FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2023/2024

Trabajo de Fin de Máster

TÍTULO: *Predicción de retraso en los vuelos comerciales de Estados Unidos debido a las condiciones climáticas*

***Alumna*: Carla Gutiérrez Quintana**

***Tutor*: Antonio Sarasa Cabezuelo**

Septiembre de 2024

UNIVERSIDAD COMPLUTENSE
MADRID

A Aida, por su ayuda incondicional.

A Tessy, por hacer posible este año.

Resumen

En un mundo globalizado, la gestión eficiente de vuelos comerciales es fundamental debido a su impacto en la vida cotidiana. Con el clima volviéndose cada vez más cambiante, anticipar los retrasos en estos vuelos se vuelve esencial para mantener la eficiencia en el transporte aéreo. Este trabajo emplea minería de datos para analizar cómo las condiciones meteorológicas afectan los retrasos de vuelos comerciales en cuatro aeropuertos de Estados Unidos durante 2023. Utilizando técnicas de aprendizaje supervisado, se desarrolla un modelo predictivo que identifica patrones en los datos históricos de vuelos y clima. El objetivo es optimizar la gestión de vuelos y reducir costes, mejorando así la experiencia del pasajero mediante herramientas precisas para anticipar y manejar los efectos del clima en la puntualidad de los vuelos.

Los resultados muestran que el método de *XGBoost* es el más eficaz para predecir retrasos. Además, se encontró que las precipitaciones y las rachas de viento tienen el mayor impacto en los retrasos de vuelos, superando a otras condiciones climáticas en su efecto sobre la puntualidad.

Palabras clave: minería de datos, retraso en los vuelos, factores climáticos, análisis predictivo, aprendizaje supervisado.

Abstract

In a globalized world, efficient management of commercial flights is essential due to its impact on daily life. With increasingly unpredictable weather patterns, anticipating flight delays becomes crucial on maintaining efficiency in air transport. This work uses data mining to analyze how weather conditions affect flight delays at four U.S. airports during 2023. By applying supervised learning techniques, a predictive model is developed to identify patterns in historical flight and weather data. The aim is to optimize flight management and reduce costs, thereby improving passenger experience through accurate tools to anticipate and manage the effects of weather on flight punctuality.

The results show that the XGBoost method is the most effective for predicting delays. Additionally, it was found that precipitations and wind gusts have the greatest impacts on flight delays, surpassing other weather conditions in their effect on punctuality.

Keywords: data mining, flight delays, weather factors, predictive analysis, supervised learning.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS.

ÍNDICE DE FIGURAS

CAPÍTULO 1. INTRODUCCIÓN

El objetivo principal de este trabajo es predecir los retrasos de vuelos comerciales en cuatro aeropuertos de Estados Unidos durante el año 2023, causados por factores climáticos.

La elección de este tema responde a la creciente necesidad de gestionar el impacto económico y operativo de los retrasos aéreos, que se ha visto exacerbado por el aumento en la frecuencia de condiciones climáticas adversas cada año. Prever estos retrasos es crucial para minimizar los costes para las aerolíneas y mejorar la experiencia del pasajero.

Los datos utilizados en este estudio provienen de fuentes oficiales. Así, datos relacionados con los vuelos han sido obtenidos de la Oficina de Estadísticas de Transporte (Bureau of Transportation Statistics, BTS), mientras que los datos climatológicos provienen de la Administración Nacional Oceánica y Atmosférica (National Oceanic and Atmospheric Administration, NOAA).

Para alcanzar el objetivo propuesto, el trabajo se organiza en varias fases. En primer lugar, se realizará una breve introducción sobre el impacto de los retrasos en los vuelos y el efecto del clima adverso en estos retrasos. A continuación, se explicará la metodología empleada en el estudio. Posteriormente, se llevará a cabo un análisis descriptivo de los datos para comprender mejor su alcance y preparar el análisis subsiguiente. Con base en la metodología descrita, se desarrollarán los objetivos del estudio. Finalmente, se presentarán las principales conclusiones alcanzadas y se propondrán posibles líneas de investigación futura.

1.1. Contextualización

En un mundo cada vez más globalizado, la demanda global de pasajeros en el transporte aéreo sigue en aumento, con un crecimiento del 36% en 2023 y una proyección del 12% adicional para 2024 (Statista, 2023). Este incremento intensifica la problemática de los retrasos en los vuelos, que representan un desafío en la cadena de suministro del transporte aéreo. No solo afectan a los pasajeros directos, sino que también se propagan a los de otros vuelos debido a la interconexión de recursos, lo que provoca más demoras en el sistema y genera costes adicionales para las aerolíneas y sus clientes, quienes sufren pérdidas en productividad, salarios y confianza (Bombelli et al., 2023).

Las consecuencias económicas de estos retrasos son notables. En 2019, las pérdidas se estimaron en 33 000 millones de dólares en Estados Unidos. En 2022, el coste promedio para las aerolíneas estadounidenses por minuto de operación de un avión, incluyendo tanto el tiempo de rodaje en tierra como el tiempo en el aire, fue de 100.8 dólares (Aviation Intelligence Portal, 2022).

El clima es la principal causa de retrasos en el Sistema Nacional del Espacio Aéreo de Estados Unidos, representando aproximadamente el 75.48% de los retrasos que duraron más de 15 minutos entre junio de 2017 y mayo de 2022 (FAA, 2024).

Además, en 2023, Estados Unidos batió un récord de 28 eventos climáticos y meteorológicos extremos, superando el récord anterior de 22 eventos en 2020. Estos eventos, que incluyeron 4 inundaciones, 1 sequía, 19 tormentas severas, 2 ciclones tropicales, 1 incendio forestal y 1 tormenta invernal, costaron al menos 92 900 millones de dólares (NCEI, 2023).

El incremento en el número y el coste de estos desastres se atribuye en parte a la intensificación de la frecuencia y severidad de ciertos eventos extremos debido al cambio climático (Quinta Evaluación Nacional del Clima de EE.UU., 2023). Este fenómeno hace que los desastres multimillonarios sean cada vez más frecuentes y costosos, amplificando su impacto económico y social. Por lo tanto, resulta indispensable prever y anticipar cuándo las condiciones meteorológicas adversas podrían causar retrasos en el transporte aéreo, con el fin de mitigar los efectos negativos de estos eventos climáticos extremos.

1.2. Estado del arte

El análisis de los retrasos en vuelos comerciales y la creación de modelos para predecir su ocurrencia no es un tema nuevo en la investigación. A lo largo de los años, diversos investigadores han abordado esta cuestión desde diferentes perspectivas, utilizando variados enfoques y técnicas. Estas investigaciones han permitido una mejor comprensión de las causas detrás de los retrasos y han ayudado a desarrollar métodos más efectivos para anticiparlos.

Entre los trabajos relevantes en esta área, algunos han empleado redes bayesianas para investigar cómo los retrasos en un aeropuerto de origen se propagan a un aeropuerto de destino. Estos modelos han demostrado la influencia de variables como el clima y las cancelaciones de vuelos en la propagación de retrasos (Xu et al., 2005).

Otros estudios han evaluado la eficacia de distintos algoritmos para predecir retrasos, y han encontrado que *Support Vector Machine* es uno de los clasificadores más efectivos para esta tarea (Moreno, 2022).

Además, otras investigaciones han analizado el impacto del cambio climático en los retrasos aeroportuarios a lo largo de los próximos cincuenta años, donde se encontró que eventos meteorológicos como tormentas, nieve y niebla incrementan significativamente la probabilidad de retrasos (Pejovic et al., 2009).

Finalmente, algunos autores han empleado algoritmos de *K-Nearest Neighbors* para predecir retrasos causados por condiciones meteorológicas adversas, demostrando que estos métodos pueden ser efectivos para clasificar si un vuelo sufrirá retraso (Priyanka, 2018).

Las principales diferencias entre los trabajos realizados hasta ahora y este Trabajo de Fin de Máster son las siguientes:

1. **Cobertura completa de datos:** A diferencia de otros estudios que emplean muestreo, este trabajo analiza la totalidad de la muestra de datos disponible, lo que permite una visión más completa y detallada de los retrasos en los vuelos.

2. **Foco en aeropuertos específicos**: El análisis de retrasos se centra exclusivamente en cuatro aeropuertos de Estados Unidos, tanto para vuelos de origen como de destino en estos aeropuertos. Esto proporciona un enfoque más detallado y localizado en comparación con estudios anteriores que abarcan una mayor variedad de ubicaciones.

3. **Diversidad de modelos estudiados:** Este trabajo explora una variedad de conjuntos de modelos para la predicción de retrasos en vuelos, a diferencia de otros estudios que se limitan a un único enfoque, como *Random Forest* o *k-NN.*

4. **Uso de modelos de ensamblado:** Se emplean modelos de ensamblado que combinan varios algoritmos de predicción para mejorar la precisión y robustez de las estimaciones de retrasos. Esto contrasta con estudios anteriores que a menudo analizan modelos de forma aislada.

5. **Incorporación de factores adicionales:** Este trabajo integra variables adicionales como si la compañía es *low cost* o no, el momento del día del vuelo, la estación del año y si es fin de semana, junto con datos climáticos.

CAPÍTULO 2. FUENTE DE DATOS

Como se mencionó en la introducción, este estudio se basa en datos obtenidos de dos fuentes principales que abarcan el periodo comprendido entre el 1 de enero y 31 de diciembre de 2023. Los datos relativos a los vuelos fueron recopilados del Departamento de Transporte de Estados Unidos a través de la Oficina de Estadísticas de Transporte (BTS). Por su parte, la información climatológica proviene de la Administración Nacional Oceánica y Atmosférica (NOAA).

Se han seleccionado cuatro aeropuertos por varias razones. En primer lugar, presentan una combinación de características que son beneficiosas para el análisis. Son aeropuertos grandes, con un alto volumen de vuelos y pasajeros, lo que proporciona una muestra significativa y representativa para el estudio. Estos aeropuertos manejan tanto vuelos nacionales como internacionales, lo que añade una capa de complejidad al análisis, permitiendo observar cómo los retrasos varían en función de la naturaleza de los vuelos y las distancias involucradas.

Además, estos aeropuertos presentan variaciones en las condiciones meteorológicas a lo largo del año, lo que permite evaluar el impacto del clima en los retrasos de vuelos en diferentes entornos. La diversidad climática entre los aeropuertos contribuye a un análisis más robusto sobre cómo las condiciones meteorológicas afectan los tiempos de retraso en distintas ubicaciones.

Los aeropuertos y el clima se pueden describir de la siguiente manera:

- **Aeropuerto Internacional O 'Hare (ORD) en Chicago**: Los veranos son calurosos y húmedos, mientras que los inviernos son fríos, nevados y con fuertes vientos. El clima tiende a ser mayormente parcialmente nublado durante todo el año.

- **Aeropuerto Internacional Hartsfield-Jackson (ATL) en Atlanta**: Los veranos son cálidos y bochornosos, y los inviernos son cortos, fríos y húmedos. El clima es también parcialmente nublado a lo largo del año.

- **Aeropuerto Internacional de Los Ángeles (LAX) en California**: Los veranos son agradables, áridos y despejados, mientras que los inviernos son largos, frescos, húmedos y parcialmente nublados.

- **Aeropuerto Internacional de Denver (DEN) en Colorado**: Los veranos son cálidos y predominantemente despejados, mientras que los inviernos son muy fríos, nevados y nublados.

En la Figura 1 se muestra la ubicación de los 4 aeropuertos.

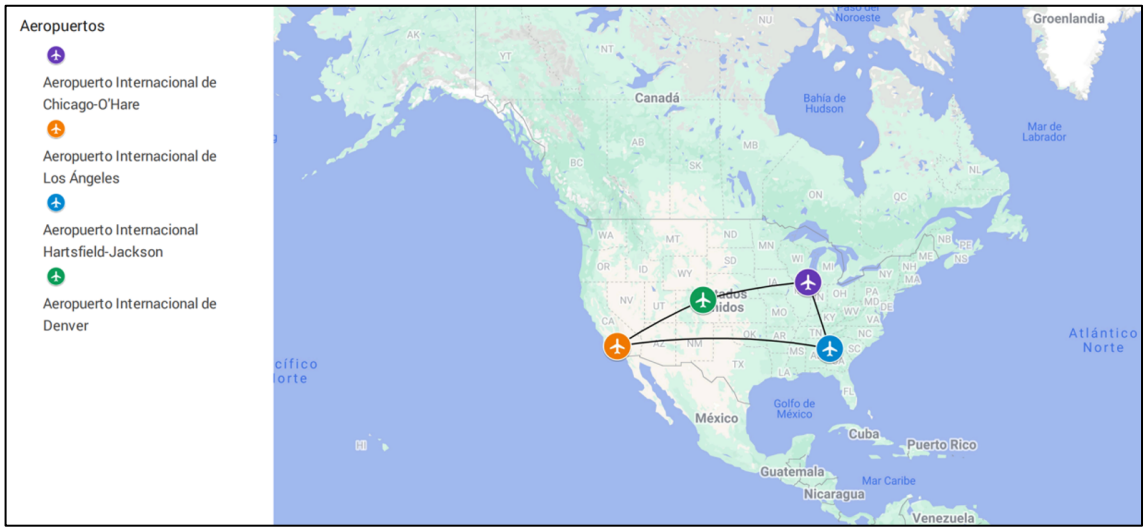


Figura 1. Ubicación de los aeropuertos de estudio

Se recopilaron registros de vuelos con origen en uno de estos cuatro aeropuertos durante el año 2023, resultando en un total de 1 083 925 vuelos. Tras filtrar los datos para incluir solo aquellos vuelos cuyo destino también era uno de los cuatro aeropuertos seleccionados, se obtuvo un conjunto de 76 433 observaciones y 12 variables.

Por otro lado, los datos climáticos incluyen el promedio diario de las condiciones meteorológicas para cada aeropuerto, con un total de 1 460 observaciones (365 por aeropuerto) y 27 variables.

Una vez reunidos ambos conjuntos de datos, se procedió a su integración. Los datos climáticos se incorporaron para los aeropuertos de origen y destino, lo que duplicó las variables climatológicas para cada vuelo. Así, se dispone de 54 variables climatológicas, además de las 12 variables relacionadas con los vuelos, lo que da un total de 66 variables.

CAPÍTULO 3. OBJETIVOS

El objetivo principal de este Trabajo de Fin de Máster es analizar la influencia de las condiciones meteorológicas en los retrasos de vuelos comerciales. Se busca determinar cómo y en qué medida las variables climáticas afectan la puntualidad de los vuelos, con el fin de predecir la probabilidad de retraso bajo diferentes escenarios meteorológicos.

Los objetivos específicos son los siguientes:

1. Evaluar la influencia de las variables meteorológicas en los retrasos de vuelos comerciales.
2. Identificar qué condiciones climáticas son más propensas a generar retrasos.
3. Desarrollar un modelo predictivo para anticipar retrasos en función del clima.
4. Explorar la influencia de factores adicionales en los retrasos de vuelos.

CAPÍTULO 4. METODOLOGÍA

La metodología aplicada en este trabajo sigue el enfoque SEMMA, desarrollada por SAS Institute, que se define como "el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocios desconocidos" (SAS Institute, 1998). SEMMA es un acrónimo que representa las cinco fases del proceso:

Muestreo (*Sample*). Es fundamental obtener un conjunto de datos que sea lo suficientemente grande para ser representativo y, al mismo tiempo, manejable para el análisis. En nuestro caso, al filtrar los datos para incluir solo cuatro aeropuertos de origen o destino, el tamaño del conjunto de datos se redujo de poco más de un millón a 76 433 registros.

Exploración (*Explore*). Esta fase implica analizar y explorar los datos mediante tablas, agrupaciones y gráficos para identificar particularidades, errores o tendencias.

Modificar (*Modify*). En esta etapa, se preparan y ajustan los datos antes de la modelización. Esto incluye el tratamiento de valores atípicos, la imputación de datos faltantes y la selección de variables relevantes para asegurar que el modelo sea lo más preciso posible.

Modelado (*Model*): Se construyen y ajustan modelos para encontrar el que mejor se adecúe a los datos, optimizando su rendimiento y precisión.

Evaluación (*Assess*): Finalmente, se evalúa el rendimiento del modelo para asegurarse de que cumple con los objetivos y requisitos establecidos.

Los modelos utilizados en la cuarta fase (modelización), son los que describen en los siguientes subapartados.

4.1. Regresión logística

La regresión logística estima la probabilidad de que ocurra un evento binario utilizando una función logística, la cual transforma la combinación lineal de las variables independientes en un rango entre 0 y 1. La fórmula básica de la regresión logística es:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots \beta_m x_m}})$$

Donde P es la probabilidad del evento de interés y βi representa el coeficiente del predictor i-ésimo. Además de proporcionar una probabilidad estimada, la regresión logística permite interpretar los efectos de las variables predictoras a través de los coeficientes de regresión. Estos coeficientes indican el cambio en el logaritmo de las probabilidades (odds) asociado con un cambio unitario en los predictores. Los odds de un evento, definidos como el cociente entre la probabilidad de que ocurra el evento y la probabilidad de que no ocurra, se expresan matemáticamente como:

$$odds\ (A) = \frac{P(A)}{1 - P(A)}$$

Esta relación facilita la interpretación de cómo cada variable predictora afecta la probabilidad del evento, proporcionando una visión más clara sobre la influencia de cada factor en el resultado binario (James et al., 2013).

4.2. Redes neuronales

Las redes neuronales son un enfoque de aprendizaje supervisado inspirado en el funcionamiento del cerebro humano, modelando de manera abstracta algunos de sus procesos. Su estructura simplificada se presenta en la Figura 2. Mientras que las redes biológicas son vastamente más complejas y adaptativas, las redes artificiales buscan replicar ciertos principios básicos de procesamiento de información y aprendizaje en un contexto computacional. Estas redes están diseñadas para reconocer patrones y realizar tareas de predicción y clasificación a partir de datos complejos.

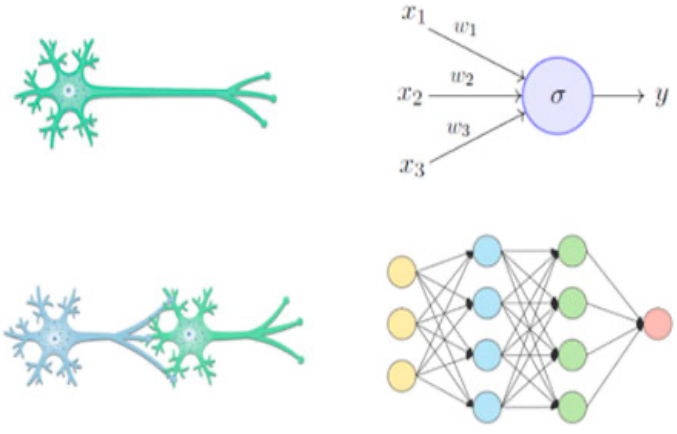


Figura 2. Esquema de una red neuronal

(Google 2024)

Una red neuronal artificial se compone de neuronas organizadas en capas: la capa de entrada, una o más capas ocultas, y la capa de salida. Cada neurona en una capa está conectada a todas las neuronas de la capa siguiente mediante pesos ajustables. El proceso de entrenamiento ajusta estos pesos para minimizar el error en las predicciones del modelo. La red recibe entradas a través de la capa de entrada, las procesa mediante funciones de activación no lineales en las capas ocultas, y produce una salida correspondiente a la predicción o clasificación solicitada (Bishop, 1995).

Las redes neuronales son versátiles y se aplican en diversas áreas. En la clasificación binaria, donde el objetivo es asignar una observación a una de dos categorías posibles, las redes neuronales proporcionan una gran herramienta para modelar y predecir. Sin embargo, la efectividad de estas redes está condicionada por varios factores, como la elección adecuada de la arquitectura y la prevención del sobreajuste.

Una de las características distintivas de las redes neuronales es su naturaleza de "caja negra". Esto se refiere a la dificultad para interpretar cómo las redes llegan a sus decisiones, debido a la complejidad de sus estructuras internas y la interrelación de los parámetros. En la clasificación binaria, esto significa que, aunque las redes neuronales pueden proporcionar predicciones precisas, entender y explicar el proceso de toma de decisiones puede ser un una tarea compleja.

Para obtener la mejor red neuronal, se ajustan tres parámetros:

> ***Size*.** Número de nodos en las capas ocultas. Un mayor número capturar patrones más complejos en los datos, lo que puede mejorar la precisión del modelo. Sin embargo, si se utilizan demasiados nodos, existe el riesgo de sobreajuste, donde el modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad de generalización para nuevos datos.
>
> ***Decay*.** La tasa de decaimiento es un parámetro de regularización que penaliza los pesos altos en la red neuronal. Al aplicar esta penalización, se evita que el modelo se vuelva excesivamente complejo, lo que da lugar a un modelo menos propenso a errores en datos no vistos.
>
> ***Maxit* .** El número máximo de iteraciones establece cuántas veces se ajustarán los pesos de la red durante el proceso de entrenamiento.

Adicionalmente, se utilizará la función de activación sigmoide, la cual introduce no linealidades en el modelo.

4.3. Modelos basados en árboles de decisión

Los modelos basados en árboles de decisión son ampliamente utilizados en minería de datos y aprendizaje automático para tareas de clasificación y regresión. Funcionan dividiendo recursivamente un conjunto de datos en subconjuntos más pequeños mediante reglas de decisión, produciendo un árbol en el que cada nodo representa una prueba sobre un atributo, las ramas muestran los resultados y los nodos hoja contienen la clasificación o predicción final. Dicha estructura está representada en la Figura 3.

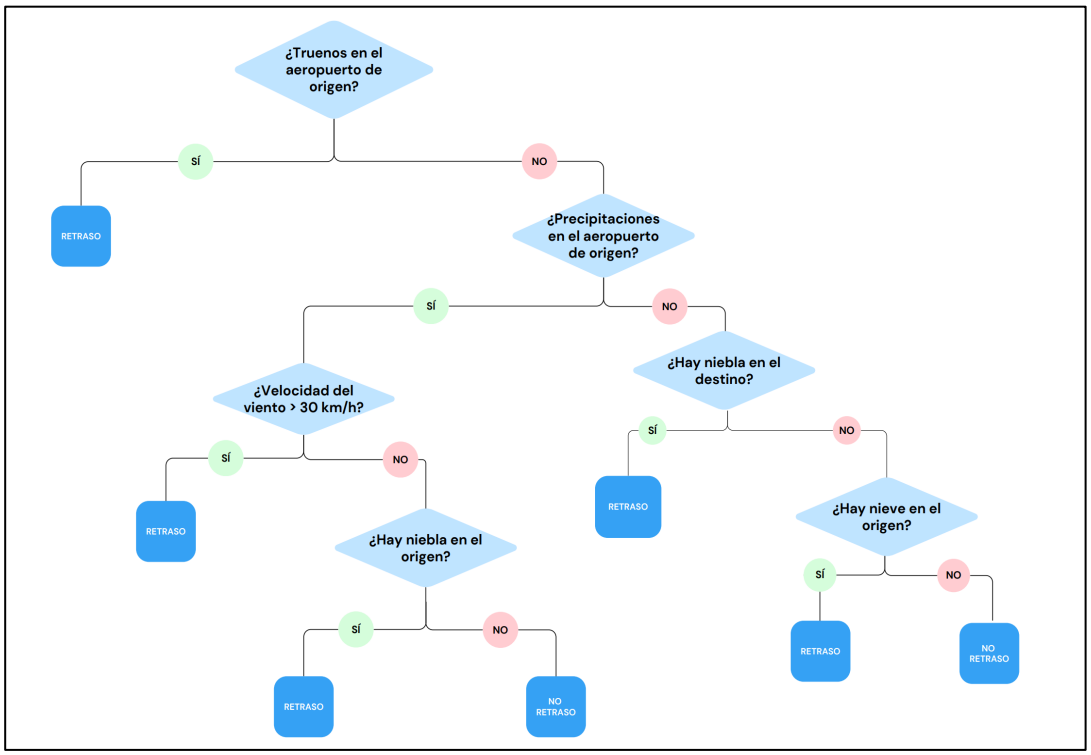


Figura 3. Esquema de un árbol de decisión

Como se puede observar, el árbol está compuesto por varios elementos en su estructura. El nodo raíz es el nodo principal del árbol y actúa como el punto de partida de todas las conexiones. Desde el nodo raíz se ramifican todos los demás componentes del árbol.

A continuación, se encuentran los nodos internos, los cuales tienen al menos un hijo y sirven para tomar decisiones intermedias. Estos nodos permiten dividir el árbol en ramas más específicas, siguiendo criterios establecidos para organizar la información de manera jerárquica.

En la parte final de cada rama están los nodos hoja, también conocidos como nodos terminales. Estos nodos representan las decisiones finales o los resultados de la clasificación y no tienen hijos adicionales. Los nodos hoja son el punto culminante de cada rama del árbol.

Las ramas del árbol, por su parte, son las conexiones que unen los nodos entre sí. Cada rama conecta un nodo con sus hijos, representando la relación entre diferentes características. Estas ramas facilitan la propagación de la información desde el nodo raíz hasta los nodos hoja, estructurando el árbol de manera que organiza y simplifica la toma de decisiones y la interpretación de datos.

4.3.1. Bagging

Es una técnica de ensamblaje diseñada para mejorar la precisión de modelos de aprendizaje automático, especialmente aquellos con alta varianza, como los árboles de decisión. Esta técnica implica crear múltiples conjuntos de datos a partir del original mediante muestreo con reemplazo, entrenar un modelo base independiente para cada

conjunto y luego combinar las predicciones obtenidas: promediando en la regresión y votando mayoritariamente en la clasificación, lo que da lugar a un modelo más robusto y preciso al minimizar el impacto de errores individuales.

