FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2022/2023

Trabajo de Fin de Máster

TITULO: Análisis de mortalidad por insuficiencia cardíaca

***Alumno*: Juan Fernando Plata Morán**

***Tutor*: David Lora Pablos**

Septiembre de 2023

UNIVERSIDAD COMPLUTENSE
MADRID

Resumen

En el contexto de las enfermedades cardíacas, y ante la necesidad de abordar nuevos escenarios que aporten una perspectiva renovada en relación con el tratamiento de las mismas, la aplicación de técnicas de *machine learning* se presenta como una alternativa contemporánea y vanguardista. A través del análisis de datos clínicos, se busca anticipar y prevenir eventos adversos, a la vez que se brinda información a los profesionales de la salud para respaldar la toma de decisiones fundamentada y así adaptar estrategias según la situación particular de cada paciente. Este proyecto explora cómo estas herramientas pueden contribuir a la creación de respuestas más efectivas.

Palabras clave

Insuficiencia cardíaca, mortalidad, *machine learning*, modelos predictivos, análisis de datos clínicos, estrategias preventivas.

Abstract

Within the realm of cardiac ailments, amidst the imperative to confront novel scenarios that infuse a revitalized outlook toward their therapeutic approaches, the utilization of advanced machine learning techniques emerges as a contemporaneous and avant-garde recourse. By means of the meticulous scrutiny of clinical data, the objective resides in the anticipation and deterrence of adverse occurrences, concomitantly furnishing healthcare practitioners with requisite information to underpin judicious decision-making, thereby customizing strategies to suit each patient's individual circumstances. This endeavor delves into the profound potential for these tools to enrich the formulation of more efficacious responses.

Keywords

Heart failure, mortality, machine learning, predictive models, clinical data analysis, preventive strategies.

Índice

Índice de figuras

Capítulo 1

Introducción

1.1. Motivación

La ingente cantidad de datos e información que se genera diariamente ha acrecentado el interés por el desarrollo de herramientas sofisticadas que permitan el tratamiento o análisis de los mismos. Los avances tecnológicos han propiciado la aparición de diferentes campos dedicados al estudio y comportamiento de conjuntos masivos de datos.

Diversos sectores empresariales están adentrándose en estas áreas con el fin de implementarlas en sus negocios. Esta inmersión es fiel reflejo de la relevancia que dichas disciplinas están adquiriendo en diferentes ámbitos y que empiezan a ser reconocidas por su gran potencial en cuanto a optimización de procesos y toma de decisiones estratégicas se refiere.

Ya no sólo consiste en la disponibilidad y recopilación de grandes volúmenes de datos, sino más bien en un correcto procesamiento de los mismos que facilite la extracción de información útil a empresas e instituciones. De esta forma, tal información se podría traducir en acciones concretas.

En este aspecto, la minería de datos es uno de los grandes campos que han venido desarrollándose a lo largo de las últimas décadas.

También conocida como *data mining*, la minería de datos es una disciplina especializada en la detección de patrones, tendencias y relaciones relevantes en grandes volúmenes de datos mediante la implementación de diferentes técnicas y algoritmos avanzados y con el objetivo de obtener información valiosa. La minería de datos abarca diversas técnicas, como el aprendizaje automático o *machine learning*, la estadística, la visualización de datos y la inteligencia artificial. Estas técnicas utilizan tecnologías de reconocimiento de patrones, permitiendo clasificar y segmentar información, realizar predicciones y crear modelos predictivos.

Esta disciplina tiene aplicaciones en una amplia gama de sectores y áreas de estudio. Más concretamente, en investigación médica, la minería de datos facilita la identificación de correlaciones entre síntomas, diagnósticos y tratamientos, a la vez que favorece el desarrollo de predicciones de riesgo de enfermedades, de progresión de condiciones médicas y de resultados de tratamientos.

Todo ello brinda la posibilidad de mejorar la compresión de enfermedades, identificar factores de riesgo y desarrollar estrategias de tratamiento más efectivas. Además, ayudan a guiar las decisiones clínicas y a adaptar los tratamientos y las intervenciones a las necesidades específicas de cada paciente, mejorando la eficacia y eficiencia de los cuidados médicos.

En definitiva, es la gran variedad de aplicaciones que ofrecen las distintas disciplinas la que dota a las organizaciones de una ventaja competitiva significativa, capacitándolas para adaptarse de manera efectiva a un entorno en constante cambio.

1.2. Objetivos

En el presente trabajo se propone el desarrollo de varios modelos predictivos que posibiliten la identificación y clasificación de pacientes con alto riesgo de mortalidad tras haber sufrido una insuficiencia cardíaca. El objetivo primordial es la consecución del modelo con mayor precisión y mejor capacidad discriminatoria, así como su aplicabilidad en futuros datos y situaciones.

Con este modelo, a su vez, se persigue:

- **Mejorar la toma de decisiones clínicas**: Al proporcionar información valiosa para adaptar tratamientos y estrategias de atención, según el riesgo de mortalidad de cada paciente.
- **Optimizar los recursos hospitalarios**: Anticipando el riesgo de mortalidad de los pacientes, mejorando así la calidad de la atención y reduciendo los costes.
- **Identificar pacientes de alto riesgo**: Especialmente relevante en este caso de enfermedad con alta tasa de mortalidad, ya que permite focalizar los esfuerzos en la prevención y el tratamiento de dichos pacientes vulnerables.
- **Evaluar la eficacia de intervenciones médicas**: Al comparar tasas de mortalidad pronosticadas con las observadas es posible medir el impacto real y ajustar las estrategias, si fuera necesario.
- **Fomentar la investigación médica**: Al analizar datos se pueden identificar nuevas tendencias y factores de riesgo que pueden guiar futuras investigaciones y contribuir al avance médico.

Capítulo 2

Estado del Arte

2.1. Investigación médica

La investigación médica hace referencia al proceso de estudio sistemático y científico que se lleva a cabo para mejorar la comprensión de las enfermedades, trastornos y condiciones médicas, así como para desarrollar nuevos enfoques de prevención, diagnóstico, tratamiento y atención médica.

Desde las prácticas basadas en la superstición y la observación empírica hasta un enfoque científico y tecnológico más sofisticado, la evolución de la investigación médica se ha extendido a lo largo de miles de años y ha sido influenciada por diversos factores sociales, culturales y científicos.

En los últimos años, la investigación médica ha logrado importantes avances en diversas áreas, tales como la neurociencia o la inmunoterapia, las cuales han supuesto una auténtica transformación sobre la forma en la que entendemos y tratamos las enfermedades.

Actualmente, las mejoras en el entendimiento de las enfermedades y la aplicación de nuevas tecnologías continúan impulsando el progreso de la investigación médica. Los avances en tecnología médica y digital han mejorado el diagnóstico y tratamiento de enfermedades. La inteligencia artificial, el aprendizaje automático y el análisis de macrodatos están siendo utilizados para el diagnóstico temprano, la predicción de enfermedades y la mejora de los resultados clínicos.

En este contexto, diversos estudios están centrando su atención en el diseño de modelos predictivos que ofrezcan una respuesta analítica ante el posible riesgo en individuos de padecer algún tipo de enfermedad. Además, también pueden ayudar a optimizar la asignación de recursos médicos al identificar a los pacientes que pueden beneficiarse más de ciertos tratamientos.

Enfermedades cardiovasculares, neurodegenerativas o respiratorias están siendo objeto de estudio desde esta nueva perspectiva, con el fin de aportar información adicional a la hora establecer un diagnóstico precoz. De esta manera, se podrá realizar un seguimiento médico del paciente en conjunto con un tratamiento personalizado.

Concretamente, enfermedades médicas graves y potencialmente mortales son las que más interés suscitan en cuanto a detección temprana se refiere, razón por la cual han sido tema de investigación en varios artículos. Así, una enfermedad como la insuficiencia cardíaca sería un claro ejemplo de estudio.

2.2. Insuficiencia cardíaca

De acuerdo con la OMS, la insuficiencia cardíaca (IC) es una condición médica caracterizada por la incapacidad del corazón para bombear sangre adecuadamente y así cubrir las demandas metabólicas del cuerpo. Esta condición se caracteriza por la presencia de síntomas típicos, tales como la disnea, la fatiga o los mareos; signos físicos, como los crepitantes o los edemas; y alteraciones hemodinámicas, tanto en los niveles de presión auricular como en los niveles de hemoglobina, entre otros.

La IC está asociada a una alta morbilidad y mortalidad (De la Cámara et al., 2012), y genera tanto un impacto significativo en la calidad de vida de los pacientes, que pueden ver limitada su capacidad física a la hora de realizar actividades diarias, como un incremento en la probabilidad de padecer graves problemas de salud.

Según un artículo publicado en la Revista Sanitaria de Investigación (Lozano Alonso et al., 2021), la IC se puede clasificar en:

- IC sistólica o IC diastólica, en función de la fracción de eyección o porcentaje de sangre que es expulsada del ventrículo izquierdo durante cada contracción cardíaca.
- IC de bajo gasto o IC de alto gasto, en función de la cantidad de sangre bombeada.
- IC aguda o IC crónica, en función de la duración y progresión de los síntomas.
- IC anterógrada o IC retrógrada, en función de la dirección del flujo sanguíneo afectada.
- IC derecha, IC izquierda o IC mixta, en función de la cavidad del corazón afectada.

Es fundamental conocer y diferenciar los distintos tipos de IC y la sintomatología que produce, para así poder proporcionar a los pacientes los mejores cuidados y tratamientos.

Por otra parte, son varios factores los que contribuyen a la aparición de la IC. Entre ellos se incluyen la edad avanzada, la presión arterial alta, la diabetes, la obesidad, el tabaquismo, el consumo excesivo de alcohol y las enfermedades del sistema circulatorio. Además, esta condición puede empeorar con el tiempo si no se trata de manera adecuada, poniendo en peligro la vida del paciente.

En los últimos años, el envejecimiento de la población y la alta prevalencia de factores de riesgo cardiovascular han provocado un aumento en el número de personas afectadas por IC. Un reciente estudio reportó, en base a una muestra de 1.189.003 españoles, una prevalencia de esta enfermedad del 1,89% en la población adulta en el año 2019, alcanzando el 9% en octogenarios. La incidencia fue de 2,78 nuevos casos por cada 1.000 personas/año (Sicras-Mainar et al., 2022).

Por esta razón, se hace imprescindible abordar esta enfermedad de manera integral, con un enfoque en la prevención, el diagnóstico temprano y el tratamiento adecuado. Esto implica, entre otros aspectos, promover estilos de vida saludables para mejorar los resultados y reducir el impacto negativo de la IC.

Al margen de los cambios en el estilo de vida, los medicamentos desempeñan un papel crucial en el control de los síntomas y la mejora de la calidad de vida de los pacientes. Los medicamentos recomendados para sobrellevar la IC incluyen los inhibidores de la enzima convertidora de angiotensina (IECA), los antagonistas de los receptores de angiotensina II (ARA-II), los betabloqueadores, los diuréticos y los anticoagulantes. Estos medicamentos actúan de diferentes maneras para reducir la carga de trabajo del corazón, mejorar la función cardíaca y prevenir complicaciones.

2.3. Líneas de investigación

A lo largo de los años, numerosos estudios han contribuido al desarrollo de diversas líneas de investigación sobre la IC, que buscan mejorar la comprensión, el diagnóstico y el tratamiento de esta enfermedad.

Se han llevado a cabo investigaciones epidemiológicas y estudios de seguimiento a largo plazo que han permitido obtener una visión más clara de la prevalencia y la incidencia de la IC en diferentes poblaciones. Al mismo tiempo, se consigue recopilar información de los pacientes sobre su estado de salud, resultados clínicos, eventos adversos, etc.

En particular, algunas investigaciones están orientadas al estudio del análisis de supervivencia. Este análisis examina el tiempo hasta que ocurre un evento clínico de interés, tratando de identificar los factores de riesgo que más influyen en la predicción de estos eventos. Estos hallazgos pueden proporcionar información relevante para la toma de decisiones clínicas y el diseño de estrategias de intervención.

Otra línea de investigación importante se ha enfocado en la aplicación de técnicas de inteligencia artificial y aprendizaje automático en cardiología, impulsando el desarrollo de nuevas herramientas y métodos que podrían tener un impacto significativo en la atención médica de los pacientes con IC. Son varios estudios los que han analizado la capacidad de estos algoritmos para realizar predicciones (Lu et al., 2021). La mayoría de los algoritmos se basan en la evaluación de los síntomas presentados por los pacientes y en el análisis de los medicamentos suministrados.

Todos estos estudios han contribuido a la identificación de nuevas estrategias terapéuticas y a la elaboración de recomendaciones clínicas basadas en evidencias. También han mejorado nuestra comprensión de la enfermedad y han permitido el desarrollo de nuevos enfoques de tratamiento, sentado las bases para futuras investigaciones en este área. Sin embargo, a pesar de estos avances, aún se pueden identificar ciertos aspectos susceptibles de mejora en la optimización del tratamiento médico de la IC.

Algunos estudios revelaron una proporción significativa de pacientes que no recibían los medicamentos adecuados (Sicras-Mainar et al., 2022), lo cual implicaría la necesidad de mejorar la implementación de las recomendaciones clínicas.

Además, conviene destacar que los estudios observacionales retrospectivos pueden presentar sesgos y limitaciones en la recopilación de datos (Barge-Caballero et al., 2022), pudiendo afectar a la generalización de los resultados a otros entornos clínicos. Para confirmar y validar los resultados obtenidos, sería necesario llevar a cabo estudios prospectivos en múltiples centros.

Otro aspecto a considerar es la interpretabilidad de los modelos de aprendizaje automático. Hay estudios que han diseñado modelos complejos y sofisticados, cuyo proceso de predicción resulta en ocasiones difícil de entender (Luo et al., 2022). En consecuencia, y como resultado de la ausencia de una comprensión clara del proceso en cuestión, es importante resaltar que la garantía de una toma de decisiones fundamentada se vería potencialmente comprometida.

Se ha de tener en cuenta también que, a menudo, los estudios se basan en conjuntos de datos limitados que no representan plenamente la diversidad de la población de pacientes con IC (Zhao et al., 2022). Asimismo, es esencial garantizar la validación externa y la reproducibilidad de los resultados de estos estudios. Esto implica probar y validar los modelos en diferentes cohortes para asegurarse de que los resultados sean generalizables y aplicables en diferentes situaciones.

Por otra parte, cabe destacar la existencia de estudios de revisión, orientados a revisitar la eficacia y la utilidad de los modelos desarrollados en otros proyectos de investigación. Algunos de ellos mencionan las limitaciones en cuanto al número y tipo de variables consideradas, los métodos de aprendizaje automático, el tamaño de la muestra, el contexto clínico y el enfoque en enfermedades individuales, en lugar de abordar la multimorbilidad de forma simultánea (Banerjee et al., 2021).

Con todas estas consideraciones, se busca destacar la importancia de seguir investigando y creando nuevas estrategias terapéuticas con el propósito de tratar las necesidades no cubiertas en el manejo de la IC y optimizar el rendimiento de estos modelos.

Capítulo 3

Metodología

En relación al presente estudio, se plantea ampliar el conocimiento sobre la IC a través del análisis de diversos modelos predictivos de mortalidad en pacientes que han sufrido un episodio de esta condición médica.

Para ello, se procederá a la extracción y utilización de datos procedentes de los siguientes artículos científicos:

- Role of biological and non biological factors in congestive heart failure mortality: PREDICE-SCORE: A clinical prediction rule (De la Cámara et al., 2012).
- The prognosis of patients hospitalized with a first episode of heart failure, validation of two scores: PREDICE and AHEAD (Ruiz-Ruiz et al., 2019).

A partir de los datos anónimos, agregados y divulgados en sendos artículos científicos de referencia, se realizó un proceso de generación, variación, distorsión aleatoria, agregación y supresión sistemática de la información que mantuvo el detalle y estructura original.

Para el primer estudio, la recopilación de datos comenzó el 1 de enero de 2003 y terminó el 31 de diciembre de 2006. Por su parte, el segundo estudio inició la recopilación de información en el año 2013 y finalizó en 2015. En ambos casos, se realizó un seguimiento de los pacientes durante un año tras el alta médica.

La población de estudio consistió en 786 pacientes, procedentes de tres hospitales españoles del Servicio Nacional de Salud (Hospital Universitario 12 de Octubre en Madrid, Hospital Universitario Virgen del Rocío y Hospital Universitario Valme en Sevilla), que ingresaron urgentemente por un primer episodio de IC.

Se establecieron los siguientes criterios de inclusión para la selección de pacientes:

- Pacientes hospitalizados con un primer episodio de IC, según los criterios de Framingham.
- Pacientes adultos, con al menos 18 años.
- Pacientes residentes en el área de influencia de cualquiera de los centros hospitalarios de estudio.

Se excluyeron del estudio a pacientes con previo diagnóstico de IC, así como a aquellos cuyo diagnóstico de IC no se encontraba registrado en el informe de alta.

El protocolo del estudio incluyó información relacionada con el paciente, como la edad y el género; datos clínicos, como la presión arterial sistólica y diastólica, así como los niveles de creatinina, sodio, potasio y hemoglobina en sangre; síntomas presentados, como cardiopatía isquémica, fibrilación auricular, crepitantes, edemas, etc. Asimismo, se recopiló información acerca de los fármacos administrados, incluyendo medicamentos como IECA, betabloqueadores, ARA II, antiagregantes, anticoagulantes, diuréticos…

Con un total de 786 observaciones, la base de datos consta de 28 variables (21 variables categóricas y 7 variables numéricas), sin presentar valores ausentes. Además, todas las variables categóricas son de carácter binario, es decir, sólo tienen dos posibles respuestas.

Más en detalle, se busca predecir la **mortalidad** a un año de pacientes que han experimentado un primer episodio de IC, basándose en la información clínica registrada durante su ingreso hospitalario. La naturaleza binaria de esta **variable objetivo** implica que los algoritmos aplicados en este contexto deben enfocarse en el desarrollo de modelos de clasificación.

Las 27 variables restantes, designadas como variables predictoras, proporcionarán la información necesaria para estimar las predicciones.

Una vez llegados a este punto, es preciso definir y explicar el funcionamiento de las distintas técnicas empleadas en el diseño de los diferentes modelos predictivos.

3.1. Regresión

Cuando se quiere evaluar la relación entre una variable que suscita especial interés (variable dependiente, que suele denominarse y) respecto a un conjunto de variables (variables independientes, que se denominan $x, x_2, \dots, x_n$) resulta adecuada la aplicación de los modelos de regresión (Moral Peláez, 2006).

Los modelos de regresión se expresan de la siguiente forma:

$$y = f(x_1, x_2, \dots) + \varepsilon$$

Uno de los objetivos principales de los modelos de regresión es estimar predicciones de una característica específica en función de otras variables relacionadas. Existen diversas opciones para construir un modelo de regresión, pero dos de los más destacados son el modelo de regresión lineal y el modelo de regresión logística. La elección del modelo de regresión depende del tipo de variable que se desea predecir. El modelo de regresión lineal se utiliza cuando la variable dependiente es continua, mientras que el modelo de regresión logística se emplea cuando la variable de interés es categórica.

Es importante evaluar diversos parámetros de los modelos de regresión para identificar el modelo idóneo. En el caso del modelo de regresión lineal, su determinación se basa en el método de los mínimos cuadrados, que busca minimizar la diferencia entre los valores predichos y los valores reales. En cambio, en la regresión logística se utiliza el método de máxima verosimilitud, que se enfoca en encontrar la función de máxima probabilidad para los datos observados.

3.1.1. Regresión logística

La regresión logística es especialmente útil cuando la variable dependiente es binaria. En este caso, se efectúa el procedimiento siguiente: se establece la función de enlace *logit*, que relaciona la variable dependiente (y) con las variables independientes (x_i). El algoritmo optimiza los parámetros y estima las probabilidades de pertenencia a cada clase.

Resulta imprescindible establecer un punto de corte para asignar cada observación a una clase. El punto de corte por defecto es 0.5, lo cual implica que si la probabilidad estimada de pertenencia a una clase determinada supera el 0.5, se asigna la observación a esa clase.

El modelo de regresión logística estima la probabilidad de evento como:

$$p_1 = P(Y = 1|x_1, x_2, \dots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

Entonces, la probabilidad de no evento se define mediante:

$$p_0 = 1 - p_1 = P(Y = 0|x_1, x_2, \dots, x_m) = \frac{e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

Por tanto, se obtiene:

$$\text{logit}(p_1) = \ln\left(\frac{p_1}{1 - p_1}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

Esta función matemática representa el logaritmo de la razón de probabilidades.

Para interpretar los resultados del modelo de regresión logística, se define el concepto de 'odds' como la relación entre la probabilidad de evento y la probabilidad de no evento:

$$Odds = \frac{p}{1 - p}$$

A su vez, podemos identificar qué variables tienen una influencia significativa en la probabilidad de evento a través del análisis de los coeficientes estimados ($\boldsymbol{\beta_i}$) en la regresión logística. Estos coeficientes nos proporcionan una medida para evaluar el riesgo asociado a cada variable en el modelo.

Por otro lado, para determinar el mejor modelo de regresión logística se realiza una comparación entre diferentes modelos utilizando el cociente de verosimilitud. Si el cociente de verosimilitud no proporciona evidencia suficiente para afirmar que un modelo es superior al otro, se considera que el modelo más adecuado es el más simple. En otras palabras, se opta por el modelo con menos variables.

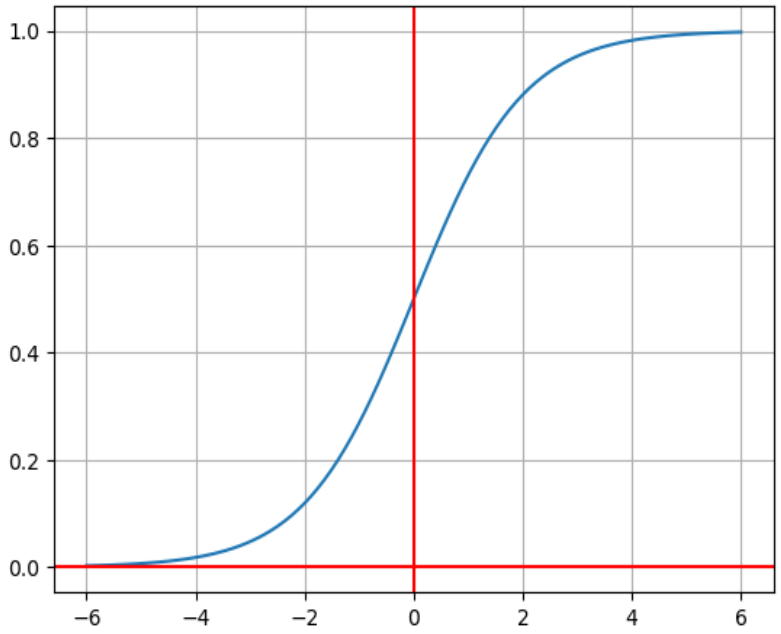

Figura 1. Función logit.

3.2. Redes neuronales

Las redes neuronales son modelos computacionales inspirados en el funcionamiento del cerebro humano, con el cual comparten muchas características debido a su similitud estructural. Estas redes tienen la habilidad de descubrir patrones y conexiones ocultas en los datos. Además, poseen la capacidad de aprender a partir de la experiencia, lo que les permite generalizar conocimientos adquiridos de casos anteriores a situaciones nuevas (Jorge Matich, 2001).

Durante el proceso de diseño de la red neuronal se generan capas internas, compuestas por unidades de procesamiento llamados nodos, que trabajan en conjunto para analizar la información de entrada.

Una red neuronal consta de una capa de entrada o capa *input*, con varios nodos de entrada o variables *input* ($X_1, X_2, \dots$); una o varias capas ocultas, compuesta cada una, a su vez, por varios nodos ocultos ($H_1, H_2, \dots$) y una capa de salida o capa *output*, que incluye el nodo de salida o variable objetivo (*Figura 2*).

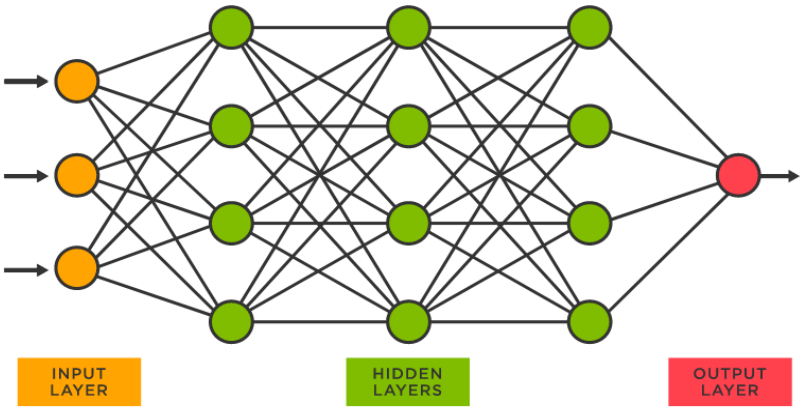

Figura 2. Esquema de una red neuronal (Google, 2021).

La información ingresa por medio de la capa *input*, se procesa a través de la capa oculta y finalmente se generan los resultados en la capa *output*.

Asimismo, en una red neuronal se encuentran los siguientes elementos:

- Pesos (w_{ij}): son parámetros que determinan la influencia de cada variable *input* en la red neuronal.
- Función de activación: es una función que permite a las redes neuronales capturar relaciones no lineales en los datos. Existen diferentes funciones de activación, entre las cuales destaca la función de activación tangente hiperbólica.

En la construcción de una red neuronal, se combinan linealmente las variables de entrada con los pesos correspondientes y se aplica una función de activación no lineal para obtener la función resultante. Utilizando la función de activación tangente hiperbólica como referencia: $\tanh(x) = 1 - \frac{2}{1+e^{2x}}$, se define la función de la red neuronal como:

$$\begin{aligned} Y = tanh\big(W_{i,out}(&tanh(w_{11}X_1 + w_{21}X_2 + w_{31}X_3 + w_{41}X_4 + b_1)) \\ &+ (tanh(w_{12}X_1 + w_{22}X_2 + w_{32}X_3 + w_{42}X_4 + b_2)) \\ &+ (tanh(w_{13}X_1 + w_{23}X_2 + w_{33}X_3 + w_{43}X_4 + b_3)) \\ &+ (tanh(w_{14}X_1 + w_{24}X_2 + w_{34}X_3 + w_{44}X_4 + b_4)) + b_{out}\big) \end{aligned}$$

Se utilizan técnicas de optimización para estimar los valores de los pesos (w_{ij}) y el sesgo (b_j) del modelo, ajustando los pesos de manera iterativa con el propósito de minimizar el error cometido.

3.3. Árboles de decisión

Los árboles de decisión, definidos como método de estimación no paramétrico, se caracterizan por su capacidad para modelar datos heterogéneos y por su buena tolerancia al ruido (Aguirre, 2019). Estos árboles se componen de nodos y forman estructuras jerárquicas, siguiendo una construcción inductiva (Rodríguez Artalejo et al., 2011). Algunos de sus componentes son:

- Nodos: divisiones del árbol que encapsulan subconjuntos de datos.
- Nodo raíz: punto de partida inicial del árbol, que engloba la totalidad de los datos.
- Nodo padre: nodo predecesor de un nodo.
- Nodo hijo: nodo sucesor de otro nodo.
- Rama: trayectoria definida entre un nodo inicial y sus nodos sucesores, que refleja una serie de decisiones y resultados en la estructura del árbol.
- Hojas: nodos terminales, que marcan el final de una rama.

A través de la aplicación secuencial de reglas de decisión simples, los árboles de decisión dividen el espacio de variables independientes en regiones distintas y no solapadas (*Figura 3*).

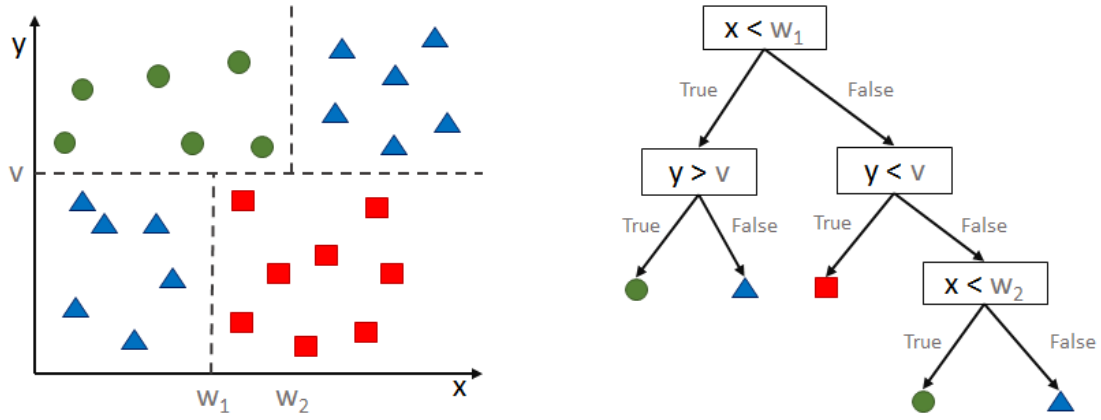

Figura 3. Esquema de un árbol de decisión (Google, 2018).

Existen varios tipos de árboles de decisión, entre los cuales destacan los árboles de clasificación y los árboles de regresión, propuestos por Leo Breiman (Breiman et al., 1984). En concreto, los árboles de regresión se emplean cuando la variable objetivo es continua, mientras que los árboles de clasificación se aplican en casos donde la variable objetivo es categórica.

Por otro lado, los árboles de decisión destacan por su capacidad para proporcionar medidas de importancia de variables y para tratar con datos faltantes o atípicos. No obstante, es importante considerar algunas de sus limitaciones, como la tendencia a desarrollar modelos demasiado complejos y, en ocasiones, difíciles de interpretar; la sensibilidad a pequeños cambios en los datos, que puede resultar en una estructura de árbol distinta; la propensión al sobreajuste, si no se controla adecuadamente los parámetros de ajuste...

Son varios los algoritmos que adoptan la configuración del árbol de decisión.

3.3.1. Bagging

El algoritmo Bagging (*Bootstrap Aggregating*) es una técnica de aprendizaje automático utilizada para mejorar la precisión y estabilidad de los modelos predictivos. Consiste en combinar múltiples modelos base para formar un modelo de predicción más robusto (Iparraguirre-Villanueva et al., 2023).

