). *Ciencia de Datos, Estadística, Programación y Machine Learning*. Obtenido de https://www.cienciadedatos.net/

Rodríguez Pinzón, E. (21 de 03 de 2014). *Los diálogos de paz en Colombia, avances y prospectiva*. Obtenido de http://www.realinstitutoelcano.org/: http://www.realinstitutoelcano.org/wps/portal/rielcano_es/contenido?WCM_GLOBAL_CONTEXT=/elcano/elcano_es/zonas_es/ari18-2014-rodriguezpinzon-+dialogos-paz-colombia-avances-prospectiva

Semana. (25 de 02 de 2018). *Semana*. Obtenido de https://www.semana.com/elecciones-congreso-2018/noticias/el-senado-de-la-republica-asi-esta-conformado-y-asi-se-elige-557839

Senado. (31 de 07 de 2020). *Senado de Colombia*. Obtenido de http://www.senado.gov.co/index.php/el-senado/funciones

Unidad de atención y reparación integral a las víctimas. (31 de 07 de 2020). *Unidad de victimas*. Obtenido de https://www.unidadvictimas.gov.co/es/registro-unico-de-victimas-ruv/37394

Valecia Delfa, J. L., & Vicente Hernanz, M. L. (2006). *Análisis Multivariante I.* Madrid: CERSA.

6. Anexos

https://github.com/anguer05/TFM