Se establecerán los siguientes parámetros:

> ***Ntree*.** Número de iteraciones que se construirán. Este parámetro controla cuántos modelos base se generarán durante el proceso. Un mayor número de árboles generalmente mejora la estabilidad y precisión del modelo, aunque con un coste adicional en términos de tiempo de computación.
>
> ***Sampsize*.** Tamaño de la muestra para cada conjunto de datos generado mediante muestreo con reemplazo. Este parámetro controla cuántas instancias se tomarán para crear cada conjunto de datos de entrenamiento.
>
> ***Mtry*.** Número máximo de variables que se consideran en cada división del árbol de decisión. En el caso particular de *Bagging*, es igual al número de variables independientes.
>
> ***Nodesize***. Tamaño mínimo de los nodos terminales (hojas) en los árboles de decisión.

4.3.2. Random Forest

Random Forest se describe como una extensión del método de *Bagging* que introduce una mayor aleatoriedad añadiendo el sorteo de variables en cada nodo. En lugar de construir múltiples árboles de decisión independientes con el mismo conjunto de variables, *Random Forest* selecciona un subconjunto aleatorio de variables en cada nodo del árbol para decidir la mejor división. Esta técnica de combinar múltiples árboles con variabilidad adicional mejora la precisión y estabilidad del modelo, reduciendo la varianza sin aumentar el sesgo.

Para ajustar un modelo *Random Forest*, se configura:

> ***Ntree*.** Número de árboles que se construyen.
>
> ***Sampsize*.** Tamaño de la muestra a sortear.
>
> ***Mtry*.** Número máximo de variables que se consideran en cada división del árbol de decisión. Un valor menor introduce más aleatoriedad, lo que puede mejorar la generalización al reducir la correlación entre los árboles.
>
> ***Nodesize***. Tamaño mínimo de las hojas en los árboles de decisión.

4.3.3. Gradient Boosting

Este algoritmo construye árboles de clasificación de manera reiterativa, con el objetivo de minimizar los residuos en cada etapa. El proceso comienza con un modelo simple, como un árbol de decisión pequeño. Luego, se ajustan modelos adicionales que se centran en los errores residuales de los modelos anteriores. Es decir, las predicciones se

actualizan en la dirección de decrecimiento indicada por el negativo del gradiente de la función de error. De este modo, las predicciones se ajustan progresivamente para asemejarse cada vez más a los datos reales.

En un modelo de *Gradient Boosting*, los principales parámetros a configurar son:

> ***Shrinkage*.** Tasa de aprendizaje. Controla la regularización del modelo, ajustando la contribución de cada nuevo árbol al modelo final. Un valor menor de *shrinkage* requiere más árboles para obtener el mismo nivel de ajuste, pero puede mejorar la generalización.
>
> ***Nminobsinnode.*** Número mínimo de observaciones por nodo. Define el número mínimo de observaciones necesarias en un nodo para que se realice una división.
>
> ***Ntree*.** Establece la cantidad total de árboles que se generarán en el proceso de *boosting*.
>
> ***Interaction.depth*.** Define la profundidad máxima de cada árbol, es decir, controla el número máximo de interacciones entre las características que cada árbol puede capturar. Para clasificación binaria, un valor comúnmente utilizado es 2.

4.3.4. XGBoost

XGBoost (*Extreme Gradient Boosting*) es una versión optimizada del anterior que mejora la velocidad y eficiencia. Utiliza técnicas avanzadas, como la paralelización de tareas y la optimización de la función de pérdida mediante el uso del hessiano (segunda derivada), para lograr un rendimiento superior. El objetivo principal es minimizar la función de pérdida, que mide la discrepancia entre las predicciones del modelo y los valores reales, aumentando así la precisión del modelo.

Implementa una regularización avanzada para combatir el sobreajuste. Ofrece dos tipos principales de regularización: L1 (Lasso), que penaliza la suma de los valores absolutos de los coeficientes para promover modelos más simples, y L2 (Ridge), que penaliza la suma de los cuadrados de los coeficientes para reducir la magnitud de los coeficientes y evitar valores extremos. Además, la técnica de poda de árboles se utiliza para optimizar el ajuste del modelo, eliminando ramas menos significativas, y la paralelización mejora la velocidad al permitir la ejecución simultánea de varias tareas.

En este modelo se ajustarán los siguientes parámetros:

> ***Nround*.** Número de iteraciones antes de detener el proceso de ajuste.
>
> ***Maxdepth*.** Controla la profundidad de los árboles de decisión.
>
> ***Eta*.** Tasa de aprendizaje del modelo (en *Bagging: Shrinkage*). Regula cuánto contribuye cada árbol al modelo final. Valores menores hacen que el aprendizaje sea más lento, lo que puede mejorar la generalización al evitar el sobreajuste.

Min_child_weight. Define el número mínimo de observaciones requeridas en un nodo para realizar una división.

Gamma. Controla el coste de regularización. Aumenta la regularización si se establecen divisiones adicionales en los árboles.

Alpha. Controla la regularización L1 sobre las ponderaciones; un aumento en alpha hace que el modelo sea más conservador al forzar a que algunas características tengan coeficientes exactamente en cero.

Lambda. Regula la regularización L2 sobre las ponderaciones; valores más altos de lambda también hacen que el modelo sea más conservador al penalizar más las grandes ponderaciones.

Lambda_bias. Ajusta la regularización L2 del sesgo en el modelo.

4.4. Support Vector Machines

*Support Vector Machine*s (SVM) es una técnica robusta en aprendizaje automático utilizada tanto para clasificación como para regresión. En el caso de la clasificación, SVM busca encontrar un hiperplano óptimo que divida las clases en un espacio multidimensional. El objetivo es maximizar el margen, es decir, la distancia entre el hiperplano y los puntos de datos más cercanos de cada clase.

Esta maximización del margen ayuda a minimizar el error de clasificación y a mejorar la capacidad del modelo para generalizar a datos nuevos no vistos. SVM es eficaz en escenarios de alta dimensión y puede manejar casos en los que las clases no son linealmente separables en el espacio original mediante el uso de funciones kernel.

4.4.1. Lineal

El kernel lineal es una función que opera en el espacio original de características sin transformar los datos. Su objetivo es encontrar una separación lineal directa entre las clases, como se presenta en la Figura 4.

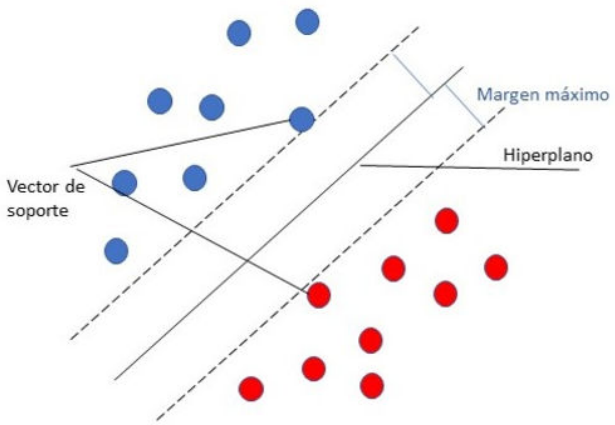


Figura 4. SVM lineal

(Google, 2024)

En esta figura, se muestra una línea que divide dos clases distintas, con los puntos de cada color representando las dos categorías de la variable dependiente. Este enfoque es adecuado cuando los datos son linealmente separables, es decir, cuando las clases pueden distinguirse mediante una línea recta en el espacio de características original.

En este algoritmo, el principal hiperparámetro que se controla es:

> $\boldsymbol{C}$. Es la constante de regularización. Este parámetro ajusta el equilibrio entre el margen del clasificador y los errores de clasificación en el entrenamiento. Un valor alto de C busca reducir los errores de clasificación al coste de un margen más estrecho, lo que puede llevar a un sobreajuste si el modelo se ajusta demasiado a los datos de entrenamiento. En contraste, un valor bajo de C permite un margen más amplio, pero a costa de permitir más errores de clasificación.

4.4.2. Polinómico

El kernel polinómico eleva las características originales a un espacio de mayor dimensión mediante una función polinómica. Esta transformación permite al SVM capturar relaciones más complejas entre los datos, facilitando la separación de clases que no son linealmente separables en el espacio original. En la Figura 5, se puede observar cómo los datos se transforman en un espacio de dimensiones superiores, donde se vuelve posible trazar una frontera de decisión no lineal que distingue mejor las diferentes clases.

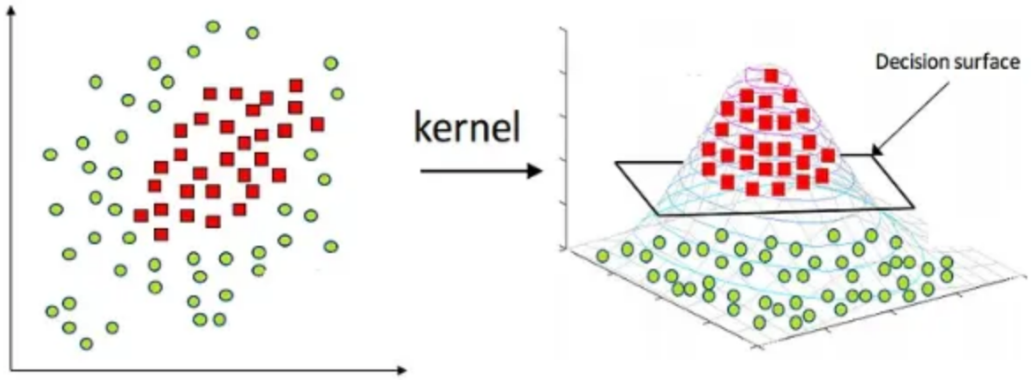


Figura 5. SVM polinómico

(Google, 2024)

En el caso del kernel polinómico, se ajustan tres hiperparámetros principales:

> $\boldsymbol{C}$. La constante de regularización.
>
> ***Degree*.** Los grados del polinomio.
>
> ***Scale*.** La escala del kernel. Este parámetro ajusta la influencia de las características en la función del kernel. Una escala mayor puede hacer que el

modelo sea más sensible a características individuales, mientras que una escala menor puede suavizar la influencia de las características.

4.4.3. Radial

El kernel radial, también conocido como Radial Basis Function (RBF) o gaussiano, utiliza una función gaussiana para transformar los datos en un espacio de dimensión infinita. Esta transformación facilita la separación de datos que están altamente entrelazados o no linealmente separables en el espacio original.

En la Figura 6 se ilustra cómo el kernel radial permite la creación de diversos límites de decisión complejos y adaptativos, lo que facilita una clasificación precisa incluso en datos que no son linealmente separables.

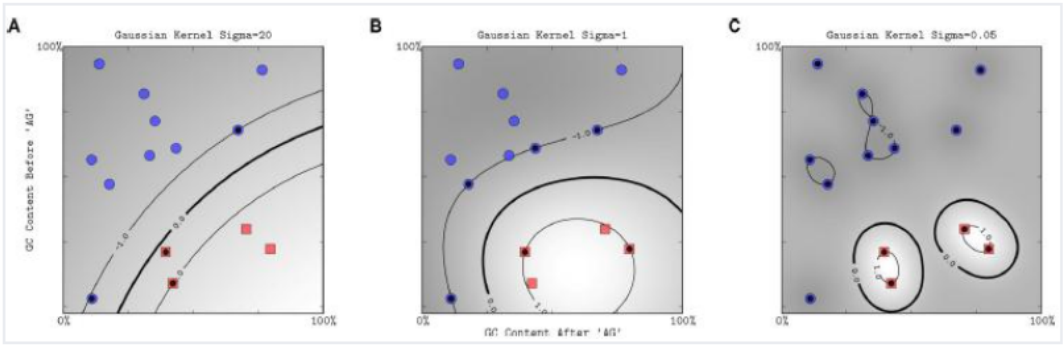


Figura 6. SVM radial

(Google, 2024)

Con kernel gaussiano se tienen en cuenta los siguientes parámetros:

C. La constante de regularización.

Sigma. Controla el comportamiento del kernel y mide la similitud entre dos puntos. Determina el ancho de la función de base radial. Un sigma mayor hace que el modelo considere puntos más distantes como similares, lo que resulta en un modelo más suave. Por otro lado, un sigma menor aumenta la sensibilidad del modelo a las diferencias entre puntos cercanos, lo que puede llevar al sobreajuste.

4.5. Métodos de ensamblado

Los métodos de ensamblado combinan las predicciones de múltiples modelos para mejorar el rendimiento y la robustez del modelo final. La idea principal es que, al combinar varios modelos, se pueden aprovechar las fortalezas de cada uno y reducir el impacto de sus debilidades individuales, lo que suele llevar a una mejor generalización y mayor precisión al conseguir modelos con menor varianza.

4.6. Técnicas evaluación de modelos

4.6.1. Remuestreo y optimización de modelos

Las técnicas de remuestreo son fundamentales en el análisis de datos y en la validación de modelos de aprendizaje automático, ya que permiten evaluar el rendimiento del modelo de manera que el azar no influya en los resultados obtenidos. En este trabajo, se utilizarán tanto la técnica de división *training* - *test* repetido como la validación cruzada repetida.

Training – Test. Esta técnica consiste en dividir el conjunto de datos en dos partes: una para entrenar el modelo (conjunto de entrenamiento o *train*) y otra para evaluarlo (conjunto de prueba o *test*) para evaluar su desempeño en datos no vistos. El modelo se ajusta usando los datos de entrenamiento y luego se evalúa su rendimiento en el conjunto de prueba, que no ha sido utilizado durante el entrenamiento. En este caso, se establece una división del 70% para *training* y el 30% restante para *test*. Esta aproximación permite una evaluación inicial del modelo, ofreciendo una medida directa de su desempeño en datos no vistos.

Validación cruzada. Esta técnica sirve para evaluar el rendimiento de un modelo al dividir el conjunto de datos en múltiples particiones. Se utiliza la validación cruzada $k-fold$, donde los datos se dividen en k subconjuntos (o *folds*). El modelo se entrena en $k-1$ de estos subconjuntos y se valida en el subconjunto restante. Este proceso se repite k veces, con cada subconjunto sirviendo como conjunto de validación una vez. Esta metodología permite que cada observación se use tanto para entrenamiento como para validación, lo que proporciona una evaluación más robusta del modelo (James et al., 2013).

Validación cruzada repetida. Es una extensión de la validación cruzada que consiste en repetir el procedimiento de validación cruzada múltiples veces para mejorar la estabilidad y precisión de la evaluación del modelo. En este caso, se realizará la validación cruzada $k-fold$ (en este tabajo, $k=4$) y se repite este procedimiento cinco veces.

4.6.2. Matriz de confusión

Para evaluar la calidad del algoritmo de clasificación desarrollado, se utilizará una matriz de confusión que permite evaluar el desempeño del modelo al comparar las predicciones realizadas con las verdaderas etiquetas de los datos.

La matriz de confusión se organiza según la Figura 7, mostrando los conteos de verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN).

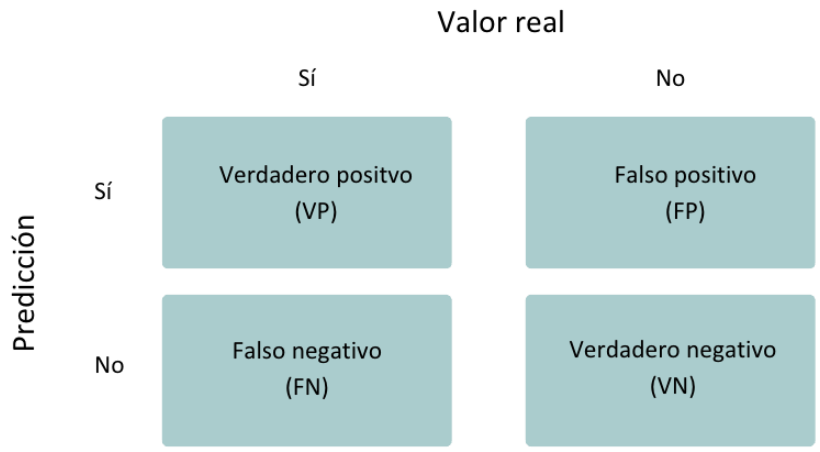


Figura 7. Matriz de confusión

A partir de la matriz de confusión, es posible calcular:

Accuracy. Mide la proporción de predicciones correctas sobre el total de predicciones realizadas.

$$Accuracy = \frac{Predicciones\ correctas}{Total\ de\ predicciones} = \frac{VP + VN}{VP + VN + FP + FN}$$

Tasa de fallos. Indica la proporción de predicciones incorrectas respecto al total de predicciones.

$$Tasa\ de\ fallos\ = \frac{FP + FN}{VP + VN + FP + FN}$$

Especificidad. Mide la proporción de negativos reales que se identifican correctamente como tales.

$$Especifididad = \frac{VN}{VN + FP}$$

Sensibilidad. Indica la proporción de positivos reales que el modelo identifica correctamente.

$$Sensibilidad\ = \frac{VP}{FN + VP}$$

4.6.3. Tasa de fallos y AUC

Los modelos resultantes de la validación cruzada repetida se compararán a través de gráficos de caja y bigotes, donde se analizará la tasa de fallos y el AUC.

El AUC (Área Bajo la Curva) evalúa la capacidad del modelo para discriminar entre las clases. Se obtiene a partir de la curva ROC (Receiver Operating Characteristic), que traza la tasa de verdaderos positivos frente a la tasa de falsos positivos para diferentes umbrales de decisión. El valor del AUC varía entre 0 y 1, donde 0 indica un modelo completamente aleatorio, y 1 representa un modelo perfecto, con una capacidad perfecta para generalizar.

Mientras que la tasa de fallos refleja la proporción de clasificaciones incorrectas realizadas por el modelo, indicando con qué frecuencia el modelo no logra predecir correctamente las clases en los datos de prueba, los gráficos que contienen esta información están disponibles en el Anexo – D: Modelización.

4.7. Software utilizado

Para el desarrollo del trabajo se han empleado los siguientes programas:

1. **Importación de datos:** Se ha utilizado SAS Base junto con el procedimiento PROC IMPORT para cargar y combinar datos provenientes de distintas fuentes.

2. **Gestión y visualización:** Microsoft Excel se empleó para la gestión de datos y la creación de tablas.

3. **Manipulación y Modelado:** El procesamiento, análisis y modelado de datos se llevaron a cabo en R Studio.

A continuación, se describen algunas de las librerías más utilizadas de RStudio:

```
# Manipulación y Procesamiento de Datos:
```
`library(dplyr)` Funciones para seleccionar, filtrar, agrupar y resumir datos.
`library(readr)` Leer datos CSV.
`library(dummies)` Convertir las variables en dummy.
`library(naniar)` Manejar los valores ausentes.
`library(ggplot2)` Crear gráficos.

```
# Modelado:
```
`library(randomForest)` Modelos basados en árboles.
`library(caret)` Herramientas para crear modelos de *Machine Learning*.
`library(nnet)` Para crear redes neuronales.

```
# Computación paralela
```
`library(parallel)` Funciones para ejecutar operaciones en paralelo, aprovechando varios núcleos de CPU para acelerar cálculos.
`library(doParallel)` Complemento de parallel, para acelerar procesos de modelado.

CAPÍTULO 5. DESCRIPCIÓN DE VARIABLES

En este apartado, se estudiarán los dos conjuntos de datos de manera independiente. Los datos climáticos incluyen 365 observaciones por aeropuerto, lo que da un total de 1 460 registros entre los cuatro aeropuertos. Al combinarlos con el conjunto de datos final, el total asciende a 76 433 observaciones, lo cual no es eficiente para analizar específicamente las variables climáticas. Por ello, se opta por tratarlos por separado para obtener un análisis más claro y detallado.

5.1. Vuelos

La base de datos de los vuelos está compuesta por 12 variables recogidas en la Tabla 1.

Variable
Aeropuerto de origen
Aeropuerto destino
Fecha
Hora salida programada
Hora de salida real
Duración del vuelo real

Variable
Duración del vuelo programada
Demora en la salida. (Diferencia entre la hora de salida programada y la hora de salida real)
Código identificador de una aerolinea
Número de vuelo
Número de matrícula de una aeronave
Tiempo desde que el avión comiena a moverse hasta que despega

Tabla 1. Variables relativas a los vuelos

Estas variables se dividen en tres categorías: aquellas que se utilizarán directamente en el análisis, que son empleadas tal como están; aquellas que no se utilizarán en el análisis, porque no influyen en los resultados; y aquellas que se emplearán para crear nuevas variables derivadas, proporcionando información adicional para el estudio.

Variables utilizadas para crear la variable dependiente

La Tabla 2 resume las tres variables que se utilizarán para generar la variable dependiente, la cual se definirá en un apartado posterior. Estas variables son cuantitativas y se miden en minutos: *Duración del vuelo real*, *duración del vuelo programada* y *demora en la salida* .

Variable	Medida	Tipo de variable
Duración del vuelo real	Minutos	Cuantitativa discreta
Duración del vuelo programada	Minutos	Cuantitativa discreta
Demora en la salida. (Diferencia entre la hora de salida programada y la hora de salida real)	Minutos	Cuantitativa discreta

Tabla 2. Variables utilizadas para crear la variable objetivo

Variables utilizadas para crear otras nuevas

La Tabla 3 presenta las variables originales que se utilizarán para generar nuevas variables; estas nuevas variables y el tipo de variable que son.

Variable original	Descripción nueva variable	Tipo de variable
Fecha	Día del mes	Cuantitativa
Fecha	Día de la semana	Cualitativa
Fecha	Mes del año	Cuantitativa
Fecha	Estación del año	Cualitiativa
Fecha	Fin de de semana o no	Binaria
Nombre de la aerolínea	Si es una compañía *low cost* o no	Binaria
Hora salida programada	Hora del día programada	Cuantitativa

Tabla 3. Variables creadas a partir de las originales

A partir de la variable de fecha, se extraerán cinco variables distintas: el día del mes, el día de la semana, el mes del año y la estación del año. La variable original de fecha, por sí sola, no es directamente predictiva, por lo que se descompone en estas nuevas variables para mejorar el análisis.

De la variable que contiene el nombre de la aerolínea del vuelo se ha creado una variable binaria para indicar si se trata de una aerolínea de bajo coste o no. Se ha clasificado de la siguiente manera:

> **Aerolíneas de bajo coste:** Frontier Airlines Inc. (F9), Spirit Airlines (NK) y Southwest Airlines Co. (WN)
> **Aerolíneas no de bajo coste:** Delta Airlines Inc. (DL), United Airlines Inc. (UA), American Airlines Inc. (AA), America West Airlines Inc. (MQ) y SkyWest Airlines Inc. (OO)

Adicionalmente, de la hora de salida programada se ha extraído únicamente la hora, descartando los minutos, creando una variable con 24 categorías para simplificar el análisis.

Seguidamente, en la Tabla 4 se presentan las variables que se incluirán en el análisis sin ninguna modificación.

Variable	Tipo de variable
Aeropuerto de origen	Cualitativa nominal
Aeropuerto destino	Cualitativa nominal
El tiempo transcurrido entre la salida de la puerta del aeropuerto de origen y el inicio del viaje (min)	Cuantitativa discreta

Tabla 4. Variables incluidas sin ninguna modificación

Por otro lado, la Tabla 5 muestra las variables que se deben descartar debido a su carácter identificativo y su falta de utilidad en la modelización. Estas variables se emplean principalmente para la identificación de observaciones y no aportan información significativa al análisis.

Variable	Tipo de variable
Número de vuelo	Cualitativa nominal
Número de matrícula de una aeronave, es un código alfanumérico distintivo asignado a cada aeronave como identificador único	Cualitativa nominal
Hora salida real	Cuantitativa

Tabla 5. Variables descartadas

Entre ellas se encuentran el número de vuelo, que tiene 3 561 categorías diferentes, y el número de matrícula del avión, que cuenta con 1 290 categorías. Además, la hora de salida real también se excluye del análisis, ya que el objetivo es predecir el retraso en función de la hora de salida programada y la hora de llegada. Estas variables no contribuyen al proceso de modelización y, por tanto, se han decidido excluir del análisis.

5.2. Clima

Como se mencionó anteriormente, los datos climatológicos abarcan el promedio diario de las condiciones meteorológicas para cada aeropuerto, resultando en 365 observaciones por año para cada aeropuerto y un total de 1 460 observaciones distribuidas en 27 variables. Estos datos climáticos se han duplicado para incluir información tanto de los aeropuertos de origen como de destino.

Durante el proceso de filtrado, se eliminaron las variables relacionadas con el nombre de la estación y la fecha, ya que esta información ya estaba disponible en el conjunto de datos de los vuelos. Además, se descartaron las variables de latitud, longitud y elevación debido a su baja capacidad predictiva. La Tabla 6 muestra las variables climatológicas descartadas.

Variable
Fecha
Nombre de la estación climatológica
Latitud de la estación
Longitud de la estación
Elevación de la estación

Tabla 6. Variables climatológicas descartadas

El resto de las variables climatológicas no han sido transformadas. La Tabla 7 presenta las variables cuantitativas incluidas en el modelo (siete continuas y cuatro discretas).