Estos algoritmos se ejecutan en paralelo y buscan aprovechar la independencia que existe entre ellos para estimar las predicciones finales mediante votación mayoritaria, en modelos de clasificación, o mediante promedio, en modelos de regresión.

Los pasos que se llevan a cabo en el proceso de implementación del algoritmo son los siguientes (*Figura 4*):

- Se crean múltiples subconjuntos de entrenamiento mediante muestreo aleatorio con reemplazo (*bootstrap*) del conjunto de datos original.
- Con cada subconjunto se entrena un modelo base.
- Las observaciones que fueron excluidas del subconjunto de entrenamiento de un modelo base (*out of the bag*) se emplean para evaluar el rendimiento del mismo.
- Las predicciones finales se determinan combinando las predicciones de todos los modelos base.

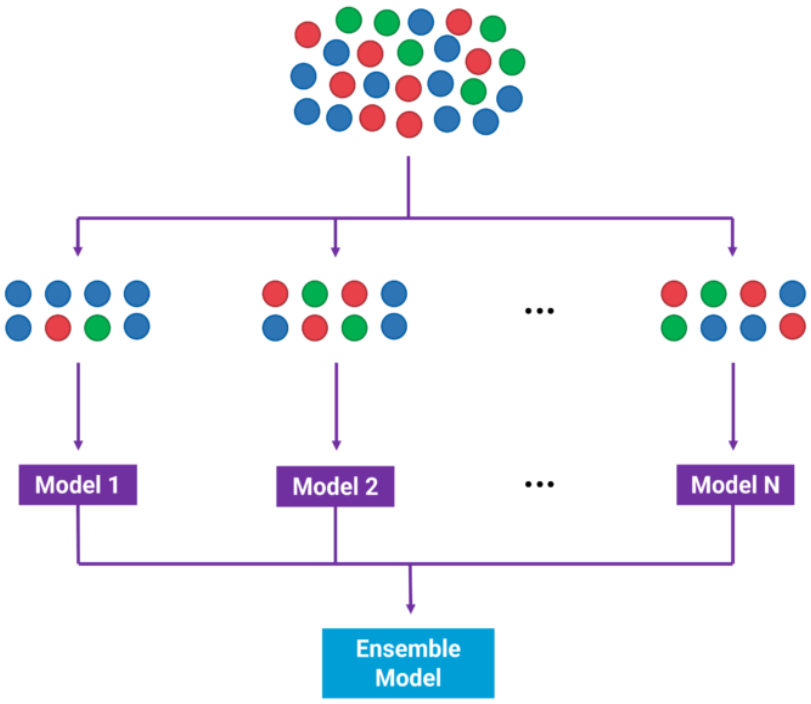

Figura 4. Esquema del algoritmo Bagging (Google, 2022a).

3.3.2. Random Forest

El algoritmo Random Forest es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión a partir de un mismo conjunto de datos. Los resultados obtenidos se combinan a fin de obtener un modelo único más sólido (Espinosa Zúñiga, 2020).

Durante la ejecución de este algoritmo, se realiza el siguiente procedimiento (*Figura 5*):

- Cada árbol generado contiene un grupo de observaciones aleatorias, elegidas mediante *boostrap*.
- Para determinar la mejor división de un nodo en un árbol, sólo se considera un subconjunto aleatorio de variables predictoras, introduciendo así mayor variabilidad entre los distintos árboles.
- Las observaciones *out of the bag* se utilizan para validar el modelo.
- Las predicciones individuales de los árboles se promedian o se votan por mayoría, dando lugar a las predicciones finales.

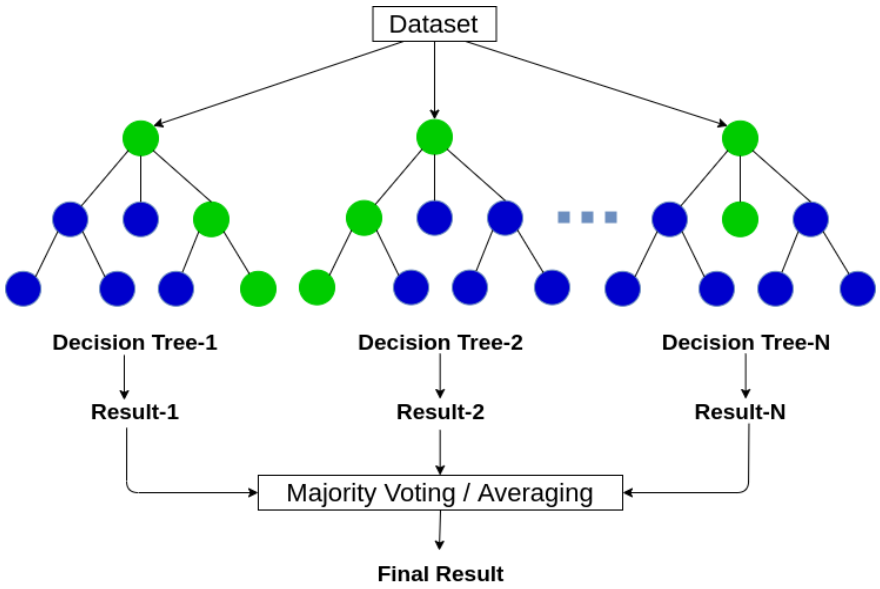

Figura 5. Esquema del algoritmo Random Forest (Google, 2022b).

Dicho algoritmo presenta diversas ventajas (Cánovas-García et al., 2017), entre las cuales destacan:

- Facilidad de entrenamiento en comparación con otras técnicas más complejas, al mismo tiempo que mantiene un rendimiento similar.
- Gran eficiencia en grandes bases de datos.
- Manejo simultáneo de numerosas variables predictoras.
- Mantenimiento del nivel de precisión en presencia de una alta proporción de datos ausentes.

Si bien, entre sus principales desventajas se encuentran:

- Visualización gráfica de los resultados difícil de interpretar.
- En presencia de ruido, tendencia al sobreajuste de ciertos grupos de datos.
- Predicciones limitadas al rango de valores del conjunto de datos de entrenamiento. En el caso de variables categóricas con diferentes clases, puede existir un resultado sesgado hacia aquellas con mayor cantidad de clases.
- Control limitado sobre el comportamiento del modelo.

3.3.3. Gradient Boosting

El algoritmo Gradient Boosting es una técnica de aprendizaje iterativa que busca construir un modelo sólido combinando múltiples clasificadores débiles de manera secuencial con el objetivo de minimizar una función de pérdida (Kuhn & Johnson, 2013). Esta función cuantifica el error cometido entre las predicciones del modelo y los valores reales del conjunto de datos.

La elección de la función de pérdida depende del tipo de problema que se esté abordando. Por ejemplo, en problemas de regresión, se suelen utilizar funciones de pérdida como el error cuadrático medio (MSE) o el error absoluto medio (MAE). Para problemas de clasificación, se puede utilizar la tasa error de clasificación.

En cada iteración, se actualizan las predicciones en la dirección de decrecimiento de la función de error, dada por el negativo del gradiente. Así, se irá mejorando gradualmente la precisión del modelo (Portela García-Miguel, 2020). No obstante, se debe tener precaución ante la posibilidad de sobreajuste, y es recomendable aplicar técnicas de regularización para mitigar este problema (*Figura 6*).

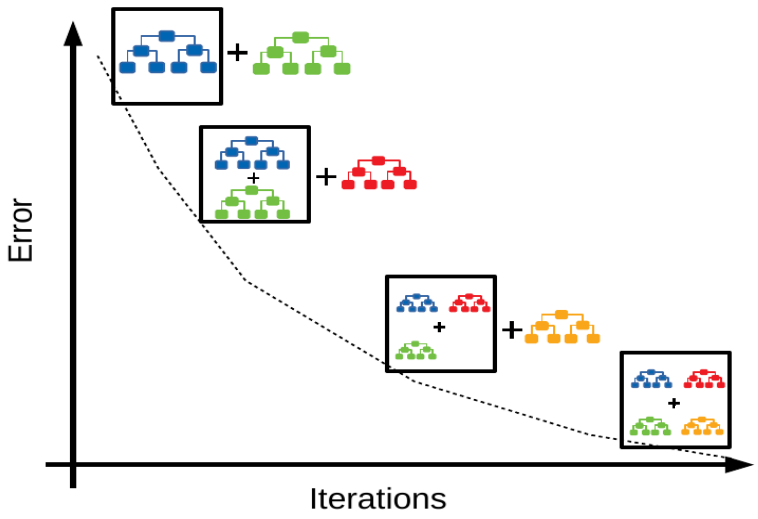

Figura 6. Funcionamiento del algoritmo Gradient Boosting (Google, 2022c).

3.3.4. Extreme Gradient Boosting

El algoritmo Extreme Gradient Boosting es una técnica de aprendizaje automático diseñada por Tianqi Chen (Chen & Guestrin, 2016), que nació con la idea de crear un sistema escalable del algoritmo Gradient Boosting (*Figura 7*).

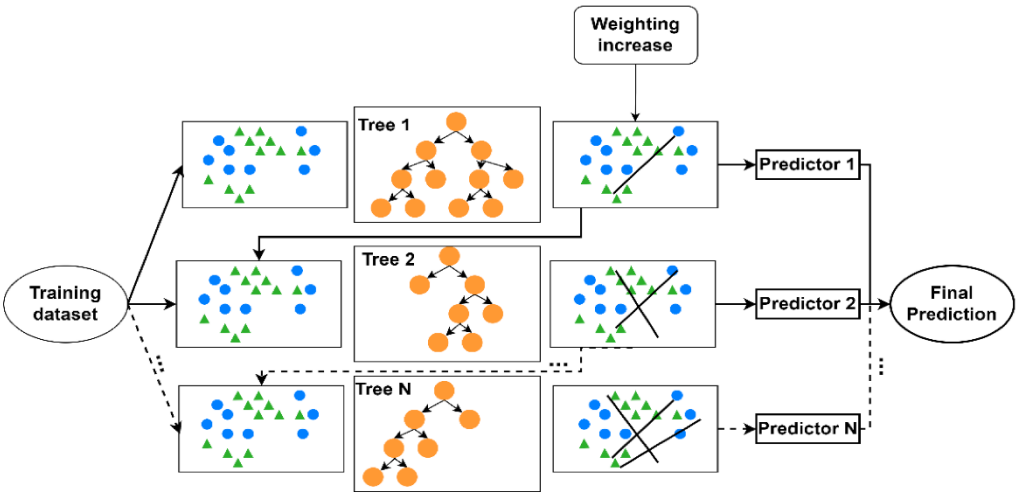

Figura 7. Estructura del algoritmo Extreme Gradient Boosting (Google, 2022d).

El funcionamiento de dicho algoritmo se detalla a continuación (Espinosa Zúñiga, 2020):

- Se genera un árbol inicial F_1 para predecir la variable objetivo y. El resultado se asocia con un error residual $\boldsymbol{y} - F_1$.
- Se utiliza el error residual previo para ajustar un nuevo árbol h_2.
- Se inserta un parámetro de penalización α_2 para evitar el sobreajuste.
- Se combinan los resultados de F_1 y h_2 para obtener el árbol F_2, donde el error cuadrático medio de F_2 será menor que el de F_1:

$$F_2\,(x) = F_1\,(x) + \alpha_2 h_2(x, r_0)$$

- Este proceso se sigue iterativamente hasta minimizar el error lo máximo posible:

$$F_m\,(x) = F_{m-1}(x) + \alpha_m h_m(x, r_{m-1})$$

Algunas de las ventajas que ofrece este algoritmo son:

- Manejo de grandes bases de datos con múltiples variables.
- Tratamiento de valores perdidos.
- Resultados muy precisos.
- Excelente velocidad de ejecución.

En contrapartida, se evidencian las siguientes desventajas:

- Elevado consumo de recursos computacionales en grandes bases de datos.
- Necesidad de calibrar adecuadamente los parámetros del algoritmo con el fin de minimizar el error de precisión y evitar el sobreajuste del modelo.
- Estandarización o normalización previa de las variables categóricas.

3.4. Support Vector Machine

El algoritmo Support Vector Machine (SVM) es una técnica de aprendizaje automático que busca encontrar el hiperplano óptimo para separar de manera efectiva dos o más clases de instancias en un conjunto de datos, maximizando la distancia entre el hiperplano de separación y los vectores soporte (Rojas et al., 2020), que son las instancias de entrenamiento más cercanas de cada clase (*Figura 8*).

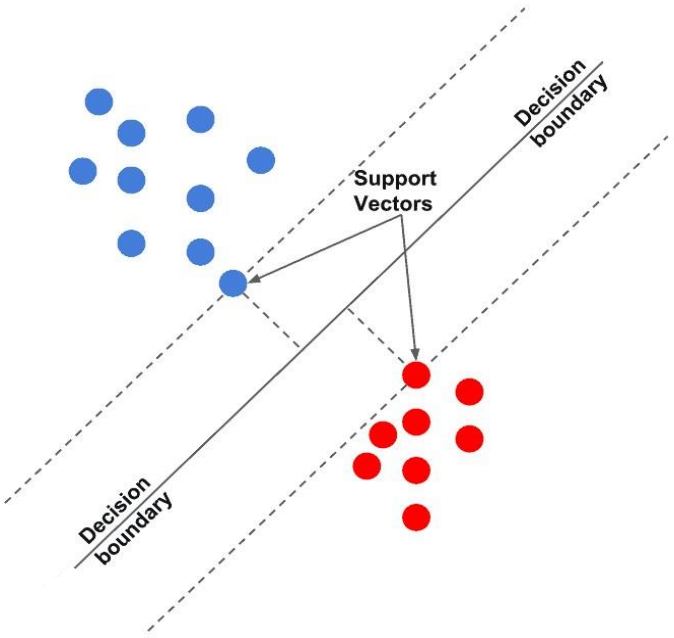

Figura 8. Criterio de clasificación del algoritmo SVM (Google, 2022e).

Aunque dicho algoritmo fue inicialmente diseñado para trabajar con conjuntos de datos separables linealmente, se puede extender fácilmente para clasificar datos no lineales utilizando funciones *kernel*. Estas funciones mapean los datos de entrenamiento a un espacio de mayor dimensión, donde se pueden separar mediante un hiperplano (Noble, 2006).

Existen multitud de *kernels* distintos, de entre los cuales se comentarán brevemente algunos de los más utilizados.

3.4.1. SVM lineal

El *kernel* lineal es el más simple y se emplea al asumir una separabilidad lineal de los datos en el espacio original. Su función consiste en llevar a cabo una transformación lineal directa de los datos hacia un espacio de mayor dimensionalidad (*Figura 9*).

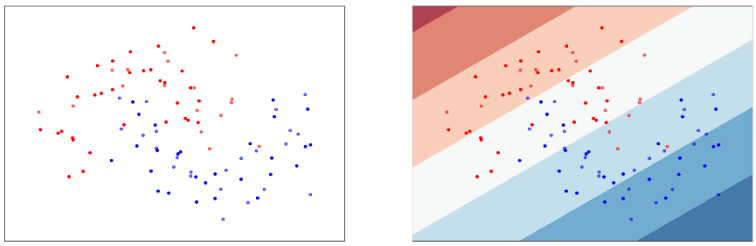

Figura 9. Separación lineal SVM (Google, 2020f).

3.4.2. SVM polinomial

El *kernel* polinomial utiliza una función polinómica para mapear los datos a un espacio de mayor dimensión, permitiendo modelar relaciones no lineales. El grado del polinomio define la complejidad de la transformación (*Figura 10*).

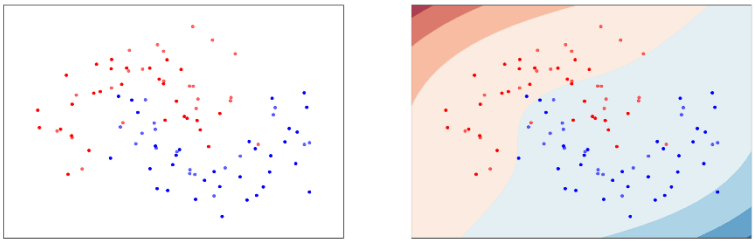

Figura 10. Separación polinomial SVM (Google, 2020f).

3.4.3. SVM radial

El *kernel* gaussiano aplica una función de base radial para mapear los datos a un espacio infinitamente dimensional, siendo capaz de modelar relaciones no lineales complejas (*Figura 11*).

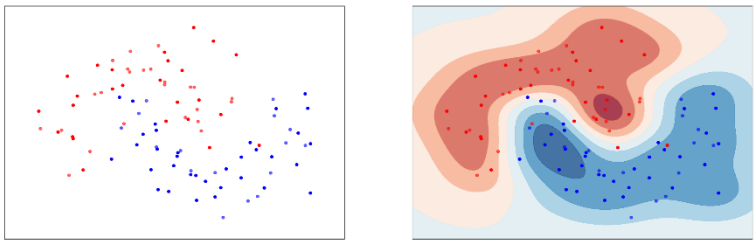

Figura 11. Separación radial SVM (Google, 2020f).

3.5. Ensamblados

El ensamblado estadístico, también conocido como ensamblaje de modelos, es un conjunto de metodologías que tienen como objetivo mejorar el rendimiento de modelos individuales, buscando aprovechar la diversidad y las fortalezas de los mismos (Recarey Fernández, 2021).

Cada modelo individual puede tener limitaciones o sesgos propios, lo que dificulta alcanzar resultados óptimos por sí solo. Sin embargo, al combinar y ponderar adecuadamente los resultados de cada uno de estos modelos, se puede establecer un modelo final robusto (Martínez Cisternas, 2018).

El ensamblado estadístico puede emplear diversas técnicas, como votación, promedio o incluso métodos más sofisticados. Estas técnicas se encargan de combinar y ponderar las predicciones de los modelos individuales para generar una predicción final más precisa y confiable.

3.6. Metodología SEMMA

La metodología SEMMA, desarrollada por *SAS Institute*, establece una lista de etapas secuenciales que guían la implementación de modelos de aprendizaje automático, garantizando la comprensión de los datos, la preparación adecuada, la construcción de modelos robustos y la evaluación del rendimiento de los mismos.

En este contexto, se describen brevemente cada una de las etapas de esta metodología:

- **Muestreo** (*SAMPLE*): selección de una muestra representativa de los datos disponibles. En la realización de nuestro estudio, y considerando el reducido número de observaciones disponibles, se opta por utilizar el conjunto completo de datos para así lograr un desarrollo lo más completo y detallado.
- **Explorar** (*EXPLORE*): exploración y visualización de los datos, para poder detectar posibles tendencias, inconsistencias, datos ausentes o anomalías.
- **Modificar** (*MODIFY*): modificación de los datos mediante la creación, selección y transformación de variables, con el objetivo de optimizar y agilizar el proceso de modelización.
- **Modelizar** (*MODEL*): generación de modelos a partir del conjunto de datos de entrenamiento, previo ajuste de ciertos parámetros específicos e inherentes a cada modelo.
- **Evaluar** (*ASSES*): evaluación de nuestras predicciones y comparación de los modelos obtenidos.

Dado su carácter estructurado, resulta oportuno utilizar esta metodología en el desarrollo de nuestro trabajo, que ha sido principalmente elaborado con el lenguaje de programación *RStudio*. Asimismo, se ha hecho uso del programa *SAS Enterprise Miner Workstation 14.1* para corroborar los resultados obtenidos.

Capítulo 4

Configuración de los datos

4.1. Descripción de variables

En primer lugar, se recogen las 21 variables binarias presentes en la base de datos, especificando las dos clases correspondientes a cada variable (*Tabla 1)*.

VARIABLE	DESCRIPCIÓN
sexo	Sexo del paciente: • hombre • mujer
cardiopatia_isquemica	¿El paciente ha sufrido una cardiopatía isquémica?: • si • no
fibrilacion_auricular	¿El paciente ha presentado fibrilación auricular?: • si • no
crepitantes	¿El paciente ha presentado crepitantes?: • si • no
tercer_tono_cardiaco	¿El paciente ha presentado tercer tono cardíaco?: • si • no
ingurgitacion_yugular	¿El paciente ha sufrido ingurgitación yugular?: • si • no
hepatomegalia	¿El paciente ha sufrido hepatomegalia?: • si • no
reflujo_hepatoyugular	¿El paciente ha sufrido reflujo hepatoyugular?: • si • no
edemas	¿El paciente ha presentado edemas?: • si • no
inhibidores_ECA	¿Se le ha suministrado medicamentos inhibidores de la ECA al paciente?: • si • no
betabloqueadores	¿Se le ha suministrado medicamentos betabloqueadores al paciente?: • si • no
antagonistas_receptores _angiotensina_II	¿Se le ha suministrado medicamentos antagonistas de los receptores de angiotensina II al paciente?: • si • no
antagonistas_calcio	¿Se le ha suministrado medicamentos antagonistas de calcio al paciente?: • si • no
antiagregantes	¿Se le ha suministrado antiagregantes al paciente?: • si • no
anticoagulantes_orales	¿Se le ha suministrado anticoagulantes (orales) al paciente?: • si • no
digoxina	¿Se le ha suministrado digoxina al paciente?: • si • no
diureticos	¿Se le ha suministrado diuréticos al paciente?: • si • no
estatinas	¿Se le ha suministrado estatinas al paciente?: • si • no
mononitrato_isosorbida	¿Se le ha suministrado mononitrato de isosorbida al paciente?: • si • no
valvulopatias	¿El paciente ha presentado valvulopatías?: • si • no
mortalidad	¿El paciente falleció al año de sufrir el primer episodio de insuficiencia cardíaca?: • Yes • No

Tabla 1. Descripción de las variables binarias, resaltando la variable objetivo: mortalidad.

De igual forma, se describen las 7 variables numéricas (*Tabla 2*).

VARIABLE	DESCRIPCIÓN
edad	Edad del paciente.
presion_arterial_sistolica	Presión arterial sistólica del paciente (mmHg).
presion_arterial_diastolica	Presión arterial diastólica del paciente (mmHg).
creatinina	Niveles de creatinina en sangre (mg/dL).
sodio	Niveles de sodio en sangre (mEq/L ~ miliequivalentes/L).
potasio	Niveles de potasio en sangre (mEq/L).
hemoglobina	Niveles de hemoglobina (g/dL).

Tabla 2. Descripción de las variables numéricas.

Por otra parte, al evaluar el balanceamiento de la variable objetivo, se puede observar que contamos con 647 casos de pacientes que no han fallecido, frente a 139 casos que sí (*Tabla A.1*). Este desbalanceamiento en la variable a predecir tiene consecuencias en el proceso de aprendizaje de los modelos, ya que no aprenden por igual de ambas clases. Si bien existen posibles soluciones para abordar este efecto, no se considera oportuno aplicarlas con el propósito de preservar el número real de observaciones en la base de datos.

4.2. Análisis exploratorio y gráfico

El análisis de las variables predictoras se realiza en base a las representaciones gráficas (*Figura B.1 - Figura B.27*) y las tablas numéricas de cada variable (que se adjunta en el enlace GitHub del Anexo).

- **edad**: La mayoría de pacientes tienen entre 60 y 89 años.
- **sexo**: Hay pocas más mujeres que hombres, aunque el porcentaje de fallecidos en ambos sexos es muy parecido.
- **cardiopatia_isquemica**: Sólo un 15% de los pacientes registrados han sufrido dicho síntoma.
- **fibrilacion_auricular**: Casi un 30% de los pacientes registrados han sufrido fibrilación auricular.
- **presion_arterial_sistolica**: Aunque los valores de presión arterial sistólica varían en función de la edad, el rango normal en adultos sanos se considera, generalmente, entre 90 y 130 mmHg. En las etapas iniciales de la insuficiencia cardíaca, la presión arterial sistólica puede estar dentro del rango normal o ligeramente elevada. Sin embargo, a medida que la enfermedad progresa y el corazón se debilita, la presión arterial sistólica tiende a disminuir. Podemos observar que gran parte de los pacientes registrados han sobrepasado este límite, lo cual es lógico al haber sufrido un episodio de insuficiencia cardíaca.

- **presion_arterial_diastolica**: De forma análoga a la variable anterior, aunque los valores de presión arterial diastólica varían en función de edad, sí que hay un cierto límite a partir del cual hablamos de hipertensión arterial: 80 mmHg. Se observa que algo menos de la mitad de pacientes registrados han sobrepasado este límite.
- **crepitantes**: Más de 2/3 de los pacientes han presentado tal síntoma.
- **tercer_tono_cardiaco**: Apenas un 5.4% de los pacientes registrados han presentado tercer tono cardíaco.
- **ingurgitacion_yugular**: Sólo un 20.9%, aproximadamente, de los pacientes registrados han sufrido dicho síntoma.
- **hepatomegalia**: Tan sólo un 11.5% de los pacientes registrados han sufrido este síntoma.
- **reflujo_hepatoyugular**: Se puede observar que los pacientes que han presentado tal síntoma no llegan al 6% de los totales registrados.
- **edemas**: Más de 2/3 de los pacientes registrados presentaron edemas.
- **creatinina**: Los niveles normales de creatinina en sangre se estiman entre 0.7 y 1.3 mg/dL en hombres y entre 0.6 y 1.1 mg/dL en mujeres. Aunque dependen de varios factores, como la edad. Niveles superiores a 1.3 mg/dL en hombres o 1.1 mg/dL en mujeres pueden ser signo de una insuficiencia cardíaca padecida. En nuestro caso, casi el 25.5% de los pacientes hombres registrados presentaron niveles de creatinina en sangre superiores a 1.3 mg/dL. Por otra parte, poco más del 38.6% de las pacientes mujeres registradas han presentado niveles de creatinina en sangre superiores a 1.1 mg/dL.
- **sodio**: Los niveles normales de sodio en sangre se estiman entre 135 y 145 mEq/L. Aquellos niveles inferiores a 135 mEq/L o superiores a 145 mEq/L pueden ser síntoma de haber padecido una insuficiencia cardíaca. Poco más del 17.8% de los pacientes registrados han presentado niveles de sodio en sangre anormales.
- **potasio**: Nuevamente, los niveles normales de potasio en sangre se estiman entre 3.7 y 5.2 mEq/L. Cualquier valor por debajo de 3.7 mEq/L o por encima de 5.2 mEq/L puede provocar alteraciones en la función cardíaca que pueden llevar a la insuficiencia cardíaca. En nuestro registro, algo más del 19.4% de los pacientes presentaron niveles anormales.
- **hemoglobina**: En este caso, los niveles normales de hemoglobina en sangre rondan de 13.2 a 16.6 g/dL en los hombres y de 11.6 a 15 g/dL en las mujeres. Si los niveles son menores a aquellos, puede ser debido al padecimiento de una insuficiencia cardíaca. En nuestros datos, algo más de la mitad de los pacientes hombres registrados presentaron unos niveles de hemoglobina en sangre menores a 13.2 g/dL y, aproximadamente el 31.4% de las pacientes mujeres registradas presentaron unos niveles de hemoglobina en sangre menores a 11.6 g/dL.
- **inhibidores_ECA**: A más de la mitad de los pacientes registrados se les suministraron medicamentos inhibidores de la ECA.
- **betabloqueadores**: Un 35.9% de los pacientes registrados aproximadamente fueron suministrados con betabloqueadores.

- **antagonistas_receptores_angiotensina_II**: En este caso, el medicamento fue suministrado al 18.2% de los pacientes registrados.
- **antagonistas_calcio**: Solamente a un 15.8% de los pacientes registrados se les suministraron medicamentos antagonistas del calcio.
- **antiagregantes**: Los antiagregantes fueron suministrados al 44.5% de los pacientes registrados.
- **anticoagulantes_orales**: Las estadísticas nos muestran que, prácticamente, al 35.5% de los pacientes registrados se les suministraron anticoagulantes orales.
- **digoxina**: Casi al 30% de los pacientes registrados se les suministró digoxina.
- **diuréticos**: En este caso, al 74% de los pacientes registrados se les suministraron diuréticos.
- **estatinas**: Las estatinas fueron suministradas a algo más del 30.5% de los pacientes registrados.
- **mononitrato_isosorbida**: Según los registros, no más del 10% de los pacientes fueron suministrados con mononitrato de isosorbida.
- **valvulopatías**: Presentaron valvulopatías casi un 30% de los pacientes registrados.

Adicionalmente, el contraste de independencia (Test χ^2) efectuado proporciona evidencias de independencia (al 95% de confianza) de ciertas variables respecto a la variable objetivo (*Tabla 3*).

VARIABLE	P-VALOR
diureticos	1.000000e+00
antagonistas_calcio	1.000000e+00
cardiopatia_isquemica	8.685562e-01
sexo	7.668782e-01
reflujo_hepatoyugular	7.489397e-01
tercer_tono_cardiaco	6.557209e-01
digoxina	4.584902e-01
fibrilacion_auricular	3.997827e-01
antagonistas_receptores_angiotensina_II	3.585522e-01
anticoagulantes_orales	3.443763e-01
antiagregantes	3.101318e-01
hepatomegalia	2.926918e-01
ingurgitacion_yugular	2.055266e-01
mononitrato_isosorbida	1.590583e-01
valvulopatias	1.133923e-01
estatinas	6.948108e-02
edemas	2.457329e-02
crepitantes	4.572004e-03
inhibidores_ECA	4.352885e-03
betabloqueadores	2.737137e-03

Tabla 3. P-valor obtenido del Test χ^2.

Asimismo, en la representación gráfica de la matriz de correlación entre variables numéricas, no existe correlación entre variables por encima de 0.65, es decir, no se aprecian variables altamente correlacionadas (*Figura 12*).