Variable	Medida	Tipo de variable
Velocidad media diaria del viento	m/s	Cuantitativa continua
Precipitación	mm	Cuantitativa continua
Temperatura media	Grados Celsius	Cuantitativa continua
Temperatura máxima	Grados Celsius	Cuantitativa continua
Temperatura mínima	Grados Celsius	Cuantitativa continua
Velocidad del viento más rápida en 2 minutos	m/s	Cuantitativa continua
Velocidad del viento más rápida en 5 segundos	m/s	Cuantitativa continua
Nieve caída	mm	Cuantitativa discreta
Profundidad de la nieve	mm	Cuantitativa discreta
Direccion del viento más rapido en 2 minutos	Grados	Cuantitativa discreta
Direccion del viento más rapido en 5 segundos	Grados	Cuantitativa discreta

Tabla 7. Variables climatológicas cuantitativas incluidas en el modelo

Finalmente, la Tabla 8 presenta las diez variables binarias, donde un valor de 1 indica la presencia de un evento meteorológico y 0 su ausencia.

Variables climatológicas binarias	Variables climatológicas binarias
Niebla, niebla de hielo o niebla helada	Escarcha o cencellada
Niebla densa o niebla helada densa	Humo o neblina
Tormenta eléctrica (truenos)	Nieve soplada o arrastrada por el viento
Pellets de hielo, aguanieve, gránulos de nieve o granizo pequeño	Polvo, ceniza volcánica, polvo soplado, arena soplada, o algún obstáculo soplado
Granizo	Tornado, tromba marina, o nube embudo

Tabla 8. Variables climatológicas binarias incluidas en el modelo

La información completa sobre todas las variables se encuentra en el Anexo – A: Variables.

5.3. Variable objetivo

La variable objetivo se ha definido a partir de la diferencia entre la duración real del vuelo y la duración programada, sumando además los minutos de demora en la salida. Este cálculo da lugar a una nueva variable denominada *Retraso*.

Con base en esta variable, se establece la variable objetivo de la siguiente manera:

- Si el valor de *Retraso* es mayor o igual a 11 minutos, la variable objetivo toma el valor de 1.
- Si el valor de *Retraso* es menor a 11 minutos, la variable objetivo toma el valor de 0.

La fórmula utilizada para calcular la variable *Retraso* es:

$$Retraso = (Duración\ Real - Duración\ Programada) + Demora\ en\ la\ salida$$

Donde:

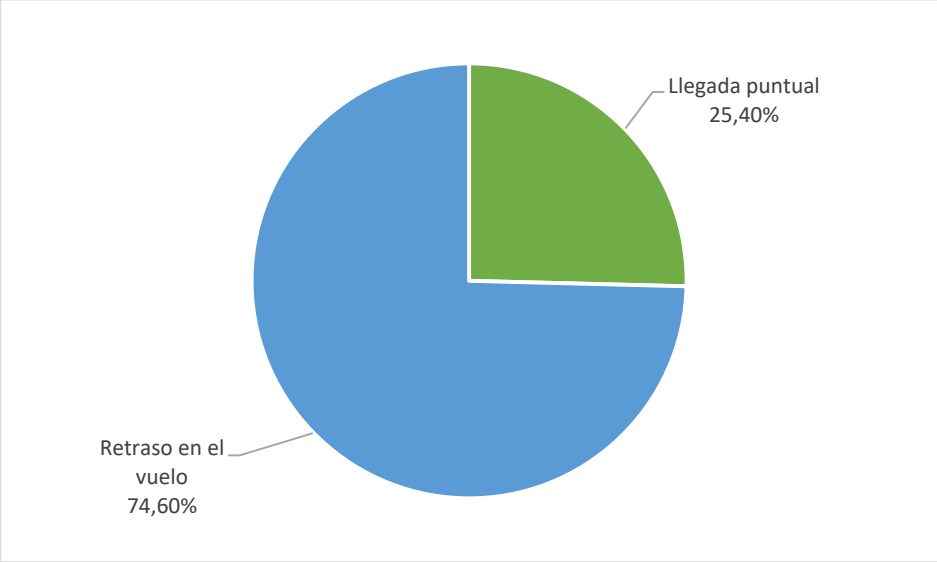
- **Duración real** es el tiempo total desde la salida hasta la llegada del vuelo.
- **Duración programada** es el tiempo estimado originalmente para el vuelo.
- **Demora en la salida** es el retraso acumulado en la salida del vuelo.

Finalmente, la variable objetivo se define como:

$$\text{Variable Objetivo} = \begin{cases} 1 & \text{si Retraso} \geq 11 \text{ minutos} \\ 0 & \text{si Retraso} < 11 \text{ minutos} \end{cases}$$

El Gráfico 1 muestra el porcentaje de eventos de la variable objetivo.

Gráfico 1. Proporción de eventos de la variable objetivo



En nuestro conjunto de datos, el 25.4% de los vuelos presentan retraso, mientras que el 74.6% de los vuelos son puntuales.

CAPÍTULO 6. ANÁLISIS EXPLORATORIO Y CORRECCIÓN DE ERRORES DETECTADOS

En este capítulo, se lleva a cabo un análisis exploratorio, nuevamente con los dos conjuntos de datos por separado para facilitar su comprensión. Se presentarán las principales estadísticas descriptivas, junto con pruebas de hipótesis y el análisis de la presencia de valores atípicos y datos faltantes. El objetivo es identificar posibles errores o inconsistencias que puedan influir en los resultados.

6.1. Detección de valores ausentes

Se identificaron un total de 904 valores ausentes en cuatro variables.

En primer lugar, se encontraron 367 valores ausentes en la variable *Nieve*. Estos datos faltantes correspondían a todos los días del año en Los Ángeles y a los días 27 y 29 de diciembre en Atlanta. Para abordar esta problemática, se consultó la fuente de datos Weather Spark, la cual confirmó que no hubo nevadas en Los Ángeles durante todo el año ni en Atlanta en las fechas mencionadas. Por lo tanto, estos valores ausentes se imputaron con 0.

Respecto a la variable *Profundidad de la nieve,* también se encontraron datos ausentes en todos los días del año en Los Ángeles y en los días 4, 7, 15, 26, 27, 29 y 30 de diciembre en Atlanta. Dado que la falta de nieve en Los Ángeles ya se había confirmado en la variable anterior y la misma fuente verificó que no hubo nevadas en Atlanta en las fechas indicadas, estos valores se imputaron con 0.

Finalmente, las variables *Dirección del viento más rápido en 5 segundos* y *Velocidad del viento más rápida en 5 segundos* presentaron valores ausentes para el 24 de enero en el aeropuerto de Denver. Para estos casos, se imputaron los datos faltantes utilizando la media de los valores del día anterior y del día posterior.

6.2. Test Chi - Cuadrado

La prueba de Chi-Cuadrado permite evaluar la independencia entre dos variables categóricas. En este análisis, se ha aplicado a todas las variables creadas a partir de la variable original *Fecha* (día de la semana, fin de semana, quincena del mes, decena del mes, mes, estación del año) y de la *hora de despegue* (hora y momento del día) con el objetivo de identificar cuáles están más asociadas con la variable objetivo y cuáles podrían ser redundantes, ayudando así a evitar problemas de multicolinealidad.

La Tabla 9 muestra el test Chi-Cuadrado de las variables creadas a partir de la hora.

Variable	Estadístico chi cuadrado	Grados de libertad	p-valor	Chi cuadrado /df
Momento del día	1.789,10	3	< 2,2e-16	596,37
Hora del día	4.041,80	23	< 2,2e-16	175,73

Tabla 9. Resultados test Chi-Cuadrado de las variables relacionadas a partir de la variable hora

Aunque la *hora del día* muestra una relación más fuerte con la variable objetivo, reflejada en un alto estadístico chi-cuadrado, la relación Chi-cuadrado/(grados de libertad) es más alta para el *momento del día* (596.37 frente a 175.73), lo que señala que el *momento del día* ofrece una relación más eficiente entre el estadístico chi-cuadrado y los grados de libertad, dando lugar a una mejor capacidad de discriminación con menos complejidad.

Por otro lado, el *momento del día* agrupa las horas en intervalos más amplios, simplificando la interpretación del modelo y reduciendo el riesgo de sobreajuste, especialmente en aquellos modelos que utilizan codificación de variables *dummy*.

Por lo tanto, pese que la variable *momento del día* puede perder algunos detalles finos, captura las principales variaciones del día con mayor claridad y de manera más manejable.

Al analizar las variables derivadas del *día del mes*, como *quincena del mes*, *decena del mes* y *día del mes*, se observan diferencias importantes en la eficacia de cada variable, que se recogen en la Tabla 10.

Variable	Estadístico chi cuadrado	Grados de libertad	p-valor	Chi cuadrado /df
Quincena del mes	41.535	1	< 2,2e-16	41.535
Decena	72.092	2	< 2,2e-16	36.046
Día del mes	223	30	< 2,2e-16	7

Tabla 10. Resultados test Chi Cuadrado de las variables relacionadas a partir de la variable día

La *quincena del mes* presenta una asociación fuerte con la variable dependiente pues se observa un estadístico Chi-cuadrado de 41 535. Además, esta variable tiene solo un grado de libertad, lo que implica una menor complejidad en el modelo.

En comparación, la *decena del mes* muestra un estadístico Chi-Cuadrado de 72, lo que también sugiere una asociación altamente significativa con la variable dependiente. Sin embargo, la decena del mes tiene 2 grados de libertad, lo que introduce una mayor complejidad en el modelo en comparación con la quincena del mes.

La variable *día del mes* muestra una menor estadístico Chi-cuadrado; de 223.4 y su relación Chi-cuadrado/df es considerablemente menor en comparación con las otras variables.

Finalmente, nos quedamos con la *quincena del mes* como la opción preferible debido a su simplicidad y asociación significativa, ya que ofrece una buena relación entre efectividad y manejabilidad en el modelo.

Lo mismo ocurre con las variables creadas a partir del mes del año, como se muestra en la Tabla 11.

Variable	Estadistico chi cuadrado	Grados de libertad	p-valor	Chi cuadrado /df
Mes del año	1.780,80	11	< 2,2e-16	161,89
Estación del año	1.220,50	3	< 2,2e-16	406,83

Tabla 11. Resultados test Chi Cuadrado de las variables relacionadas a partir de la variable año

Aunque el *mes del año* tiene un estadístico Chi-Cuadrado más alto, su complejidad y el gran número de niveles lo hacen menos práctico. En cambio, la *estación del año*, con un menor estadístico Chi-cuadrado pero mejor ratio Chi-cuadrado/df, ofrece una simplificación que reduce la complejidad del modelo y el riesgo de sobreajuste, resultando en una opción más manejable.

Por último, la Tabla 12 muestra los resultados de las variables derivadas del día.

Variable	Estadistico chi cuadrado	Grados de libertad	p-valor	Chi cuadrado /df
Día de la semana	227,43	6	< 2,2e-16	37,91
Fin de semana	17.351	1	0,19	17.351

Tabla 12. Resultados test Chi-Cuadrado de las variables relacionadas a partir de la variable día

El p-valor para la variable *fin de semana* es 0.19, lo que indica que su asociación con la variable dependiente no es estadísticamente significativa, ya que es mayor que el umbral de 0.05. En contraste, el *día de la semana* presenta un p-valor significativamente bajo, indicando una asociación estadísticamente significativa con la variable dependiente. Por lo tanto, se prefiere *el día de la semana* por su clara y relevante asociación con la variable objetivo.

6.2.1. V de Cramér

Seguidamente se calcula la V de Cramér, que mide la asociación entre dos variables categóricas y se basa en el estadístico Chi-Cuadrado. Se calcula a través de la fórmula:

$$V = \sqrt{\frac{\chi^2/_n}{\min\,(k-1, r-1)}}$$

Donde, χ^2 es el estadístico Chi-cuadrado obtenido en el test, n el número total de muestra, k es el número de categorías de una de las variables, r el número de categorías de otra variable y $\min\,(k-1, r-1)$, es el menor de los grados de libertad de las dos variables

Su valor varía entre 0 y 1:

0 indica que no hay asociación entre las variables.

1 indica una asociación perfecta.

La V de Cramér es útil para cuantificar la fuerza de la asociación en tablas de contingencia, especialmente cuando las variables tienen más de dos categorías. Ajusta el estadístico Chi-cuadrado por el tamaño de la tabla de contingencia, proporcionando una medida estandarizada de la fuerza de la asociación y facilitando la interpretación práctica de los resultados del test Chi-cuadrado.

Los resultados se recogen en la Tabla 13.

V de Cramér	
Mes del año	0.1526374
Momento del día	0.152993
Hora del día	0.2299561
Día del mes	0.0540626
Decena del mes	0.03071162
Quincena del mes	0.02331137
Estación del año	0.1263645
Día semana	0.05454852
Fin de semana	0.004764568

Tabla 13. Resultados V de Cramér

La *hora del día* es la variable con la asociación más significativa con la variable categórica, mostrando un valor de 0.23. Por otro lado, variables como el *mes del año*, el *momento del día* y la *estación del año* también presentan asociaciones con la variable categórica, aunque con valores más bajos, en torno a 0.15.

A modo de resumen se presenta la Tabla 14, que muestra las variables que se han decidido mantener en el modelo y aquellas que se han descartado.

Variables incluidas
Momento día
Quincena del mes
Estación del año
Día de la semana

Variables descartadas
Hora del día
Decena del mes
Día del mes
Mes del año
Fin de semana

Tabla 14. Variables descartadas e incluidas en el modelo

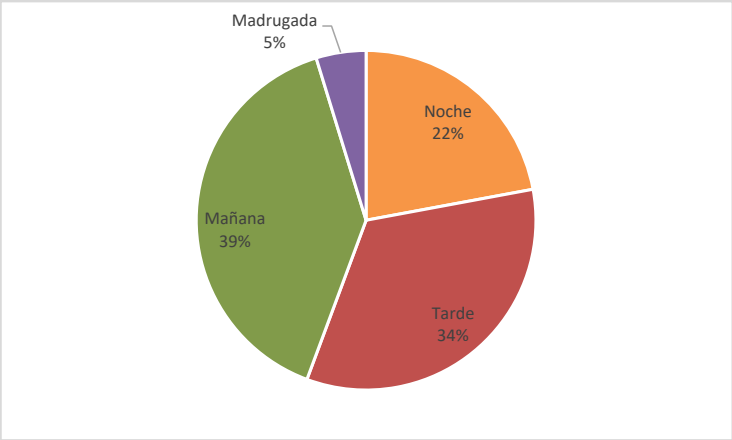
6.3. Análisis descriptivo

A continuación, se presenta un análisis descriptivo de las variables, destacando los gráficos y tablas más relevantes para entender los datos, antes de pasar a su modelización.

Número de vuelos según el momento del día

En el Gráfico 2 se observa el porcentaje de vuelos en función del *momento del día*.

Gráfico 2. Número de vuelos según el momento del día



Por la mañana (6:00-11:59), se observa que es el período de mayor actividad, con un 39% de los vuelos registrados en este tramo. Durante la tarde (12:00-17:59), aunque la actividad disminuye ligeramente en comparación con la mañana, sigue siendo un período de alta actividad aérea, con un 34% de los vuelos.

En el transcurso de la noche (18:00-23:59), la frecuencia de vuelos disminuye en comparación con las franjas matutina y vespertina, con solo un 22% de los vuelos registrados. Durante la madrugada (00:00-5:59), se registra la menor cantidad de vuelos de 2023, representando solo un 5% lo que refleja la reducción en la demanda y de la oferta.

Número de vuelos por estación del año

La Tabla 15 muestra el número de vuelos según la estación del año.

ESTACIÓN	N DE VUELOS
Invierno	17.883
Otoño	19.047
Primavera	19.429
Verano	20.074

Tabla 15. Número de vuelos según la estación del año

Se observa que el *verano* es la estación con el mayor número de vuelos, seguida de cerca por *primavera*, *otoño*, e *invierno*, que tienen menos vuelos en comparación. El aumento gradual en el número de vuelos desde el invierno hasta el verano refleja una mayor demanda de viajes durante los meses más cálidos y las vacaciones.

Número de vuelos con retraso en función del tipo de compañía

La Tabla 16 confirma que las aerolíneas de bajo coste enfrentan más problemas de puntualidad en comparación con las aerolíneas tradicionales.

Tipo de aerolinea	Nº de vuelos	Nº de vuelos con retraso	% de retraso
Low cost	19.243	13.351	69,38%
Normal	57.190	6.083	10,64%

Tabla 16. Número de vuelos con retraso según el tipo de compañía

Las compañías *low cost* presentan un porcentaje de retraso del 69.38%, frente al 10.64% de las aerolíneas no de bajo coste, a pesar de las condiciones climáticas. Aunque la variable considerada no sea climática, se incluirá en el análisis para proporcionar una visión más completa de los factores que afectan la puntualidad.

Diferencia entre la duración real y programada

Hoy en día, es común que un vuelo tenga una duración programada y que, a pesar de posibles retrasos en el despegue, se llegue a destino antes del tiempo previsto. Esta estrategia permite a las aerolíneas asegurar que sus vuelos lleguen puntuales o incluso antes de lo anunciado, salvo excepciones.

Para evaluar esta práctica, se ha calculado la diferencia entre la duración programada y la real de los vuelos. Los resultados de la Tabla 17 muestran que, en promedio, los vuelos duran 9.1 minutos menos de lo programado, lo que indica que las aerolíneas tienden a sobreestimar la duración de sus vuelos. Aunque la mayoría de las diferencias están cerca de este promedio, existen algunos casos con desviaciones notables. Esto confirma que, en general, los vuelos tienden a ser más cortos de lo previsto.

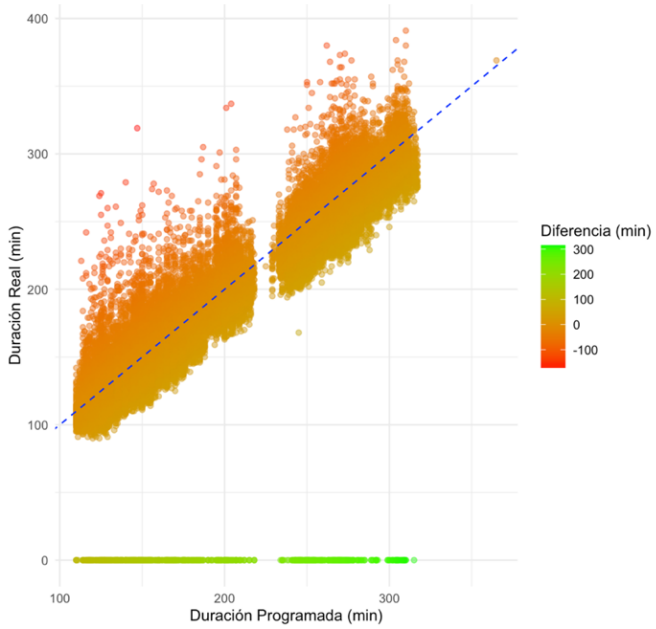
Variable	Mínimo	Primer cuartil	Mediana	Media	Tercer cuartil
Diferencia entre duracion programada y duracion real	-172	0	9	9,108	16

Tabla 17. Diferencia entre la duración programada y real de los vuelos

Si nos fijamos en el tercer cuartil (el valor que separa el 75% de los datos más bajos del 25% de los datos más altos) esta diferencia es de 16 minutos. El 25% superior de los vuelos llegó 16 minutos o más antes del tiempo programado, mostrado que, para la mayoría de los vuelos, la diferencia entre la duración real y programada es positiva.

Estos resultados se presentan de manera complementaria en el Gráfico 17.

Gráfico 3. Diferencia entre la duración programada y real de los vuelos



La línea discontinua azul en el gráfico representa la igualdad entre la duración programada y la real. Los puntos ubicados sobre esta línea corresponden a vuelos cuya duración real coincidió exactamente con la programada.

Los puntos en verde situados en la parte inferior del gráfico representan vuelos que fueron cancelados, mientras que los puntos en la parte superior indican vuelos que experimentaron mayores retrasos.

Los vuelos que se encuentran por debajo de la línea azul son aquellos que llegaron antes de lo previsto. La presencia de vuelos con duraciones reales menores a las programadas confirma que las aerolíneas tienden a programar tiempos ligeramente más largos como estrategia para mejorar su puntualidad.

Además, la variabilidad en la duración real, particularmente en los vuelos más largos, indica las dificultades de las aerolíneas para predecir con precisión las duraciones de vuelos más complejos, lo que podría estar relacionado con factores como el tráfico aéreo, condiciones meteorológicas y otros imprevistos.

Temperaturas máximas por aeropuerto

Los cuatro aeropuertos presentan valores máximos de temperatura similares según se observa en la Tabla 18. Sin embargo, las temperaturas mínimas muestran una mayor variación.

Temperaturas (ºC)			
Aeropuerto	**Máxima**	**Mínima**	**Media diaria**
Atlanta	37,2	-2,1	18,5
Chicago	37,8	-18,2	12,25
Denver	37,2	-23,8	10,42
Los Ángeles	32,8	5	16,59

Tabla 18. Resumen de las temperaturas según el aeropuerto

La temperatura media diaria varía en aproximadamente 2 grados entre los aeropuertos, con la temperatura media más baja en Denver (10.42°C) y la más alta en Atlanta (18.5°C).

Chicago y Denver experimentan temperaturas mínimas muy bajas, -18.2°C y -23.8°C respectivamente. Por otro lado, Los Ángeles presenta un clima más moderado, con temperaturas menos extremas tanto en máximas (32.8°C) como en mínimas (5°C). Su temperatura media diaria es más alta que la de Chicago y Denver, aunque algo inferior a la de Atlanta.

Eventos climatológicos por aeropuerto

La Tabla 19, muestra la distribución porcentual de los distintos fenómenos meteorológicos por aeropuerto.

AEROPUERTO	ATLANTA	CHICAGO	DENVER	LOS ÁNGELES
Niebla (%)	34.8	38.4	29.3	42.8
Niebla densa (%)	2.2	2.7	7.9	6.6
Tormenta eléctrica (%)	16.4	10.1	20.0	1.9
Granizo pequeño (%)	0	1.9	0	0
Granizo intenso (%)	0	2.2	1.4	0
Escarcha o cencellada (%)	0	1.4	0.6	0
Polvo, ceniza volcánica, polvo, arena o algún objeto soplado (%)	0	17.5	10.4	22.8
Humo o neblina (%)	0	0.5	0.6	0
Nieve soplada (%)	0	0	0.3	0
Tornado, tromba marina (%)	0	0	0.3	0

Tabla 19. Porcentaje de eventos climatológicos por aeropuerto

La *niebla* es el evento meteorológico más frecuente en todos los aeropuertos analizados, con una incidencia que varía entre el 29.3% y el 42.8%. En comparación, la *niebla densa* ocurre con menor frecuencia, con un rango del 2.2% al 7.9%, siendo más común en Denver. Los *truenos* (tormenta eléctrica) se presentan principalmente en Denver, donde ocurren en el 20% de los casos, mientras que en Los Ángeles su frecuencia es mucho menor, con solo un 1.9%. El *granizo* es un evento raro, registrado únicamente en Chicago con una frecuencia del 2.2%, mientras que *granizo intenso* se presencia en Chicago (2.2%) y Denver (1.4%).

Por otro lado, el *humo o neblina* se da en Los Ángeles con una frecuencia del 0.5%, seguido por Chicago con un 0.6%. La *nieve soplada* y los *tornados* solo se presentan en Denver, con una frecuencia del 0.3%.

Precipitaciones por aeropuerto

La Tabla 20 ofrece un resumen de las estadísticas de precipitaciones para cada ciudad. En general, se observa que en la mayoría de los días no se registran lluvias.

Precipaciones por aeropuerto (mm)						
Aeropuerto	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
Atlanta	0	0	0	2,858	0,5	51,8
Chicago	0	0	0	2,349	1,3	85,1
Denver	0	0	0	1,321	0,3	74,2
Los Ángeles	0	0	0	1,765	0	51,8

Tabla 20. Precipitaciones por aeropuerto

Chicago destaca por tener las precipitaciones más intensas, con un máximo de 85.1 mm, aunque su promedio es de 2.3 mm, lo que indica lluvias intensas, pero poco frecuentes. Atlanta, con una máxima de 51.8 mm, tiene una mayor regularidad en las lluvias, pues presenta el promedio de precipitaciones más alto con 2.82 mm. Los Ángeles muestra una máxima de 51.8 mm similar a la de Atlanta, pero su promedio es más bajo, con 1.71 mm, indicando que, aunque las lluvias pueden ser fuertes, son menos frecuentes.

Por otro lado, Denver tiene el promedio de precipitaciones más bajo con 1.33 mm, aunque también puede experimentar lluvias intensas, con un máximo de 74.2 mm. En todas las ciudades, los valores mínimos de 0 mm reflejan que hay días sin lluvia, lo cual es común en todas ellas.

En resumen, mientras que Chicago y Denver pueden experimentar lluvias más intensas, Atlanta tiene una mayor regularidad en las precipitaciones. Pese a que Atlanta y Chicago tienen la misma cantidad máxima de lluvia (51.8 mm), estas se produjeron en fechas diferentes: el 24 de febrero en Atlanta y el 22 de enero en Los Ángeles.

Nieve por aeropuerto

Al observar los datos de nieve de la Tabla 21, de nuevo, se observan diferencias entre los aeropuertos.

Nieve por aeropuerto (mm)						
Aeropuerto	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
Chicago	0	0	0	1,378	1,3	91
Denver	0	0	0	2,559	0	135

Tabla 21. Nieve por aeropuerto

Chicago y Denver son las ciudades donde se registran cantidades significativas de nieve. Denver destaca con un máximo de 135 mm de nieve y un promedio de 2.6 mm, reflejando que las nevadas son más intensas y frecuentes en esta ciudad en comparación con Chicago, que tiene un máximo de 91 mm y un promedio de 1.4 mm.