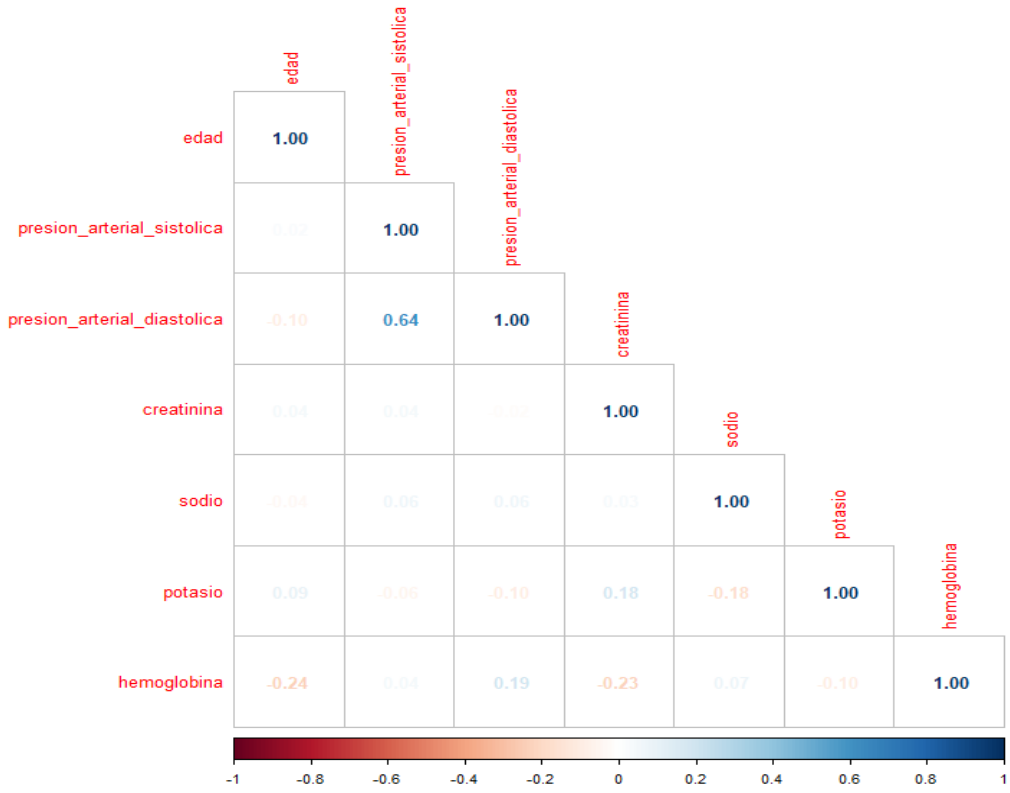

Figura 12. Gráfico de matriz de correlación entre variables numéricas.

4.3. Modificación de los datos

En base al análisis exploratorio previo de las variables predictoras, se plantean las siguientes modificaciones:

- Eliminación de clases con menos de 20 observaciones para evitar sobreajuste.
- Normalización por rango a las variables numéricas, facilitando su comparación y análisis.
- Creación de variables *dummy*, simplificando el análisis y la interpretación de los datos.

4.4. Selección inicial de variables

Al modelar diversos algoritmos, es de singular importancia determinar qué variables son las más relevantes de cara a estimar predicciones precisas de la variable objetivo. Una cuidadosa selección inicial de variables contribuye a evitar el sobreajuste, mejorando la capacidad de generalización del modelo y reduciendo el ruido presente en los datos.

En este caso, los métodos de selección de variable que se van a implementar son:

- **SBF**: Emplea técnicas de búsqueda y evaluación de subconjuntos para identificar un subespacio óptimo de características relevantes en un conjunto de variables, basándose en la idea de que las características relevantes pueden estar correlacionadas entre sí y formar un subespacio en el espacio de características.
- **RFE**: Elimina iterativamente las variables menos importantes en un modelo, evaluando el rendimiento y seleccionando las características más relevantes hasta obtener un subconjunto óptimo.
- **AIC**: Evalúa diferentes modelos con diferentes subconjuntos de variables y selecciona aquellos que minimizan el AIC, un criterio de información que penaliza la complejidad del modelo.
- **BIC**: Similar al método anterior, evalúa diferentes modelos con distintas combinaciones de variables y selecciona aquellos que minimizan el BIC, un criterio de información bayesiano que penaliza la complejidad del modelo.
- **Boruta**: Compara la importancia de las variables originales con un conjunto de variables aleatorias, determinando cuáles son las variables realmente significativas y descartando las que no aportan información relevante al modelo.
- **MMPC**: Utiliza técnicas de aprendizaje de redes bayesianas para encontrar las variables más relevantes en función de sus relaciones de dependencia con otras variables, seleccionando aquellas que tienen un mayor impacto en el fenómeno de interés.
- **SES**: Selecciona variables que ofrecen información estadísticamente equivalente, simplificando el conjunto de variables sin perder información relevante.
- **AIC repetido**: Evalúa iterativamente el método AIC.
- **BIC repetido**: Análogamente, evalúa iterativamente el método BIC.

Una vez se hayan seleccionado las variables mediante cada uno de estos métodos (*Tabla A.2*) y eliminando aquellos que seleccionen las mismas variables, se avanzará a las fases de modelización y evaluación de los diversos algoritmos.

Capítulo 5

Modelización y evaluación

La creación y análisis de los diferentes modelos predictivos se lleva a cabo en las fases de modelización y evaluación. Previamente, se sugiere ajustar los parámetros correspondientes a cada modelo con el objetivo de lograr la configuración óptima. En última instancia, se valorará y comparará el rendimiento de cada modelo generado mediante técnicas de validación y métricas de evaluación.

Se denomina tuneado al proceso de optimización de los parámetros de los diferentes modelos de aprendizaje automático. Implica explorar diferentes combinaciones de valores para los parámetros y evaluar cómo afectan al desempeño del modelo. Aquel conjunto de valores que proporcione un resultado equitativo entre una alta precisión y un bajo error conformará la combinación idónea.

Por otro lado, la partición de los datos en conjuntos de entrenamiento y prueba se lleva a cabo mediante validación cruzada repetida. Con diferentes divisiones de los datos en cada repetición, el modelo se entrena en un subconjunto de entrenamiento y se evalúa en el subconjunto de prueba, obteniéndose la tasa de fallos y el AUC. Esta técnica ayuda a reducir el impacto del azar en la evaluación del modelo y proporciona una estimación más confiable del rendimiento promedio del modelo en diferentes conjuntos de datos.

En este contexto, la tasa de fallos representa el porcentaje de observaciones clasificadas incorrectamente en relación al total de observaciones. Una tasa de fallos baja indica un buen rendimiento del modelo, ya que ha logrado clasificar correctamente la mayoría de las observaciones. Por ende, una tasa de fallos alta indica una baja precisión del modelo y la necesidad de mejoras en su capacidad predictiva.

Paralelamente, el AUC (área bajo la curva ROC) es una métrica que evalúa la capacidad de discriminación del modelo para distinguir entre clases. Su valor varía entre 0.5 y 1, donde un valor cercano a 1 indica un modelo con alta capacidad de discriminación y buena capacidad para clasificar correctamente las observaciones.

Por tanto, la tasa de fallos y el AUC obtenidos por los diferentes modelos permiten contrastar el rendimiento de los mismos. En concreto, a través de validación cruzada repetida, se registran varias medidas de tasa de fallos y AUC para cada modelo. Estos resultados se representan en gráficos boxplot, que reflejarán su distribución y variabilidad. De esta forma, se puede realizar las comparaciones pertinentes entre los diferentes modelos desarrollados.

5.1. Modelos de regresión logística

Con cada conjunto de variables seleccionado se genera un modelo de regresión logística. Se comparan las tasas de fallos y el AUC obtenido entre los distintos modelos (*Figura 13 y 14*).

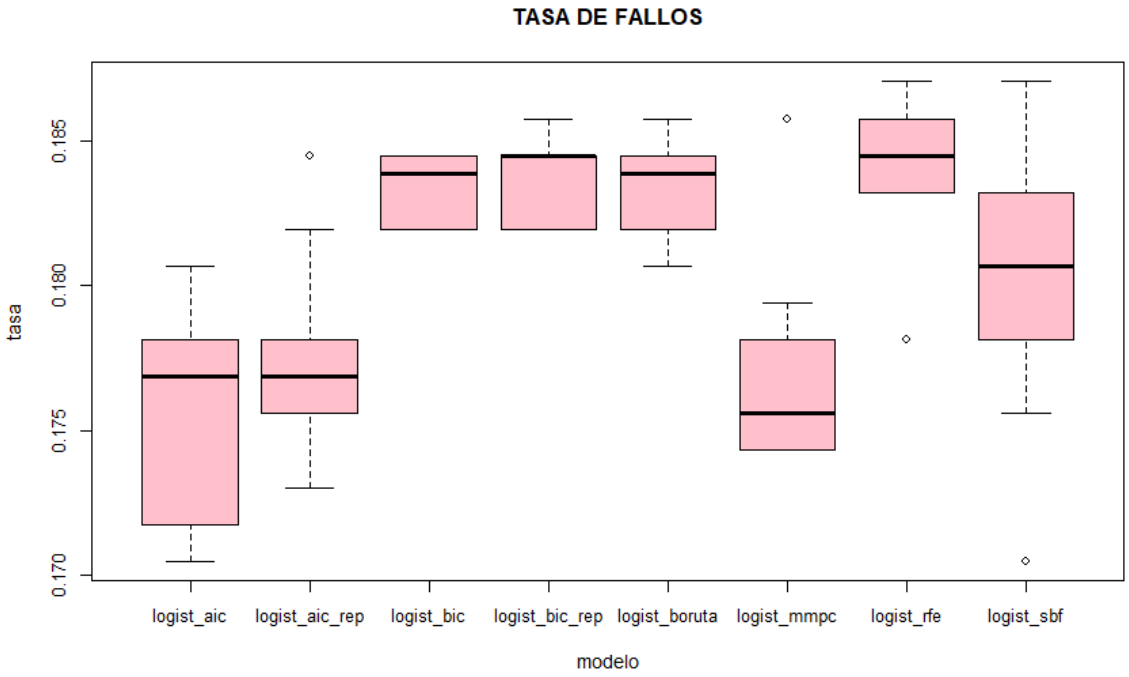

Figura 13. Tasa de fallos de los modelos de regresión logística.

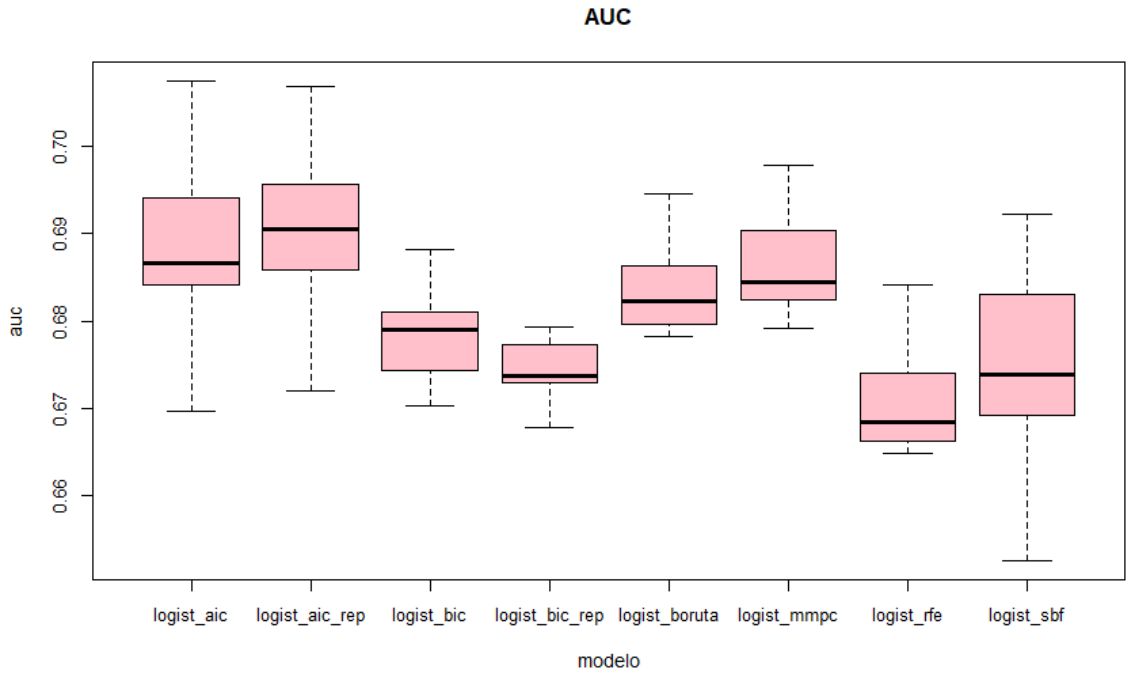

Figura 14. AUC de los modelos de regresión logística.

En base al análisis conjunto de la tasa de fallos, la varianza de los resultados y el AUC obtenido por cada modelo, se considera óptimo el modelo logist_mmpc.

Por otra parte, conviene destacar que el conjunto de variables asociado al método MMPC será implementado en la modelización de los algoritmos que no proporcionen un orden de importancia de variables propio.

5.2. Modelos de redes neuronales

En el proceso de modelización del algoritmo de red neuronal, resulta necesario estimar los valores óptimos de los parámetros:

- **size**: número de nodos ocultos en cada capa oculta de la red neuronal.
- **decay**: Coeficiente de decaimiento que controla el ajuste a los pesos de la red neuronal durante el proceso de entrenamiento.
- **maxit**: número máximo de iteraciones permitidas durante el entrenamiento de la red neuronal. Cada iteración corresponde a un ciclo completo en el que se ajustan los pesos de la red para minimizar la función de pérdida.

Particularmente, el valor máximo del parámetro *size* se determina mediante la siguiente expresión:

$$h = \frac{n^o parámetros\ máx. - 1}{k + 2}$$

Donde:

- h: número máximo de nodos ocultos en una capa específica de la red neuronal.
- k: número de variables *input*: 6 variables.
- $n^\circ\ parámetros\ máx. = \frac{n^\circ\ obs.\ clase\ minoritaria\ de\ la\ variable\ objetivo}{n^\circ\ obs./parámetro}$

Con 139 observaciones de la clase minoritaria de la variable objetivo y considerando 22 observaciones por parámetro, se llega a:

$$n^\circ\ parámetros\ máx. = \frac{139}{22} \approx 6$$

Por tanto:

$$h\ = \frac{6 - 1}{6 + 2} \approx 1\ nodos\ máx.$$

Este resultado obtenido, no obstante, no arroja un valor lógico, por lo que se sugiere establecer 5 nodos como mínimo aceptable para la configuración de la red neuronal. Esta elección busca mejorar el rendimiento y evitar un sobreajuste del modelo.

Respecto al resto de parámetros, se propone tunearlos como se observa (*Figura 15*).

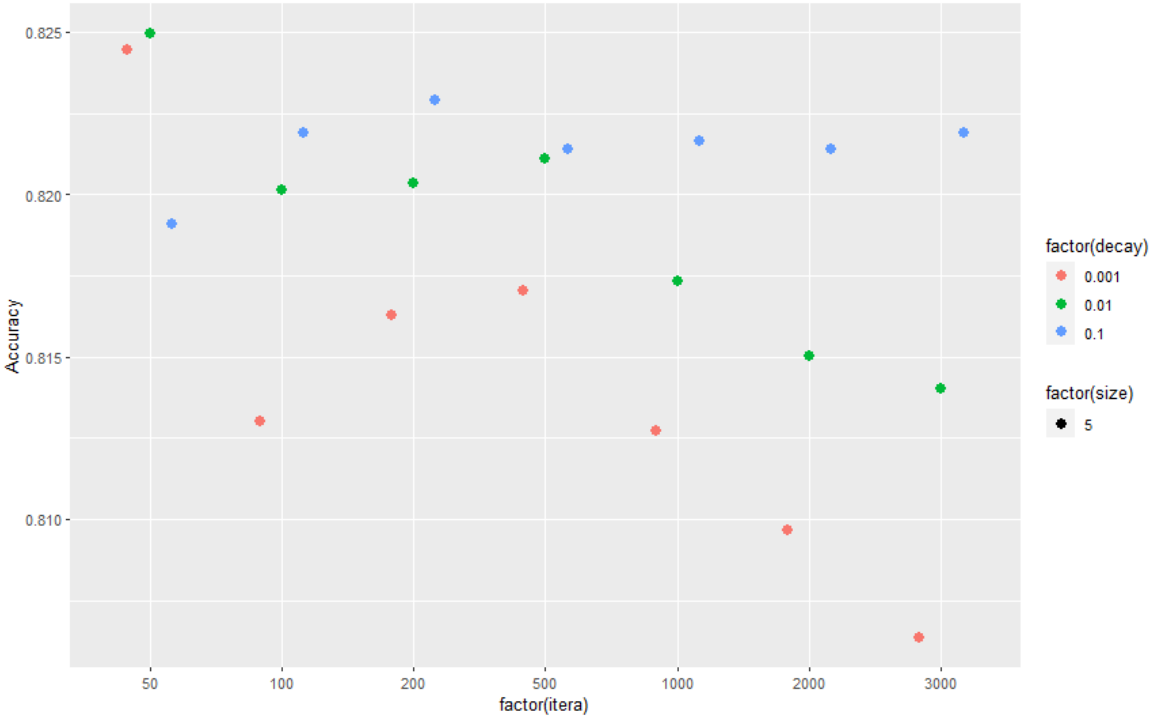

Figura 15. Accuracy asociado a cada modelo de red neuronal.

Tras evaluar el Accuracy, y sin perder capacidad de generalización, la configuración idónea corresponde al modelo con parámetros: **size** = 5, **decay** = 0.01 y **maxit** = 100.

5.3. Modelos Bagging

En este caso, los modelos base mencionados en la teoría se corresponderán con árboles de decisión. Además, este algoritmo proporciona orden de importancia de variables, que permitirá identificar aquellas variables de mayor relevancia en el modelo (*Figura B.28*), propiciando así una selección de variables propia para modelos Bagging y Random Forest.

Este algoritmo cuenta con los parámetros:

- **mtry**: número de variables predictoras que se consideran al seleccionar la mejor división en cada nodo de un árbol base. Según el orden de importancia de variables, se seleccionan 6 variables.
- **ntree**: número de árboles de decisión que se modelizarán.
- **sampsize**: tamaño de cada subconjunto.
- **nodesize**: número mínimo de observaciones requeridas en las hojas de cada árbol. Un valor 10 se considera adecuado, buscando la precisión del modelo y sin perder capacidad de generalización.

En primer lugar, el cambio de la tasa de error respecto al parámetro *ntree* se ilustra en la *Figura B.29*.

El valor máximo del parámetro *sampsize* debe ser menor que el número de observaciones de entrenamiento, y se determina mediante la siguiente expresión:

$$n° \, max. \; sampsize = \frac{k-1}{k} \cdot \; n° \, obs. totales$$

Donde:

- k: número de grupos en los que se dividirá la muestra por validación cruzada.

Con 786 observaciones totales en nuestra base de datos y dividiendo la muestra en 4 grupos para realizar validación cruzada, se llega a:

$$n° \, max. \; sampsize = \frac{3}{4} \cdot \; 786 \approx 589$$

En consecuencia, se plantea generar tantos modelos como valores considerados del parámetro *sampsize* (*Figura 16 y 17)*.

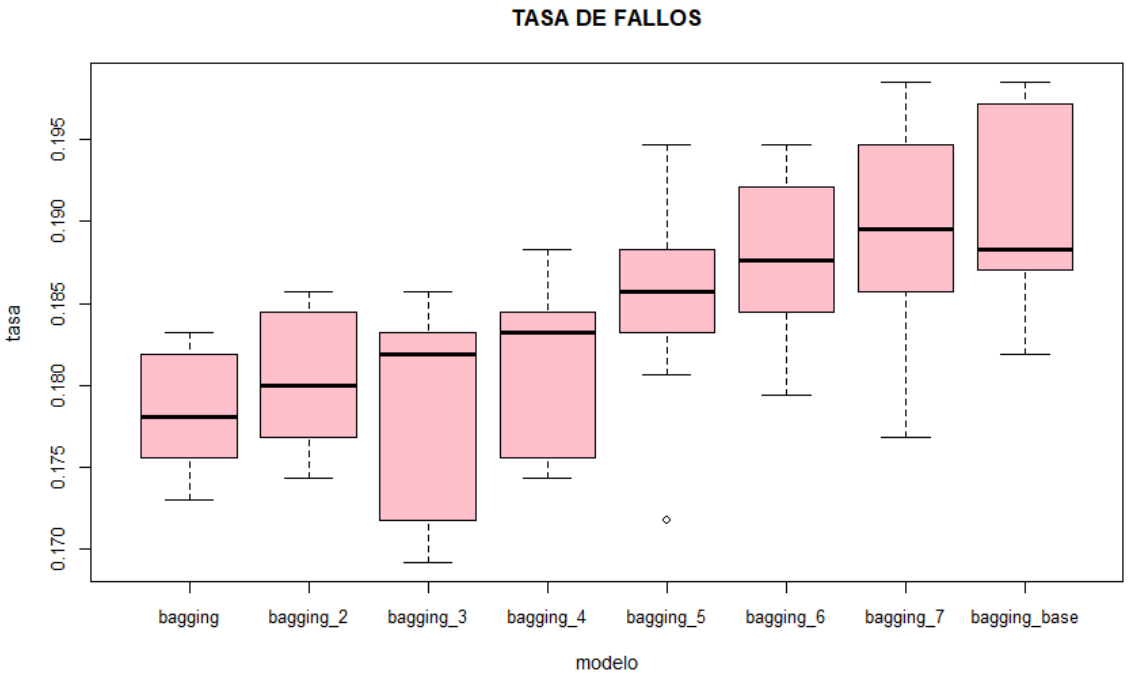

Figura 16. Tasa de fallos de los modelos Bagging.

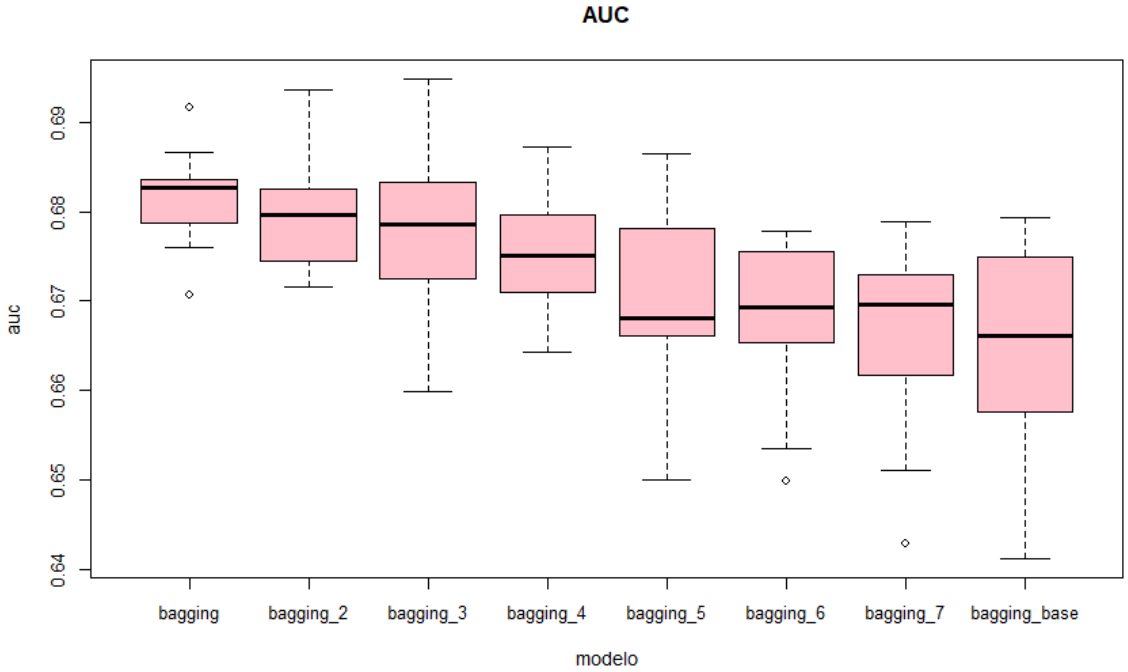

Figura 17. AUC de los modelos Bagging.

El rendimiento óptimo se corresponde con el modelo bagging, que cuenta con los parámetros siguientes: **mtry** = 6, **ntree** = 300 y **sampsize** = 50.

5.4. Modelos Random Forest

El algoritmo Random Forest incluye los siguientes parámetros a optimizar:

- **mtry**: número de variables predictoras que se consideran al seleccionar la mejor división en cada nodo de un árbol.
- **ntree**: número de árboles de decisión que se modelizarán.
- **sampsize**: tamaño de cada subconjunto.
- **nodesize**: número mínimo de observaciones requeridas en las hojas de cada árbol. Un valor 10 parece adecuado, buscando la precisión del modelo sin perder capacidad de generalización.

El parámetro *mtry* se tunea con los valores registrados en la *Tabla A.3*. De igual modo, en la *Figura B.30* se representa la variación de la tasa de error en función del parámetro *ntree*.

Por su parte, el parámetro *sampsize* se ajusta de igual forma al caso anterior. Los resultados representan las tasas de fallos y AUC de los distintos modelos (*Figura 18 y 19*).

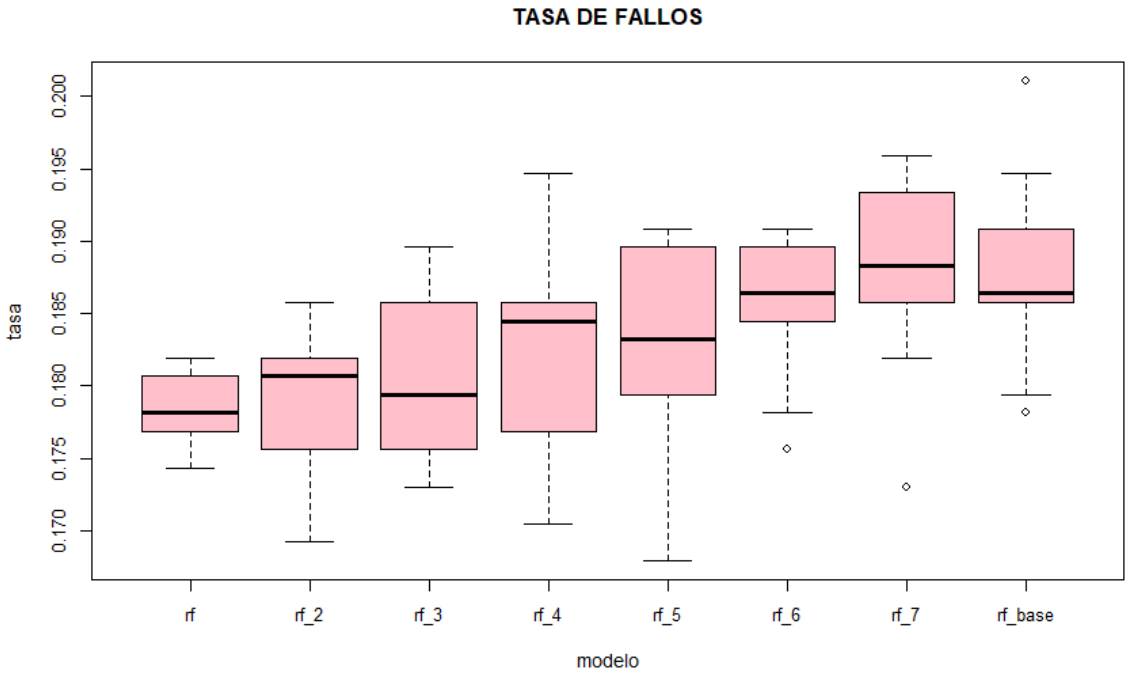

Figura 18. Tasa de fallos de los modelos Random Forest.

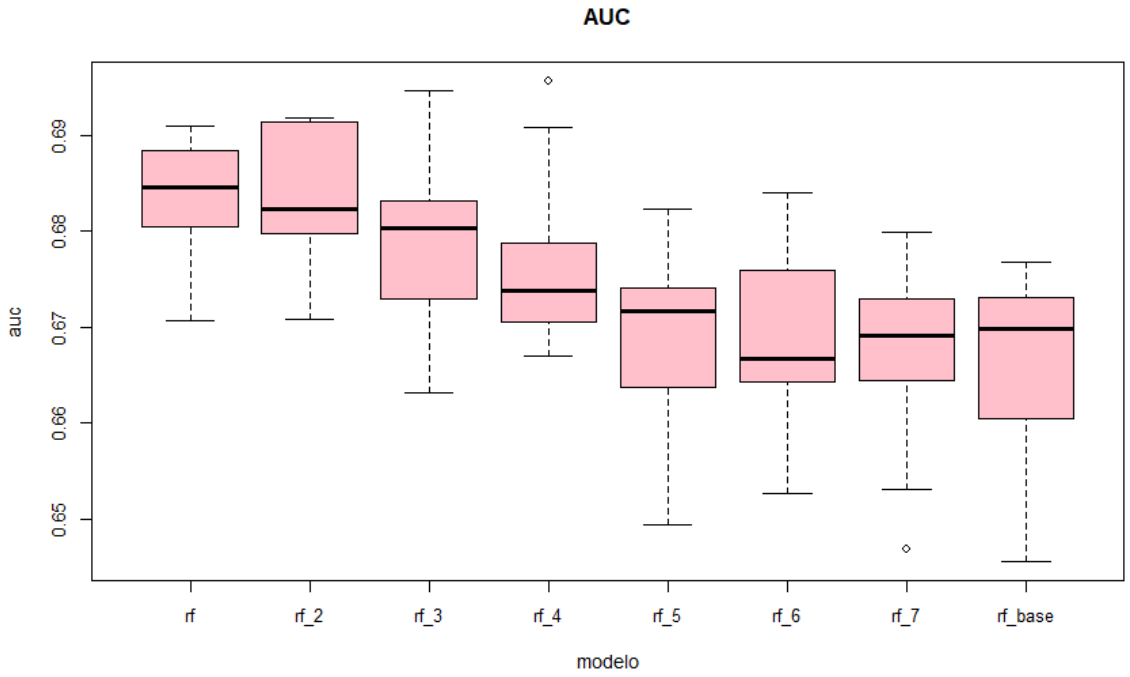

Figura 19. AUC de los modelos Random Forest.

El modelo rf, caracterizado por los parámetros: **mtry** = 4, **ntree** = 200 y **sampsize** = 50, muestra un rendimiento óptimo, entre los diferentes modelos evaluados.

5.5. Modelos Gradient Boosting

Este algoritmo presenta orden de importancia de variables, favoreciendo así una selección de variables más específica (*Figura B.31*). En algoritmos Gradient Boosting se estiman los parámetros:

- **shrinkage**: controla la contribución de cada árbol al modelo final.
- **n.minobsinnode**: número mínimo de observaciones requeridas en las hojas de cada árbol.
- **n.trees**: número de árboles de decisión que se modelizarán.
- **bag.fraction**: fracción de observaciones que se utilizarán para ajustar cada árbol. Un valor 1 es aceptable, dada las pocas observaciones en nuestro conjunto de datos.
- **interaction.depth**: controla la profundidad máxima de cada árbol en el modelo. Por simplicidad, y pudiendo ser tuneado, se probará un nivel 2 de profundidad.