Por otro lado, tanto Atlanta como Los Ángeles no presentan acumulaciones de nieve, y por lo tanto se han omitido en la tabla, con valores mínimos, máximos y promedios de 0 mm.

6.4. Detección y tratamiento de datos atípicos

Un valor atípico es un dato en una variable que se distingue significativamente de los demás valores en un conjunto de datos. Estos valores se consideran excepciones porque se comportan de manera diferente en comparación con el comportamiento promedio

de las demás observaciones. La presencia de valores atípicos puede ser problemática, ya que pueden sesgar o alterar los resultados de los análisis en los que se incluye la variable que contiene dichos valores.

Primero se analizarán los valores atípicos de forma univariante. Se considera un valor atípico (Figura 8) a cualquier observación que se encuentre fuera del rango de 1.5 veces el Rango Intercuartílico (IQR), que es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1).

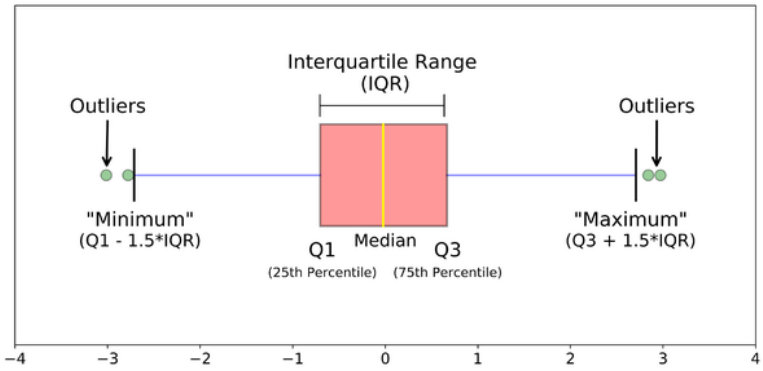


Figura 8. Descripción de datos atípicos

(Google, 2024)

6.4.1. Enfoque univariante

A continuación, se presenta un resumen del número de datos atípicos encontrados en las variables climatológicas, recogidos en la Tabla 22, así como su proporción en la muestra total.

Variable	Nº ATÍPICOS	% MUESTRA ATÍPICOS	MIN ATÍPICO	MAX ATÍPICO
Precipitación	309	21.16%	0.8	85.1
Nieve	46	3.15%	3	135
Profundidad de la nieve	69	4.73%	30	150
Temperatura media	11	0.75%	-20.1	-8.6
Temperatura mínima	8	0.55%	-23.8	-18.2
Temperatura máxima	18	1.23%	-15.5	-2.7
Velocidad media diaria del viento	46	3.15%	6.8	11.7
Velocidad del viento más rápida en 2 min	6	0.41%	14.8	21.5
Velocidad del viento más rápida en 5 sec	4	0.27%	20.6	32.2

Tabla 22. Número de atípicos por variable

En términos generales, las mediciones extremas de precipitación son bastante frecuentes, ya que la variable *Precipitación* muestra la mayor proporción de datos atípicos, alcanzando un 21.16% de la muestra.

La variable *Profundidad de la nieve* también muestra una alta proporción de atípicos y extremos, representando un 4.73%. En contraste, la variable *Nieve* presenta una menor cantidad de datos atípicos, con 46 observaciones en total. Además, se observa que los eventos de temperaturas altas extremas son más frecuentes, pues hay una mayor

cantidad de datos atípicos para las temperaturas máximas en comparación con las temperaturas mínimas.

Por otro lado, las variables relacionadas con la velocidad del viento, como la *velocidad del viento más rápida en 2 minutos* y *en 5 segundos*, tienen una menor incidencia de valores atípicos.

6.4.2. Enfoque bivariante

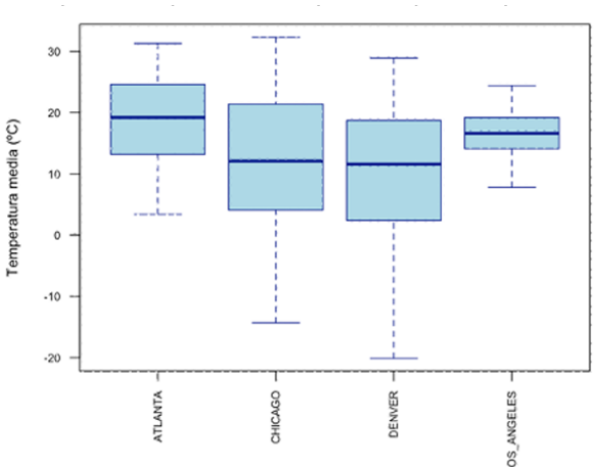
A continuación, se realiza un análisis conjunto de los atípicos. En este enfoque, observamos que el número de atípicos para algunas variables cambia al considerar su relación con otras variables. Es posible que los atípicos identificados en un análisis univariante se modifiquen, se mantengan o incluso aumenten al incorporar estas nuevas dimensiones.

El análisis bivariante, que compara una variable continua con una variable categórica, nos permite visualizar cómo varía una variable en función de otra. Para ello, utilizamos diagramas de cajas que facilitan la identificación de atípicos dentro de cada categoría de la variable categórica. En esta sección se presentan los resultados más interesantes, el resto del análisis se puede consultar en el Anexo – B: Tabla y Gráficos.

Temperatura media

Cuando se analiza la *temperatura media* de forma univariante, se identifican 11 valores atípicos debido a temperaturas que se desvían significativamente de la distribución general por condiciones meteorológicas inusuales. Sin embargo, en el análisis bivariante por aeropuerto, estos atípicos desaparecen, como muestra el Gráfico 4.

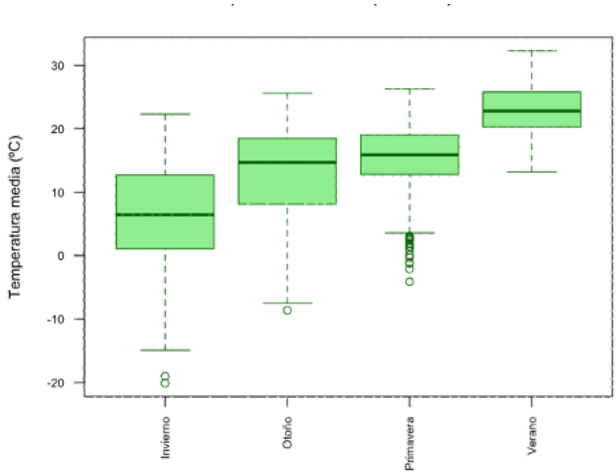
Gráfico 4. Boxplot de la temperatura por aeropuerto



Esto puede ocurrir porque la temperatura media es relativamente homogénea en cada aeropuerto o porque el efecto de agrupación diluye la influencia de los valores extremos.

En contraste, al analizar la temperatura media por estación del año, representado en el Gráfico 5, se encuentran 20 atípicos.

Gráfico 5. Boxplot de la temperatura por estación del año



Este aumento se debe a la variabilidad estacional, donde ciertas estaciones presentan más desviaciones debido a fenómenos climáticos específicos que no se capturan en el análisis univariante o por aeropuerto.

Temperatura mínima

Cuando se analizaba la temperatura mínima, se detectaban 11 atípicos. Sin embargo, al estudiarla conjuntamente con los aeropuertos, estos desaparecen, pero al estudiarla por estación del año aumentan hasta 20. Dicha información queda recogida en la Tabla 23.

Atípicos de la temperatura media(ºC)		
En la variable	**Por aeropuerto**	**Por estación**
11	0	20

Tabla 23. Atípicos en las temperatura mínima por estación

En este caso, los valores atípicos se han reducido considerablemente. Se ha comprobado que los atípicos de *Temperatura media* corresponden a temperaturas inusualmente bajas, que ocurren principalmente durante el invierno en Denver.

Velocidad del viento más rápida en 5 segundos

En el análisis univariante de la *velocidad del viento más rápida en 5 segundos*, se identificaron 4 valores atípicos, recogidos en la Tabla 24.

Atípicos en la velocidad del viento más rápida en 5 sec			
Atlanta	**Chicago**	**Denver**	**Los Ángeles**
10	7	1	20

Tabla 24. Atípicos en la Velocidad del viento más rápida en 5 segundos por aeropuerto

Cuando se analiza la *velocidad del viento más rápida en 5 segundos* en función de los diferentes aeropuertos, se identifican 38 atípicos. Esto indica que, al considerar la variable *aeropuerto*, hay una mayor cantidad de valores que se desvían significativamente de los valores típicos para cada aeropuerto en particular.

En el análisis bivariante por estación del año, se identifican 50 atípicos, como se muestra en la Tabla 25.

Atípicos en la velocidad del viento más rápida en 5 sec			
Invierno	Otoño	Primavera	Verano
8	10	9	23

Tabla 25. Atípicos en la Velocidad del viento más rápida en 5 segundos por estación

Este aumento en el número de atípicos comparado con el análisis univariante indica que, al considerar la estación del año, se detectan más valores que se desvían de lo esperado, lo que puede indicar que las condiciones extremas del viento son más prevalentes en ciertas estaciones.

Una vez concluido el análisis conjunto de variables continuas con categóricas, procedemos a examinar la relación entre dos variables continuas. Para ello, se ha elaborado una matriz de correlaciones, que se presenta en la Figura 9 con las variables correspondientes al origen y destino. Esta matriz muestra cómo se relacionan las variables continuas entre sí, proporcionando una visión cuantitativa de la fuerza y la dirección de sus asociaciones lineales. Los valores en la matriz varían entre -1 y 1, donde:

- 1 indica una correlación positiva perfecta.
- -1 indica una correlación negativa perfecta.
- 0 indica ninguna correlación lineal.

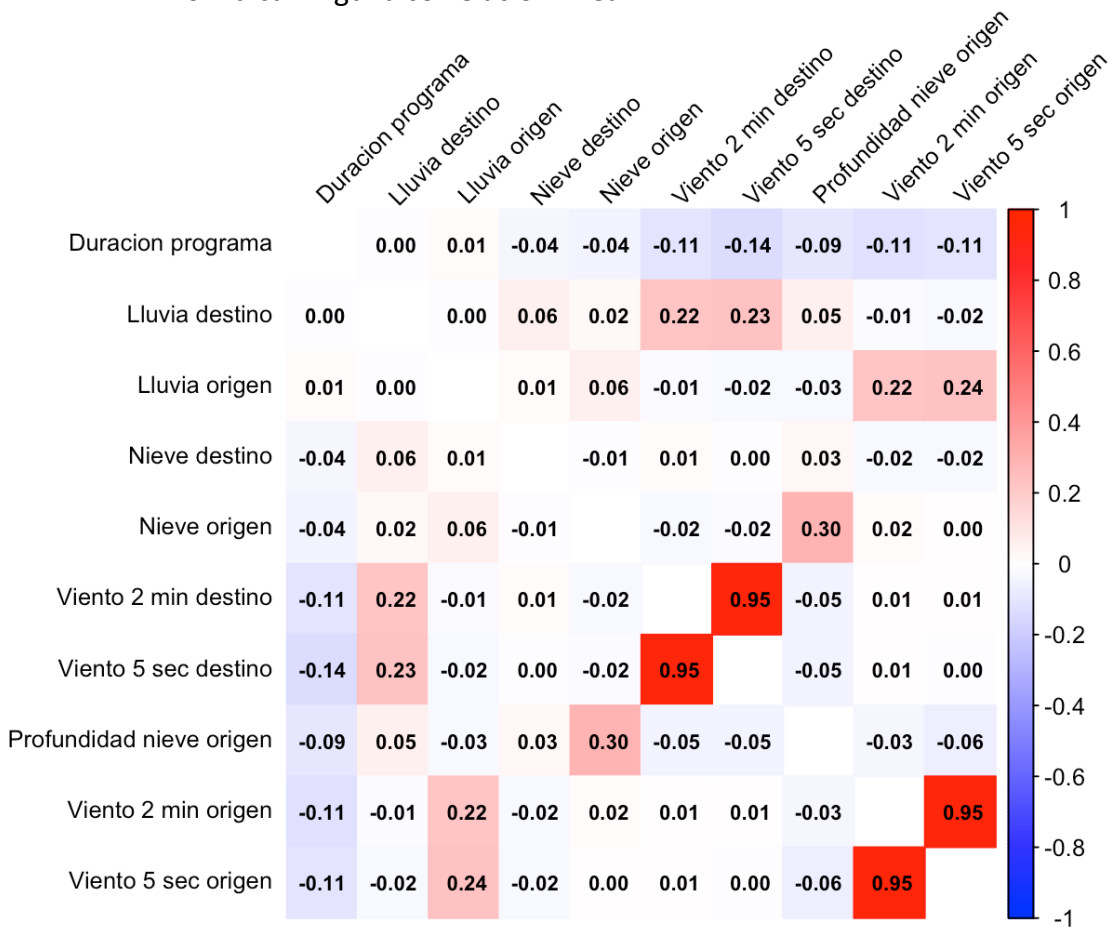


Figura 9. Matriz de correlaciones entre variables continuas

De la matriz de correlaciones, se observa que la *velocidad máxima del viento en 2 minutos* y *en 5 segundos* en el destino está correlacionada con la cantidad de lluvia en el destino con coeficientes de 0.22 y 0.23, respectivamente. Esto indica que rachas de viento más intensas tienden a estar asociadas con mayores precipitaciones. De manera similar, la *velocidad máxima del viento en 2 minutos* y en *5 segundos* en el origen muestra correlaciones con la lluvia en el origen de 0.22 y 0.24, respectivamente.

Además, existe una correlación positiva entre la profundidad de la nieve y la cantidad de nieve, lo cual es esperable.

Por otro lado, la velocidad del viento muestra una relación negativa con la duración programada, reflejando que a medida que aumenta la velocidad del viento, la duración programada tiende a disminuir.

Para una exploración más detallada de estas relaciones, los gráficos de dispersión de todas las combinaciones de las variables continuas están disponibles en el Anexo – B: Tablas y Gráficos.

Para concluir el análisis de datos atípicos, se deciden mantener estos valores, dado que el objetivo principal de este estudio es evaluar el impacto del clima en los retrasos de los aviones, y que en muchos casos el porcentaje de atípicos en los datos supera el 2.5 convirtiéndolos así en *datos típicos*, se ha decidido incluir estas observaciones extremas. La inclusión de estos datos, que representan condiciones climatológicas extremas, permitirá obtener una visión más completa y precisa del efecto del clima en los retrasos.

CAPÍTULO 7. Selección de variables

Seleccionar las variables adecuadas antes de modelizar es un paso necesario para lograr un equilibrio óptimo entre la complejidad del modelo y su capacidad de ajuste a los datos. Una selección de variables eficaz ayuda a prevenir problemas como el sobreajuste y el subajuste, garantizando que el modelo se ajuste bien a los datos de entrenamiento y también generalice de manera efectiva a nuevos datos.

En esta sección, se presentan los cuatro métodos utilizados para la selección de variables:

Método *Stepwise*. Este enfoque iterativo busca el modelo óptimo añadiendo y eliminando variables en función de su impacto en el criterio de selección. Los criterios de selección utilizados son el Criterio de Información de Akaike (AIC) que se enfoca en equilibrar la bondad de ajuste del modelo con la simplicidad, penalizando ligeramente la inclusión de más parámetros, lo que puede llevar a elegir modelos más complejos cuando se prioriza un buen ajuste. El Criterio de Información Bayesiano (BIC), en cambio, introduce una penalización más fuerte basada en el tamaño de la muestra, favoreciendo modelos más simples y evitando el sobreajuste. AIC es preferible para predicciones en nuevos datos, mientras que BIC es mejor para identificar el modelo más probable y evitar la sobreparametrización (Hastie et al., 2009).

Método Boruta. Utiliza un modelo de *Random Forest* para identificar todas las características relevantes en un conjunto de datos. Genera características aleatorias, conocidas como *shadow features*, para evaluar la importancia de las variables originales. Si una variable es significativamente más importante que sus copias aleatorias, se clasifica como relevante. Este proceso se repite hasta que todas las características son clasificadas como importantes o no importantes, asegurando que solo se seleccionen las características útiles (SpringerLink, 2024).

Método MMPC (Max-Min Parent Child). Este método identifica un subconjunto mínimo de variables que tienen el mayor impacto en la variable dependiente. Comienza evaluando la relevancia de todas las variables mediante pruebas de independencia condicional y selecciona las más informativas, evitando redundancias. El resultado es un conjunto optimizado de variables que mejora la precisión y eficacia del modelo (Tsamardinos et al., 2006).

Método SES (*Subset Evaluation and Selection*). Identifica las variables más relevantes para una variable dependiente mediante la evaluación de dependencias condicionales y selecciona un subconjunto óptimo de variables que son las más informativas para predecir o explicar la variable dependiente.

Método RFE (Selección Recursiva de Características). Es una técnica iterativa que selecciona las características más relevantes para un modelo eliminando gradualmente las menos importantes. Se ha utilizado con criterio *Random Forest* y *Naive Bayes*. Cuando se utiliza con Random *Forest*, se basa en la medida de importancia de las características proporcionada por el modelo, mientras que con *Naive Bayes*, se enfoca en evaluar cómo la eliminación de características individuales afecta el rendimiento global del modelo.

La Tabla 26 presenta el número de variables seleccionadas en común por diferentes combinaciones de métodos y, además, el número total de variables que han sido seleccionadas por cada uno de los métodos de selección.

La información completa sobre todas las variables que ha seleccionado cada método está disponible en el Anexo – C: Selección de variables.

Nº de variables seleccionado por cada método	
Todos los métodos (7)	4
6 métodos	10
5 métodos	4
4 métodos	7
3 métodos	5
2 métodos	9
1 método	8
Ningún método	3

Nº de variables seleccionado por cada método	
Boruta RF	43
Stepwise AIC	35
RFE NB	27
Stepwise BIC	23
RFE RF	7
MXM	21
SES	21

Tabla 26. Número de variables seleccionadas en común y por cada método

En la Figura 10 se presenta un Diagrama de Venn, una herramienta visual que muestra las intersecciones entre diferentes conjuntos, comparando los métodos de selección de

variables dos a dos. Cabe destacar que los métodos MXM y SES han seleccionado las mismas variables, por lo que se representan conjuntamente en el diagrama.

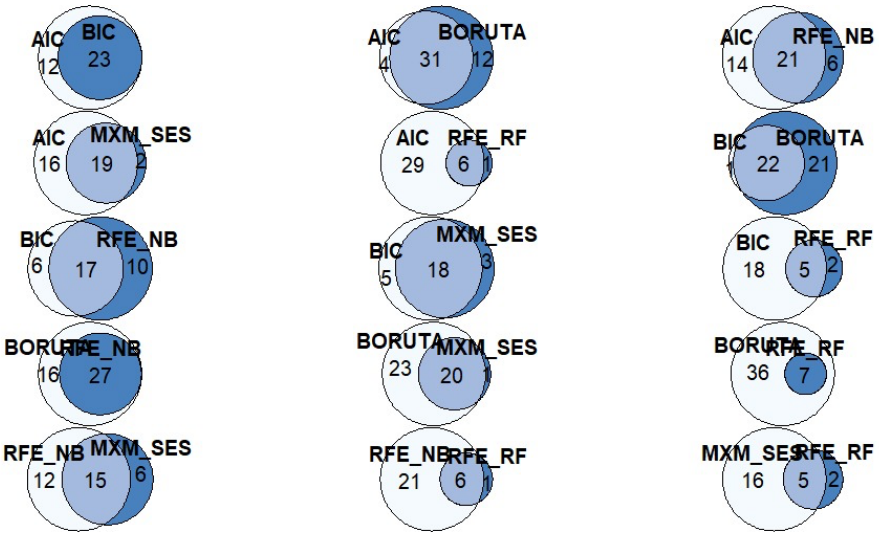


Figura 10. Diagrama de Ven

Al observar los resultados, se aprecia que el método *stepwise* con *AIC* seleccionó las mismas 23 variables que *BIC*, además de 12 variables adicionales. De manera similar, *Boruta* y *RFE* con criterio *NB* muestran coincidencias: *Boruta* selecciona las mismas 27 variables que *NB*, pero incluye 16 variables adicionales. Del mismo modo, cuando se compara *Boruta* con *RFE* utilizando *RF* como criterio, se observa que ambos métodos comparten 7 variables presentes en *RF*, con *Boruta* seleccionando 36 variables adicionales. Por otro lado, los métodos que menos variables tienen en común son aquellos que incluyen RFE con RF, ya que este último solo selecciona 7 variables.

La selección de variables concluye con la elección de aquellas que han sido seleccionadas por al menos 4 de los métodos, resultando en un total de 25 variables y recogidas en la Tabla 27.

Variables seleccionadas	Aeropuerto
Aeropuerto	Origen + Destino
Duración programada del vuelo	N.A
Momento del día del vuelo	N.A
Día de la semana	N.A
Quincena del mes	N.A
Estación del año	N.A
Compañía low cost	N.A
Precipitaciones	Origen + Destino
Niebla	Origen + Destino
Tormenta eléctrica	Origen + Destino
Nieve	Origen + Destino
Profundidad de la nieve	Origen
Velocidad del viento más rápida en 5 sec	Origen + Destino
Escarcha o cencellada	Origen + Destino
Humo o neblina	Origen + Destino
Velocidad del viento más rápida en 2 min	Origen + Destino

Tabla 27. Variables utilizadas para modelizar.

Se observa que se ha reducido el número de variables originales a casi la mitad. Las variables descartadas se encuentran en el Anexo – C: Selección de variables.

CAPÍTULO 8. MODELIZACIÓN

En este apartado se presentan, tras haber realizado el estudio de los métodos de modelización, el análisis descriptivo y la selección de variables, los principales resultados obtenidos durante el proceso de modelización.

8.1. Regresión logística

En esta sección se presentan los resultados del modelo de regresión logística separados en función de variables no significativas, variables significativas de los vuelos y variables significativas del clima.

Variables no significativas

Las variables con un valor p superior a 0.05, o que son marginalmente significativas ($p < 0.1$), no se consideran estadísticamente significativas al nivel de significancia del 5%. Se identificaron nueve variables que no son significativas y cuatro variables con significancia marginal. Todas estas variables se presentan en la Tabla 28.

Variables no significativas					
Variable	**Estimate**	**Std, Error**	**z value**	**Pr(>z)**	**OR**
Apto destino Los Ángeles	-0,0578604	0,0369295	-1,567	0,12	0,9438
Día semana: lunes	-0,0379238	0,0387955	-0,978	0,33	0,9628
Día semana: miércoles	-0,0185342	0,389623	-0,476	0,63	0,9816
Día semana: sábado	-0,0129864	0,039962	-0,325	0,75	0,9871
Día semana: viernes	0,026204	0,0383399	0,683	0,49	1,0266
Velocidad del viento más rápida en 2 min destino	-0,0024899	0,0135219	-0,184	0,85	0,9975
Día semana: jueves	0,0648929	0,0381515	1,701	0,09	1,0670
Segunda quincena	0,0386044	0,0211071	1,829	0,07	1,0394
Escarcha o cencellada en destino	0,2499706	0,1527969	1,636	0,10	1,2840
Humo o neblina en destino	0,0511751	0,0293362	1,744	0,08	1,0525
Duración programada del vuelo	-0,0006268	0,0002551	-2,457	0,01	0,9994

Tabla 28. Variables no significativas en la regresión logística

Variables significativas: vuelos y fecha

La Tabla 29 presenta los coeficientes de las variables relacionadas con el aeropuerto de origen y destino, así como las variables derivadas de la fecha y la hora del vuelo.

Coeficientes las variables significativas relacionadas con los vuelos obtenidos en la regresión logística					
Variable	**Estimate**	**Std, Error**	**z value**	**Pr(>z)**	**OR**
Apto origen Chicago	-0,241	0,032	-7,480	7,44E-14	0,786
Apto origen Denver	-0,290	0,038	-7,670	1,72E-14	0,749
Apto origen Los Ángeles	-0,280	0,035	-8,074	6,82E-16	0,756
Apto destino Denver	-0,340	0,035	-9,582	< 2e-16	0,712
Apto destino Chicago	-0,198	0,033	-6,017	1,78E-09	0,820
Compañía low cost	0,432	0,024	18,227	< 2e-16	1,541
Momento día: mañana	0,402	0,062	6,497	8,17E-11	1,494
Momento día: noche	1,223	0,063	19,509	< 2e-16	3,398
Momento día: tarde	1,017	0,062	16,498	< 2e-16	2,765
Día semana: martes	-0,287	0,041	-7,082	1,42E-12	0,751
Estación: otoño	-0,396	0,034	-11,551	< 2e-16	0,673
Estación: primavera	0,179	0,032	5,657	1,54E-08	1,196
Estación: verano	0,212	0,033	6,367	1,93E-10	1,236

Tabla 29. Variables significativas de los vuelos en la regresión logística

De esta tabla se derivan las siguientes conclusiones:

Los coeficientes negativos asociados a los vuelos que tienen como origen Chicago, Denver o Los Ángeles indican que estos vuelos tienen entre un 21% y un 25% menos de probabilidades de sufrir retraso en comparación con los vuelos que salen de Atlanta. Mientras que los vuelos con destino a Denver o Chicago tienen un 28.8% y 18% menos de probabilidad de retraso en comparación con Atlanta o Los Ángeles.

Los vuelos operados por una compañía *low cost* tienen un 54% más de probabilidad de sufrir retraso en comparación con los vuelos operados por otras compañías.

Los vuelos que parten durante la noche tienen las mayores probabilidades de retraso, siendo hasta 3.39 veces más propensos a experimentar demoras en comparación con aquellos que salen de madrugada.