Los parámetros *shrinkage* y *n.minobsinnode* se ajustan según se observa en la *Figura B.32*. Asimismo, en la *Figura B.33* se representa la variación del Accuracy en función del parámetro *n.trees*.

En definitiva, el conjunto de parámetros: **shrinkage** = 0.05, **n.minobsinnode** = 20 y **n.trees** = 75, conforman la configuración óptima del modelo.

4.5. Modelos Extreme Gradient Boosting

En el desarrollo de modelos Extreme Gradient Boosting, el orden de importancia de variables establece un criterio adecuado para la selección de variables (*Figura B.34*). Por su parte, los parámetros a tunear son:

- **min_child_weight**: número mínimo de observaciones requeridas en las hojas de cada árbol.
- **eta**: controla la contribución de cada árbol al modelo final.
- **nrounds**: número de árboles de decisión que se modelizarán.
- **max_depth**: controla la profundidad máxima de cada árbol en el modelo. Un valor de 6 limita a estos niveles la profundidad de cada árbol.
- **gamma**: regula la cantidad mínima de pérdida requerida para que se realice una partición en un nodo. Se propone un valor 0, por lo que no se aplicará ninguna restricción en la reducción de pérdida y todas las particiones que mejoren incluso mínimamente la función de pérdida serán consideradas.
- **colsample_bytree**: fracción de variables que se utilizarán para construir cada árbol. Dada la selección de variables previa, se sugiere tomar todas las variables, por lo que un valor 1 es correcto.
- **subsample**: fracción de observaciones que se utilizarán para ajustar cada árbol. Un valor 1 es aceptable, dado las pocas observaciones en nuestro conjunto de datos.

La calibración de los parámetros *min_child_weigth* y *eta* figuran en la *Figura B.35*, mientras que la *Figura B.36* muestra la relación entre el Accuracy y el parámetro *nrounds*.

Entre los posibles resultados, la combinación de los parámetros: **min_child_weight** = 20, **eta** = 0.05 y **nrounds** = 75, da lugar al modelo óptimo.

5.6. Modelos SVM lineal

Para modelos SVM lineal, se busca optimizar el siguiente parámetro:

- **C**: controla el nivel de regularización del modelo para lograr un buen equilibrio entre la minimización del error y la maximización del margen.

Si bien no se aprecia variación del Accuracy con respecto al parámetro *C*, se prefieren valores intermedios con el fin de no concurrir en sobreajuste ni pérdida de generalización del modelo (*Figura 20*).

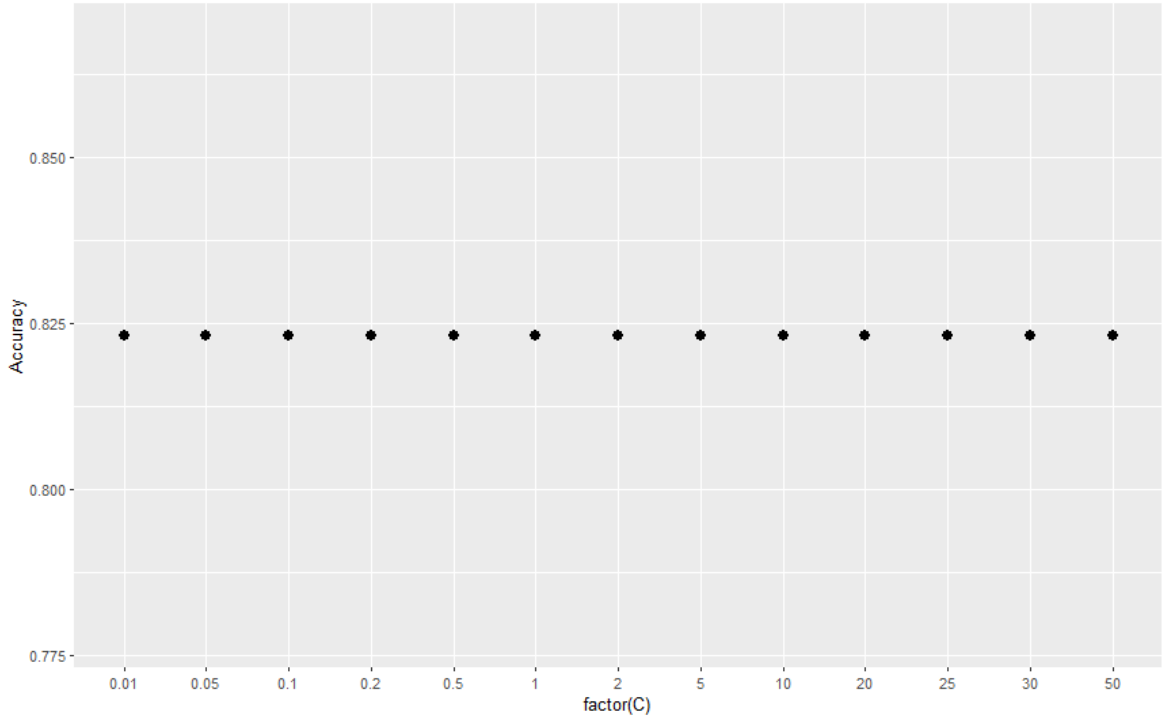

Figura 20. Accuracy asociado a cada modelo SVM lineal.

Por esta razón, el modelo con parámetro: **C** = 0.1, puede considerarse apropiado.

5.7. Modelos SVM polinomial

Modelos basados en el algoritmo SVM polinomial permiten ajustar los parámetros:

- **C**: controla el nivel de regularización del modelo para lograr un buen equilibrio entre la minimización del error y la maximización del margen.
- **degree**: grado del polinomio utilizado en la función del *kernel* polinomial. Define la complejidad de la superficie de decisión polinomial. Con motivo de no complicar excesivamente el modelo, se propone una función de 2º grado.
- **scale**: factor de escala polinomial aplicado a los datos que influye en la importancia relativa de cada variable.

El Accuracy correspondiente a distintas configuraciones de los parámetros mencionados permite discernir cuál es la opción más conveniente (*Figura 21*).

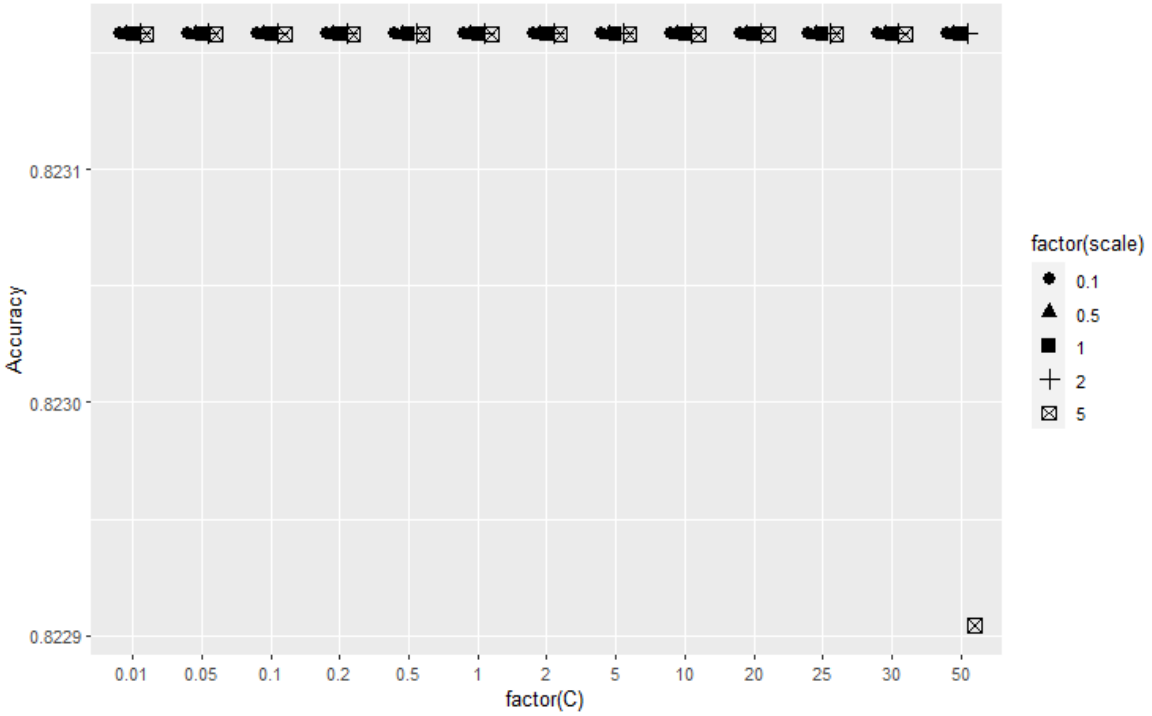

Figura 21. Accuracy asociado a cada modelo SVM polinomial.

Ante la falta de una clara opción, se opta por elegir los parámetros: **C** = 0.5 y **scale** = 5, para conformar el modelo.

5.8. Modelos SVM radial

En modelos SVM radial, los parámetros a optimizar son:

- **C**: controla el nivel de regularización del modelo para lograr un buen equilibrio entre precisión y capacidad de generalización.
- **sigma**: controla la influencia que cada instancia de entrenamiento tiene en la construcción de los límites de decisión.

La representación gráfica de las distintas combinaciones de parámetros nos permite identificar cuáles proporcionan un mejor Accuracy del modelo (*Figura 22*).

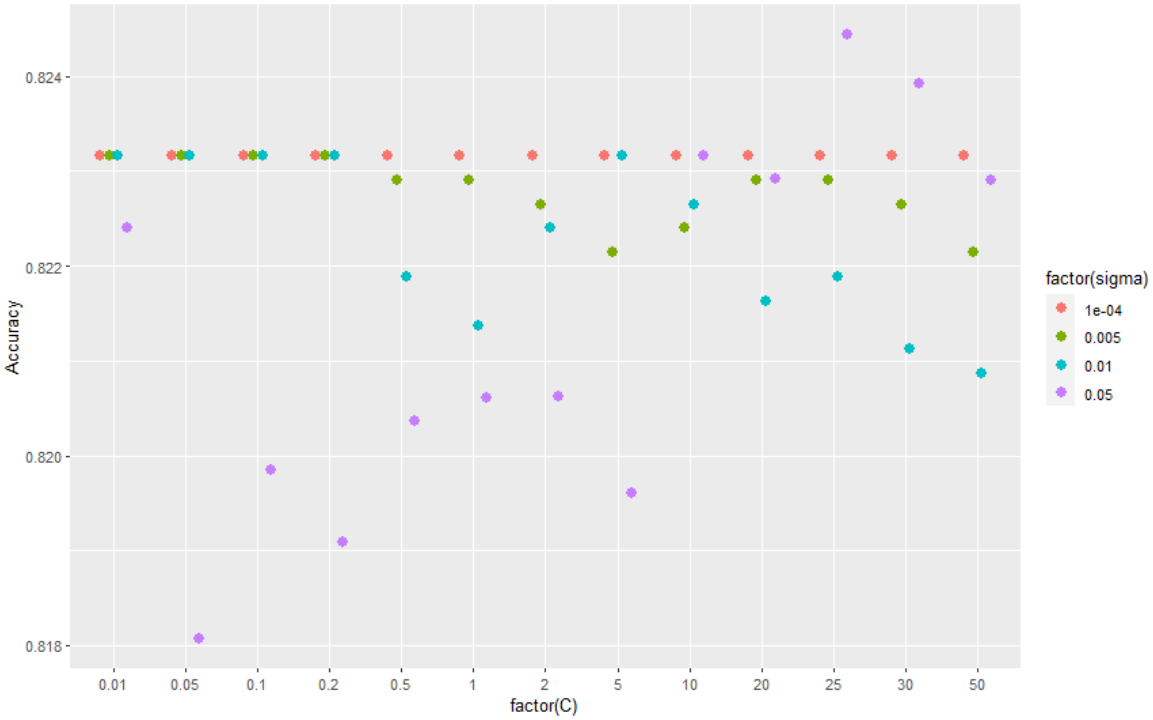

Figura 22. Accuracy asociado a cada modelo SVM radial.

En este caso, los parámetros que optimizan el rendimiento del modelo a la vez que pretenden evitar un sobreajuste son: **C** = 2 y **sigma** = 0.005.

5.9. Modelos ensamblados

Identificados los mejores modelos de cada algoritmo, se procede al ensamblado de varios modelos mediante promedio. Se combinan todos los modelos por pares, e incluso en grupos de tres, cuatro o cinco, en función del rendimiento de los mismos.

Se evaluaron hasta 82 combinaciones diferentes (predi10, …, predi91), de las cuales fueron consideradas aquellas que demostraron un desempeño más satisfactorio en términos de tasa de fallos y AUC.

El proceso de evaluación concluye con la comparación y análisis global de todos los modelos seleccionados (*Figuras 23 y 24)*. Cada modelo es evaluado con los parámetros óptimos (*Tabla 4*).

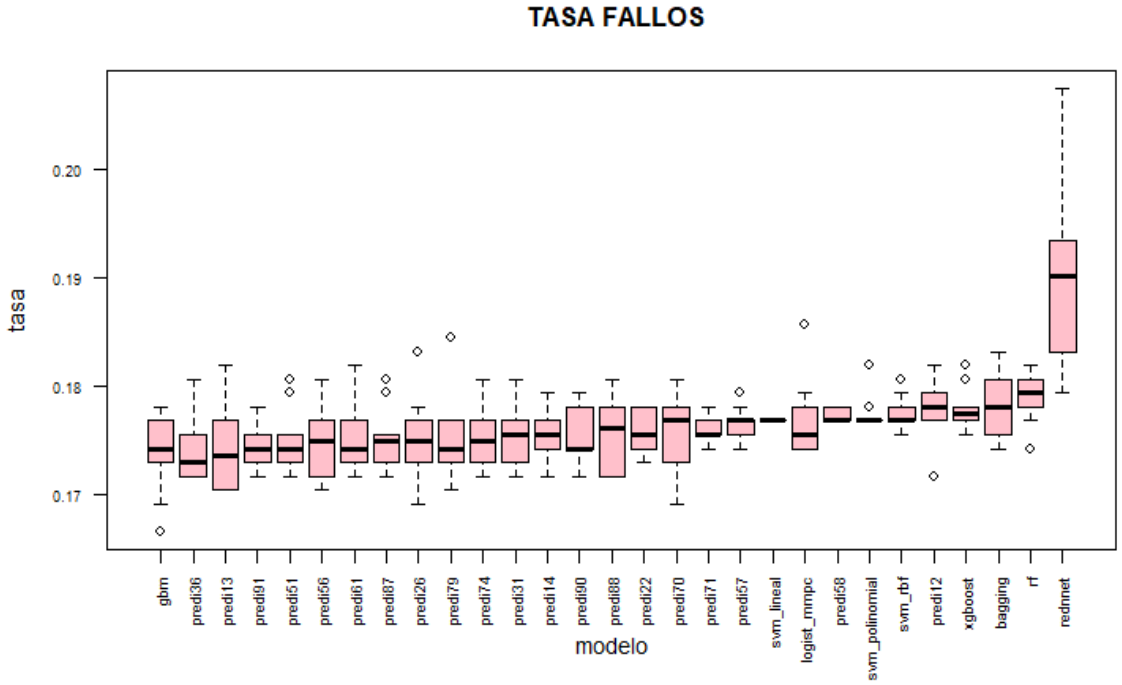

Figura 23. Tasa de fallos de los modelos finales.

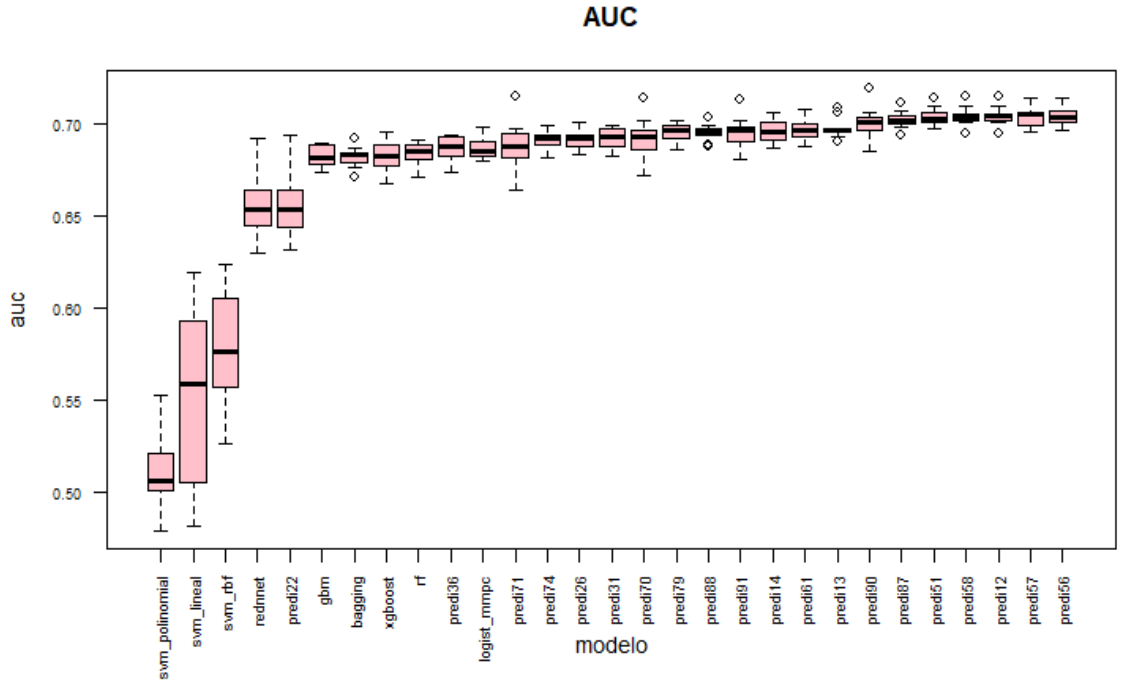

Figura 24. AUC de los modelos finales.

MODELO	DEFINICIÓN	PARÁMETROS
logist_mmpc	modelo de Regresión Logística	
rednnet	modelo de Red Neuronal	size = 5, decay = 0.01 y maxit = 100
bagging	modelo de Bagging	mtry = 6, ntree = 300, sampsize = 50 y nodesize = 10
rf	modelo de Random Forest	mtry = 4, ntree = 200, sampsize = 50 y nodesize = 10
gbm	modelo de Gradient Boosting	shrinkage = 0.05, n.minobsinnode = 20, n.trees = 75, bag.fraction = 1 e interaction.depth = 2
xgboost	modelo de Extreme Gradient Boosting	min_child_weight = 20, eta = 0.05, nrounds = 75, max_depth = 6, gamma = 0, colsample_bytree = 1 y subsample = 1
svm_lineal	modelo de SVM lineal	C = 0.1
svm_polinomial	modelo SVM polinomial	C = 0.5, degree = 2 y scale = 5
svm_rbf	modelo SVM radial	C = 2 y sigma = 0.005
predi12	promedio de los modelos logist_mmpc y rf	
predi13	promedio de los modelos logist_mmpc y gbm	
predi14	promedio de los modelos logist_mmpc y xgboost	
predi22	promedio de los modelos rednnet y svm_lineal	
predi26	promedio de los modelos bagging y gbm	
predi31	promedio de los modelos rf y gbm	
predi36	promedio de los modelos gbm y xgboost	
predi51	promedio de los modelos logist_mmpc, bagging y gbm	
predi56	promedio de los modelos logist_mmpc, rf y gbm	
predi57	promedio de los modelos logist_mmpc, rf y xgboost	
predi58	promedio de los modelos logist_mmpc, rf y svm_lineal	
predi61	promedio de los modelos logist_mmpc, gbm y xgboost	
predi70	promedio de los modelos rf, rednnet y xgboost	
predi71	promedio de los modelos rf, rednnet y svm_lineal	
predi74	promedio de los modelos rf, bagging y gbm	
predi79	promedio de los modelos rf, gbm y xgboost	
predi87	promedio de los modelos logist_mmpc, rf, bagging y gbm	
predi88	promedio de los modelos rf, bagging, xgboost y gbm	
predi90	promedio de los modelos logist_mmpc, rf, bagging, xgboost y rednnet	
predi91	promedio de los modelos rf, bagging, xgboost, gbm y rednnet	

Tabla 4. Modelos finales seleccionados.

Es importante tomar decisiones prudentes en la selección del mejor modelo, valorando tanto la calidad de ajuste como la adecuación a la realidad de los datos disponibles. A partir de los resultados obtenidos, y en base a una minuciosa revisión de los mismos, considerando además la complejidad de los modelos diseñados, se determina que el **modelo logist_mmpc** es la elección más apropiada y efectiva.

Capítulo 6

Resultados

La elección del modelo **logist_mmpc** sugiere que el algoritmo de regresión logística será el más adecuado para conseguir un mejor ajuste a nuestros datos y la obtención de resultados óptimos. En consecuencia, se procede a estimar las predicciones, cuyos resultados se plasman en la subsiguiente matriz de confusión (*Tabla 5*) y curva ROC (*Figura 25*).

	REFERENCIA	
PREDICCIÓN	Yes	No
Yes	8	8
No	131	639

Tabla 5. Matriz de confusión con punto de corte 0.5.

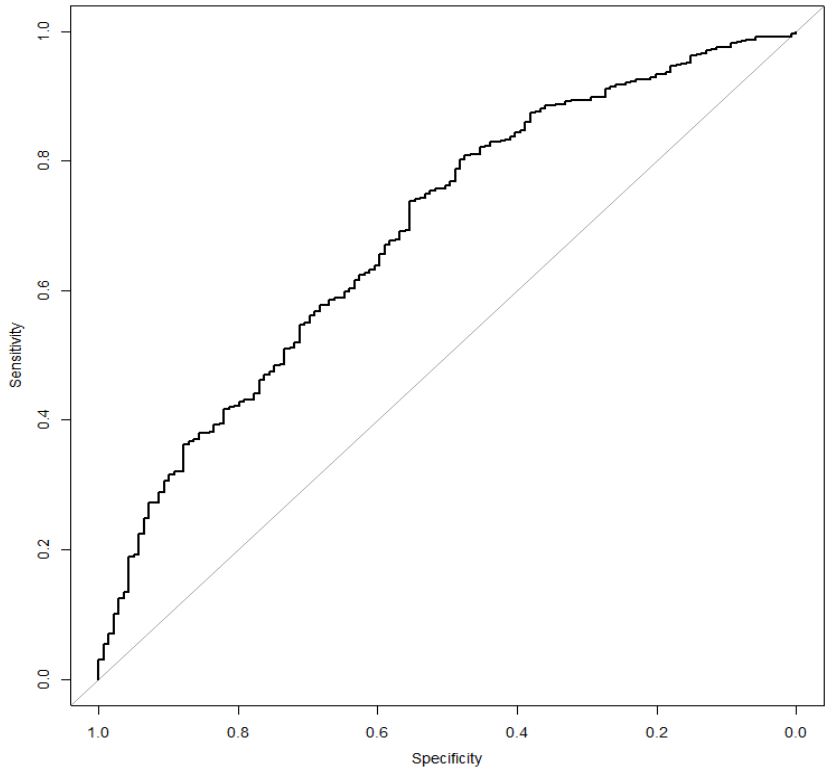

Figura 25. Curva ROC del modelo logist_mmpc.

Aunque es cierto que se logra una correcta clasificación de la mayor parte de los pacientes, no podemos obviar la notable dificultad en la correcta detección de aquellos que han fallecido.

Conforme a las estimaciones previas, se constatan los siguientes valores:

- Sensibilidad o tasa de verdaderos positivos: 0.0576.
- Especificidad o tasa de verdaderos negativos: 0.9876.
- Accuracy: 0.8232.
- Tasa de fallos: 0.1768.
- Precisión: 0.5000.
- AUC: 0.6896.

Si bien estos resultados aportan información válida en relación a la clasificación de pacientes con IC, uno de los principales objetivos al concebir el modelo es lograr una correcta identificación de pacientes de alto riesgo, en la medida de lo posible. Con este propósito y con la intención de mejorar la sensibilidad del modelo, se plantea ajustar el punto de corte. Para ello, y con objeto de determinar el punto de corte óptimo que maximice la capacidad de discriminación del modelo, se recurrirá al método Youden, que busca un equilibrio adecuado entre la sensibilidad y la especificidad.

Con la implementación del nuevo punto de corte (0.2), se procederá a realizar una reevaluación de las predicciones y así determinar si ha conferido mejoras al desempeño del modelo (*Tabla 6*).

	REFERENCIA	
PREDICCIÓN	Yes	No
Yes	77	182
No	62	465

Tabla 6. Matriz de confusión con punto de corte 0.2.

Se observa una notable mejora en la sensibilidad a costa de un empeoramiento de la especificidad. Paralelamente, se ve que la precisión de nuestro modelo ha disminuido ligeramente ya que, al disminuir el punto de corte, ha disminuido la proporción de pacientes fallecidos clasificados correctamente sobre el total de pacientes identificados como fallecidos. También nuestro Accuracy ha disminuido ya que hay menos pacientes totales clasificados correctamente.

En resumen:

- Sensibilidad o tasa de verdaderos positivos: 0.5540.
- Especificidad o tasa de verdaderos negativos: 0.7187.
- Accuracy: 0.6896.
- Tasa de fallos: 0.1768.
- Precisión: 0.2973.
- AUC: 0.6896.

A pesar de que estos resultados podrían parecer menos favorables en comparación con los anteriores, se ha logrado atenuar la tasa de falsos positivos, al mismo tiempo que se demuestra prudencia respecto a pacientes con una menor probabilidad de riesgo.

También resulta conveniente determinar algunos parámetros de nuestro modelo de regresión logística (*Tabla 7*).

VARIABLES	ESTIMATE	STD. ERROR	Z VALUE	PR(>|z|)	SIGNIF. CODES
(Intercept)	-2.2126	0.7565	-2.925	0.00345	**
edad	3.4261	0.7526	4.552	5.31e-06	***
presion_arterial_sistolica	-1.8427	0.7815	-2.358	0.01838	*
creatinina	2.4478	0.7609	3.217	0.00130	**
sodio	-1.7254	0.7179	-2.403	0.01624	*
crepitantes_no	-0.5462	0.2507	-2.179	0.02934	*
inhibidores_ECA_si	-0.4875	0.1974	-2.470	0.01351	*

Tabla 7. Tabla de parámetros de la regresión logística.

Los coeficientes estimados evidencian la significativa relevancia estadística de todas las variables predictoras, destacando la edad y la creatinina. Los signos de los coeficientes sugieren que la mayoría de las variables, salvo la edad y creatinina, están asociadas a una disminución en la probabilidad de fallecimiento.

6.1. Modelo de árbol de decisión

Como último recurso, se procede a modelar un árbol de decisión simple con el fin de ilustrar el efecto de cada variable seleccionada en este modelo, así como de informar de la existencia o no de interacciones entre las mismas (*Figura 26*).

Para ello, se ajustarán los parámetros:

- **minbucket**: número mínimo de observaciones requeridas en las hojas de cada árbol. Análogamente a los modelos Bagging y Random Foret, se fijan 10 observaciones mínimas.
- **cp**: medida de penalización por la complejidad del árbol. Se probará un valor 0, por lo que no se aplicará ninguna restricción por complejidad al construir el árbol y se incluirán todas las divisiones posibles.

Cada nodo, a su vez, contendrá la información siguiente:

- Número de observaciones en cada nodo (en porcentaje).
- Número de observaciones de la clase minoritaria de la variable objetivo en cada nodo (en porcentaje decimal).
- Número de observaciones de la clase mayoritaria de la variable objetivo en cada nodo (en porcentaje decimal).

Por ejemplo, en el nodo raíz del árbol siguiente el 100% corresponde al número de pacientes presentes en dicho nodo. El 0.18 hace referencia a los pacientes fallecidos, mientras que el 0.82 a los no fallecidos.

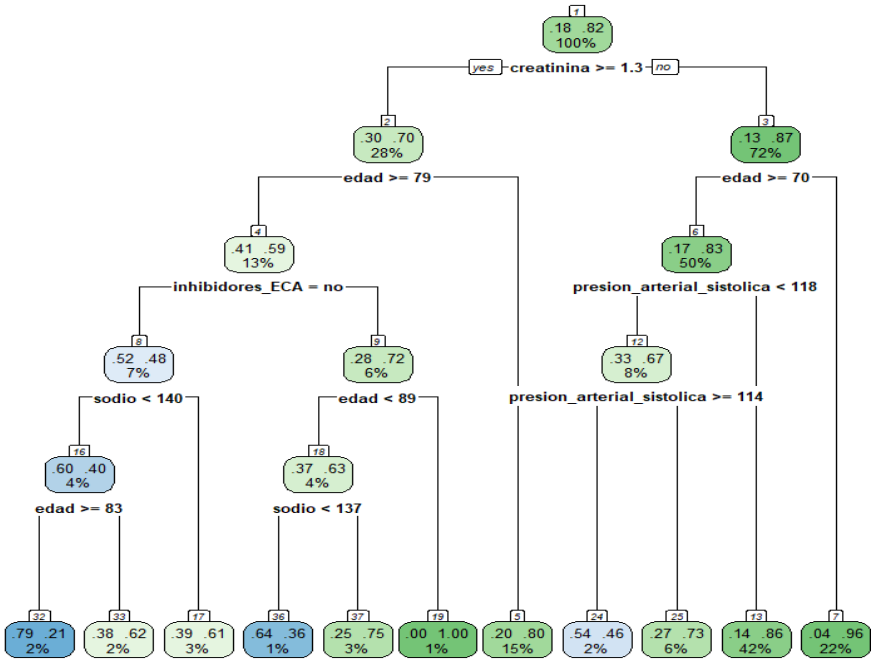

Figura 26. Modelo árbol de decisión.

Se aprecia una clara correspondencia con los parámetros de nuestro modelo de regresión logística: pacientes de mayor edad o con niveles de creatinina superiores a 1.3 mg/dL ven aumentada su probabilidad de fallecimiento, mientras que aquellos a los que se les ha administrado inhibidores de la ECA o cuentan con niveles de sodio superiores a 137 mEq/L disminuyen su riesgo.

6.2. Gráficos

Para finalizar este análisis, se exponen gráficos que permiten visualizar la capacidad predictiva y separabilidad de datos en nuestro modelo (*Figura 27 - Figura 31*).

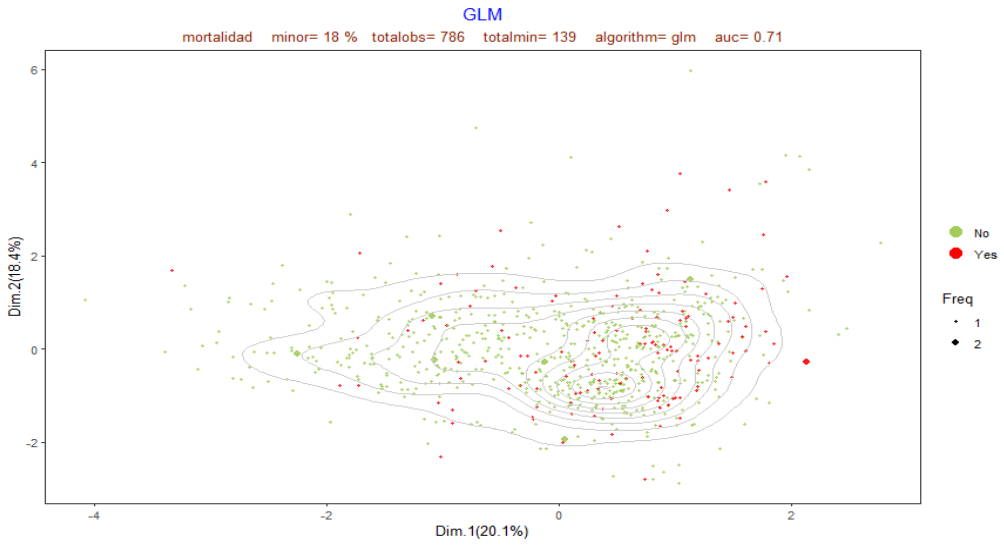

Figura 27. Distribución de los pacientes en el espacio de la FAMD (modelo logist_mmpc).

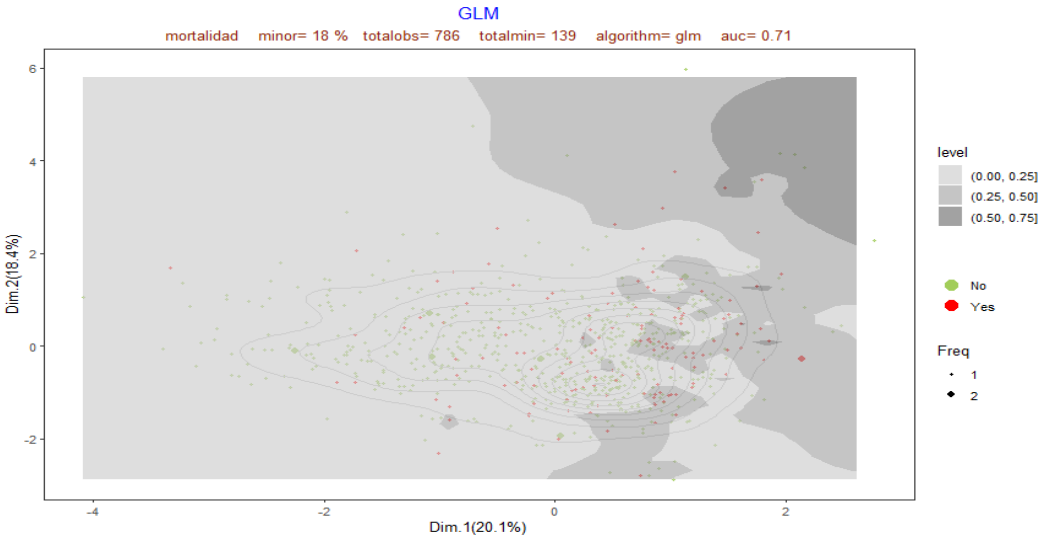

Figura 28. Curvas de contorno en el espacio de la FAMD (modelo logist_mmpc).

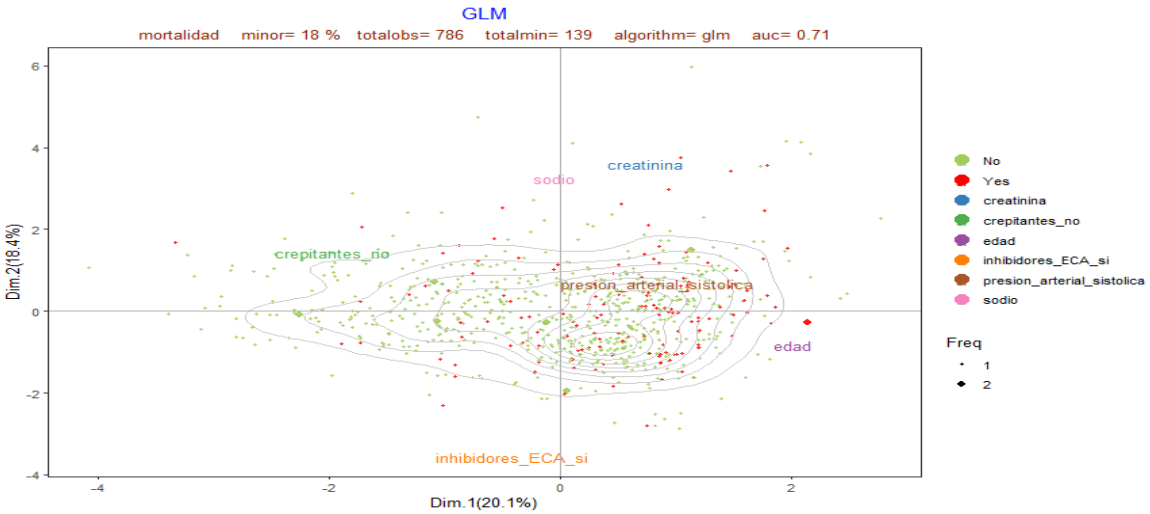

Figura 29. Distribución de las variables predictoras en el espacio de la FAMD (modelo logist_mmpc).

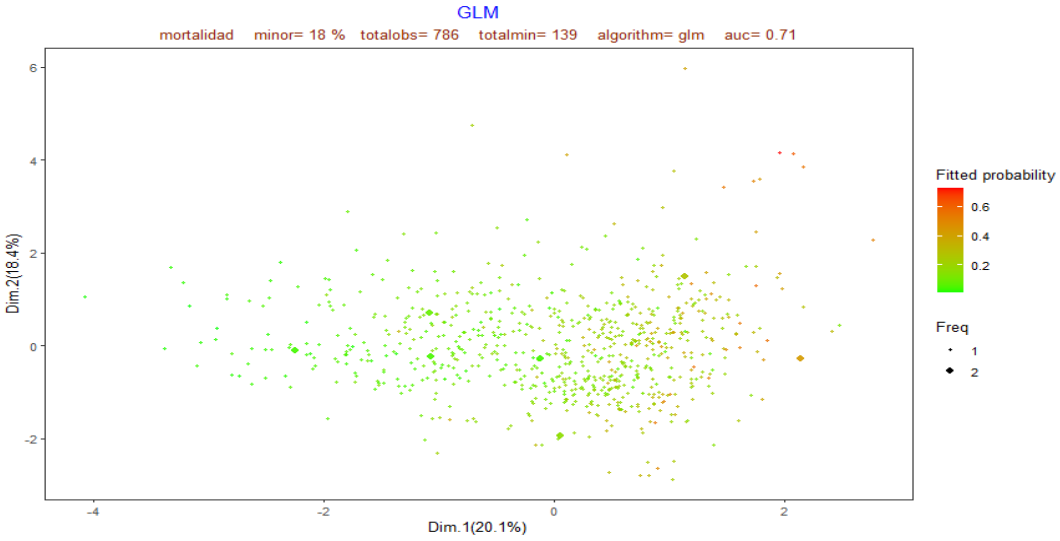

Figura 30. Probabilidad de pertenencia de los pacientes en el espacio FAMD (modelo logist_mmpc).

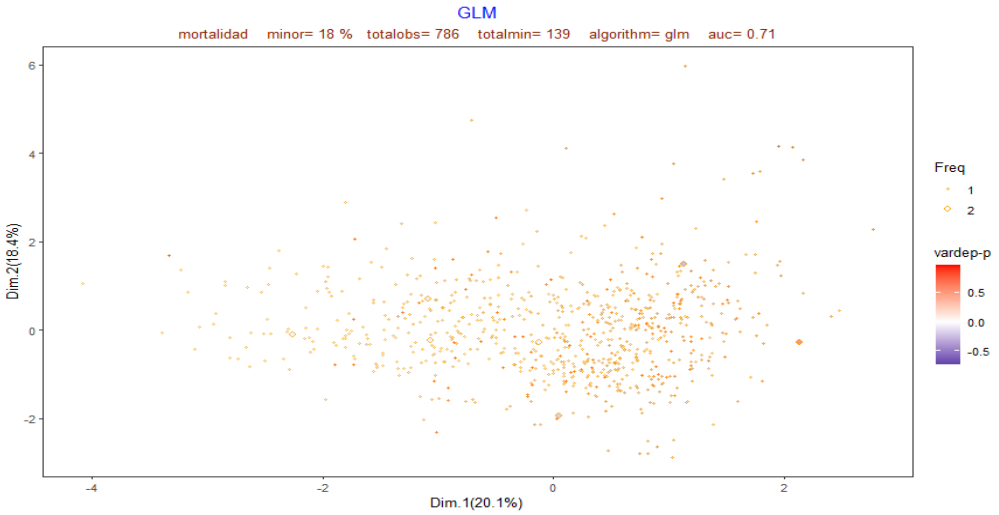

Figura 31. Diferencia entre el valor real y el valor pronosticado en el espacio de la FAMD (modelo logist_mmpc).

En estos gráficos cada paciente es representado mediante un punto en el espacio del Análisis de Correspondencia Múltiple Factorial (FAMD). Más específicamente, en los tres gráficos iniciales, los pacientes que han fallecido se destacan a través de puntos de color rojo, mientras que aquellos que no han fallecido se representan con puntos verdes.

- El primer gráfico ilustra la distribución de los pacientes y presenta contornos de densidad que indican la frecuencia de pacientes en el espacio FAMD.
- En el segundo gráfico, las curvas de contorno agregadas reflejan la capacidad discriminativa del algoritmo predictivo, diseñadas en función de las probabilidades estimadas de nuestro modelo logist_mmpc.
- El tercer gráfico incorpora las variables seleccionadas, cuya proyección en el espacio FAMD ayuda a comprender la influencia de las mismas en la distribución de los pacientes.
- El cuarto gráfico muestra la probabilidad de pertenencia de cada paciente, estimada por el modelo. Tonos rojos indican alta probabilidad de pertenecer a la clase de pacientes fallecidos, mientras que tonos verdes indican baja probabilidad.
- El último gráfico exhibe la discrepancia entre el valor de referencia de la variable objetivo y la probabilidad estimada por nuestro modelo. Las diferentes tonalidades permiten visualizar la magnitud de la discrepancia.

En líneas generales, y a pesar del desbalanceamiento de la variable objetivo y de la limitada segregación de los datos, se observa que ciertas áreas muestran una mayor claridad en la separación de las instancias en nuestra base de datos. Es en estas regiones principalmente donde el modelo demuestra una sólida capacidad predictiva al realizar una discriminación adecuada.

6.3. SAS Entreprise Miner

Con objeto de corroborar los resultados obtenidos, en esta sección se ha llevado a cabo un procedimiento análogo al anterior, utilizando el software *SAS Enterprise Miner*, y donde el diagrama de flujo ilustra la secuencia de todas las fases desarrolladas (*Figura 32*).

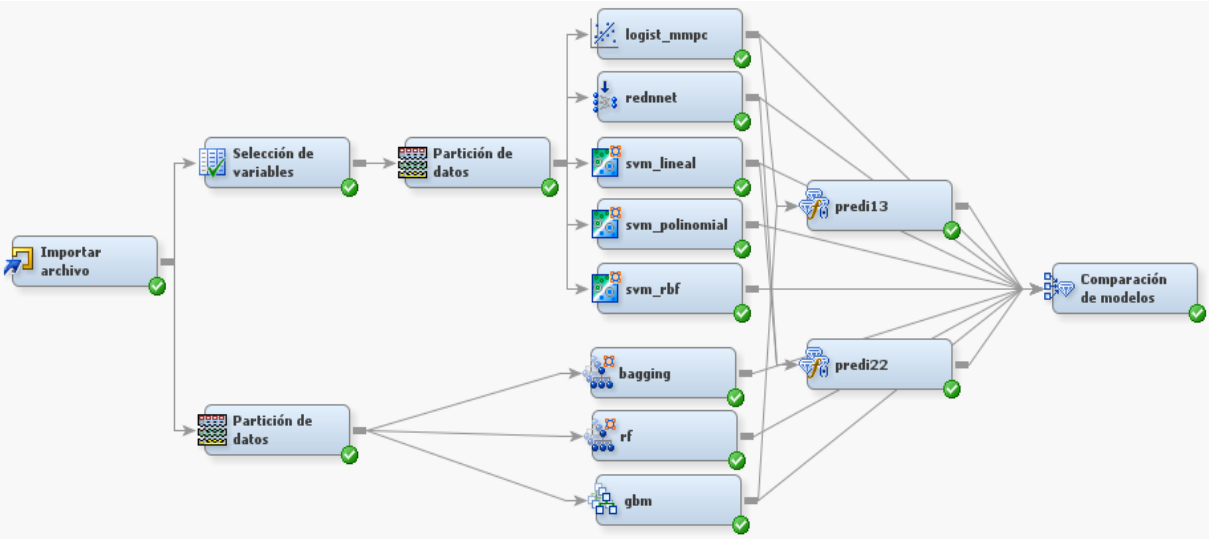

Figura 32. Diagrama de flujo en SAS Miner.

Desde una perspectiva gráfica, la curva ROC supone un estimador visual del rendimiento y capacidad discriminatoria de cada modelo. En consecuencia, la clave consiste en identificar el modelo asociado a la curva ROC óptima en el conjunto de datos *TEST* (*Figura 33*).

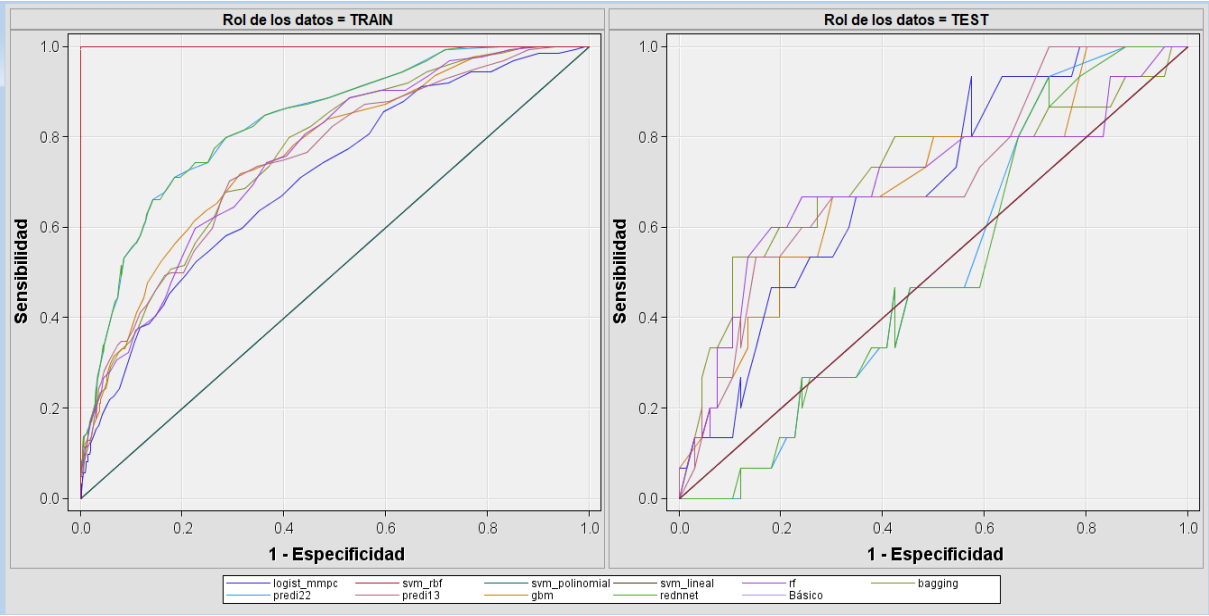

Figura 33. Curva ROC de los modelos en SAS Miner.

Por su parte, los estadísticos de ajuste dan cuenta de los resultados numéricos (*Tabla 8*).

MODELO	TRAIN: ASE	TRAIN: TASA DE CLASIFICACIÓN ERRÓNEA	TRAIN: ÍNDICE ROC	TEST: ASE	TEST: TASA DE CLASIFICACIÓN ERRÓNEA	TEST: ÍNDICE ROC
logist_mmpc	0.132	0.174	0.705	0.140	0.173	0.693
rednnet	0.109	0.156	0.831	0.203	0.27	0.517
bagging	0.128	0.176	0.760	0.137	0.185	0.723
rf	0.129	0.176	0.753	0.135	0.185	0.718
gbm	0.125	0.163	0.762	0.139	0.185	0.689
svm_lineal	0.176	0.176	0.706	0.185	0.185	0.697
svm_polinomial	0.176	0.176	0.710	0.185	0.185	0.393
svm_rbf	0.001	0.001	1.000	0.151	0.185	0.505
predi13	0.127	0.169	0.749	0.138	0.185	0.700
predi22	0.126	0.176	0.831	0.173	0.185	0.517

Tabla 8: Estadísticos de ajuste de los modelos en SAS Miner.

En concordancia con los resultados previos y considerando el índice ROC, la tasa de fallos y tasa de clasificación errónea tanto del conjunto *TRAIN* como del conjunto *TEST*, además de tener presente las implicaciones derivadas de la complejidad de cada modelo, se corrobora que el modelo logist_mmpc es una elección adecuada.

Capítulo 7

Conclusiones y Trabajo Futuro

En la elaboración del presente proyecto, el énfasis principal se ha puesto en la formulación y estructuración de un modelo predictivo destinado a evaluar la mortalidad en pacientes que han experimentado un primer episodio de insuficiencia cardíaca. El resultado obtenido, respaldado por una selección previa de variables, convenientemente fundamentada, sugiere que el algoritmo de regresión logística es la elección óptima para conseguir tal propósito. Señalar, a estos efectos, que la importancia de las variables es semejante en todos los modelos y es probable que la no existencia de interacciones ayude a que el modelo logístico sea el mejor.

Se ha realizado un análisis completo que engloba múltiples componentes médicos, mediante la incorporación de datos clínicos, síntomas específicos y medicamentos administrados, que permite importantes mejoras en la precisión de las predicciones. Se ha pretendido que la posible implementación del modelo mencionado pueda servir como herramienta eficaz en la práctica médica.

Es oportuno destacar que, de las 27 variables predictoras disponibles en la base de datos, se ha optado por seleccionar sólo 6 para la construcción del algoritmo. Entre ellas, resulta evidente que la edad es un factor que influye negativamente en la evolución del paciente. Asimismo, otros factores han mostrado también un impacto desfavorable en la predicción de mortalidad, como los niveles elevados de creatinina, la omisión de inhibidores de la ECA, la presencia de crepitantes o los bajos niveles de sodio.

Centrándonos en un aspecto más específico, en el trabajo se han ensayado ajustes adicionales sobre el propio modelo para intentar optimizar el nivel de precisión del mismo. La elección de un punto de corte alternativo ha arrojado resultados más prometedores, en concordancia con la idea inicial de desarrollar un modelo con la capacidad de detectar correctamente pacientes con un alto riesgo de mortalidad.

En la etapa final del análisis llevado a cabo se ponen de manifiesto las dificultades que más afectan a la efectividad del modelo, debido, en parte, al tamaño limitado de la base de datos de partida. En particular, se evidencia una significativa escasez de observaciones de la clase minoritaria de la variable objetivo.

Se observa también la falta de segregación de los datos, como puede comprobarse en los gráficos comentados anteriormente (*Figura 27 - Figura 31*). Esta problemática repercute en la capacidad del modelo para establecer límites de decisión precisos y claramente definidos, afectando directamente a su rendimiento predictivo.

Por todo ello, sería aconsejable considerar la incorporación de otras variables y la ampliación del número de registros de pacientes, permitiendo así una expansión del marco analítico. Del mismo modo, y para garantizar la robustez del modelo, se recomienda alcanzar un equilibrio sólido entre las dos clases de la variable objetivo, lo que contribuiría a evitar el sesgo inherente a la desproporción de datos.

Por último, y aun teniendo en cuenta los resultados satisfactorios obtenidos, es conveniente señalar que este estudio y la utilización del modelo propuesto, no son incompatibles con la consideración de otros modelos alternativos, que puedan enriquecer la comprensión del fenómeno bajo análisis con perspectivas adicionales que necesariamente quedan fuera del objeto de este trabajo particular.

7.1. Aplicaciones y futuras líneas de investigación

Las implicaciones de este proyecto van más allá de la simple capacidad de prever eventos. Su impacto se extiende al ámbito clínico, pretendiendo ofrecer a los profesionales de la salud una herramienta que ayude a guiarlos en la toma de decisiones.

Algunas de las aplicaciones derivadas de este estudio son las siguientes:

- **Optimización de tratamientos**. El modelo puede implementarse con el propósito de sugerir ajustes en el tratamiento farmacológico, basándose en la información clínica, los síntomas del paciente y la evaluación de la probabilidad de mortalidad. Esto permitiría una atención médica más personalizada y efectiva.
- **Evaluación de intervenciones preventivas**. El modelo contribuye a evaluar la calidad de las intervenciones y estrategias preventivas al anticipar cómo afectarían a la probabilidad de mortalidad de los pacientes.
- **Gestión de recursos**. Hospitales y sistemas de atención médica pueden utilizar el modelo para prever la gestión de sus recursos, en función de las estimaciones de mortalidad.
- **Educación médica**. El modelo puede integrarse en plataformas de educación sanitaria para proporcionar a los expertos médicos información actualizada acerca de las tendencias en mortalidad y los tratamientos terapéuticos más exitosos.
- **Estudios clínicos**. Al identificar a los pacientes con mayor probabilidad de mortalidad, el modelo se puede emplear para seleccionar candidatos adecuados para ensayos clínicos y estudios de investigación, mejorando su eficacia y relevancia.
- **Investigación científica**. Los resultados y patrones identificados por el modelo pueden inspirar investigaciones más profundas en el campo de la cardiología y la atención médica en general.

De cara al futuro, un proyecto de esta naturaleza genera posibles líneas para la investigación. Algunas de ellas podrían ser:

- **Recopilación de nuevos datos**: Considerando, como ya se ha comentado, la incorporación de otros datos que permitan mejorar la capacidad predictiva y comprender mejor los factores de riesgo.
- **Análisis de segmentación**: Explorando si una sofisticación del modelo podría incluso identificar subgrupos específicos de pacientes con diferentes niveles de riesgo de mortalidad, lo que podría llevar a un tratamiento aún más personalizado.
- **Introducción de variables temporales**: Agregando variables de seguimiento temporal que reflejen la evolución de los pacientes a lo largo del tiempo y permita establecer un marco temporal más concreto.
- **Utilización de técnicas avanzadas**: Como el modelo de competencia de riesgos de Cox o series temporales.
- **Generalización a diferentes poblaciones**: Probando el modelo en poblaciones más diversas y con diferentes perfiles médicos para asegurarse de su aplicabilidad en diversos contextos.
- **Ampliación del tiempo de estudio en pacientes con menor riesgo**: Determinando con mayor precisión la evolución de estos pacientes y la manifestación tardía de factores influyentes, para así anticipar resultados a largo plazo.

En conclusión, este proyecto trasciende los aspectos técnicos para adoptar una atención médica más eficaz y centrada en el paciente. Su desarrollo pretende ser una contribución concreta al avance de la medicina y una muestra de cómo la innovación tecnológica y el bienestar humano pueden ir de la mano.

Bibliografía

Aguirre, C. (2019, May 26). *ML Part 1: Introducción a los arboles de decisión*. Cristobal-Aguirre.Com. https://www.cristobal-aguirre.com/arboles-de-decision

Banerjee, A., Chen, S., Fatemifar, G., Zeina, M., Lumbers, R. T., Mielke, J., Gill, S., Kotecha, D., Freitag, D. F., Denaxas, S., & Hemingway, H. (2021). Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. *BMC Medicine*, *19*(1). https://doi.org/10.1186/s12916-021-01940-7

Barge-Caballero, E., Barge-Caballero, G., Paniagua-Martín, M. J., Couto-Mallón, D., Pardo-Martínez, P., Sagastagoitia-Fornie, M., Barrios, V., Escobar, C., Cosín-Sales, J., Muñiz, J., Vázquez-Rodríguez, J. M., & Crespo-Leiro, M. G. (2022). Valor pronóstico de un nuevo modelo de evaluación clínica de pacientes ambulatorios con insuficiencia cardíaca. *REC: CardioClinics*, *57*(2), 76–84. https://doi.org/10.1016/j.rccl.2021.06.004

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees* (1st ed.). Routledge. https://www.perlego.com/book/1579805/classification-and-regression-trees-pdf

Cánovas-García, F., Alonso-Sarría, F., Gomariz-Castillo, F., & Oñate-Valdivieso, F. (2017). Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery. *Computers and Geosciences*, *103*, 1–11. https://doi.org/10.1016/j.cageo.2017.02.012

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

De la Cámara, A. G., Guerra Vales, J. M., Tapia, P. M., Esteban, E. A., del Pozo, S. V. F., Sandubete, E. C., Ortega, F. J. M., Puerto, A. N., & Marín-León, I. (2012). Role of biological and non biological factors in congestive heart failure mortality: PREDICE-SCORE: A clinical prediction rule. *Cardiology Journal*, *19*(6), 578–585. https://doi.org/10.5603/CJ.2012.0108

Espinosa Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología*, *21*(3), 1–16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Google. (2018, April 16). *Wieviele Trainungsbeispiele benötigen Lernverfahren? (1/2)*. Https://Data-Science-Blog.Com/. https://data-science-blog.com/de/blog/tag/decision-tree/

Google. (2020, May 4). *How Does Support Vector Machine (SVM) Algorithm Works In Machine Learning?* Https://Www.Analyticssteps.Com.

Google. (2021, May 27). *What is a Neural Network?* Https://Www.Tibco.Com. https://www.tibco.com/reference-center/what-is-a-neural-network

Google. (2022a, January 19). *Le Bagging en Machine learning, de quoi s'agit-il ?* Https://Kobia.Fr. https://kobia.fr/le-bagging-en-machine-learning-de-quoi-sagit-il/

Google. (2022b, January 21). *Inteligencia Artificial aplicada a la medicina. En busca de un envejecimiento saludable.* Https://Cenie.Eu. https://cenie.eu/es/blogs/tecnologia-y-longevidad/inteligencia-artificial-aplicada-la-medicina-en-busca-de-un

Google. (2022c, June 2). *XGBoost-DNN Mixed Model for Predicting Driver's Estimation on the Relative Motion States during Lane-Changing Decisions: A Real Driving Study on the Highway*. Https://Www.Mdpi.Com. https://www.mdpi.com/2071-1050/14/11/6829

Google. (2022d, June 27). *What is Text Classification*. Https://Www.Exxactcorp.Com. https://www.exxactcorp.com/blog/Deep-Learning/What-is-Text-Classification

Google. (2022e, November 5). *Top 10 Interview Questions on Gradient Boosting Algorithms*. Https://Www.Analyticsvidhya.Com/. https://www.analyticsvidhya.com/blog/2022/11/top-10-interview-questions-on-gradient-boosting/

Jorge Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Universidad Tecnológica Nacional | Facultad Regional Rosario.

Kuhn, M., & Johnson, K. (2013). Regression Trees and Rule-Based Models. In *Applied Predictive Modeling* (1st ed., pp. 173–220). Springer.

Lozano Alonso, S., Sisamón Marco, I., García Andrés, I., Ponce Lázaro, M. J., Delgado Guerrero, B., & Muñoz Solera, C. (2021). Clasificación de la insuficiencia cardíaca. *Revista Sanitaria de Información*. https://revistasanitariadeinvestigacion.com/clasificacion-de-la-insuficiencia-cardíaca/

Lu, J., Wang, L., Bennamoun, M., Ward, I., An, S., Sohel, F., Chow, B. J. W., Dwivedi, G., & Sanfilippo, F. M. (2021). Machine learning risk prediction model for acute coronary syndrome and death from use of non-steroidal anti-inflammatory drugs in administrative data. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-97643-3

Luo, C., Zhu, Y., Zhu, Z., Li, R., Chen, G., & Wang, Z. (2022). A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. *Journal of Translational Medicine*, *20*(1). https://doi.org/10.1186/s12967-022-03340-8

Martínez Cisternas, I. A. (2018). *Clasificación de disrupciones nucleares con métodos ensamblados en dispositivo Tokamak*. Pontificia Universidad Católica de Valparaíso.

Moral Peláez, I. (2006). Modelos de regresión: lineal simple y regresión logística. In *Métodos estadísticos para enfermería nefrológica* (pp. 195–214). SEDEN.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*, 1565–1567. https://doi.org/https://doi.org/10.1038/nbt1206-1565

Portela García-Miguel, J. (2020). Técnicas de Machine Learning. *Facultad de Estudios Estadísticos, Universidad Complutense de Madrid*.

Recarey Fernández, R. (2021). *Métodos de ensamblado en Machine Learning*. Universidade de Santiago de Compostela.

Rodríguez Artalejo, M., González Calero, P. A., & Gómez Martín, M. A. (2011). Árboles. In *Estructuras de datos. Un enfoque moderno*. Editorial Complutense.

Rojas, M. G., Carballido, J. A., Olivera, A. C., & Vidal, P. J. (2020). Optimización de Support Vector Machine mediante metaheurísticas para clasificación de retinopatia diabética. *XXI Simposio Argentino de Inteligencia Artificial (ASAI 2020) - JAIIO 49 (Modalidad Virtual)*, 73–86.