En términos de días de la semana, los vuelos que salen los martes tienen una probabilidad de retraso un 24.95% menor que los vuelos que parten en otros días de la semana.

Según la estación del año, los vuelos son más propensos a sufrir retraso. Los vuelos en otoño presentan un 32% menos de probabilidad de retraso en comparación con los vuelos en invierno. Mientras que los vuelos en verano y primavera tienen un 23.6% y 19.6%, respectivamente más de probabilidad de retraso comparados con los vuelos en invierno.

Variables significativas: clima

Por otro lado, la Tabla 30 muestra con los coeficientes de las variables relacionadas con las condiciones meteorológicas en la regresión logística.

Coeficientes las variables significativas relacionadas con el clima obtenidos en la regresión logística					
Variable	**Estimate**	**Std, Error**	**z value**	**Pr(>z)**	**OR**
Niebla en destino	0,122	0,025	4,843	1,28E-06	1,129
Tomenta eléctrica en destino	0,291	0,036	8,078	6,59E-16	1,337
Niebla en origen	0,263	0,025	10,559	< 2e-16	1,301
Tomenta eléctrica en origen	0,393	0,035	11,198	< 2e-16	1,482
Escarcha o cencellada en origen	0,645	0,144	4,476	7,61E-06	1,906
Humo o neblina en origen	0,134	0,029	4,638	3,53E-06	1,144
Precipitaciones destino	0,013	0,002	8,208	2,25E-16	1,013
Precipitaciones origen	0,009	0,002	5,434	5,51E-08	1,009
Nieve en destino	0,009	0,001	6,987	2,81E-12	1,009
Nieve en origen	0,018	0,001	13,197	< 2e-16	1,018
Profundidad de la nieve en origen	0,003	0,001	4,803	1,56E-06	1,003
Velocidad del viento más rápida en 2 min origen	-0,049	0,014	-3,597	0,000322	0,952
Velocidad del viento más rápida en 5 sec en origen	0,055	0,010	5,693	1,25E-08	1,056

Tabla 30. Variables significativas de las condiciones climáticas en la regresión logística

Las condiciones de visibilidad reducida a causa de la *niebla* en el aeropuerto de origen aumentan la posibilidad de retraso hasta un 30.1%. De manera similar ocurre en el aeropuerto de destino, donde un incremento en la niebla está asociado con un aumento del 12.9% en la probabilidad de retraso.

Las *tormentas eléctricas* en el aeropuerto de origen están asociadas con una probabilidad de retraso considerablemente mayor. En presencia de tormentas eléctricas, la probabilidad de retraso aumenta en un 48.2%. En el aeropuerto de destino, las tormentas eléctricas también contribuyen a un incremento en la probabilidad de retraso, aunque el aumento es de un 33.7%.

La presencia de *escarcha o cencellada* en el aeropuerto de origen tiene el mayor impacto en la probabilidad de retraso. Este factor incrementa la probabilidad de retraso en un 90.6%.

La *nieve* tiene un *impacto* menos pronunciado en los retrasos en comparación con otros factores meteorológicos. La nieve en el aeropuerto de origen incrementa la probabilidad de retraso en un 1.8%, mientras que en el aeropuerto de destino el incremento es inferior al 1%. Adicionalmente, la *profundidad de la nieve* en el aeropuerto de origen aumenta la probabilidad de retraso en un 0.34%, indicando que, aunque la nieve puede influir en los retrasos, su impacto es relativamente menor.

Finalmente, la *velocidad del viento más rápida en 2 minutos* en el aeropuerto de origen está asociada con una disminución de la probabilidad de retraso en un 4.7%. En cambio, la velocidad del viento más rápida en 5 segundos en el aeropuerto de origen incrementa

la probabilidad de retraso en un 5.6%, reflejando un efecto contrario en comparación con la velocidad del viento en intervalos más largos.

8.2. Redes Neuronales

Es importante destacar que, a diferencia de otros modelos como la regresión logística, las redes neuronales no realizan automáticamente una estandarización interna de los datos. Por esta razón, se llevó a cabo un preprocesamiento de los datos que incluyó la estandarización de las variables continuas y la transformación de las variables categóricas en variables dummy. Para las variables categóricas nominales, se crearon nuevas variables binarizadas, asignando valores de 0 y 1 para representar cada categoría, dando lugar a un total de 38 variables independientes.

Para obtener la mejor red neuronal de clasificación binaria, se han ajustado varias configuraciones del modelo, descritas en el apartado de metodología y se ha utilizado validación cruzada con 4 grupos y cinco repeticiones para la evaluación.

Se ha determinado el número óptimo de nodos en la capa oculta (*size*) para equilibrar la minimización del error y la prevención del sobreajuste. Para calcular el número ideal de nodos ocultos, se emplea la fórmula:

$$h\,(k+1)+h+1=\frac{obs}{p}$$

Donde, h es el número de nodos ocultos, k el número de nodos input, obs el número de observaciones y p el número de observaciones por parámetro. Con 53 502 observaciones y 38 variables, se tiene:

$$h(38+1)+h+1=\frac{53.504}{30}$$

$$h\simeq 44{,}56$$

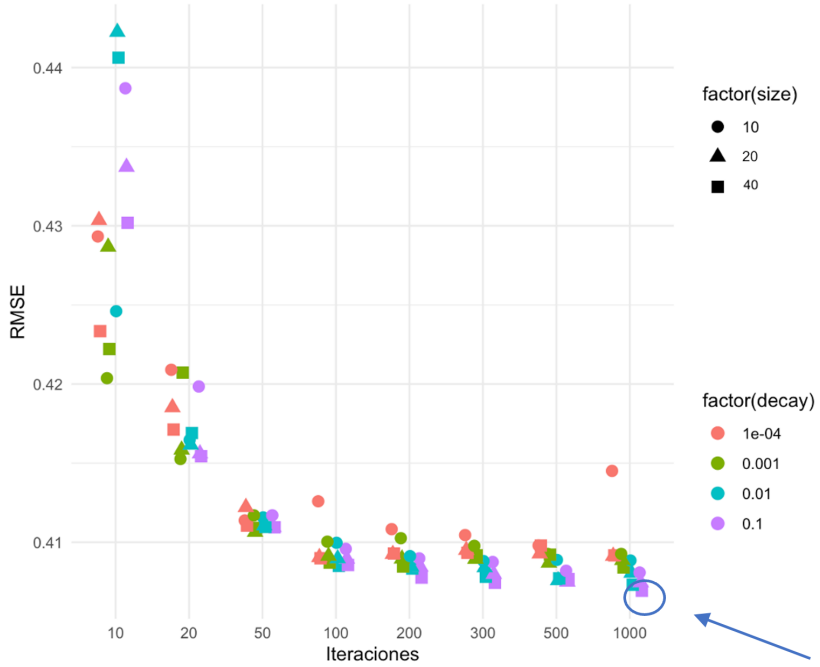
Por lo que se probará configuraciones de 10, 20 y 40 nodos ocultos, ajustando en grupos de cinco nodos para evitar el sobreajuste.

Además, se han evaluado varios valores de tasa de decaimiento (*decay*) para controlar el sobreajuste. Se probaron tasas de decaimiento de 0.01 hasta 0.0001 para encontrar el balance óptimo entre rapidez y precisión.

El número máximo de iteraciones se ha ajustado variando de 10 a 1 000. Un rango amplio de iteraciones permite identificar el equilibrio adecuado entre el tiempo de entrenamiento y el rendimiento del modelo. El objetivo es determinar el punto en el que el modelo ofrece el mejor rendimiento, medido por el error cuadrático medio (RMSE). El algoritmo detendrá el entrenamiento si encuentra la convergencia antes de alcanzar el límite de iteraciones especificado.

Finalmente, los resultados se ordenan según el error cuadrático medio (RMSE) y se presentan en el Gráfico 6 para visualizar el rendimiento del modelo en función de las diferentes configuraciones probadas.

Gráfico 6. Resultados del tuneo de la red neuronal



En el gráfico anterior, se señala la combinación de parámetros que presenta el menor valor de RMSE. Por lo tanto, en la sección final de comparación de modelos, se realizará validación cruzada repetida de la red neuronal configurada con 40 nodos, un valor de *decay* de 0.1 y 1 000 iteraciones.

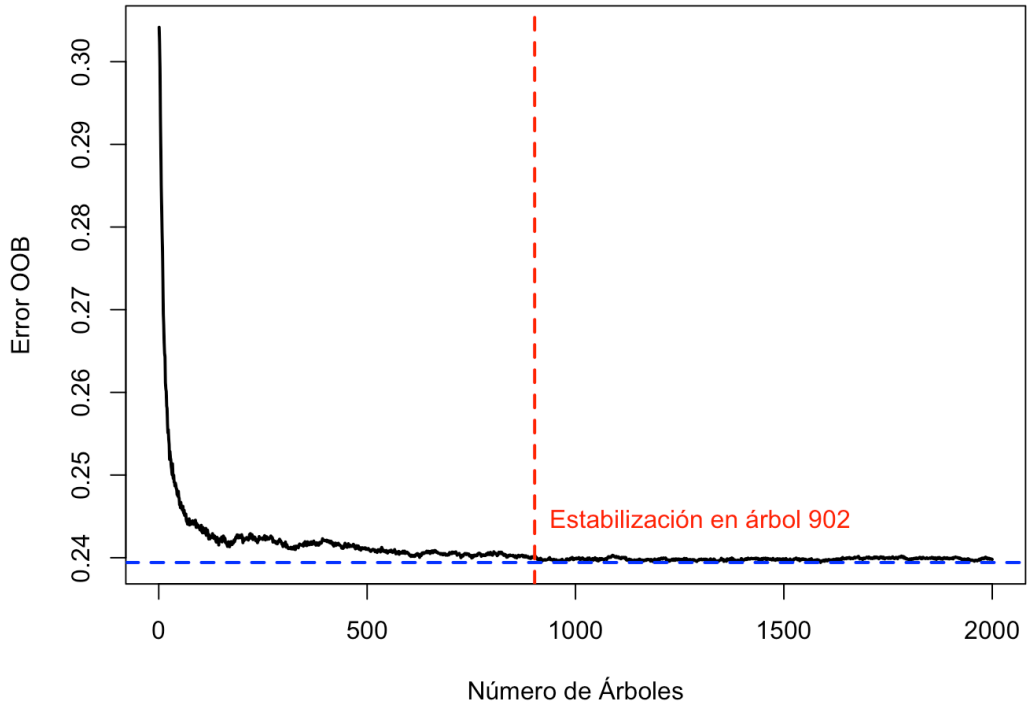
8.3. Bagging

Para evaluar el rendimiento en *Bagging*, se analiza el *Error Out Of Bag* (OOB) en función del número de iteraciones. (*ntree*). El error OOB es una medida de la capacidad de generalización del modelo, calculada utilizando las observaciones que no fueron incluidas en el entrenamiento de cada árbol durante el proceso de *Bagging*.

En este análisis, se estableció que el número de variables consideradas en cada división sería el total disponible (*mtry*=25). Para identificar con mayor precisión el punto en el que el error se estabiliza, se ajustó el umbral de error a 0.00005.

El Gráfico 7 muestra cómo el error OOB cambia con el número de árboles en el modelo de *Bagging*. En la gráfica, se observa que el error OOB tiende a estabilizarse alrededor de los 900 árboles. Este punto indica que agregar más árboles no mejora significativamente el rendimiento del modelo, permitiendo así una estimación más precisa de la capacidad de generalización sin necesidad de incrementar innecesariamente el tamaño del modelo.

Gráfico 7. Evolución del error OOB en Bagging



A continuación, se elige el tamaño de la muestra (*sampsize*). Existen dos opciones para definir este parámetro:

1. Utilizar *sampsize=TRUE*, lo que significa que se emplean todas las observaciones disponibles en el conjunto de datos.
2. Aplicar la fórmula:

$$\boldsymbol{Sampsize} = \left(\mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{k}}\right) * \boldsymbol{N}$$

donde k es el número de grupos de validación cruzada (en nuestro caso, 4), y N es el número total de observaciones. Aplicando la fórmula, se tiene:

$$\boldsymbol{Sampsize} = \left(\mathbf{1} - \frac{\mathbf{1}}{\mathbf{4}}\right) * \mathbf{53504} = \mathbf{40.128}$$

Una vez establecidos los dos tamaños de muestra (*sampsize*), se evaluaron diferentes tamaños de nodos (*nodesize*) en configuraciones de 20, 50, 100, 250 y 500, junto con los valores de *mtry* y *ntree* previamente definidos.

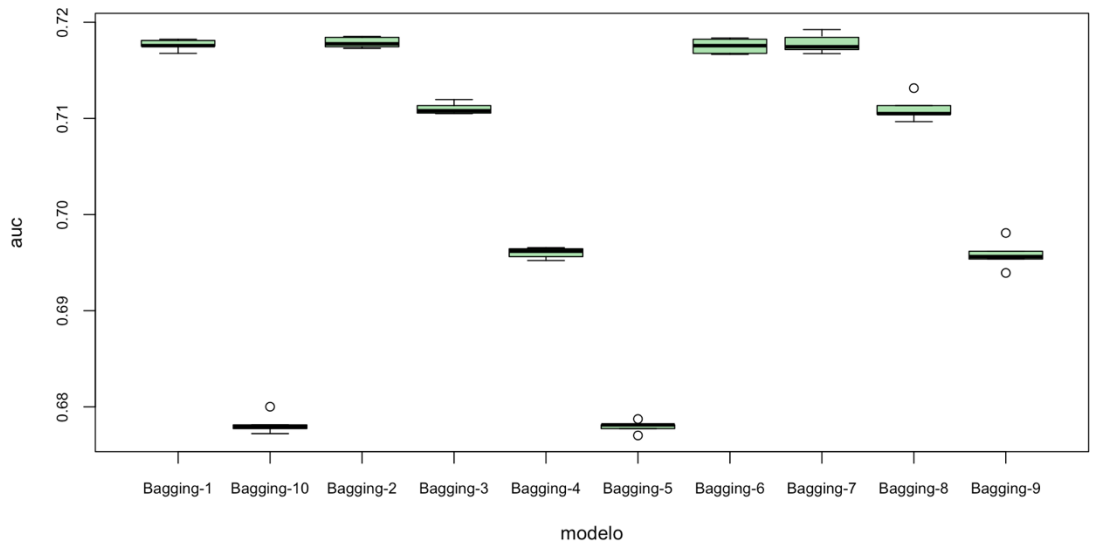
Estas configuraciones se detallan en la Tabla 31.

Modelo	Nodesize	sampsize	N tree	mtry
Bagging-1	20	Máx	900	25
Bagging-2	50	Máx	900	25
Bagging-3	100	Máx	900	25
Bagging-4	250	Máx	900	25
Bagging-5	500	Máx	900	25
Bagging-6	20	40128	900	25
Bagging-7	50	40128	900	25
Bagging-8	100	40128	900	25
Bagging-9	250	40128	900	25
Bagging-10	500	40128	900	25

Tabla 31. Modelos Bagging evaluados en validación cruzada repetida

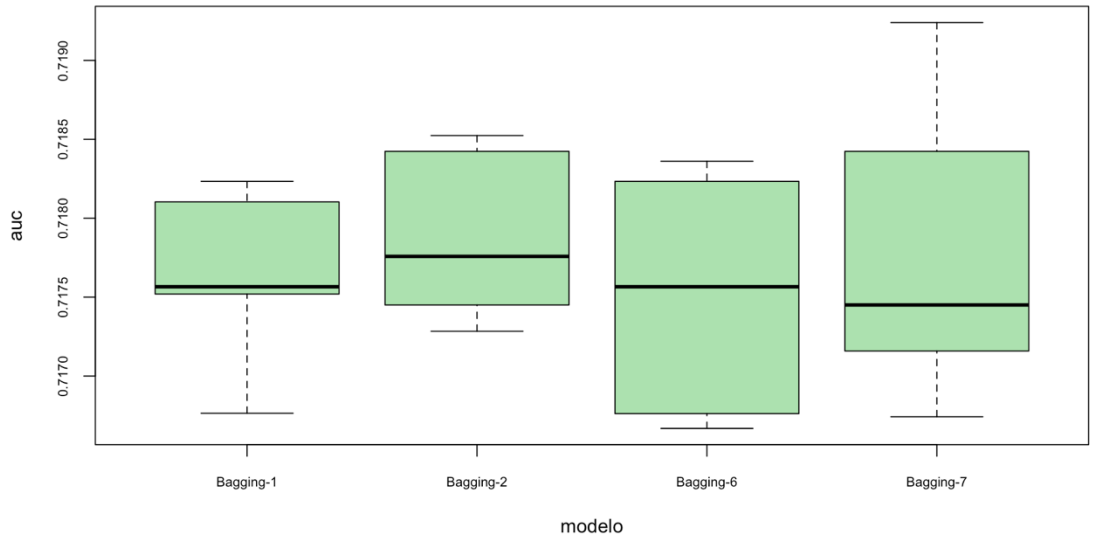
El Gráfico 8 presenta la media de AUC obtenida mediante validación cruzada repetida. En este gráfico, las cajas más altas, son los modelos con valores más altos de AUC y por lo tanto, son los mejores. Además, es importante considerar la altura de la caja, que representa la varianza; cuanto más pequeña sea la caja, mejor, ya que indica mayor consistencia en los resultados.

Gráfico 8. AUC de los modelos Bagging



Entre los candidatos a mejor modelo se encuentran *Bagging*-1, *Bagging*-2, *Bagging*-6 y *Bagging*-7, cuyos resultados se presentan en el Gráfico 9 mediante un boxplot que muestra la distribución de los valores de AUC para estos cuatro modelos de *Bagging*.

Gráfico 9. AUC de los cuatro mejores modelos Bagging



Bagging-1 tiene una mediana ligeramente inferior a 0.7180, con una dispersión más baja, lo que se traduce en resultados predecibles y estables, con la mayoría de sus valores de AUC concentrados cerca de la mediana. Por otro lado, *Bagging*-2 y *Bagging*-6 presentan medianas ligeramente superiores, alrededor de 0.7185, pero con una mayor variabilidad en sus resultados, lo que implica que, aunque pueden ofrecer un rendimiento ligeramente mejor, sus predicciones son menos consistentes. *Bagging*-7, en cambio, es descartado porque presenta la mediana más baja y la mayor dispersión, lo que indica un rendimiento menos confiable.

En conclusión, dado que se busca maximizar el AUC, aunque con cierto riesgo de variabilidad, se considera que ***Bagging*-2** es el mejor modelo de *Bagging*. Este modelo se caracteriza por utilizar 50 nodos y un tamaño de muestra (*sampsize*) igual al número total de observaciones.

8.4. Random Forest

A diferencia de *Bagging*, el método *Random Forest* requiere seleccionar un número específico de variables para cada división del árbol de decisión.

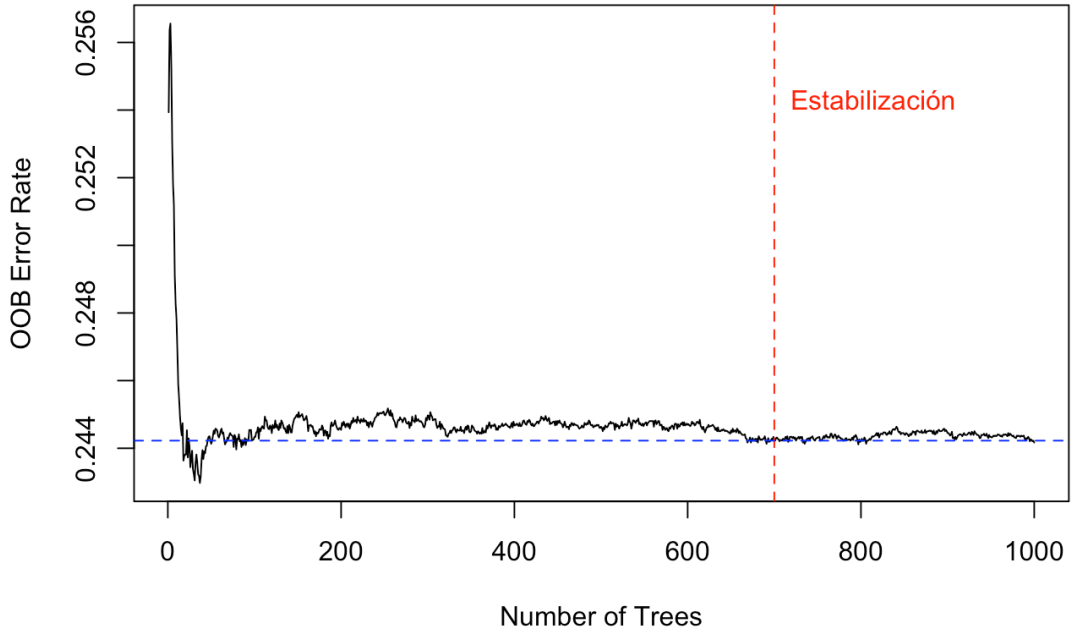
Los parámetros a ajustar en *Random Forest* incluyen:

- Número de variables sorteadas en cada división (*mtry*),
- Número de árboles generados (*ntree*),
- Número de observaciones por nodo (*nodesize*),
- Tamaño de la muestra utilizada en *Bagging* (*sampsize*).

El primer paso consiste en determinar el valor óptimo de *mtry*. Esto se realiza probando valores en el rango de 1 a 25, utilizando un ntree de 2 000 y estableciendo un mínimo de 10 observaciones por nodo.

El Gráfico 10 muestra que a partir de 700 árboles el error OOB se estabiliza, indicando que añadir más iteraciones no mejora el rendimiento del modelo.

Gráfico 10. Evolución del error OOB en Random Forest



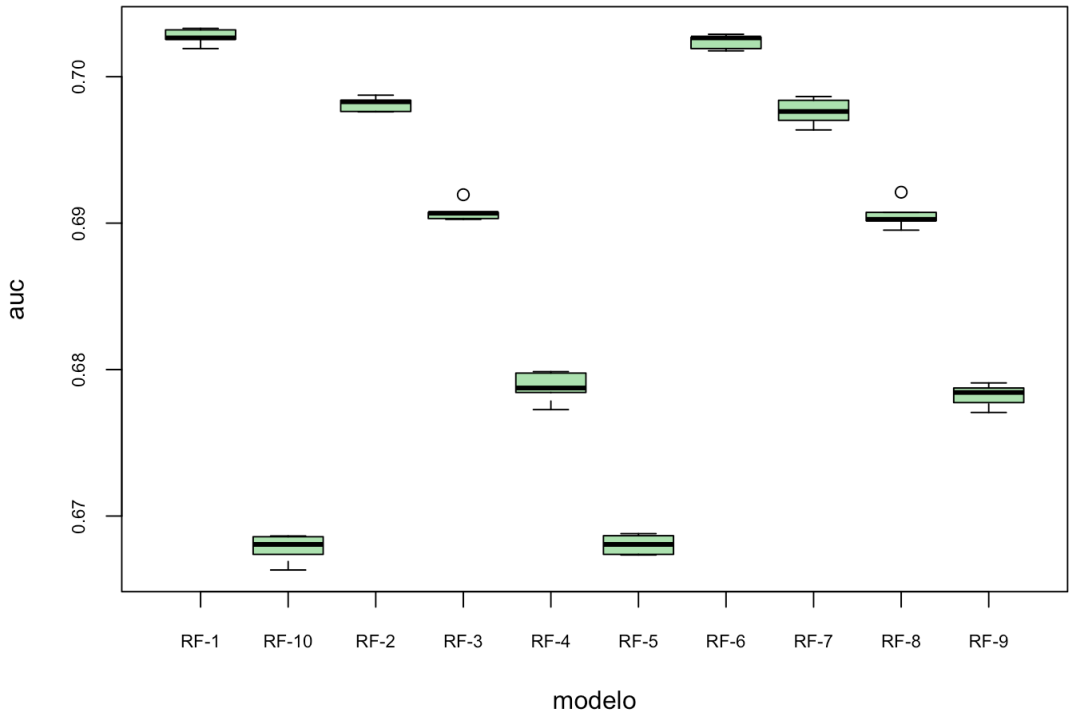
Con estos parámetros ajustados, se procede a realizar validación cruzada repetida con los modelos seleccionados, tal como se detalla en la Tabla 32.