Ruiz-Ruiz, F., Menéndez-Orenga, M., Medrano, F. J., Calderón, E. J., Lora-Pablos, D., Navarro-Puerto, M. A., Rodríguez-Torres, P., & de la Cámara, A. G. (2019). The prognosis of patients hospitalized with a first episode of heart failure, validation of two scores: PREDICE and AHEAD. *Clinical Epidemiology*, *11*, 615–624. https://doi.org/10.2147/CLEP.S206017

Sicras-Mainar, A., Sicras-Navarro, A., Palacios, B., Varela, L., & Delgado, J. F. (2022). Epidemiology and treatment of heart failure in Spain: the HF-PATHWAYS study. *Revista Española de Cardiología (English Edition)*, *75*(1), 31–38. https://doi.org/10.1016/j.rec.2020.09.033

Zhao, X., Sui, Y., Ruan, X., Wang, X., He, K., Dong, W., Qu, H., & Fang, X. (2022). A deep learning model for early risk prediction of heart failure with preserved ejection fraction by DNA methylation profiles combined with clinical features. *Clinical Epigenetics*, *14*(1). https://doi.org/10.1186/s13148-022-01232-8

Anexo

A.1. Anexo de tablas

MORTALIDAD	N	PORC
No	647	82.31552
Yes	139	17.68448

Tabla A.1. Balance de la variable objetivo.

MÉTODO	VARIABLES SELECCIONADAS
SBF	edad, presion_arterial_sistolica, presion_arterial_diastolica, creatinina, sodio, potasio, hemoglobina, crepitantes_no, edemas_no, inhibidores_ECA_si, betabloqueadores_si
RFE	edad, creatinina, presion_arterial_sistolica, presion_arterial_diastolica, sodio, betabloqueadores_si, hemoglobina, anticoagulantes_orales_si
AIC	edad, creatinina, presion_arterial_sistolica, inhibidores_ECA_si, sodio, crepitantes_no, ingurgitacion_yugular_si, hepatomegalia_si, edemas_no, betabloqueadores_si, valvulopatias_no, digoxina_si
BIC	edad, creatinina, presion_arterial_sistolica
Boruta	edad, presion_arterial_sistolica, creatinina, sodio
MMPC	edad, presion_arterial_sistolica, creatinina, sodio, crepitantes_no, inhibidores_ECA_si
SES	edad, presion_arterial_sistolica, creatinina, sodio, crepitantes_no, inhibidores_ECA_si
AIC repetido	edad, creatinina, inhibidores_ECA_si, ingurgitacion_yugular_si, presion_arterial_sistolica, crepitantes_no, estatinas_si, edemas_no, hepatomegalia_si, sodio
BIC repetido 1	edad, creatinina
BIC repetido 2	edad, creatinina, presion_arterial_sistolica

Tabla A.2. Variables seleccionadas por cada método de selección de variables. SES y BIC repetido 2 no modelizarán algoritmos por ofrecer conjuntos de variables ya seleccionados.

MTRY	ACCCURACY	KAPPA
2	0.8170474	0.1053831
3	0.8122108	0.1083567
4	0.8076345	0.1040433
6	0.8066231	0.1082336

Tabla A.3. Tuneado del parámetro mtry en Random Forest.

B.1. Anexo de figuras

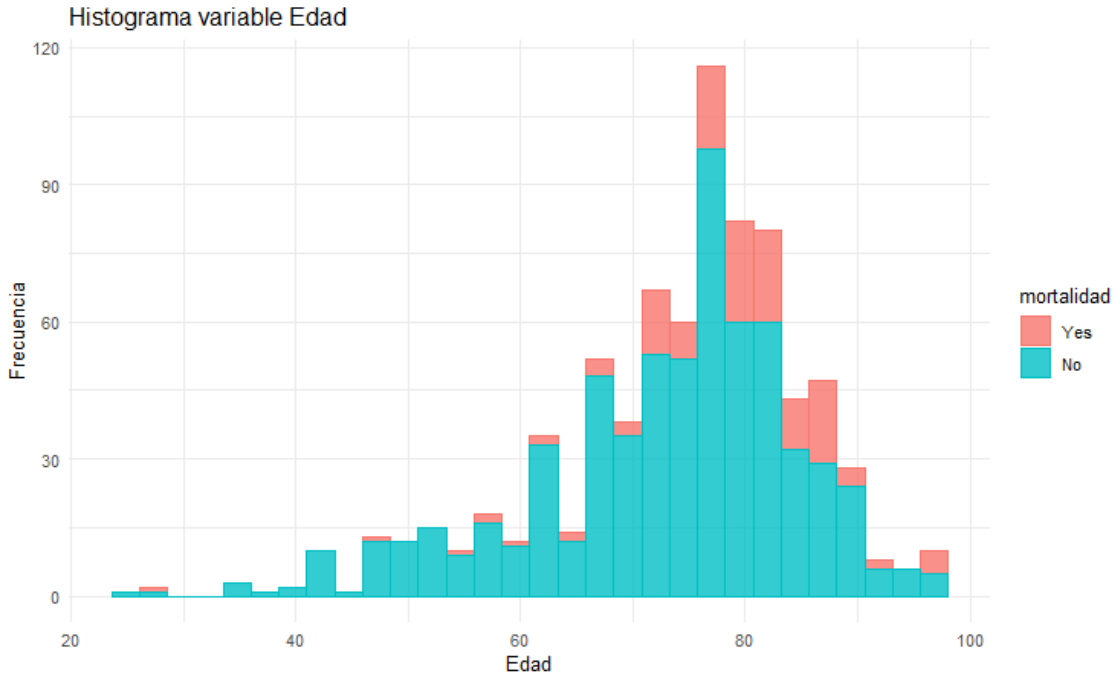

Figura B.1. Histograma de la variable edad.

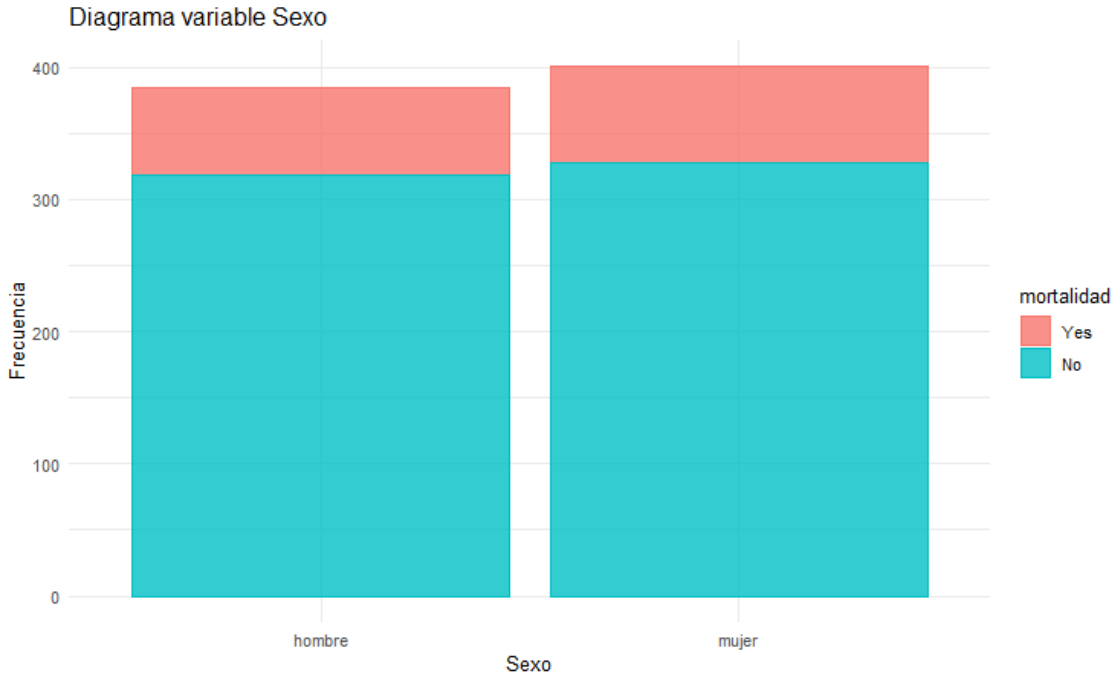

Figura B.2. Diagrama de barras de la variable sexo.

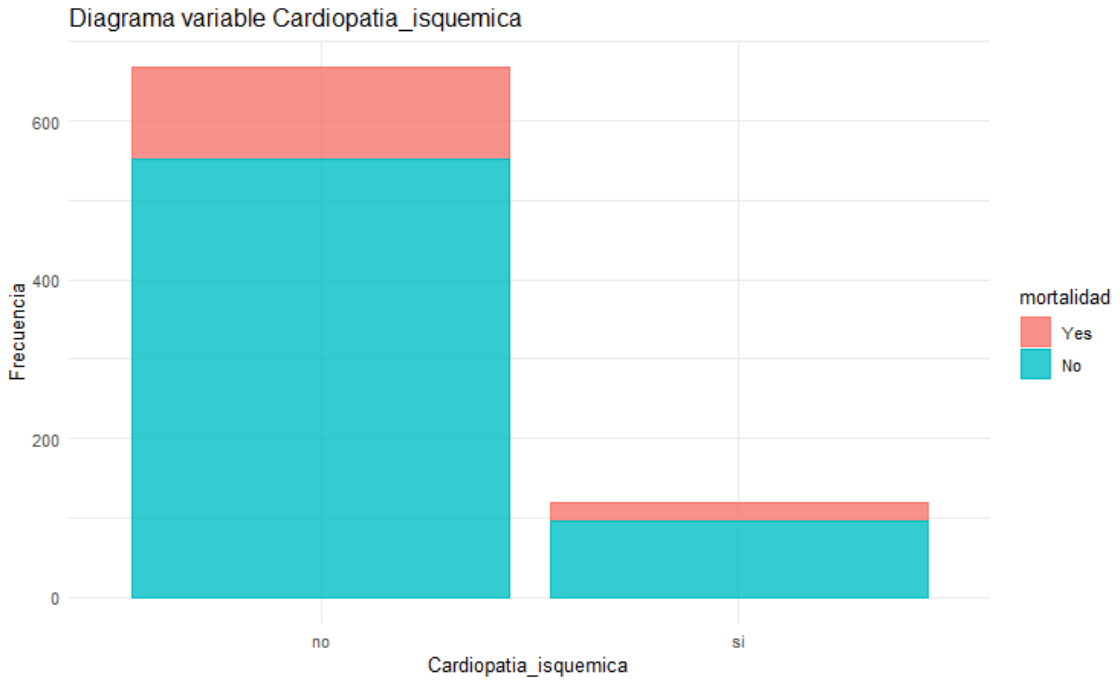

Figura B.3. Diagrama de barras de la variable cardiopatía isquémica.

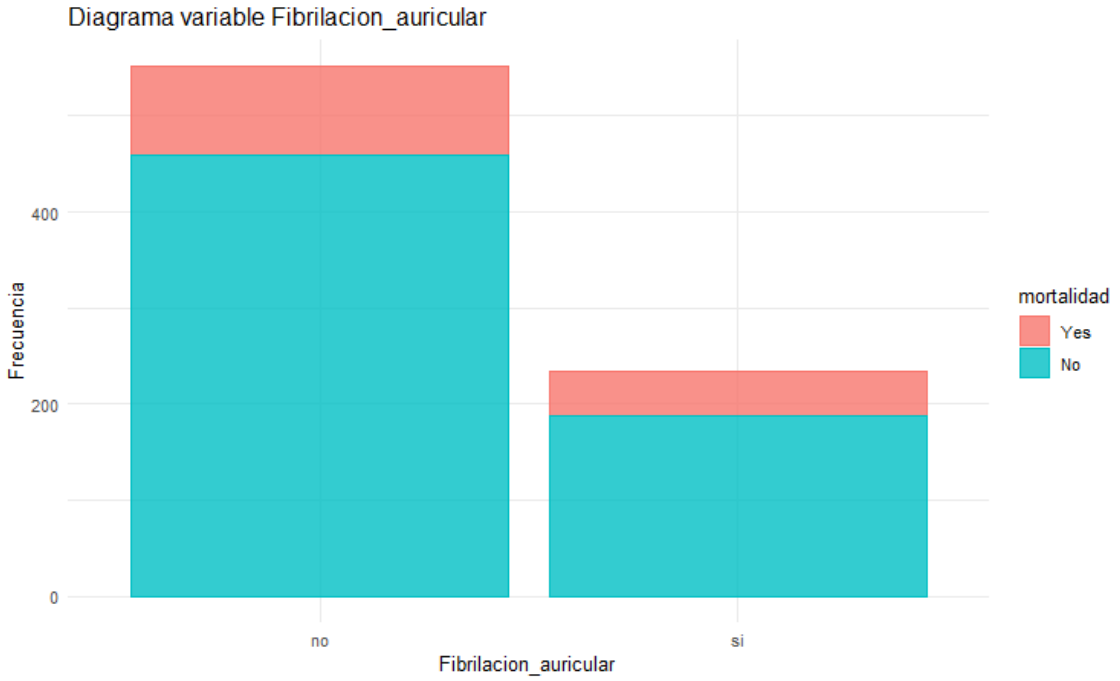

Figura B.4. Diagrama de barras de la variable fibrilación auricular.

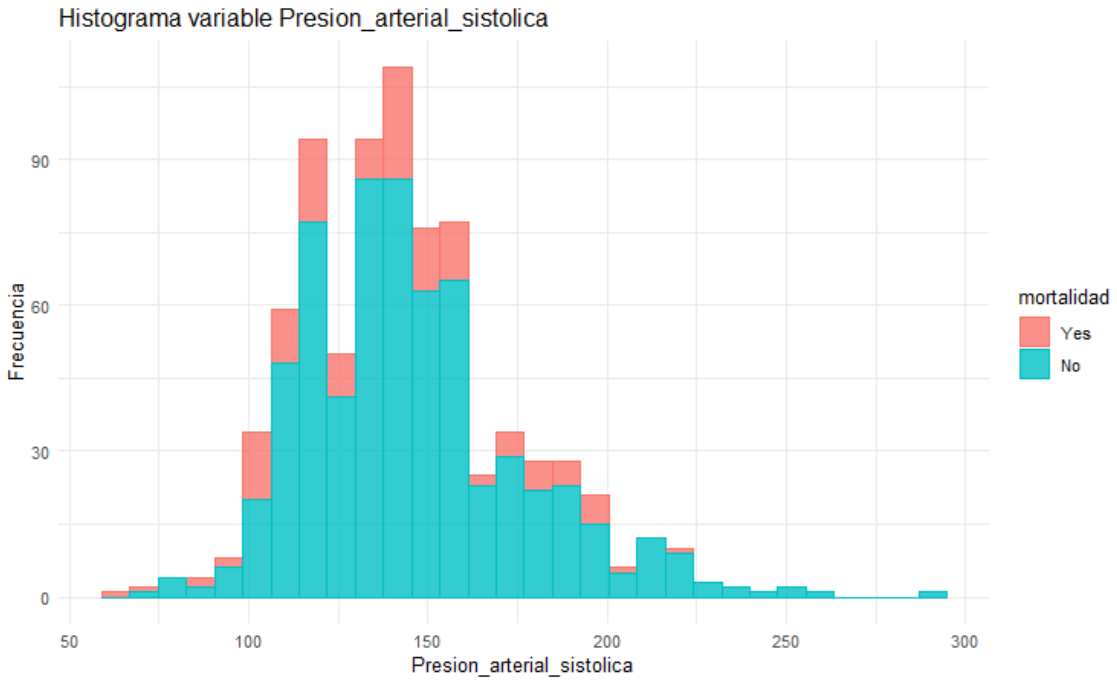

Figura B.5. Histograma de la variable presión arterial sistólica.

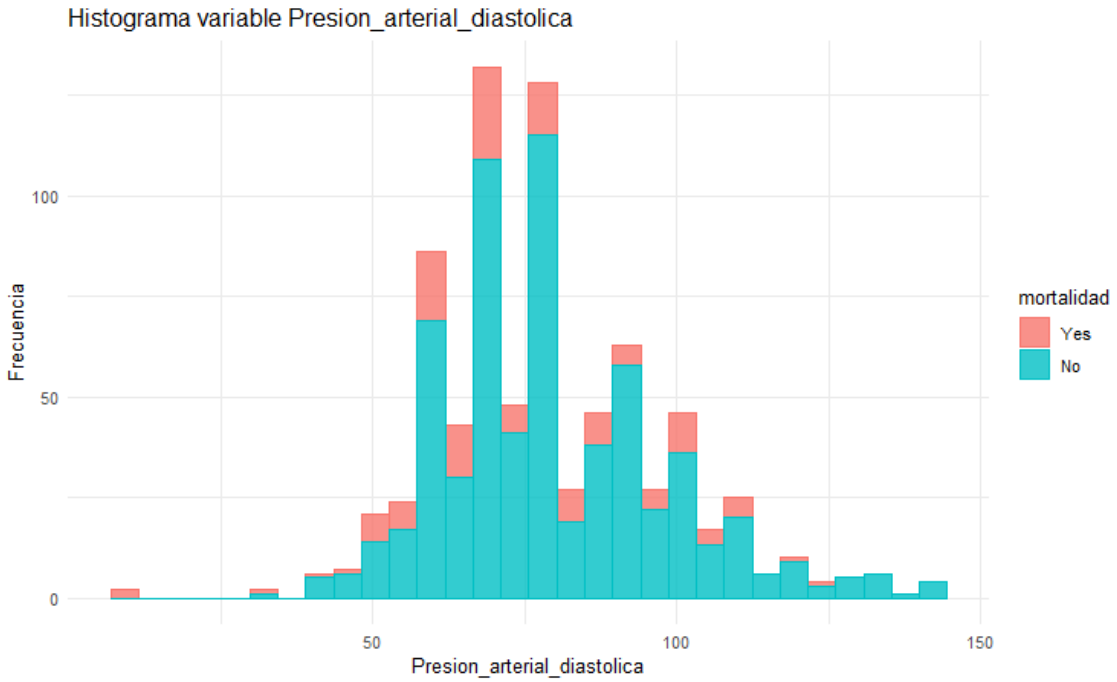

Figura B.6. Histograma de la variable presión arterial diastólica.

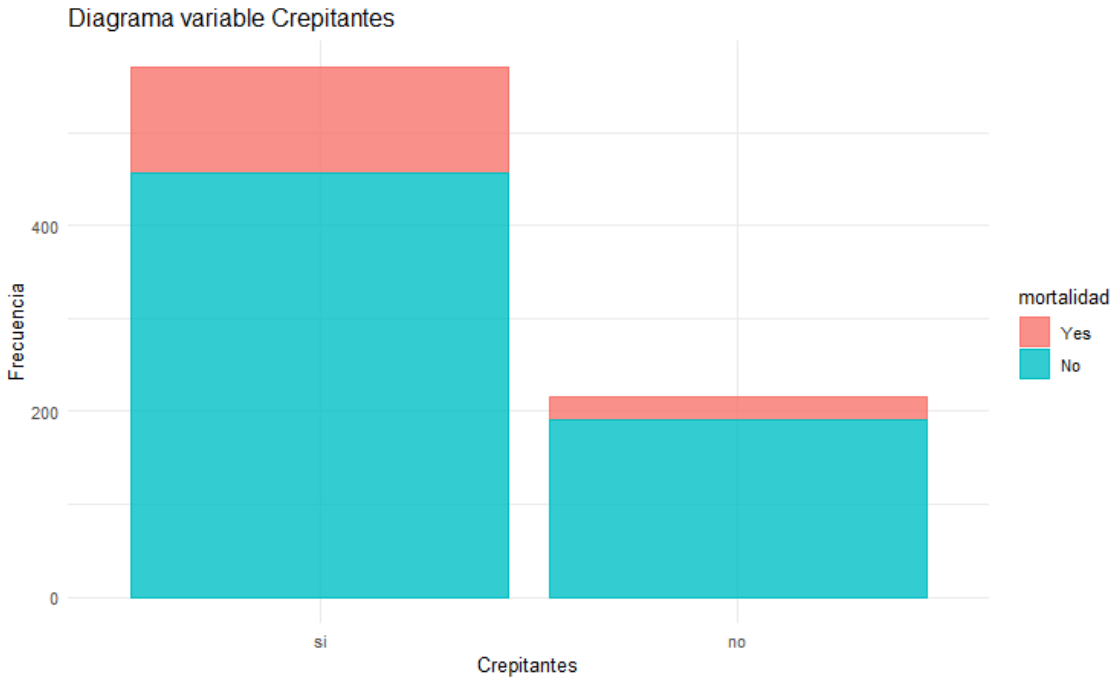

Figura B.7. Diagrama de barras de la variable crepitantes.

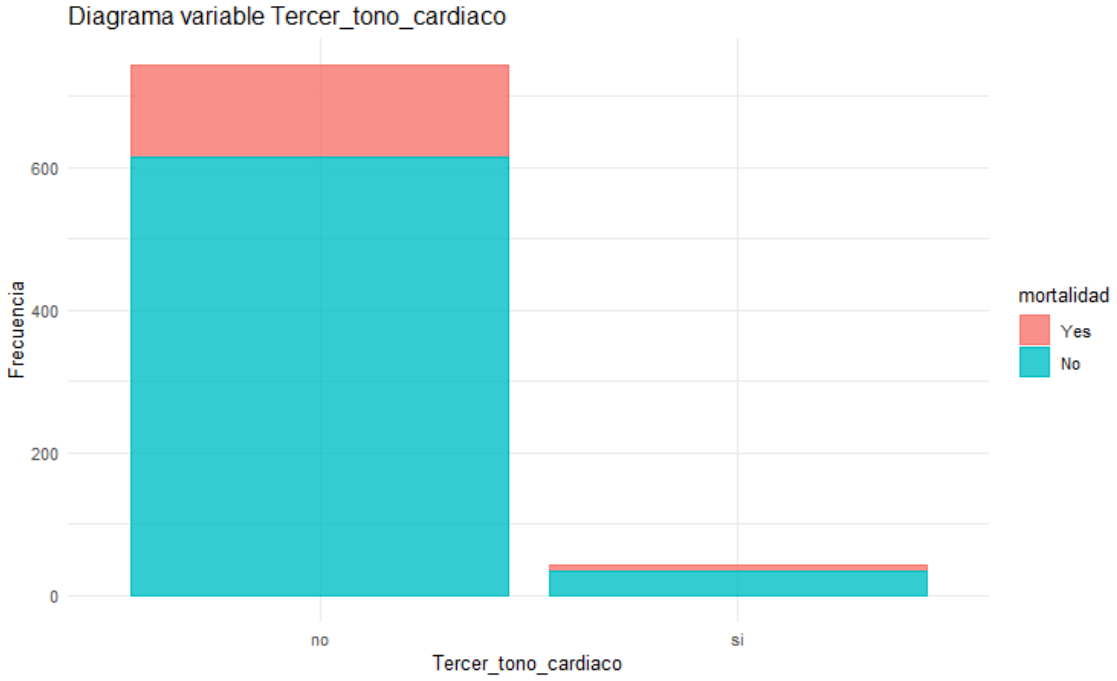

Figura B.8. Diagrama de barras de la variable tercer tono cardíaco.

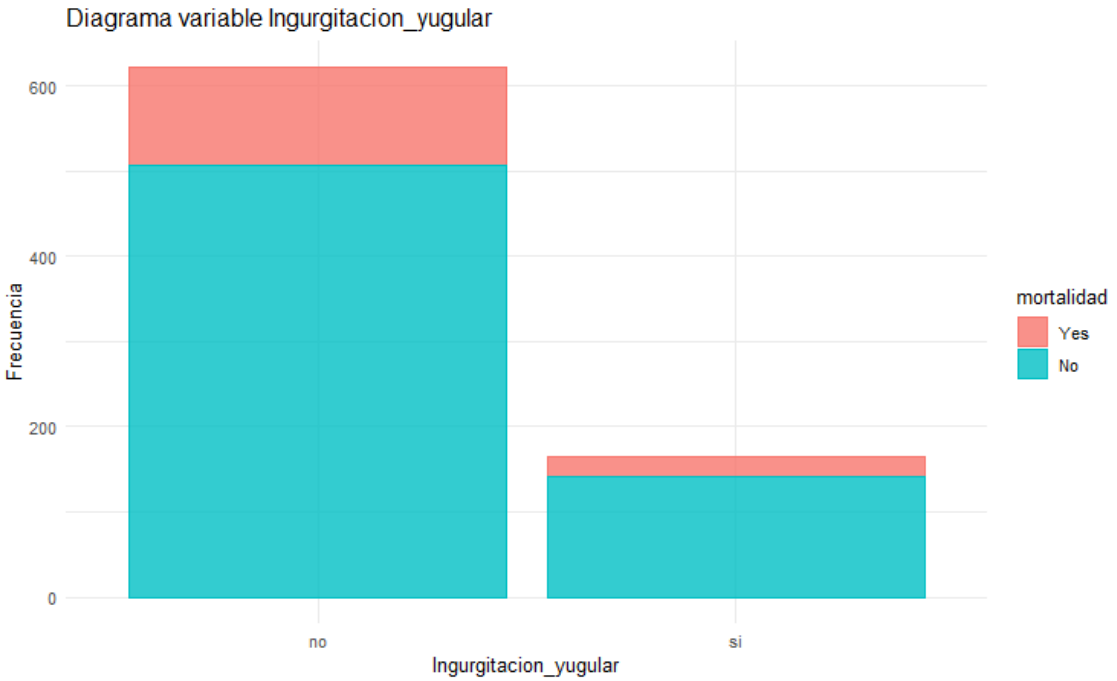

Figura B.9. Diagrama de barras de la variable ingurgitación yugular.

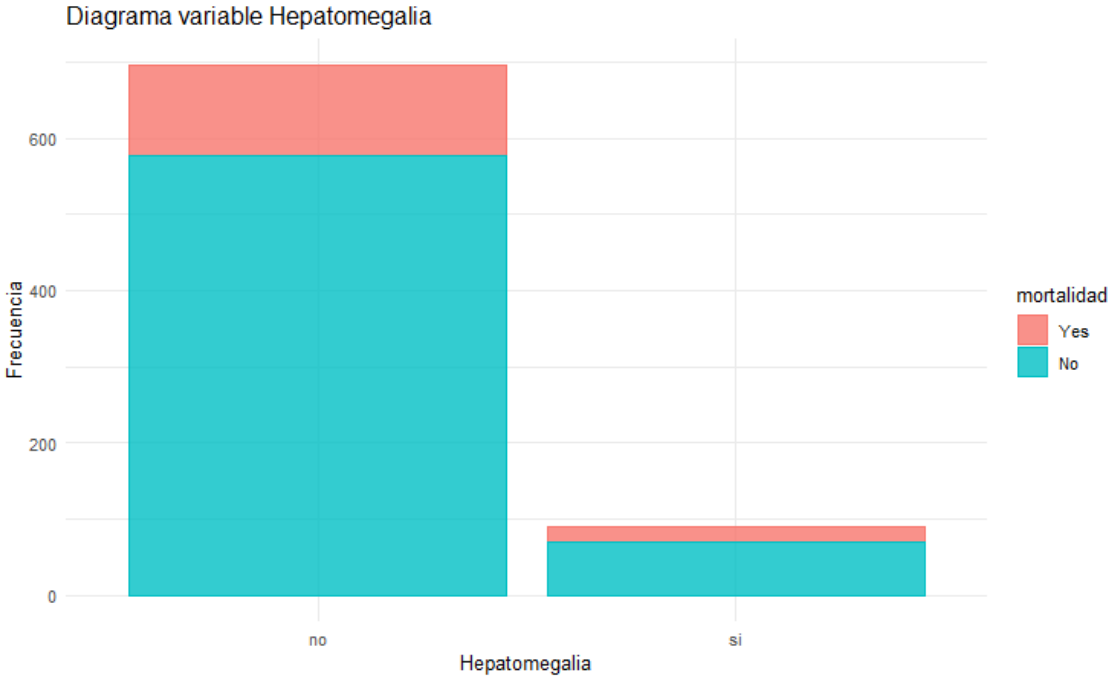

Figura B.10. Diagrama de barras de la variable hepatomegalia.

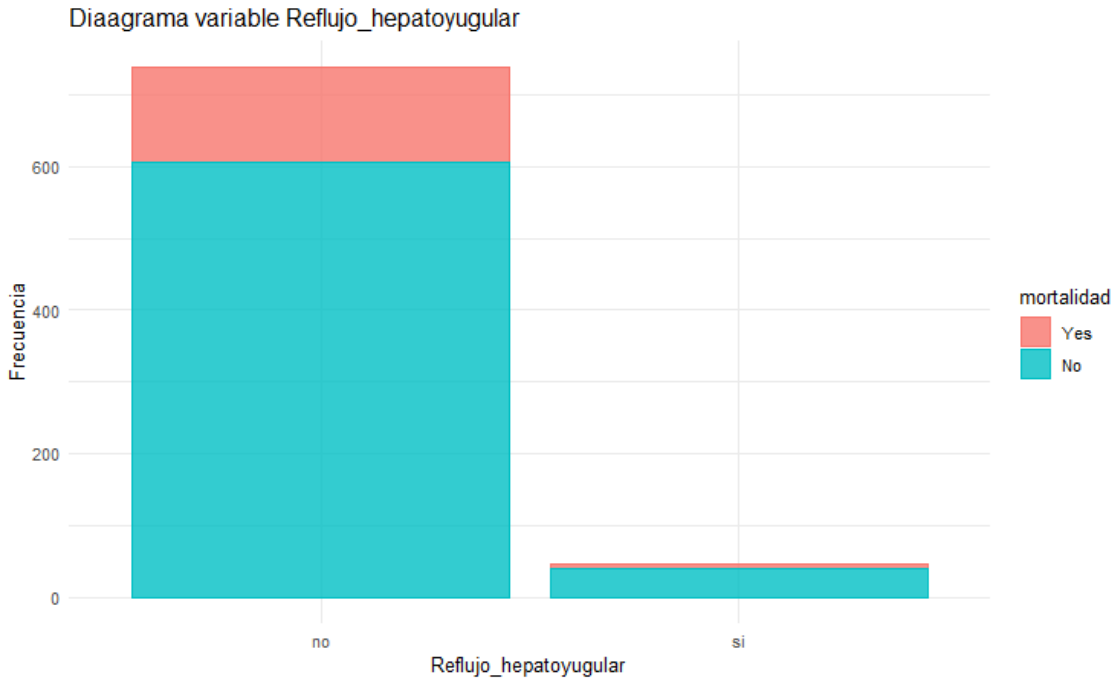

Figura B.11. Diagrama de barras de la variable reflujo hepatoyugular.

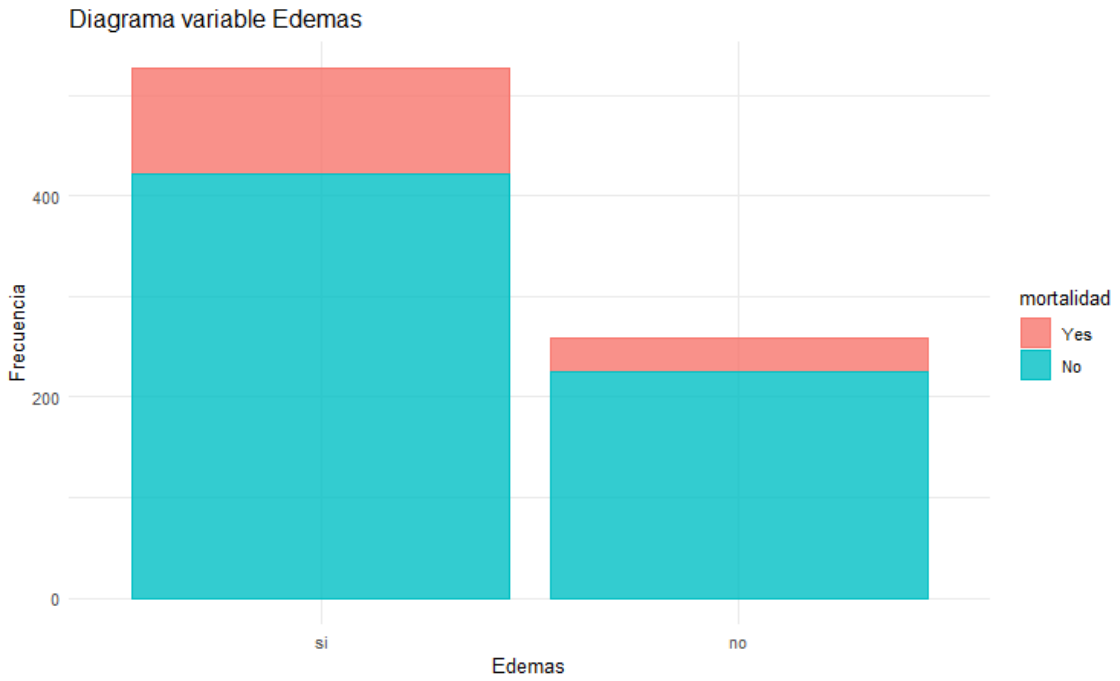

Figura B.12. Diagrama de barras de la variable edemas.

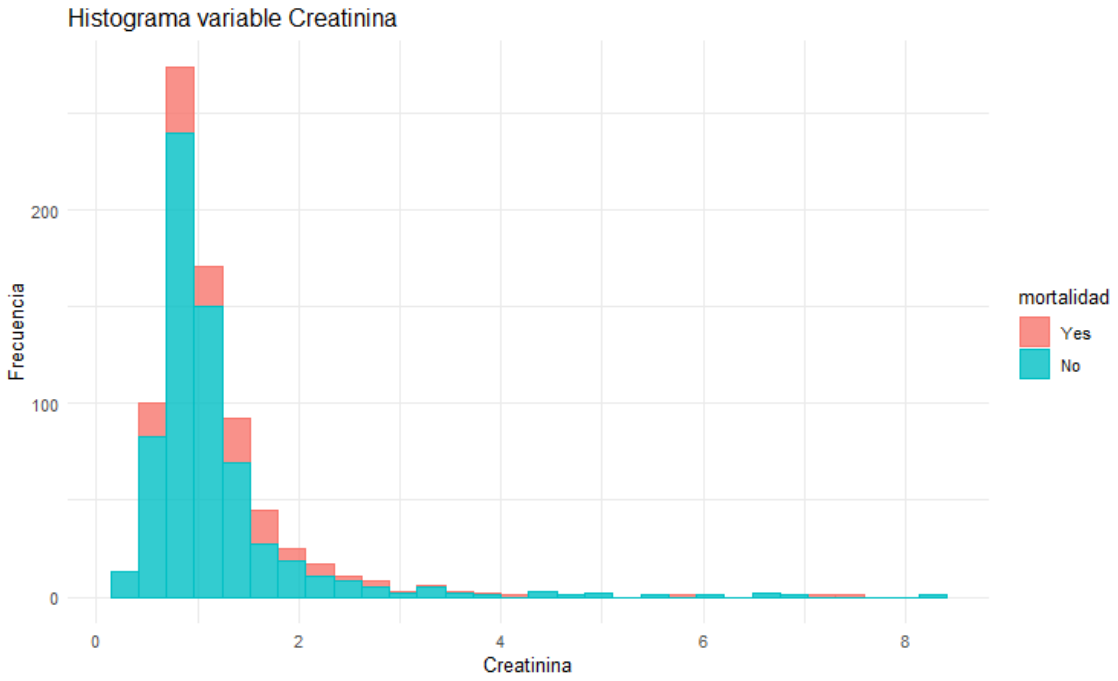

Figura B.13. Histograma de la variable creatinina.

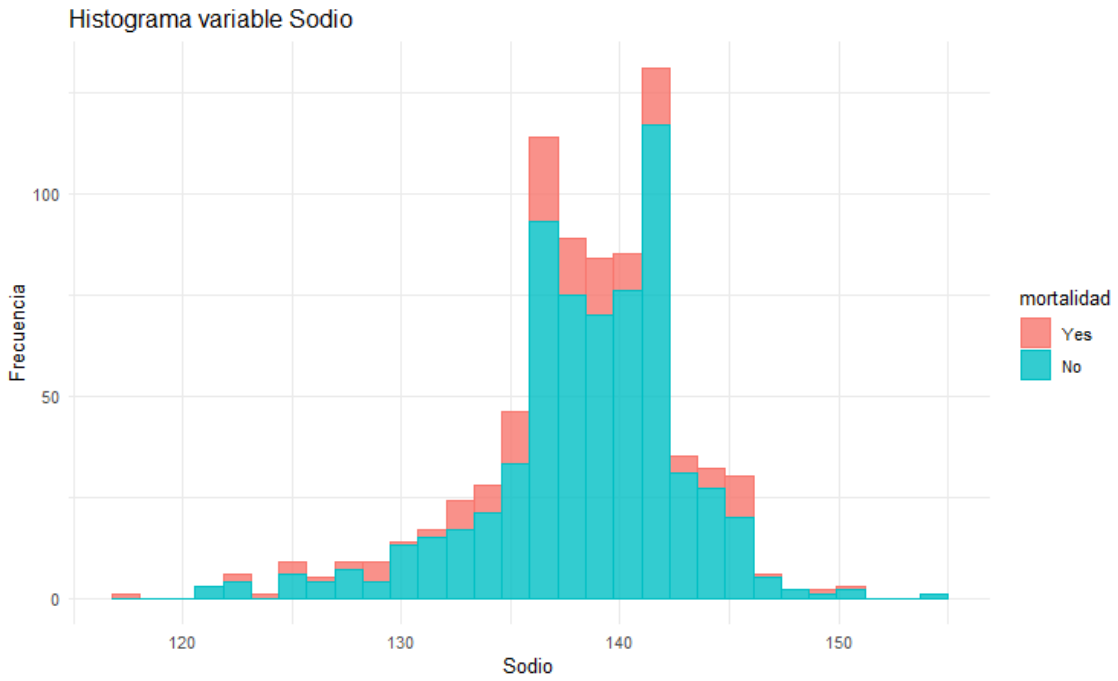

Figura B.14. Histograma de la variable sodio.

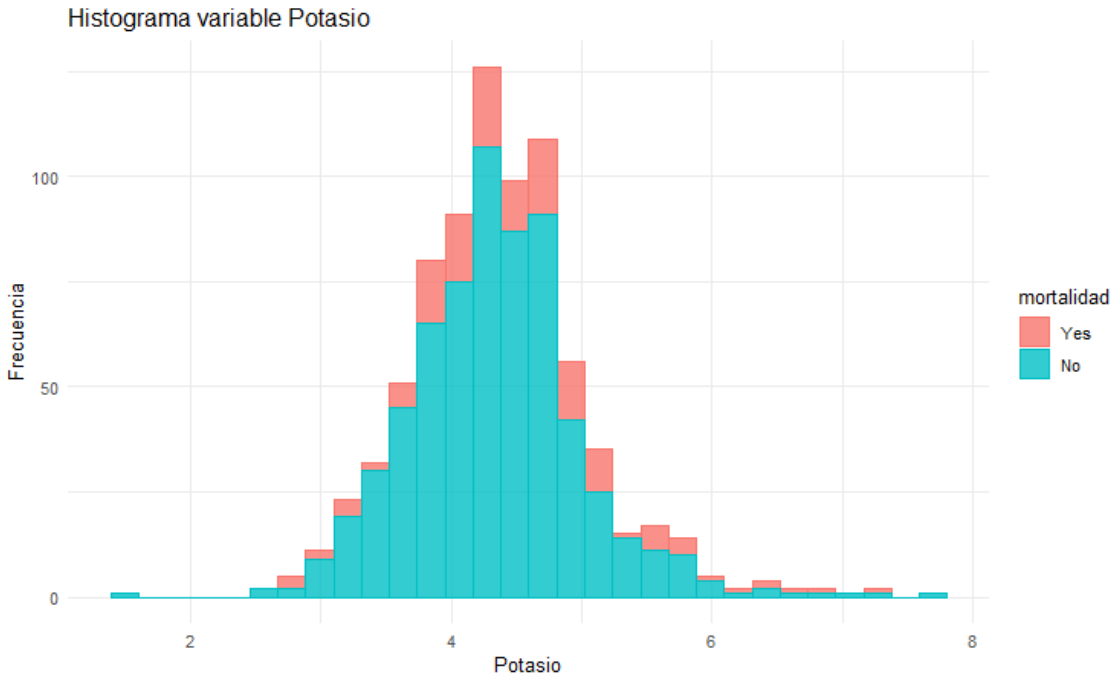

Figura B.15. Histograma de la variable potasio.

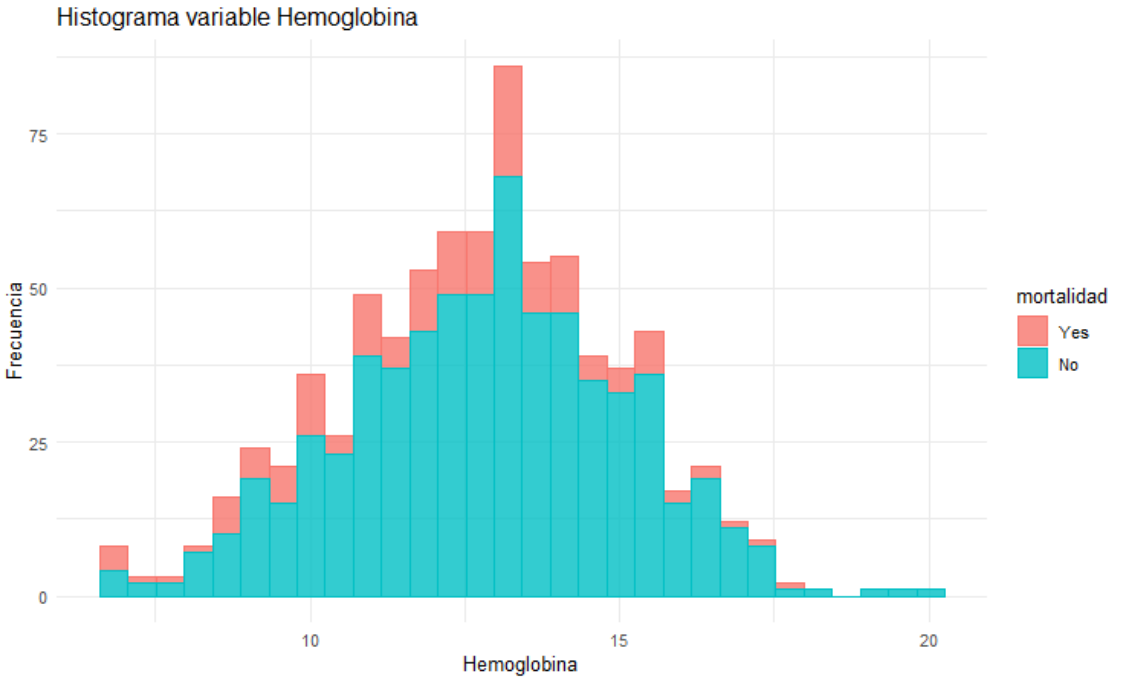

Figura B.16. Histograma de la variable hemoglobina.

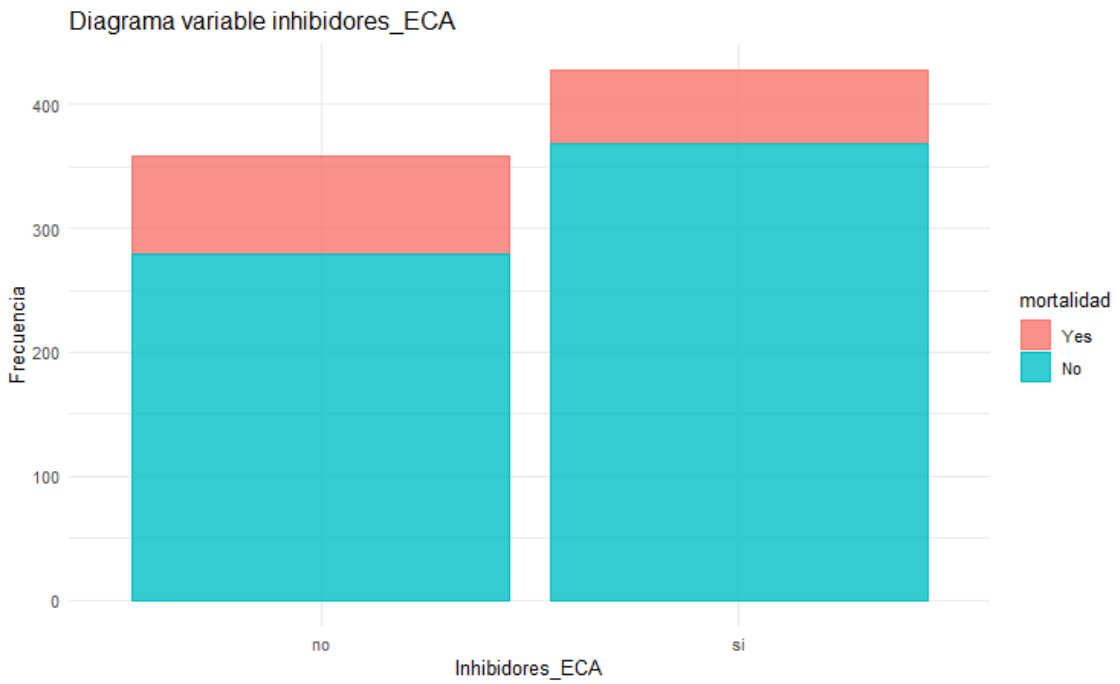

Figura B.17. Diagrama de barras de la variable inhibidores de la ECA.

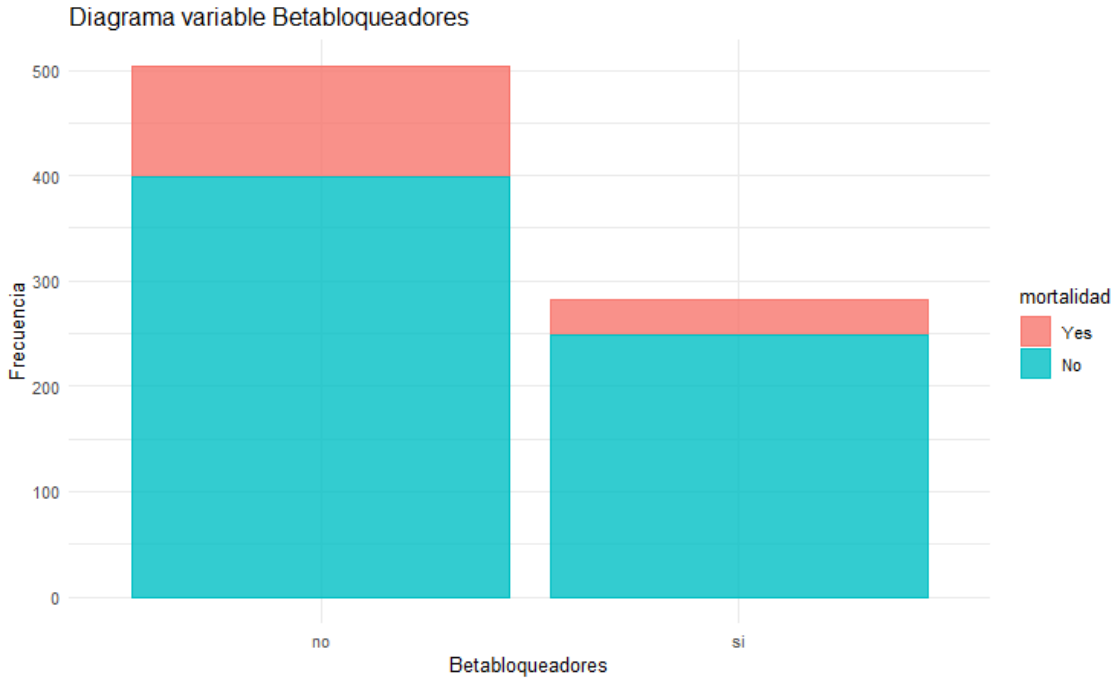

Figura B.18. Diagrama de barras de la variable betabloqueadores.

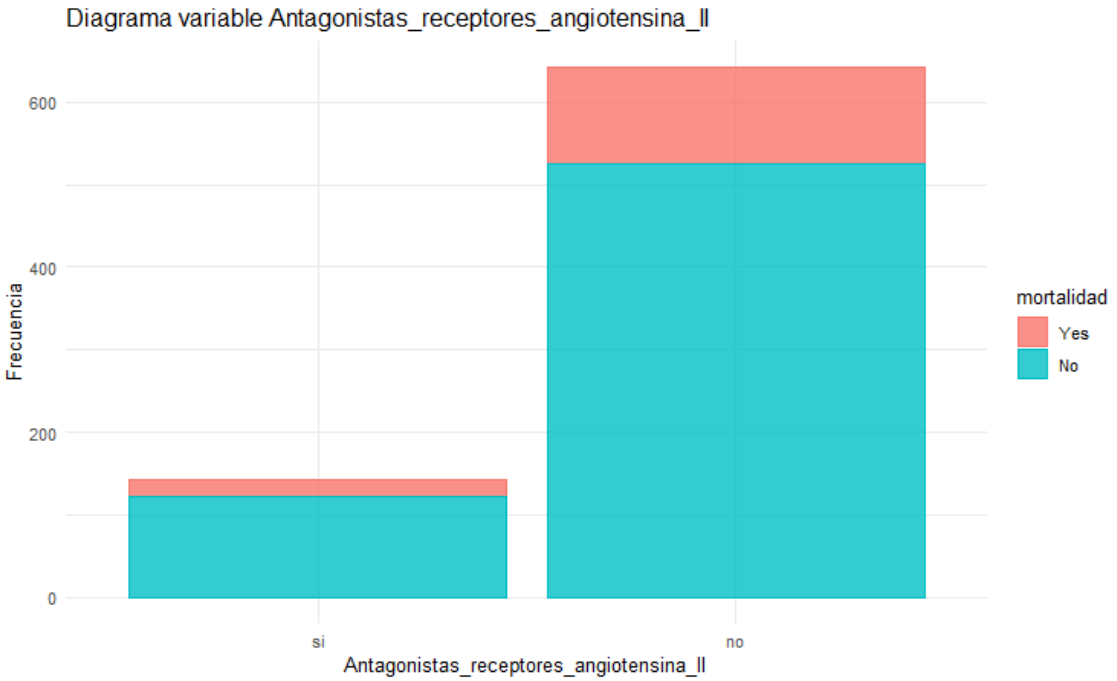

Figura B.19. Diagrama de barras de la variable antagonistas de los receptores de la angiotensina II.

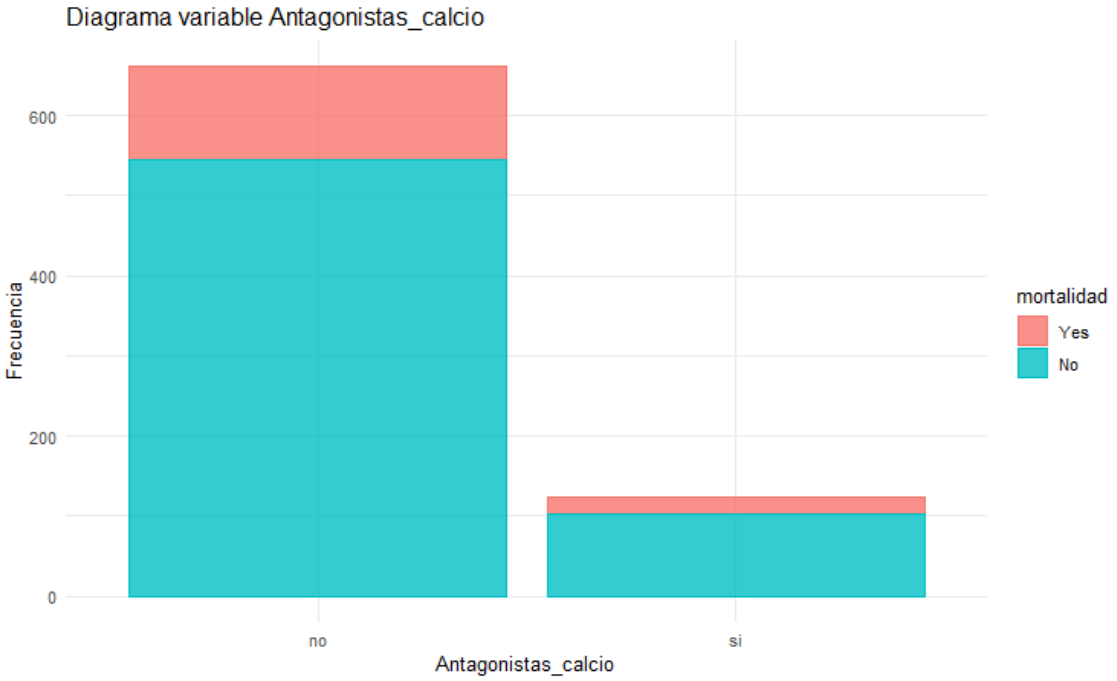

Figura B.20. Diagrama de barras de la variable antagonistas del calcio.

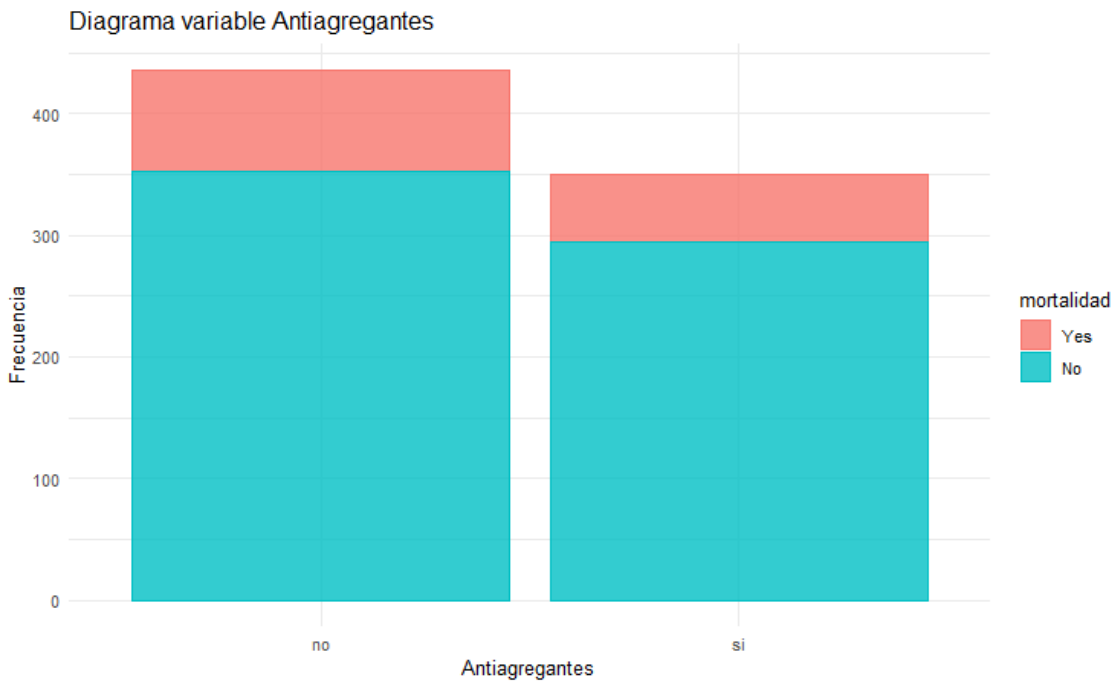

Figura B.21. Diagrama de barras de la variable antiagregantes.

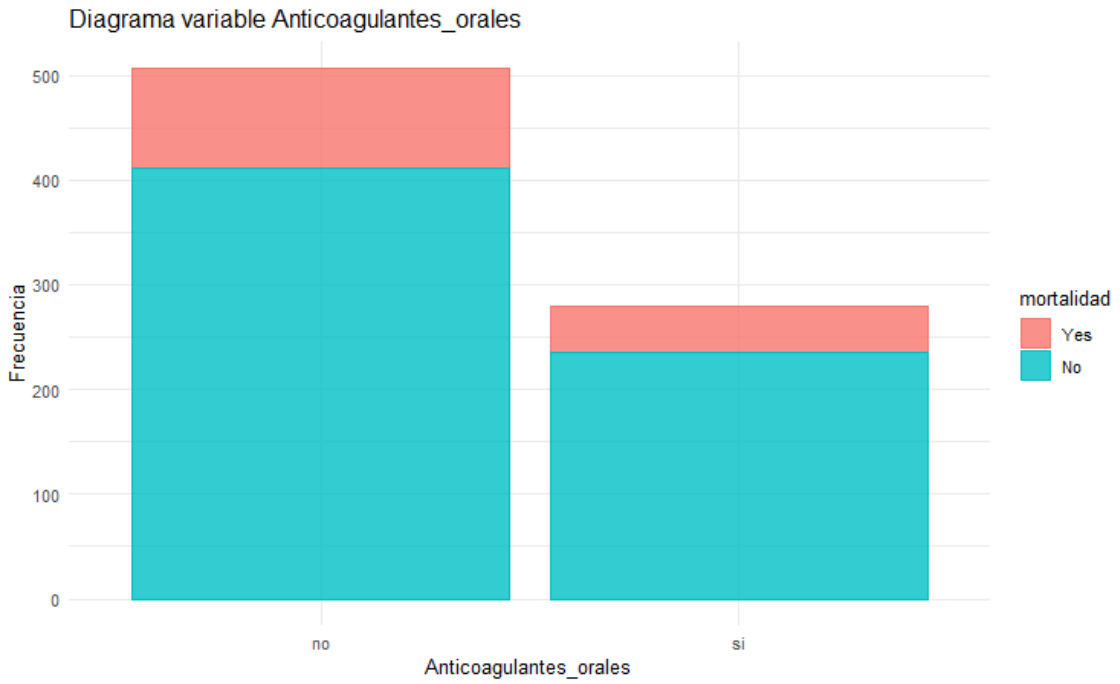

Figura B.22. Diagrama de barras de la variable anticoagulantes orales.

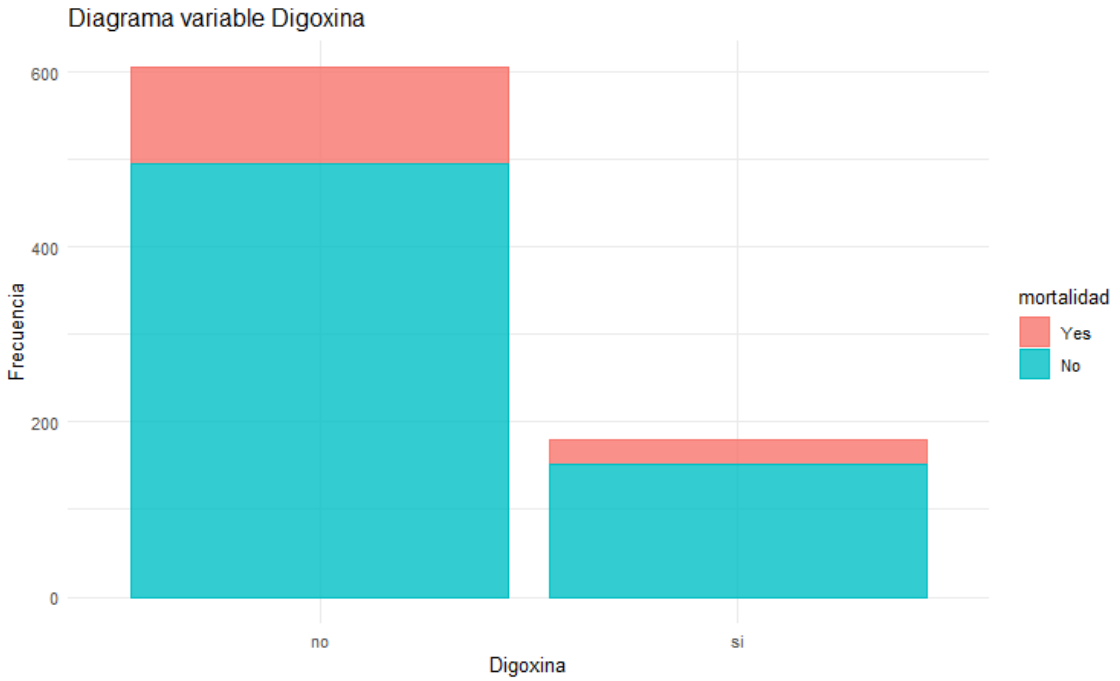

Figura B.23. Diagrama de barras de la variable digoxina.