Modelo	Nodesize	Sampsize	N tree	mtry
RF-1	20	Máx	700	2
RF-2	50	Máx	700	2
RF-3	100	Máx	700	2
RF-4	250	Máx	700	2
RF-5	500	Máx	700	2
RF-6	20	40.128	700	2
RF-7	50	40.128	700	2
RF-8	100	40.128	700	2
RF-9	250	40.128	700	2
RF-10	500	40.128	700	2

Tabla 32. Modelos Random Forest evaluados en validación cruzada repetida

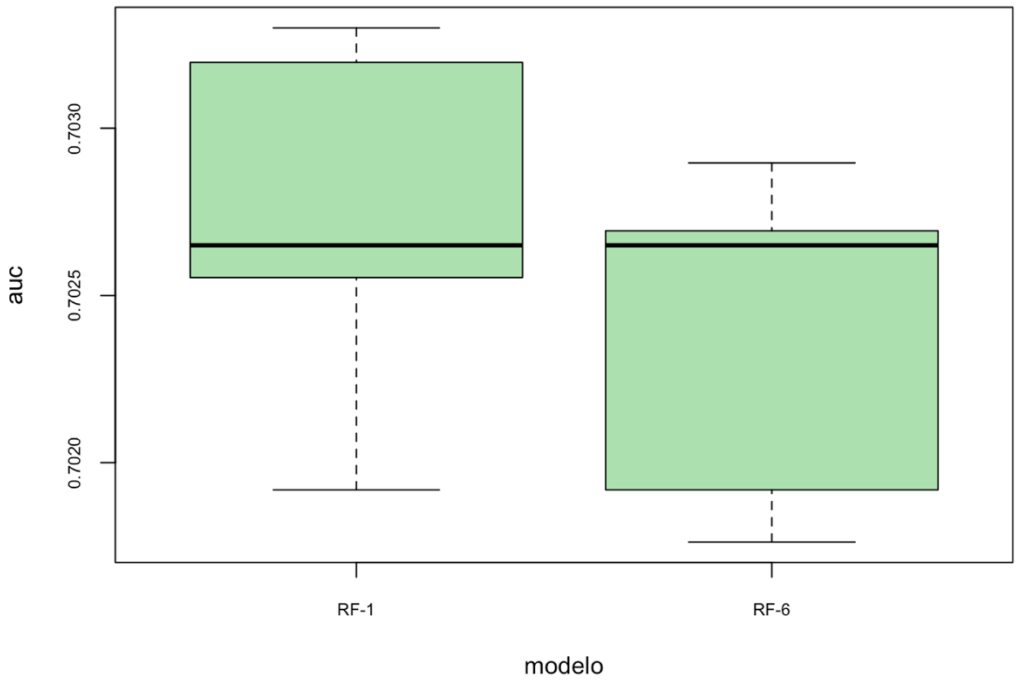
Los resultados obtenidos para estos modelos se presentan en el Gráfico 11.

Gráfico 11. AUC de los modelos Random Forest



Entre los modelos evaluados, los candidatos más prometedores son RF-1 y RF-7, los cuales se analizan individualmente en el Gráfico 12.

Gráfico 12. AUC de los cuatro mejores modelos Random Forest



El modelo RF-1 muestra una mediana justo por debajo de 0.703, acompañada de una dispersión más amplia. En contraste, el modelo RF-6 presenta una mediana ligeramente inferior, pero con una dispersión menor en comparación con RF-1.

En resumen, **RF-1** se consolida como el mejor modelo de *Random Forest*, con 20 nodos y el *sampsize* máximo (igual al número de observaciones).

8.5. Gradient Boosting

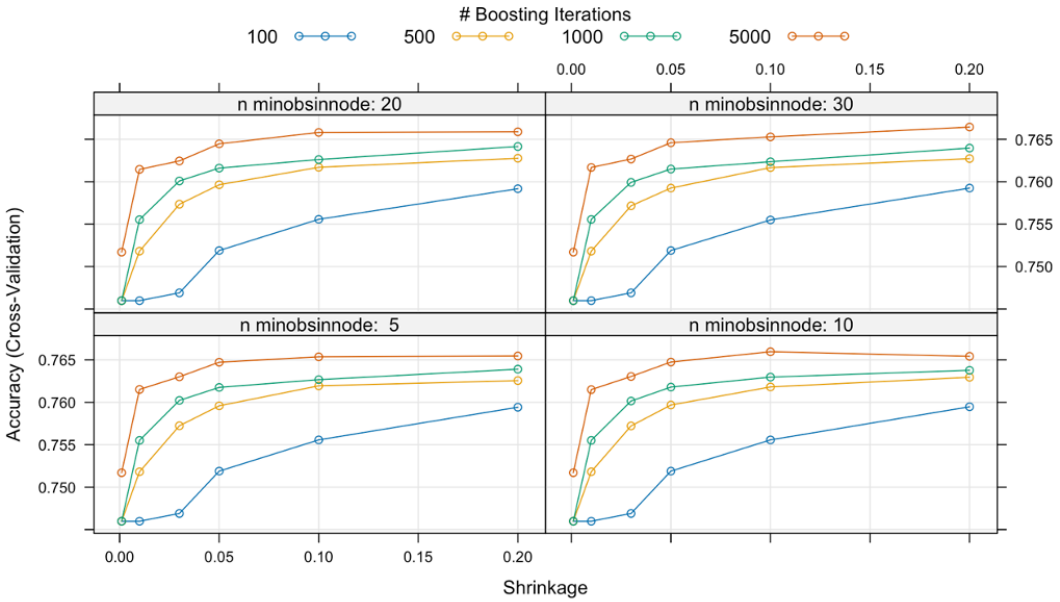
En este algoritmo primero se define la constante de regularización (*shrinkage*), que controla la velocidad a la que el algoritmo converge. Se probaron diferentes valores de *shrinkage* entre 0.2 y 0.001. Un valor mayor de *shrinkage* permitirá que el algoritmo converja más rápidamente, pero si es demasiado alto, podría comprometer la precisión del modelo. Por el contrario, si el valor es demasiado bajo, el algoritmo necesitará más iteraciones para alcanzar la convergencia.

Además, se consideran otros dos parámetros: *n.minobsinnode*, que indica el número mínimo de observaciones que deben estar presentes en un nodo terminal, con valores probados de 5, 10, 20 y 30; y *n.trees*, que representa el número total de árboles a construir, con valores posibles de 100, 500, 1 000 y 5 000.

Finalmente, dado que la variable objetivo es binaria, el número de hojas finales (*interaction.depth*) se fija en 2, lo que determina la profundidad máxima de las interacciones entre las variables en cada árbol.

A partir del Gráfico 13 se observan los valores de *shrinkage* óptimos, a través del *early stopping.*

Gráfico 13. Early Stopping en Gradient Boosting



Del gráfico anterior vemos como para *shrinkage* = 0,001, independientemente del número de árboles (n.*trees*) y del tamaño mínimo de nodo (*n.minobsinnode*), la precisión es baja, lo que indica que el modelo no está realizando una buena clasificación.

Al aumentar el shrinkage a 0.01 y luego a 0.03, la precisión mejora, especialmente con un mayor número de árboles (5 000), lo que indica que el modelo está haciendo mejores predicciones.

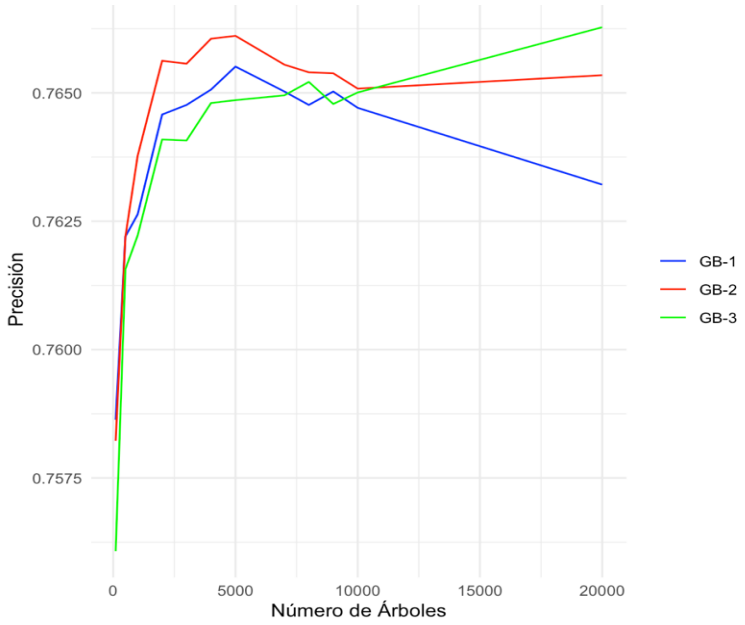
Con base en los resultados anteriores, se procede a estudiar el *early stopping* utilizando las configuraciones presentadas en la Tabla 33. Estas configuraciones han obtenido las mayores precisiones y están ordenadas de mayor a menor rendimiento.

Modelo	Shrinkage	N.minobsinnode
GB-1	0,2	30
GB2	0,2	20
GB-3	0,1	30

Tabla 33. Comparativa de configuraciones de shrinkage y número de árboles

El Gráfico 14 muestra los resultados del *early stopping* aplicados a las configuraciones seleccionadas.

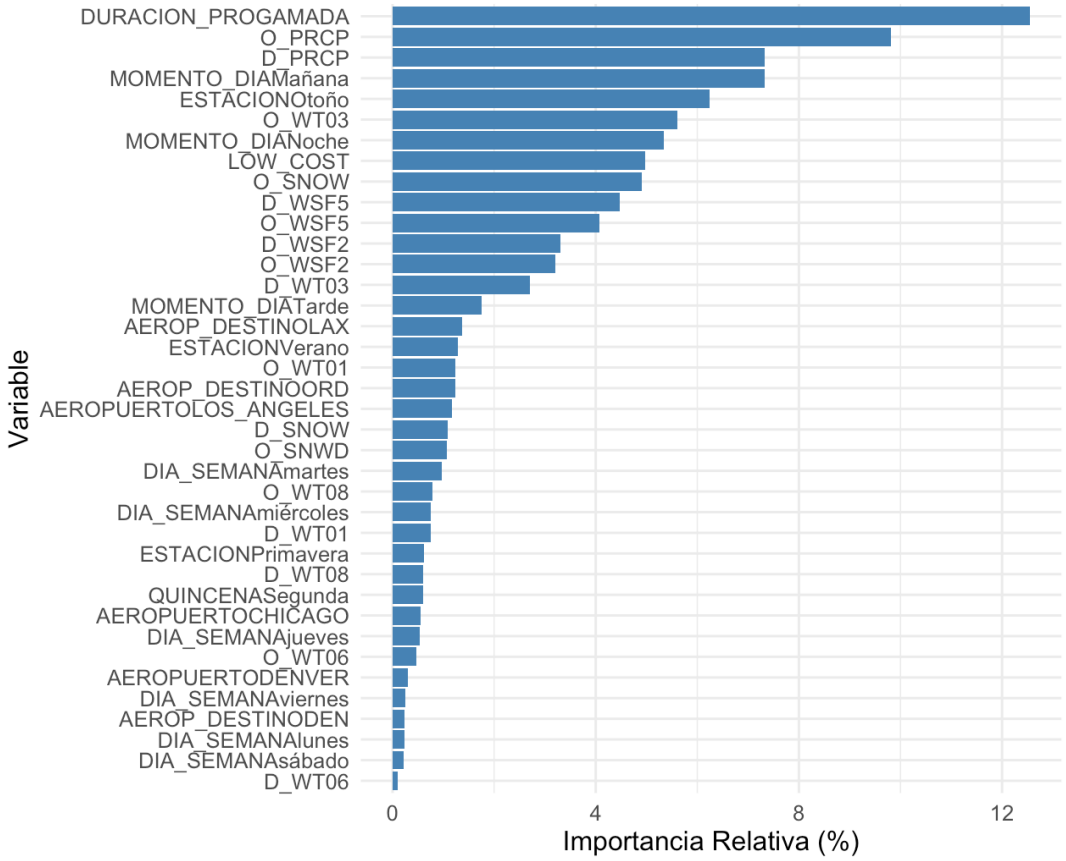
Gráfico 14 . Optimización del modelo a través de Early Stopping



En el gráfico anterior se observa que se estabiliza a partir de 10 000 árboles, para todos los sets configurados de *Gradient* boosting.

Antes de pasar a realizar validación cruzada repetida, el Gráfico 15 muestra la importancia relativa de las variables en este modelo

Gráfico 15. Importancia de variables en el modelo Gradient Boosting



La variable que más contribuye a la capacidad predictiva del modelo es la *duración programada*, con una importancia superior al 12%. A continuación, las *precipitaciones en el origen* tienen casi un 10% de importancia. En tercer lugar, se encuentran las *precipitaciones en el destino* y el *momento del día (mañana)*, ambos con cerca del 8% de importancia. En quinto lugar, la *estación de otoño* aporta más del 6%, y la *tormenta eléctrica* (truenos en el origen) representa algo menos del 6% de la importancia total del modelo.

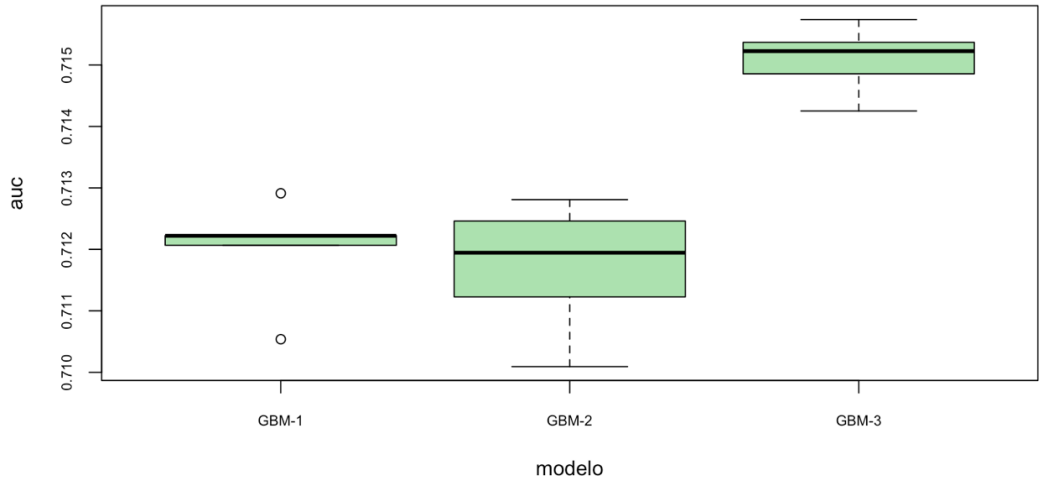
La Tabla 34 muestra los tres modelos de *Gradient* Boosting que se evaluarán mediante validación cruzada repetida para seleccionar el mejor modelo.

Modelo	Shrinkage	N.minobsinnode	Ntree
GB-1	0,2	30	10.000
GB2	0,2	20	10.000
GB-3	0,1	30	10.000

Tabla 34. Modelos Gradient Boosting evaluados en validación cruzada repetida

En el Gráfico 16 se muestran los resultados de la validación cruzada repetida para los mejores modelos de *Gradient Boosting*.

Gráfico 16. AUC de los cuatro mejores modelos Gradient Boosting



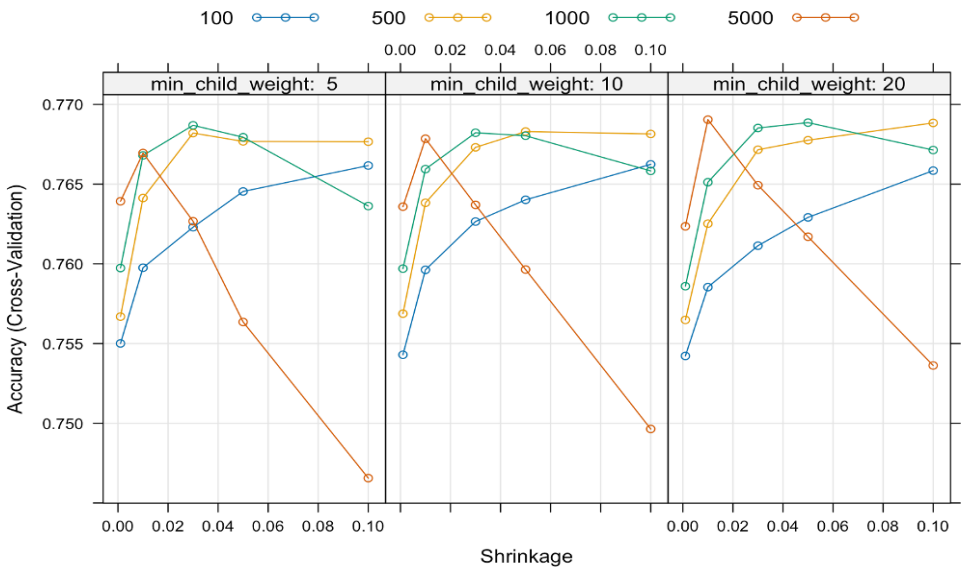
El análisis de *Gradient Boosting* revela que el modelo más eficaz es el **GMB-3**, con una configuración de *shrinkage* = 0.1 y *n.minobsinnode* = 30.

8.6. XGBoost

XGBoost es una versión avanzada de *Gradient Boosting* que optimiza el rendimiento y la flexibilidad del modelo. Para su ajuste, se consideran varios parámetros. La tasa de aprendizaje (*eta*) varía entre 0.3 y 0.001, el número mínimo de observaciones por nodo (*min_child_weight*) se ajusta entre 20 y 30, y el número de árboles (*nrounds*) se explora de 100 a 10 000. Además, se ajusta *gamma* para controlar la mínima reducción de pérdida necesaria para una partición, *alpha* para la regularización L1 en 3, *lambda* para la regularización L2 de ponderaciones en 6 , y *lambda_bias* para la regularización L2 del sesgo en 10

Los resultados del ajuste muestran la combinación optima de parámetros para mejorar su rendimiento en el Gráfico 17.

Gráfico 17. Early Stopping en XGBoost



El modelo con una tasa de aprendizaje (*eta*) de 0.05, un número mínimo de observaciones por nodo (*min_child_weight*) de 10 y 5 000 iteraciones (*nrounds*) obtuvo la mayor precisión, con un Accuracy de 0.7683. Le sigue el modelo con una tasa de aprendizaje de 0.010, un número mínimo de observaciones por nodo de 5 y 5 000 iteraciones, que logró una precisión de 0.7678. Finalmente, el modelo con una tasa de aprendizaje de 0.010, un número mínimo de observaciones por nodo de 5 y 1 000 iteraciones tuvo una precisión de 0.7668.

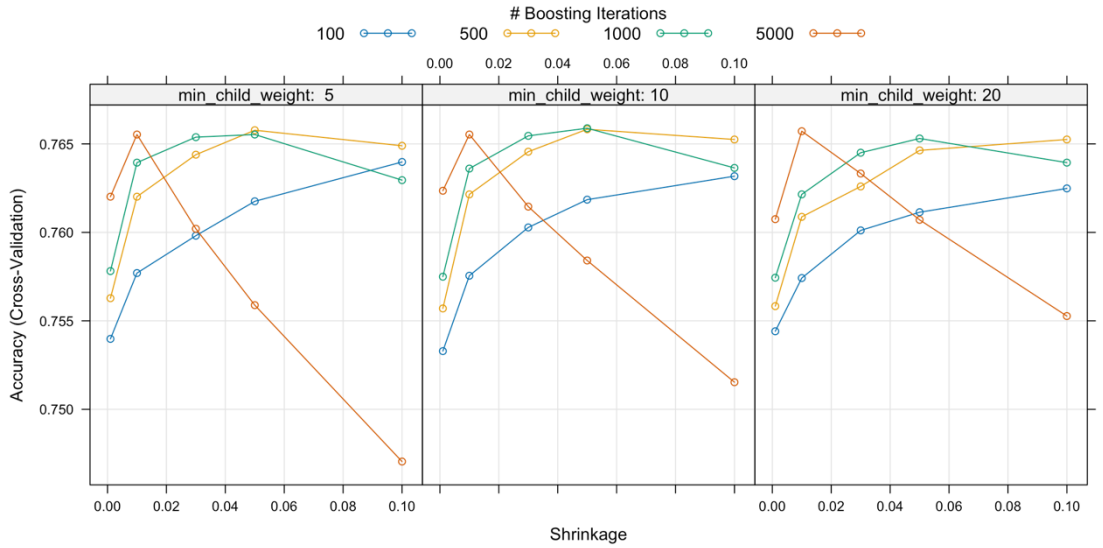
Seguidamente se evaluó la importancia de las variables para comprender cuáles tienen mayor impacto en las predicciones, recogidas en el Gráfico 18.

Gráfico 18. Importancia de variables en el modelo XGBoost

En este caso, la importancia de las variables en *XGBoost* se reporta en una escala absoluta, donde la variable más relevante (*duración programada del vuelo*), tiene una importancia máxima de 100. Aunque las variables seleccionadas como importantes son en su mayoría consistentes entre los dos métodos, se presentan algunas novedades. Por ejemplo, las *precipitaciones en el origen* tienen una importancia de 60.165 en *XGBoost*, lo que es comparable a su relevancia en *Gradient Boosting*. Sin embargo, en *XGBoost*, la *velocidad del viento más rápida en 5 segundos* en el destino muestra una importancia de 43.989, mientras que en el origen es de 43.637.

En consecuencia, se procederá a repetir el proceso de tuneo de parámetros utilizando solo las variables más importantes. Estos resultados se muestran en el Gráfico 19.

Gráfico 19. Early Stopping en XGBoost con variables importantes



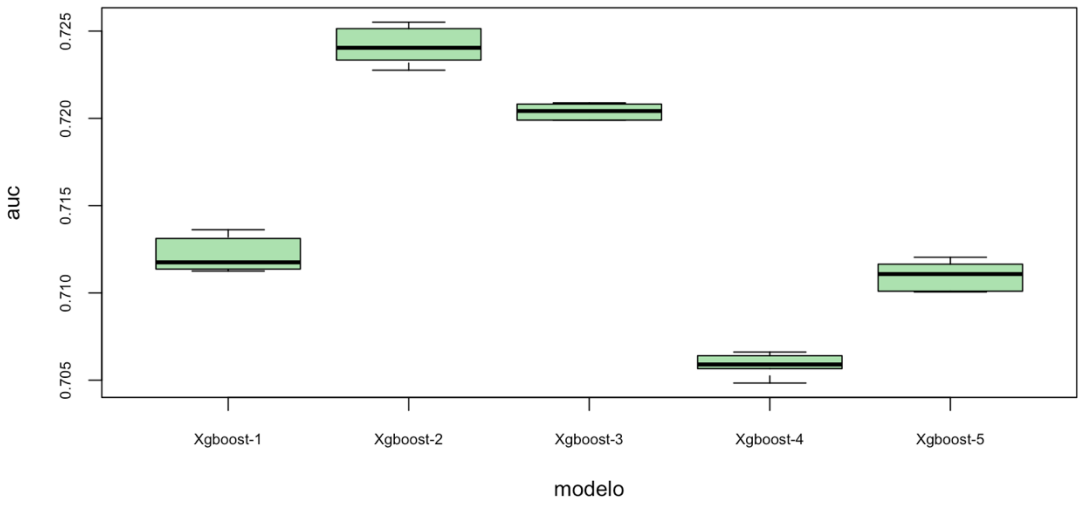
El análisis de los resultados muestra dos combinaciones óptimas de parámetros. La primera configuración, que incluye una tasa de aprendizaje (*eta*) de 0,05, un mínimo de observaciones por nodo (*min_child_weight*) de 10 y 5 000 iteraciones (*nrounds*), logró la mejor precisión, alcanzando un *accuracy* de 0.7659. La segunda combinación, que también ofrece la misma precisión de 0.7659, utiliza 5 000 iteraciones.

Por lo tanto, recapitulando, la Tabla 35 muestra las combinaciones de parámetros analizadas mediante validación cruzada repetida y el Gráfico 20 los resultados.

Nombre modelo	Variables utilizadas	shrinkage	min_child_weight	nround
Xgboost-1	Todas (25)	0,05	10	5.000
Xgboost-2	Todas (25)	0,01	5	5.000
Xgboost-3	Todas (25)	0,01	5	1.000
Xgboost-4	Importantes (16)	0,05	10	1.000
Xgboost-5	Importantes (16)	0,05	10	5.000

Tabla 35. Modelos XGBoost evaluados en validación cruzada repetida

Gráfico 20. AUC de los modelos XGBoost



Se observa que el mejor modelo es ***XGBoost-2*** y es el que utiliza *mtry* = 25, un mínimo de 5 observaciones por nodo, una tasa de regularización de 0.01 y 5 000 iteraciones.

8.7. SVM

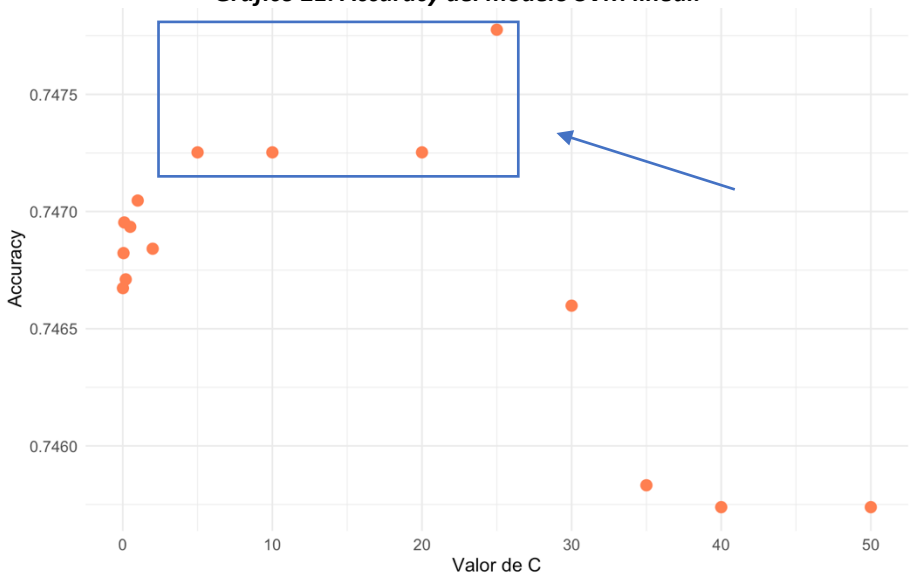
En este apartado se ajustan los tres tipos de kernel de SVM: lineal, polinómico y radial (gaussiano). Una vez obtenidos los mejores resultados para cada tipo de kernel, se realiza una validación cruzada para determinar el mejor modelo entre ellos.