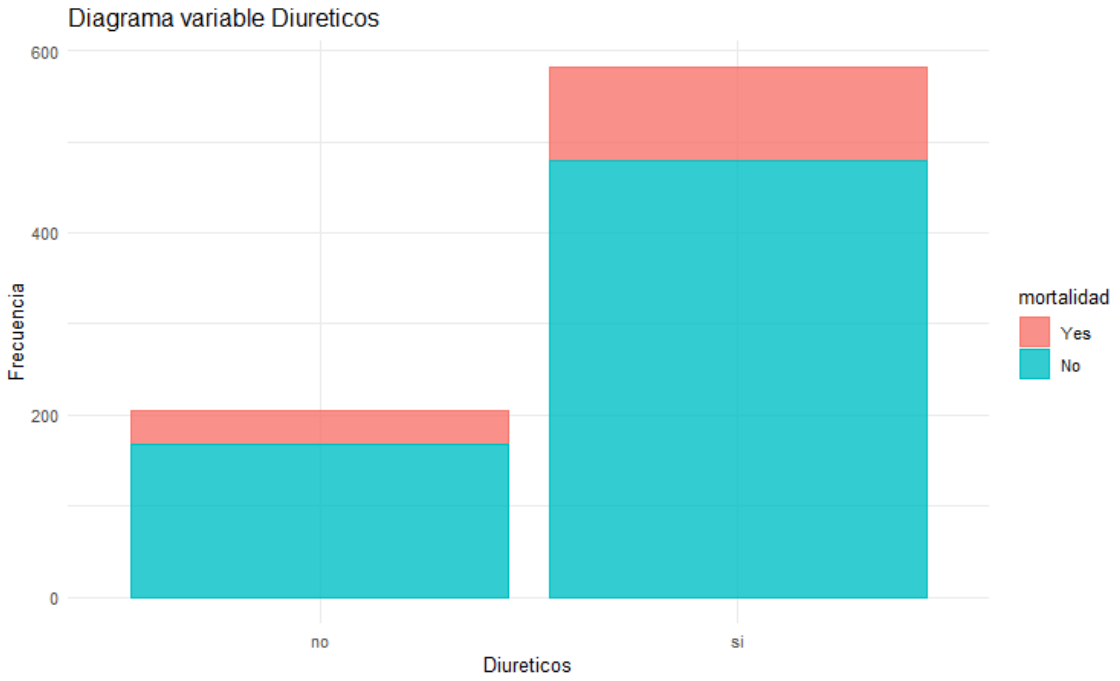

Figura B.24. Diagrama de barras de la variable diuréticos.

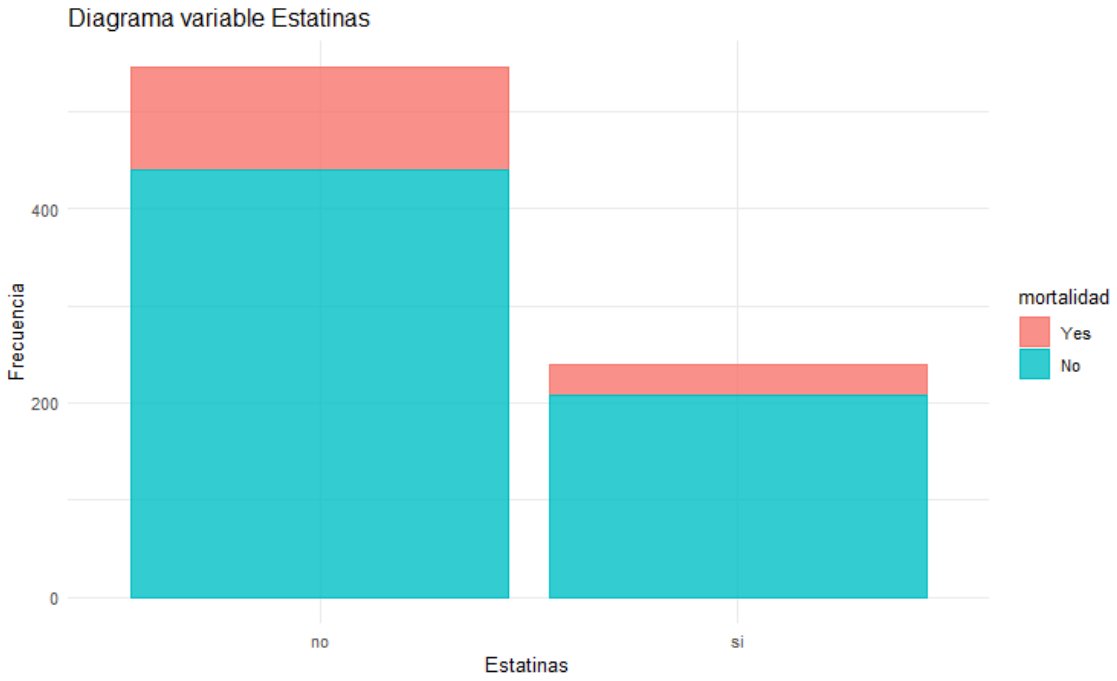

Figura B.25. Diagrama de la variable estatinas.

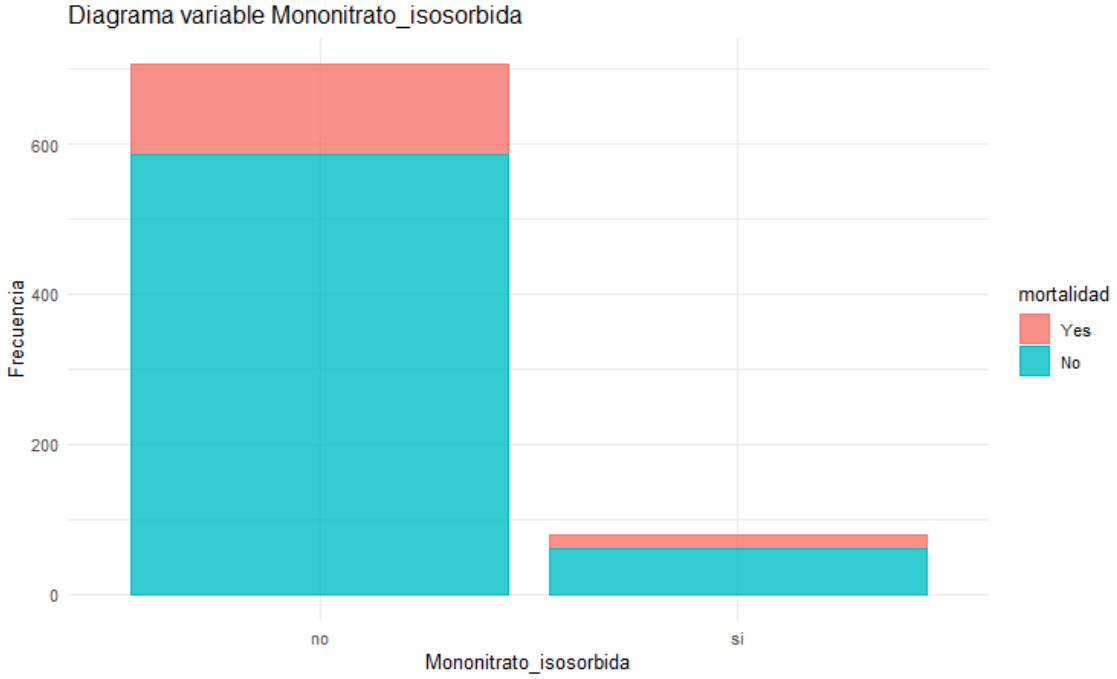

Figura B.26. Diagrama de barras de la variable mononitrato de isosorbida.

Figura B.27. Diagrama de barras de la variable valvulopatías.

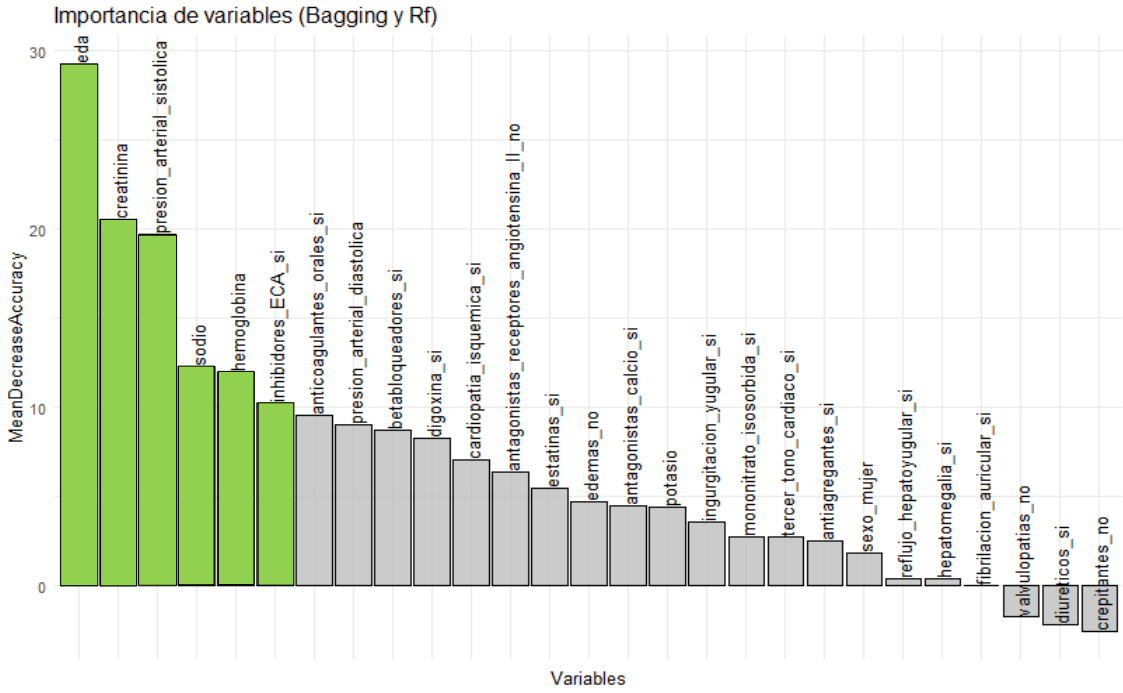

Figura B.28. Orden de importancia de variables (Bagging y Random Forest) (en verde las variables seleccionadas).

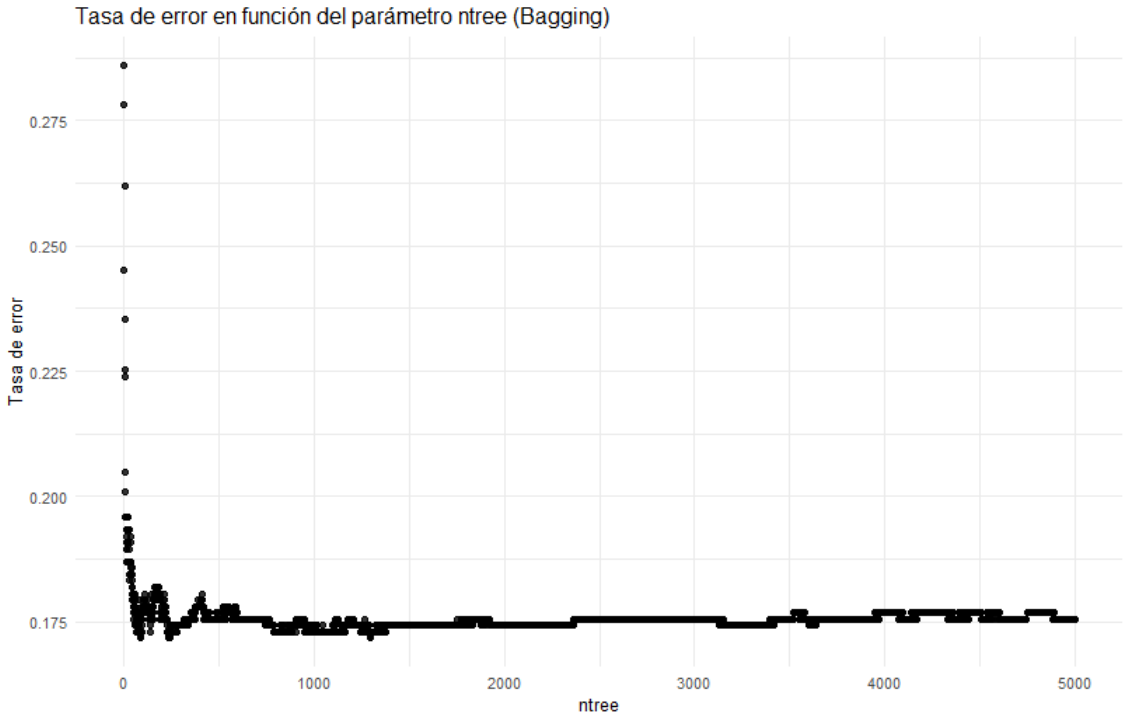

Figura B.29. Variación de la tasa de error en función del parámetro ntree (Bagging).

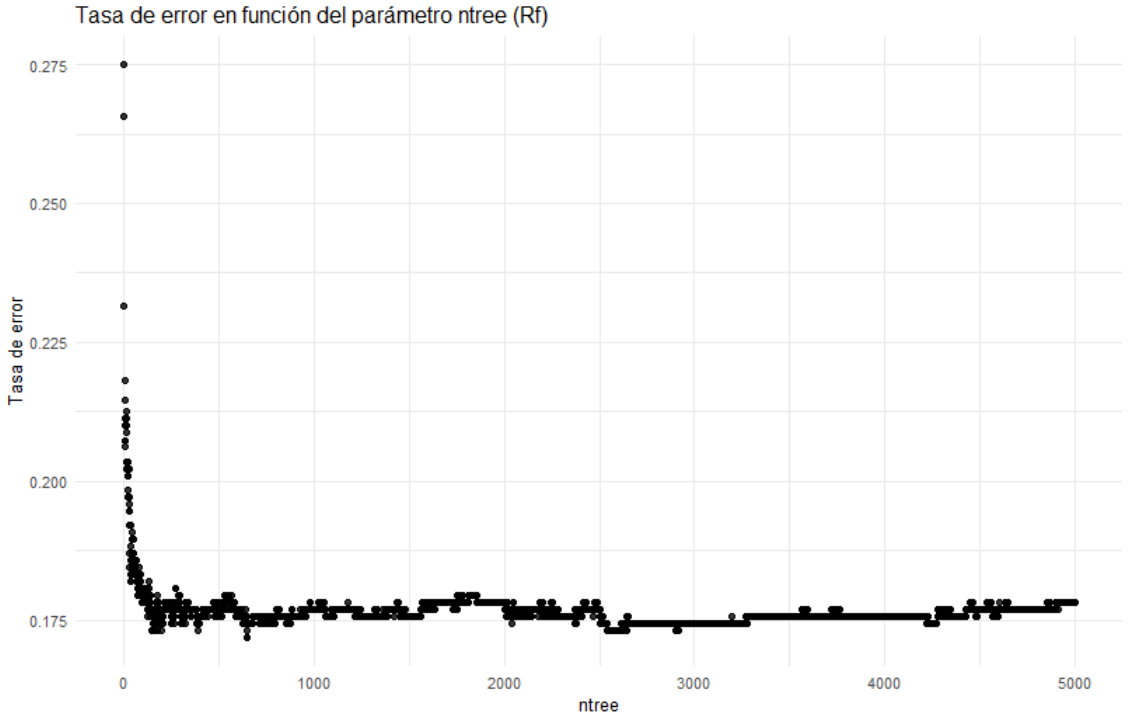

Figura B.30. Variación de la tasa de error en función del parámetro ntree (Random Forest).

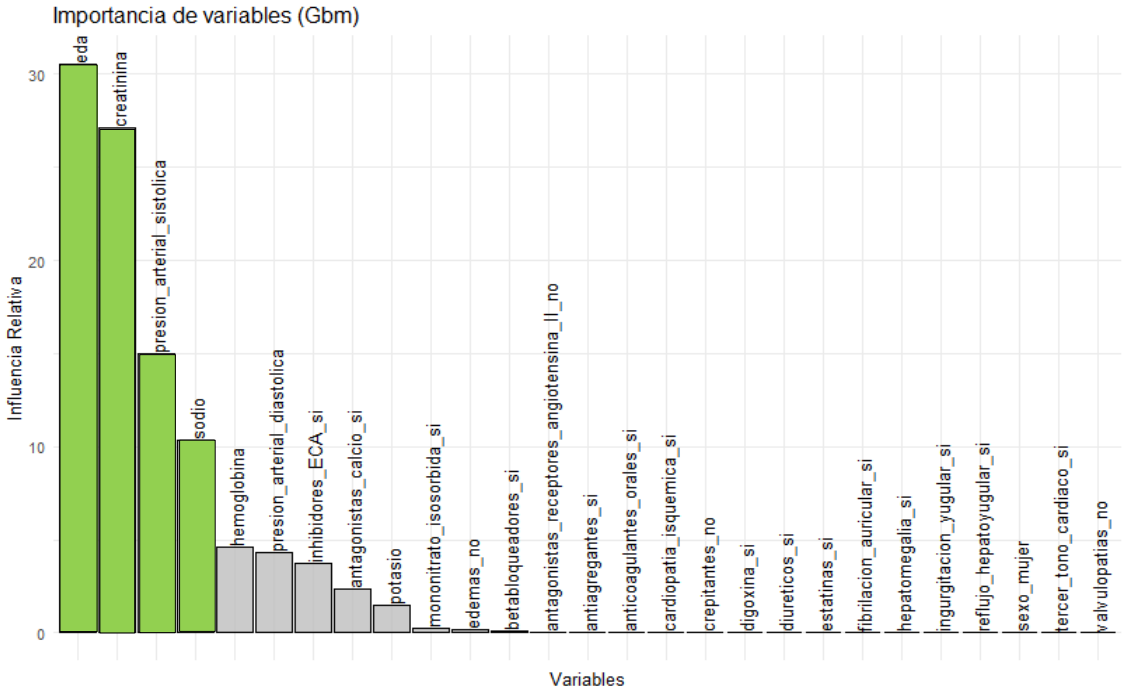

Figura B.31. Orden de importancia de variables (Gradient Boosting) (en verde las variables seleccionadas).

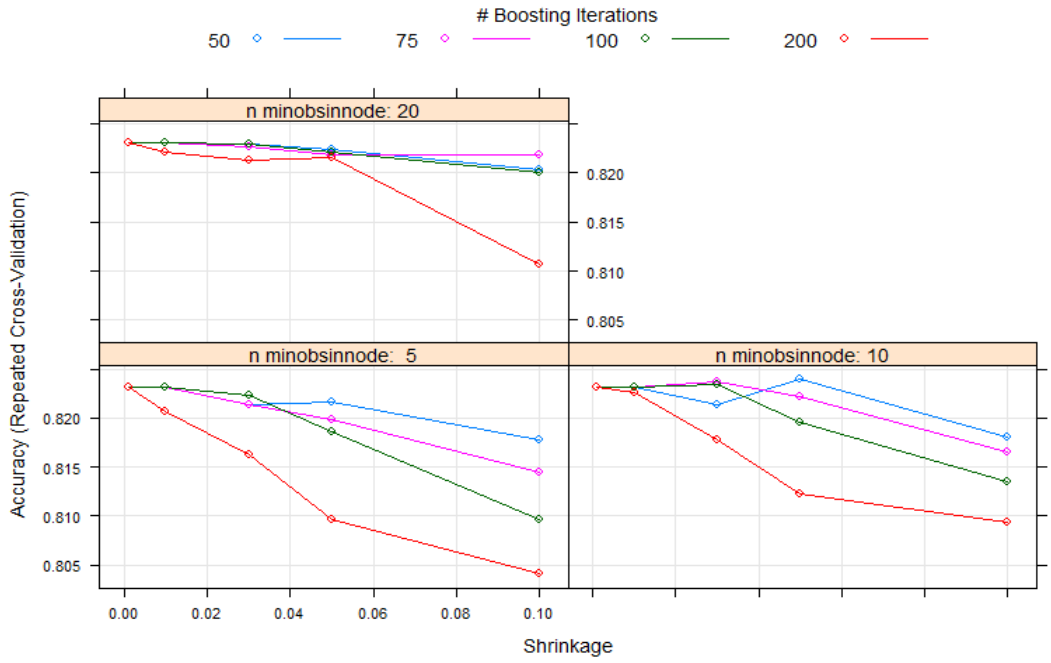

Figura B.32. Variación del Accuracy en función de los parámetros shrinkage y n.minobsinnode (Gradient Boosting).

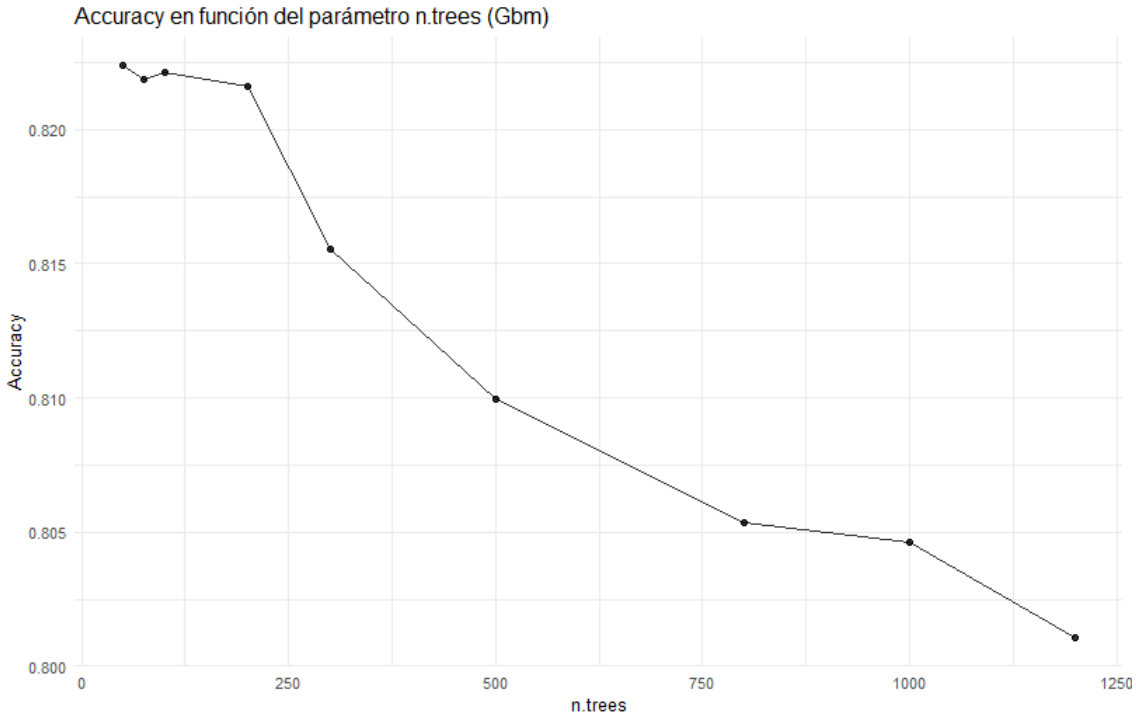

Figura B.33. Variación del Accuracy en función del parámetro n.trees (Gradient Boosting).

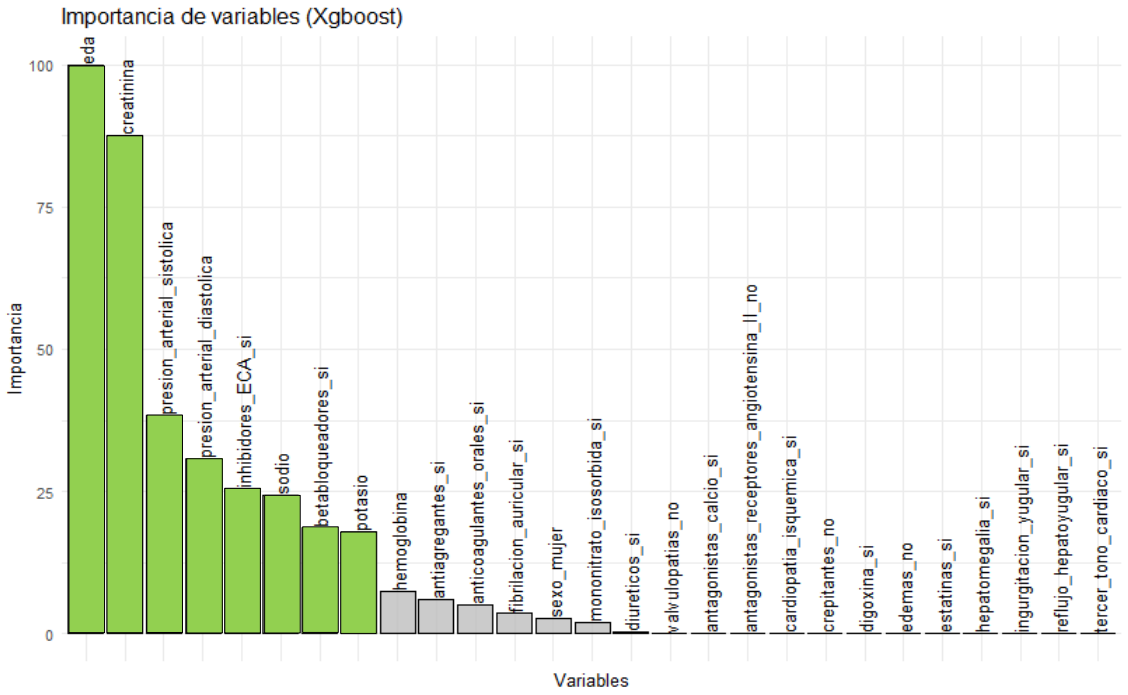

Figura B.34. Orden de importancia de variables (Extreme Gradient Boosting) (en verde las variables seleccionadas).

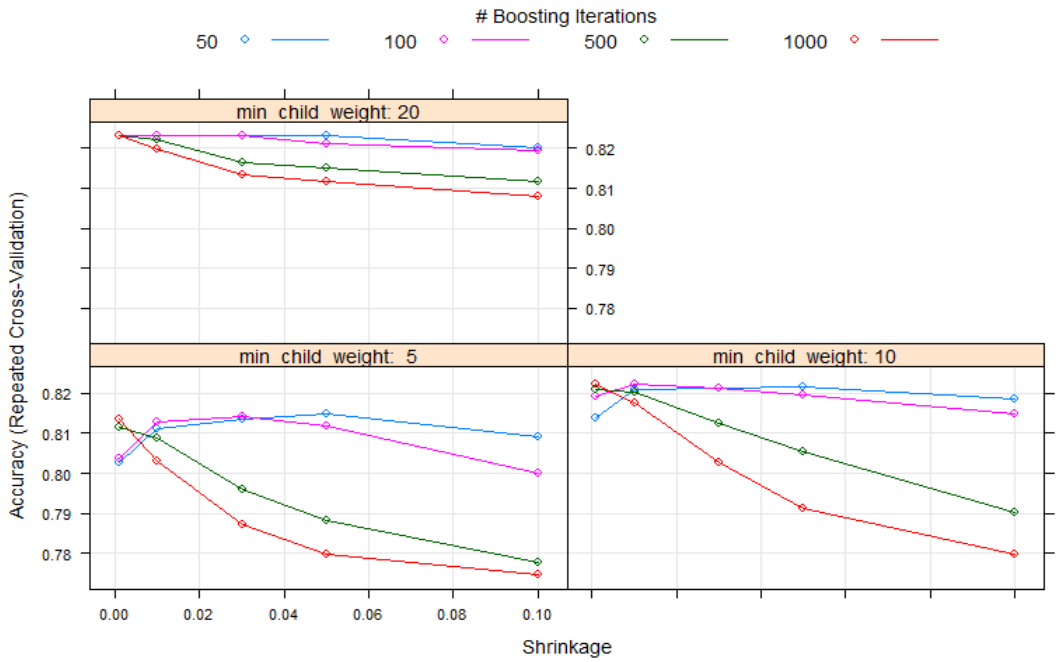

Figura B.35. Variación del Accuracy en función de los parámetros min_child_weigth y eta (Extreme Gradient Boosting).

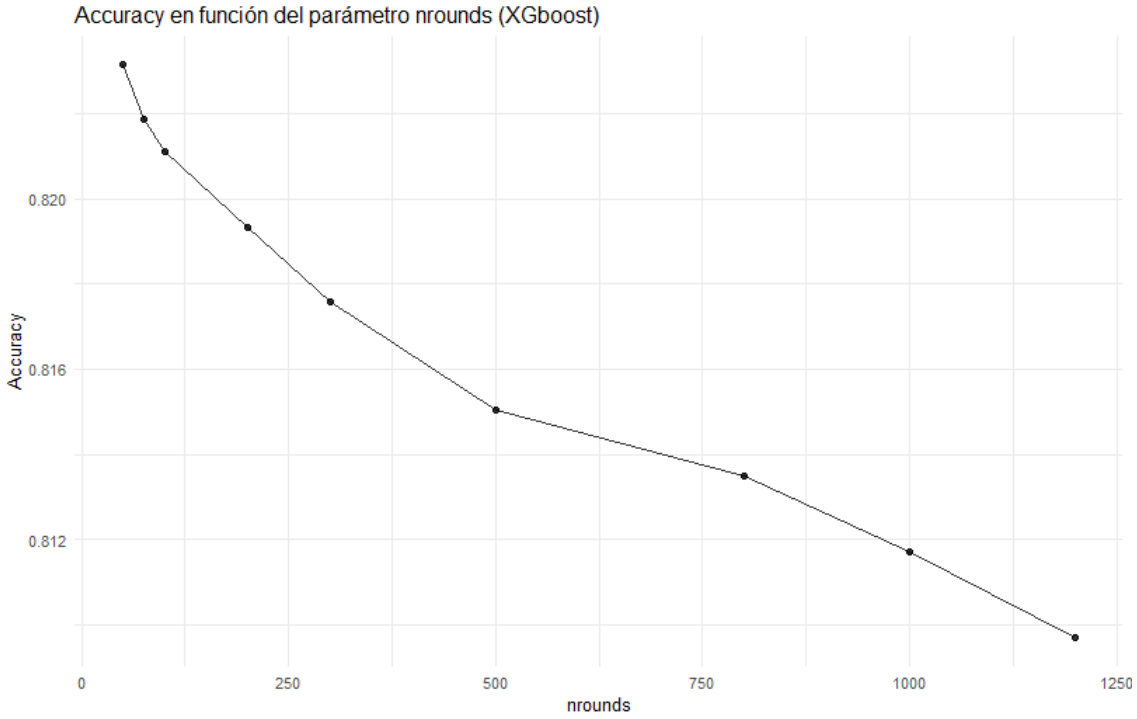

Figura B.36. Variación del Accuracy en función del parámetro nrounds (Extreme Gradient Boosting).

Para realizar consultas más específicas de código y desarrollo del trabajo, se adjunta a continuación un QR con enlace al repositorio GitHub, donde se encuentra toda la información necesaria y utilizada.