8.7.1. SVM lineal

El único parámetro ajustable del modelo SVM con un kernel lineal es la constante de regularización C. Los valores de C considerados fueron desde 0.01 hasta 50.

El Gráfico 21 muestra el desempeño del modelo, medido en términos de *accuracy*, en función de diferentes valores de C. Estos resultados también se detallan en la Tabla 36.

Gráfico 21. Accuracy del modelo SVM lineal.



C	Accuracy
25	0.7477759
10	0.7472526
20	0.7472526
5	0.7472525
1	0.7470469
0,1	0.7469535
0,5	0.7469348

C	Accuracy
2	0.7468414
0,05	0.7468227
0,2	0.7467105
0,01	0.7466732
30	0.7465984
35	0.7458321
40	0.7457386

Tabla 36. Accuracy de las distintas combinaciones de SVM lineal

La tabla presenta los resultados ordenados de manera descendente según el *accuracy*. Se observa que valor máximo se obtiene con $C = 25$. Le siguen $C = 10$, $C = 20$ y $C = 5$ con valores similares

8.7.2. SVM polinómico

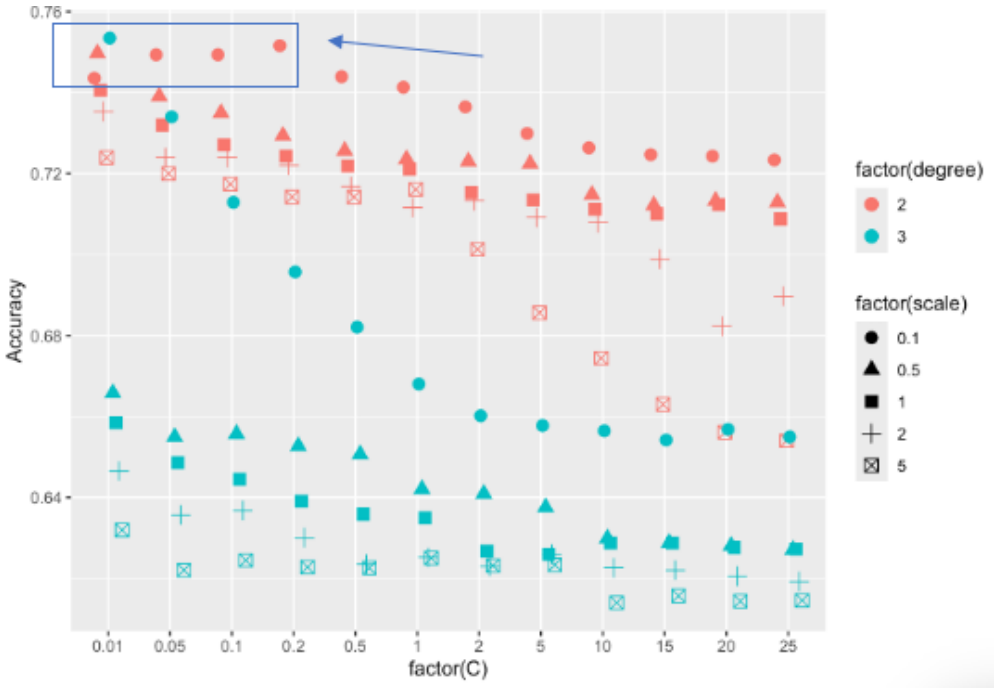
En el modelo de SVM con kernel polinómico, además de ajustar la constante de regularización C como en el SVM con kernel lineal, se incorporan dos parámetros específicos del kernel polinómico:

- **Grado del Polinomio**: Se exploran diferentes grados del polinomio, 2 y 3. Este parámetro determina la complejidad del modelo polinómico, permitiendo capturar interacciones no lineales entre las características.

- **Escala del Polinomio**: Se prueba con diferentes valores de escala, que son 0.5, 1, 2 y 5. La escala ajusta la influencia de los términos polinómicos en el modelo.

En el Gráfico 22, se observan los mayores niveles de precisión en función del accuracy. Por su parte la Tabla 37 muestra los diez mejores resultados ordenados de manera descendente.

Gráfico 22. Accuracy del modelo SVM polinómico



C	degree	scale	Accuracy
0.01	3	0.1	0.7534575
0.20	2	0.1	0.7515882
0.01	2	0.5	0.7497193
0.05	2	0.1	0.7493456
0.10	2	0.1	0.7493453
0.50	2	0.1	0.7439247
0.01	2	0.1	0.7435511
1.00	2	0.1	0.7413076
0.01	2	1.0	0.7405601
0.05	2	0.5	0.7390643

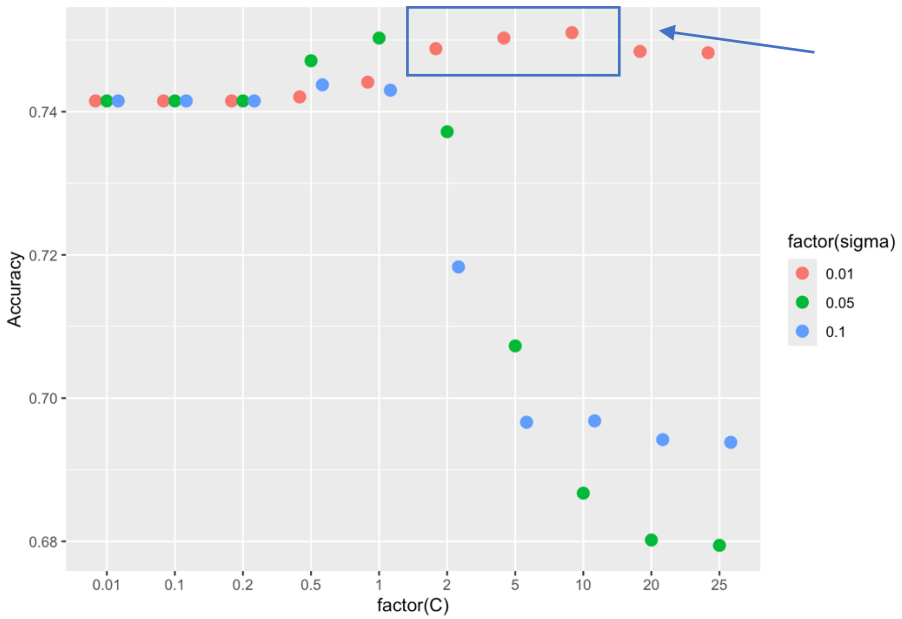
Tabla 37. Mejor accuracy de las combinaciones de parámetros con kernel polinómico

8.7.3. SVM radial

En SVM con kernel radial se tunean de nuevo diferentes valores del parámetro C. Además, se determina el valor de *sigma*, que controla la influencia que un solo punto de entrenamiento tiene en la clasificación. Valores más bajos de sigma hacen que la función de kernel sea más amplia, reduciendo así el riesgo de sobreajuste.

En la Tabla 38 se presentan los cinco mejores resultados ordenados de mayor a menor precisión (*accuracy*), y el Gráfico 23 ofrece una visualización de estos resultados.

Gráfico 23. Accuracy SVM radial



C	sigma	Accuracy
6	0,01	0.7553269
7	0,01	0.7551404
5	0,01	0.7551402
8	0,01	0.7538322
10	0,01	0.7528976

Tabla 38. Mejor accuracy de las combinaciones de parámetros con kernel radial

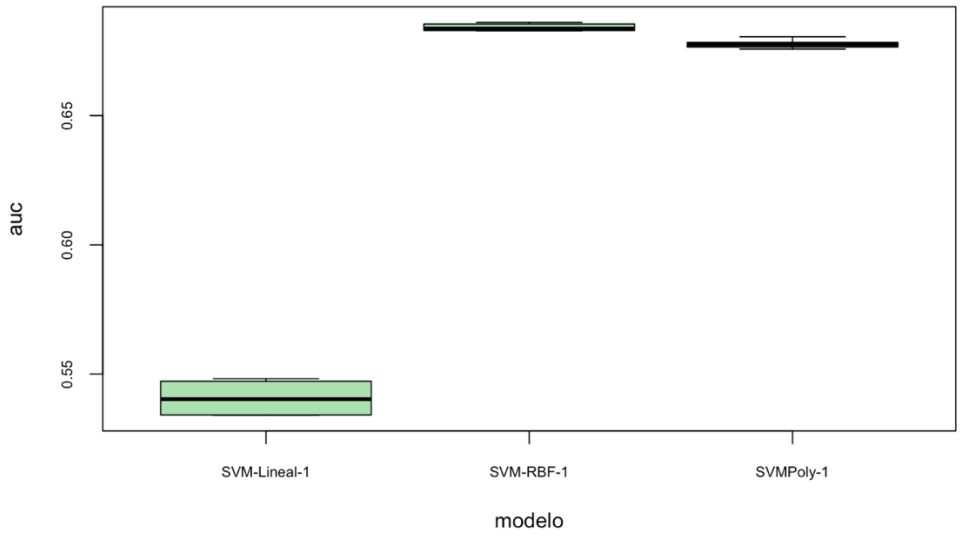
El mejor SVM radial es aquel que tiene $C = 6$ y *sigma*=0.01

Finalmente, para concluir la sección sobre SVM, se realiza una comparativa a través de validación cruzada repetida el mejor modelo en cada tipo de kernel. La Tabla 39 resume los tres mejores modelos de SVM, con kernel lineal, polinómico y radial que se probarán mediante validación cruzada y en el Gráfico 24 se presentan estos resultados.

Modelo SVM	C	Grado	Scale	Sigma
Lineal	25			
Polinómico	0.01	2	0.1	
Radial	10			0.01

Tabla 39. Modelos SVM e valuados en validación cruzada repetida

Gráfico 24. AUC de los mejores modelos SVM



Se observa que los tres modelos en general obtuvieron una precisión menor de lo que se anticipaba inicialmente. No obstante, el modelo con mejor rendimiento el radial, por lo que será el seleccionado para realizar comparaciones en el apartado final.

CAPÍTULO 9. RESULTADOS

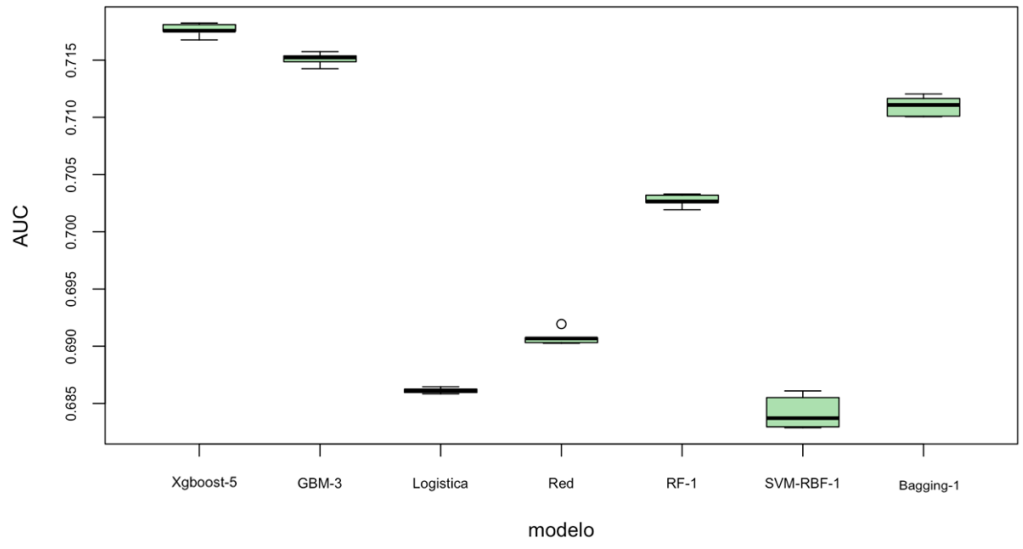
9.1. Comparación de modelos

La Tabla 40 resume las características de los modelos con mejor rendimiento obtenidos hasta ahora. Estos modelos se ilustran en el Gráfico 25, y se selecciona el modelo más adecuado para predecir retrasos en los vuelos.

Modelo	Características			
Logística	21 variables significativas			
Red neuronal	nodesize=20	Decay=0,1	Maxit=1000	
Bagging	Sampsize=max	nodesize=20	ntree=900	mtry=25
Random Forest	Sampsize=max	nodesize=20	ntree=700	mtry=2
Gradient Boosting	Sampsize=max	shrinkage=0,1	n.minobsinnode=30	ntree=100000
Xgboost	Sampsize=max	shrinkage=0,01	min_child_weight=5	nround=5000
SVM radial	C=6		sigma=0,1	

Tabla 40: Resumen de las Características de los modelos finalistas

Gráfico 25. AUC del mejor modelo de cada algoritmo



Se observa que el modelo con el mejor rendimiento es ***XGBoost***, seguido por *Gradient Boosting* y *Bagging*. El modelo Random Forest muestra una capacidad predictiva ligeramente inferior. En posiciones más bajas se encuentran la red neuronal y el modelo de regresión logística, mientras que en el último lugar se sitúa el SVM con kernel radial.

9.2. Ensamblado de modelos

Para reducir aún más el error de los modelos más prometedores, se han implementado técnicas de ensamblado. Es decir, se combinan múltiples modelos con el fin de mejorar el rendimiento general. En la Tabla 41 se detallan las combinaciones de modelos que se han probado mediante validación cruzada, utilizando un esquema de 4 grupos y 10 repeticiones.

Tipo de Combinación	Nombre modelo	Fórmula
Modelos 2 a 2	predi1	$\frac{bag + rf}{2}$
	predi2	$\frac{bag + gbm}{2}$
	predi3	$\frac{bag + xgb}{2}$
	predi4	$\frac{rf + gbm}{2}$
	predi5	$\frac{rf + xgb}{2}$
	predi6	$\frac{gbm + xgb}{2}$
Modelos 3 a 3	predi7	$\frac{bag + rf + gbm}{3}$
	predi8	$\frac{bag + rf + xgb}{3}$
	predi9	$\frac{bag + gbm + xgb}{3}$
	predi10	$\frac{rf + gbm + xgb}{3}$
Modelos 4 a 4	predi11	$\frac{bag + rf + gbm + xgb}{4}$

Tipo de Combinación	Nombre modelo	Fórmula
Ponderaciones	predi12	$0.6bag + 0.4gbm$
	predi13	$0.8bag + 0.2xgb$
	predi14	$0.7rf + 0.3gbm$
	predi15	$0.6rf + 0.4xgb$
	predi16	$0.8gbm + 0.2xgb$
	predi17	$0.7bag + 0.3rf$
	predi18	$0.4bag + 0.4rf + 0.2xgb$
	predi19	$0.6bag + 0.2gbm + 0.2xgb$
	predi20	$0.5rf + 0.3gbm + 0.2xgb$
	predi21	$0.5bag + 0.25rf + 0.25gbm$
	predi22	$0.25bag + 0.25rf + 0.25gbm + 0.25xgb$
	predi23	$0.3bag + 0.3rf + 0.3gbm + 0.1xgb$
	predi24	$0.4bag + 0.3rf + 0.2gbm + 0.1xgb$

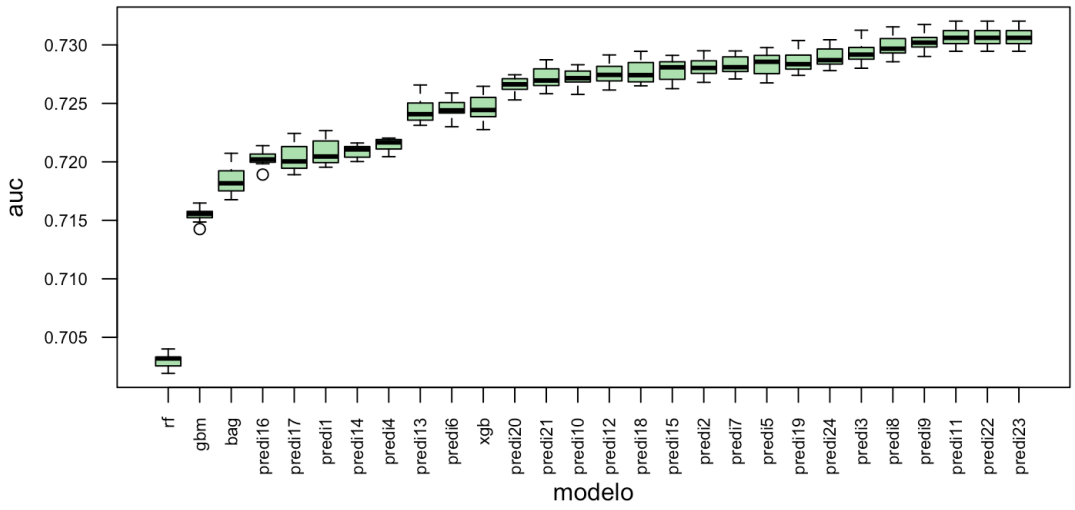
Tabla 41. Modelos ensamble

En la tabla se presentan diferentes enfoques para combinar las predicciones de varios modelos, detallados de la siguiente manera:

- **Modelos 2 a 2.** Se combinan las predicciones de dos modelos diferentes mediante un promedio simple.
- **Modelos 3 a 3.** En este caso, se combinan las predicciones de tres modelos distintos, también utilizando un promedio simple.
- **Modelos 4 a 4:** Aquí, se juntan las predicciones de los cuatro modelos mediante un promedio simple.
- **Ponderaciones.** Se asignan diferentes pesos a los modelos al combinarlos. Las ponderaciones reflejan la importancia relativa de cada modelo en la combinación final.

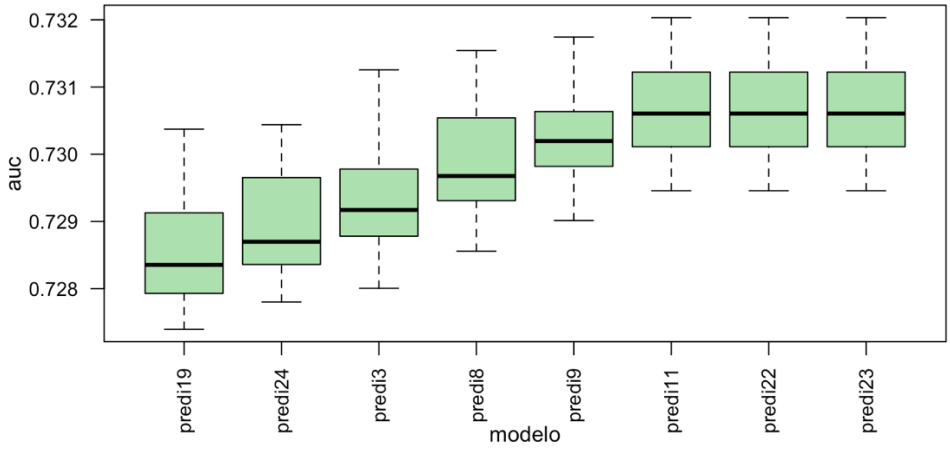
Los resultados se muestran por orden ascendente en el Gráfico 26.

Gráfico 26. AUC de los modelos de ensamble



Para una mejor visualización, en el Gráfico 27 se presentan los ocho mejores resultados obtenidos.

Gráfico 27. AUC de los mejores modelos de ensamble



A continuación, se destacan los tres modelos con mejor rendimiento:

1. **Mejor modelo**

$$Predi23 = 0.3bag + 0.3rf + 0.3gbm + 0.1xgb$$

2. **Segundo mejor modelo**

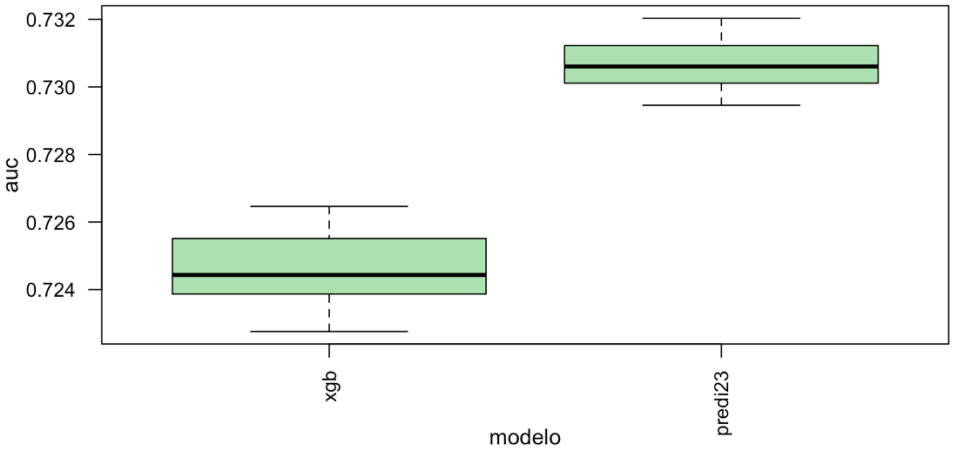
$$Predi22 = 0.25bag + 0.25rf + 0.25gbm + 0.25xgb$$

3. **Tercer mejor modelo**

$$Predi11 = \frac{bag + rf + gbm + xgb}{4}$$

Seguidamente, en el Gráfico 28 se procede a comparar el mejor modelo de ensamblado (*Predi23*) con el mejor modelo individual (*XGBoost*).

Gráfico 28. AUC de XGBoost y el mejor modelo de ensamble



La selección del modelo óptimo se basa en varias consideraciones que hacen de XGBoost (XGBM) la opción preferida en lugar de utilizar un ensamblado. Aunque XGBoost presenta un menor AUC, no tiene diferencias significativa comparado con *Predi23*.

La simplicidad de *XGBoost* elimina la necesidad de realizar ensamblajes complejos, lo que no solo reduce el coste computacional, sino que también simplifica el proceso de modelado.

Además, uno de los principales beneficios de XGBoost es su capacidad de regularización. Las constantes de regularización que ofrece el modelo permiten ajustar el enfoque con mayor o menor agresividad, lo que ayuda a controlar la varianza y a prevenir el sobreajuste. Otro aspecto relevante es su habilidad para generar gráficos de importancia de variables, una funcionalidad que resulta especialmente útil en este caso, ya que facilita la identificación de las variables que más influyen en el retraso de los vuelos.

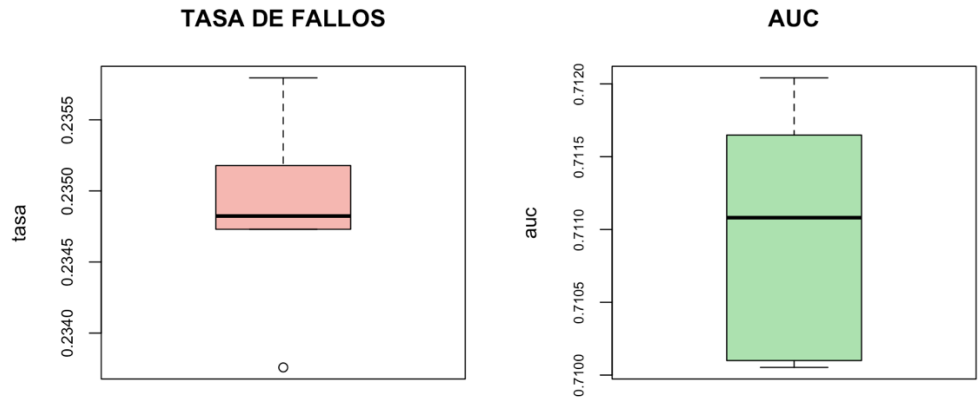
Aunque el modelo *Predi23* también es una opción válida, especialmente si se busca un equilibrio entre complejidad y precisión, *XGBoost* ofrece una solución más sencilla y eficaz, maximizando las ventajas mencionadas.

CAPÍTULO 10. DISCUSIÓN

Después de analizar todos los modelos, en este capítulo se procederá a analizar e interpretar los resultados obtenidos.

Se ha clasificado *XGBoost* como el mejor modelo para predecir si un vuelo sufrirá retraso. El Gráfico 30 muestra el AUC y tasa de fallos.

Gráfico 29. Tasa de fallos y AUC del modelo ganador



El gráfico de la izquierda muestra un diagrama de caja de la tasa de fallos del modelo que representa el porcentaje de predicciones incorrectas hechas por el modelo. La mediana de la tasa de fallos está alrededor de 0.245, lo que significa que el modelo está fallando en aproximadamente una cuarta parte de las predicciones. El rango intercuartílico es relativamente pequeño, lo que indica que la variabilidad en la tasa de

fallos no es muy alta. Además, se observa un valor atípico más bajo que indica una ejecución del modelo con un rendimiento mejor que las demás.

En el gráfico de la derecha se muestra el AUC. La mediana está alrededor de 0.711 lo cual es un indicador sólido de que el modelo tiene una buena capacidad de discriminación. El modelo predice con precisión la clase correcta en aproximadamente el 79.3% de los casos. Al igual que con la tasa de fallos, el rango intercuartílico es relativamente estrecho, lo que implica consistencia en el rendimiento del modelo en cuanto a esta métrica.

10.1. Matriz de confusión y análisis del modelo ganador

Como se ha adelantado, el modelo más adecuado para predecir si un vuelo sufrirá un retraso, basándose en las condiciones climatológicas del aeropuerto de origen o destino, es el modelo de *XGBoost*. La Figura 11 presenta la matriz de confusión correspondiente a este modelo.

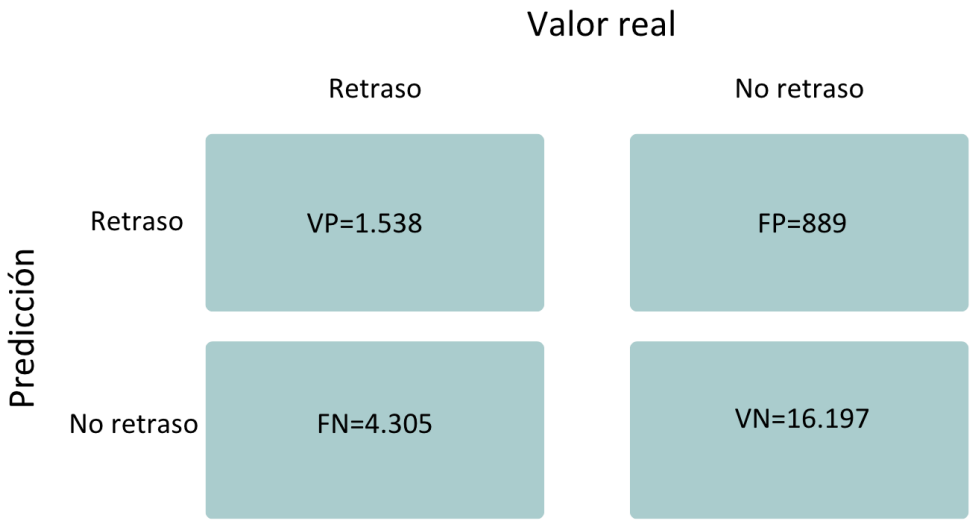


Figura 11. Matriz de confusión del modelo ganador

A partir de la matriz de confusión se derivan los siguientes resultados:

- **Precisión del modelo (*accuracy*):** El modelo tiene una precisión del 77.34%, lo que significa que predice correctamente en el 77.34% de las ocasiones. Se calcula utilizando la fórmula:

$$Accuracy = \frac{1.538 + 16.197}{1.538 + 16.197 + 889 + 4.305} = \frac{17.735}{22.929} = 0.7734$$

- **Tasa de fallos:** La tasa de fallos es del 22.65%, indicando que el modelo comete errores en el 22.65% de las predicciones. Se calcula con la fórmula:

$$Tasa\ de\ fallos\ = \frac{889 + 4.305}{1.538 + 16.197 + 889 + 4.305} = \frac{5.194}{22.929} = 0.2265$$

- **Especificidad:** La especificidad del modelo es del 94.79%, indicando que el modelo es muy eficaz en identificar correctamente los casos negativos. Se calcula de la siguiente manera:

$$Especifididad = \frac{16.197}{16.197 + 889} = \frac{16.197}{17.086} = 0.9479$$

- **Sensibilidad:** La sensibilidad del modelo es del 22.65%, lo que indica que tiene dificultades para identificar correctamente los casos positivos. Esta métrica se calcula como:

$$Sensibilidad = \frac{1.538}{1.538 + 4.305} = \frac{1.538}{5.843} = 0.2265$$

Como la sensibilidad es considerablemente baja mientras que la especificidad es alta, se ajusta el punto de corte para mejorar la detección de casos positivos y lograr un mejor equilibrio entre ambas métricas. Se prueban los puntos de corte de 0.3 y 0.4 Los resultados de precisión, especificidad y sensibilidad para estos puntos de corte se presentan en la Tabla 42.

Punto de Corte	Accuracy	Sensibilidad	Especificidad
0.3	72.36%	55.16%	78.24%
0.4	76.25%	39.22%	88.91%
0.5	77.34%	22.65%	94.79%

Tabla 42: Resultados de precisión, especificidad y sensibilidad para diferentes puntos de corte

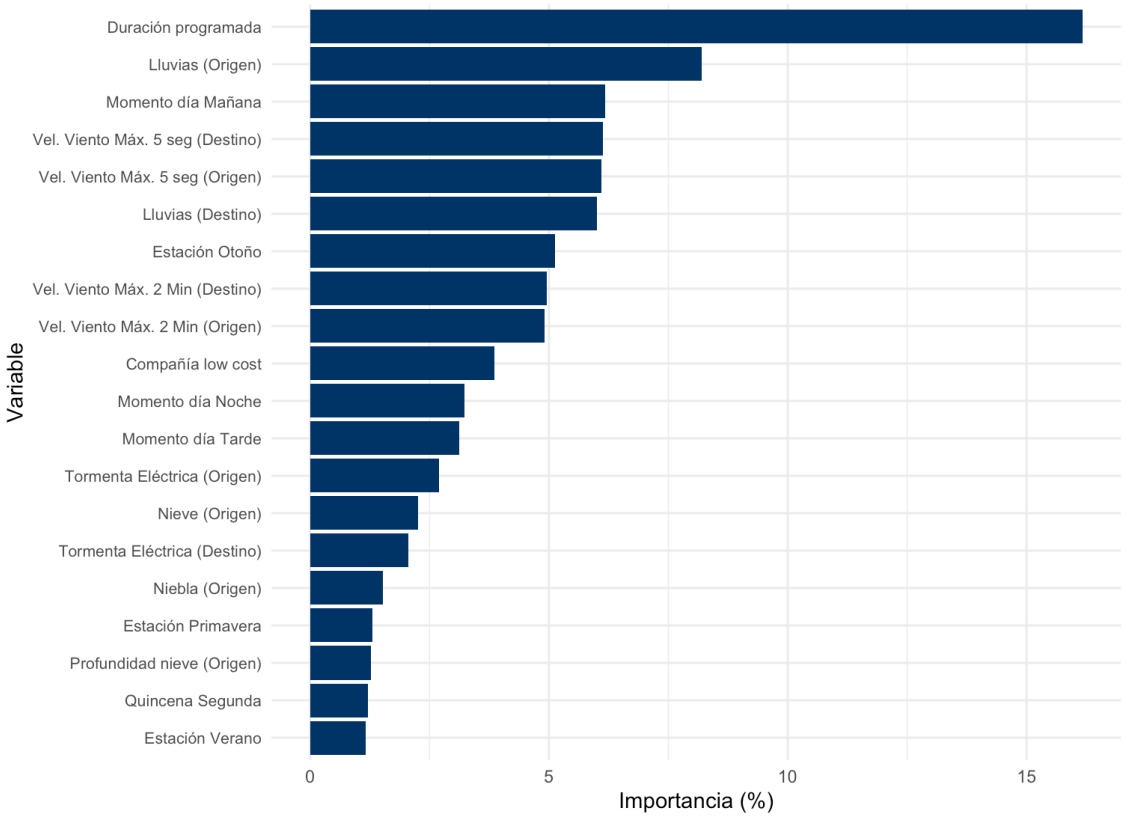
Con un punto de corte de 0.3, el modelo clasifica correctamente el 72.36% de las observaciones en general. La sensibilidad es del 55.16% y la especificidad es del 78.24%.

Al ajustar el punto de corte a 0.4, la exactitud del modelo mejora al 76.25%. Este aumento en la exactitud se debe a una reducción en los errores generales del modelo. La sensibilidad es 39.22%, pero la especificidad aumenta significativamente al 88.91%.

La elección del punto de corte depende de los objetivos específicos del modelo. En este trabajo, se prioriza identificar una mayor proporción de casos positivos, por lo que un punto de corte de 0.3 puede ser preferible a pesar de una reducción en del *accuracy*.

Para concluir el análisis del modelo ganador, se presenta el Gráfico 31, que muestra la importancia de las variables en el modelo utilizando los datos de prueba.

Gráfico 30. Importancia de las variables en el modelo ganador XGBoost



Se observa que la **duración programada del vuelo** es la variable más influyente en la predicción de retrasos. Esto puede indicar que los vuelos más largos tienden a experimentar más retrasos, posiblemente debido a una mayor probabilidad de acumulación de retrasos en el trayecto o a una mayor complejidad en la gestión de esos vuelos.

Sin embargo, esto plantea la cuestión de si los retrasos están directamente relacionados con las condiciones meteorológicas locales o si la correlación observada podría deberse a otros factores. Existen dos posibles escenarios a considerar:

- **Retraso debido a otras razones:** Los retrasos podrían no estar directamente vinculados con las condiciones meteorológicas locales, sino que podrían estar relacionados con factores diferentes, como las condiciones meteorológicas durante el vuelo o problemas operativos. En este caso, sería útil investigar si las condiciones meteorológicas durante el trayecto del vuelo influyen significativamente en los retrasos.
- **Influencia de aeropuertos con condiciones meteorológicas adversas:** Otra posibilidad es que los trayectos más largos entre aeropuertos puedan incluir aeropuertos con condiciones meteorológicas desfavorables. Esto podría explicar la correlación observada entre retrasos y condiciones meteorológicas locales, ya que los trayectos más largos podrían pasar por aeropuertos con peores condiciones.

Adicionalmente, las variables climatológicas con mayor impacto en la probabilidad de retraso, ordenadas de mayor a menor influencia, son las siguientes:

1. **Precipitaciones en el origen:** 8.19%
2. **Velocidad del viento más rápida en 5 segundos en el origen y destino** : 6.1 %
3. **Precipitaciones en el destino:** 6%
4. **Velocidad del viento más rápida en 2 minutos en el destino y origen**: 4.9%
5. **Tormenta eléctrica (truenos) en el origen:** 2.69 %
6. **Nieve en el origen:** 2.25%
7. **Tormenta eléctrica (truenos) en el destino:** 2.05 %
8. **Niebla en el origen:** 1.526%
9. **Profundidad de la nieve en el origen:** 1.26%

También se consideran variables no climáticas, como:

1. **Estación otoño:** 5.13%
2. **Compañía *low cost*:** 3.85%
3. **Momento del día: tarde** 3.11%
4. **Estación del año: primavera**: 1.31%

Estos resultados indican que las condiciones meteorológicas, particularmente las *precipitaciones* y la *velocidad del viento*, tienen un efecto significativo en la probabilidad de retraso de un vuelo. La velocidad del viento en intervalos cortos tanto en el destino como en el origen, juega un papel crucial en los retrasos. Las precipitaciones en ambos lugares también influyen notablemente en los retrasos, además de algunos factores no climáticos que también afectan la probabilidad de retraso.

Por otro lado, la *estación del año* presenta cierta ambigüedad en cuanto a su clasificación como variable climática. Aunque puede estar relacionada con condiciones meteorológicas, como temperaturas más bajas en otoño o mayores precipitaciones en primavera, se ha excluido del conjunto de variables climáticas, como medida de precaución.

Compañías low cost y el *momento del día* reflejan la importancia de los factores operativos en la predicción de retrasos. Las aerolíneas de bajo coste suelen tener modelos de negocio que priorizan la reducción de gastos, lo que puede implicar ciertas limitaciones operativas. Por ejemplo, suelen operar con menos márgenes de tiempo entre vuelos (tiempo de giro más corto), lo que significa que cualquier retraso pequeño puede tener un impacto acumulativo. Además, la falta de personal o la dependencia de infraestructuras aeroportuarias más limitadas en aeropuertos secundarios puede agravar los retrasos.

Por otro lado, el *momento del día* es otro factor importante. Los vuelos que salen en la tarde tienen más probabilidades de experimentar retrasos posiblemente debido a la acumulación de retrasos en vuelos anteriores a lo largo del día.

CAPÍTULO 11. CONCLUSIONES

Este estudio se propuso investigar cómo las condiciones climatológicas en los aeropuertos de origen y destino influyen en los retrasos de los vuelos. Para abordar esta cuestión, se seleccionaron datos de vuelos y clima de cuatro aeropuertos estadounidenses con climas diversos: Atlanta, Chicago, Denver y Los Ángeles. Se combinó esta información con datos meteorológicos específicos de cada aeropuerto, resultando en una base de datos compuesta por 76 433 registros y 65 variables.

Posteriormente, se llevó a cabo una depuración y análisis de los datos, seguido de una selección de variables relevantes. Se aplicaron diversos modelos de aprendizaje supervisado, y se compararon los mejores modelos de cada técnica para determinar el óptimo para predecir retrasos en función de las condiciones climatológicas.

El modelo de *XGBoost* se destacó como el clasificador más efectivo, optimizando su rendimiento mediante el ajuste de distintos criterios. En particular, se utilizó el tamaño de muestra completo para la construcción de cada árbol, se estableció un mínimo de 5 observaciones por nodo terminal, y se realizaron 5 000 iteraciones para asegurar un entrenamiento adecuado del modelo.

Las principales conclusiones del estudio son:

1. **Eficiencia del modelo de *XGBoost*:** El modelo de *XGBoost* mostró un rendimiento superior en comparación con otros enfoques, con un *accuracy* de 0.7236, una sensibilidad de 0.5516 y una especificidad de 0.7824 en datos no vistos.

2. **Efecto de las precipitaciones:** Las precipitaciones en el aeropuerto de origen son la variable climática más influyente, con un impacto del 8.19% en la predicción de retrasos. En el destino, las precipitaciones también juegan un papel importante, aunque menor, con una influencia del 6%.

3. **Impacto de la velocidad del viento:** Las variables que miden la velocidad del viento en intervalos cortos, tanto en el origen como en el destino, tienen una influencia significativa en los retrasos, con un peso del 6.1% en la predicción.

4. **Tormentas eléctricas.** Las tormentas eléctrica influyen en la probabilidad de retrasos, con una importancia del 2.69% en el origen y del 2.05% en el destino.

5. **Influencia de variables no climáticas:** Aunque tienen un impacto menor, factores como si el vuelo es por la tarde o si la aerolínea es *low cost* también contribuyen a la probabilidad de retrasos.

6. **Otras variables climáticas con menor influencia:** La presencia de nieve y la profundidad de la misma en el origen, aunque con menos impacto juegan un papel en la probabilidad de retrasos.

11.1. Líneas futuras de investigación

A pesar de los avances logrados con el modelo *XGBoost* para la predicción de retrasos en vuelos, existen posibles líneas futuras de investigación que podrían mejorar aún más la precisión y utilidad de las predicciones.

Una opción es incorporar más variables no climatológicas, como factores operacionales, tales como la eficiencia en el manejo de equipaje y la disponibilidad de pistas de aterrizaje. Estos elementos pueden proporcionar una visión más completa del fenómeno de los retrasos.

No obstante, si el objetivo es analizar exclusivamente el impacto del clima, podría ser sería recomendable excluir variables no meteorológicas para una evaluación más precisa de cómo las condiciones climáticas afectan los retrasos.

Otra línea de investigación interesante sería clasificar los retrasos en la salida, el destino o durante el trayecto. Al incorporar factores adicionales, como la presencia de turbulencias u otros eventos en vuelo, se podría mejorar la precisión de las predicciones de retrasos.

Estas recomendaciones buscan no solo mejorar la precisión del modelo, sino también ampliar su utilidad en diversos contextos operativos, contribuyendo a una gestión de vuelos más eficaz y una experiencia de viaje más satisfactoria para los pasajeros.

BIBLIOGRAFÍA

Aviation Intelligence Portal. (2022). *Aircraft operating costs*. ANS Performance. Recuperado de: https://ansperformance.eu/economics/cba/standard-inputs/chapters/cost_of_delay.html.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bombelli, A., & Sallan, J. M. (2023). Analysis of the effect of extreme weather on the US domestic air network. A delay and cancellation propagation network approach. *Journal of Transport Geography*, *107*, 103541.

Bureau of Transportation Statistics. (2023). Recuperado de: https://www.bts.gov/.

Federal Aviation Administration. (2024). *Weather Delay Information*. Recuperado de: https://www.faa.gov/nextgen/programs/weather/faq.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical + learning: With applications in R*. Springer.

Moreno Maderuelo, A. M. (2022). Análisis y clasificación del retraso de los vuelos comerciales de Estados Unidos en 2019.

National Centers for Environmental Information. (2023). *Billion-Dollar Weather and Climate. Disasters: Events. Recuperado de:* https://www.ncei.noaa.gov/access/billions/events/US/2023?disasters[]=all-disasters.

NOAA. (2023). National Oceanic and Atmospheric Administration. Recuperado de: https://www.noaa.gov/.

Pejovic, T., Williams, V. A., Noland, R. B., & Toumi, R. (2009). Factors affecting the frequency and severity of airport weather delays and the implications of climate change for future delays. *Transportation research record*, *2139*(1), 97-106.

Pich, S., Kulkarni, P., Paul, R., & Durga, M. (2024). A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. *Functional & Integrative Genomics*. Recuperado de: https://doi.org/10.1007/s10142-024-01415-x.

Priyanka, G. (2018). Prediction of airline delays using K-nearest neighbor algorithm. *Int. J. Emerg. Technol. Innov. Eng*, *4*(5), 87-90.

Quinta Evaluación Nacional del Clima de EE.UU. (2023). *Informe de evaluación nacional sobre el clima*.

SAS Institute. (1998). *Data mining and the case for sampling*. Recuperado de: http://www.sasenterpriseminer.com/documents/SAS-SEMMA.pdf.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research,* 15(1), 1929-1958.

Statista. (2023). Growth of global air traffic passenger demand. Recuperado de: https://www.statista.com/statistics/193533/growth-of-global-air-traffic-passenger-demand/.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). *The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm*. *Machine Learning*, 31-78.

Weatherspark. *Clima promedio en Estados Unidos. Recuperado de: https://es.weatherspark.com/*

Xu, N., Donohue, G., Laskey, K. B., & Chen, C. H. (2005, June). Estimation of delay propagation in the national aviation system using Bayesian networks. In 6th USA/Europe Air Traffic Management Research and Development Seminar. FAA and Eurocontrol Baltimore.

ANEXO

A. Variables

Nombre de las variables en RStudio y su definición

Nombre variable en Rstudio	Variable
AEROPUERTO	Aeropuerto de origen
AEROP_DESTINO	Aeropuerto destino
DURACION_PROGRAMADA	Duración del vuelo programada
DURACION_REAL	Duración del vuelo real
HORA_SALIDA_PROGRAMADA	La hora programada a la que una aeronave debe despegar del aeropuerto de origen.
HORA_SALIDA_REAL	La hora de la salida real es la hora en que el piloto libera el freno de estacionamiento del avión
DEMORA_SALIDA	Demora en la salida
FECHA	Fecha
Código_IATA	Código identificador de una aerolinea
TIEMPO_RODAJE	El tiempo transcurrido entre la salida de la puerta del aeropuerto de origen y el inicio del viaje (min)
HORA_DESPEGUE	Hora de despegue en la que el avión deja de tocar el suelo
N_VUELO	Número de vuelo
TAIL_NUMBER	Número de matrícula de una aeronave, es un código alfanumérico distintivo asignado a cada aeronave como identificador único
O_AWND	Velocidad media diaria del viento en el origen
D_AWND	Velocidad media diaria del viento en el destino
O_PRCP	Precipitación en el origen
D_PRCP	Precipitación en el destino
O_TAVG	Temperatura media en el origen
D_TAVG	Temperatura media en el destino
O_TMAX	Temperatura máxima en el origen
D_TMAX	Temperatura máxima en el origen
O_TMIN	Temperatura mínima en el origen
D_TMIN	Temperatura mínima en el destino
O_WSF2	Velocidad del viento más rápida en 2 minutos en el origen
D_WSF2	Velocidad del viento más rápida en 2 minutos en el destino
O_WSF5	Velocidad del viento más rápida en 5 segundos en el origen
D_WSF5	Velocidad del viento más rápida en 5 segundos en el destino
O_SNOW	Nieve caída en el origen
D_SNOW	Nieve caída en el destino
D_SNOW	Nieve caída en el destino
O_SNWD	Profundidad de la nieve en el origen
D_SNWD	Profundidad de la nieve en el destino

Nombre variable en Rstudio	Variable
O_WDF2	Direccion del viento más rapido en 2 minutos en el origen
D_WDF2	Direccion del viento más rapido en 2 minutos en el destino
O_WDF5	Direccion del viento más rapido en 5 segundos en el origen
D_WDF5	Direccion del viento más rapido en 5 segundos en el destino
O_WT01	Niebla, niebla de hielo o niebla helada en el origen
D_WT01	Niebla, niebla de hielo o niebla helada en el destino
O_WT02	Niebla densa o niebla helada densa en el origen
D_WT02	Niebla densa o niebla helada densa en el destino
O_WT03	Tormenta eléctrica (truenos) en el origen
D_WT03	Tormenta eléctrica (truenos) en el destino
O_WT04	Pellets de hielo, aguanieve, gránulos de nieve o granizo pequeño en el origen
D_WT04	Pellets de hielo, aguanieve, gránulos de nieve o granizo pequeño en el destino
O_WT05	Granizo en el origen
D_WT05	Granizo en el destino
O_WT06	Escarcha o cencellada en el origen
D_WT06	Escarcha o cencellada en el destino
O_WT07	Polvo, ceniza volcánica, polvo soplado, arena soplada, o algún obstáculo soplado en el origen
D_WT07	Polvo, ceniza volcánica, polvo soplado, arena soplada, o algún obstáculo soplado en el destino
O_WT08	Humo o neblina en el origen
D_WT08	Humo o neblina en el destino
O_WT09	Nieve soplada o arrastrada por el viento en el origen
D_WT09	Nieve soplada o arrastrada por el viento en el destino
O_WT10	Tornado, tromba marina, o nube embudo en el origen
D_WT10	Tornado, tromba marina, o nube embudo en el destino

Variables , tipo de variable, unidades y tratamiento.

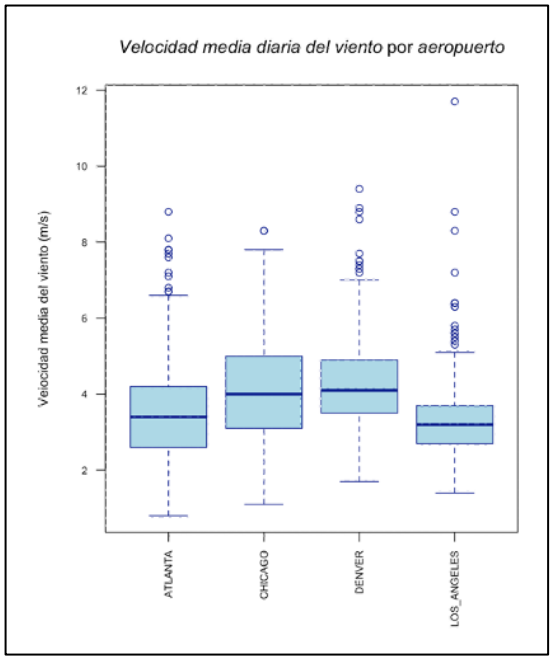
Variable	Unidades	Tipo de Variable	Tratamiento de la variable
Precipitación	mm	Cuantitativa continua	Se mantiene
Temperatura media	Grados Celsius	Cuantitativa continua	Se mantiene
Temperatura máxima	Grados Celsius	Cuantitativa continua	Se mantiene
Temperatura mínima	Grados Celsius	Cuantitativa continua	Se mantiene
Velocidad del viento más rápida en 2 minutos	m/s	Cuantitativa continua	Se mantiene
Velocidad del viento más rápida en 5 segundos	m/s	Cuantitativa continua	Se mantiene
Nieve caída	mm	Cuantitativa discreta	Se mantiene
Profundidad de la nieve	mm	Cuantitativa discreta	Se mantiene
Direccion del viento más rapido en 2 minutos	Grados	Cuantitativa discreta	Se mantiene
Direccion del viento más rapido en 5 segundos	Grados	Cuantitativa discreta	Se mantiene

Variable	Unidades	Tipo de Variable	Tratamiento de la variable
Fecha	YYYY/MM/DD	Cuantitativa	Se elimina
Nombre de la estación climatológica	n.a	Cualitativa	Se elimina
Latitud de la estación	Grados (°)	Cuantitativa continua	Se elimina
Longitud de la estación	Grados (°)	Cuantitativa continua	Se elimina
Elevación de la estación	Metros sobre el nivel del mar	Cuantitativa continua	Se elimina
Niebla, niebla de hielo o niebla helada (puede incluir niebla densa)	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Niebla densa o niebla helada densa (no siempre distinguida de la niebla)	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Tormenta eléctrica (truenos)	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Pellets de hielo, aguanieve, gránulos de nieve o granizo pequeño	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Granizo (puede incluir granizo pequeño)	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Escarcha o cencellada	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Polvo, ceniza volcánica, polvo soplado, arena soplada, o algún obstáculo soplado	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Humo o neblina	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Nieve soplada o arrastrada por el viento	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Tornado, tromba marina, o nube embudo	1: evento 0:no evento	Cualitativa binaria	Se mantiene
Velocidad media diaria del viento	m/s	Cuantitativa continua	Se mantiene

B. Tablas y gráficos

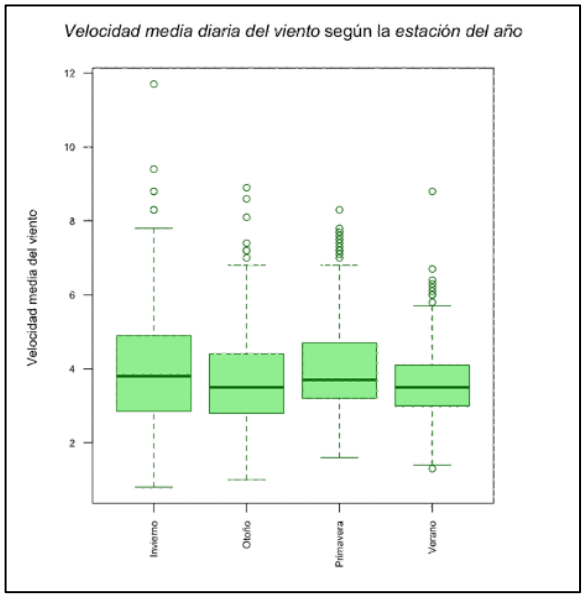
Atípicos

Atípicos por <u>aeropuerto</u> y <u>velocidad media diaria del viento</u>



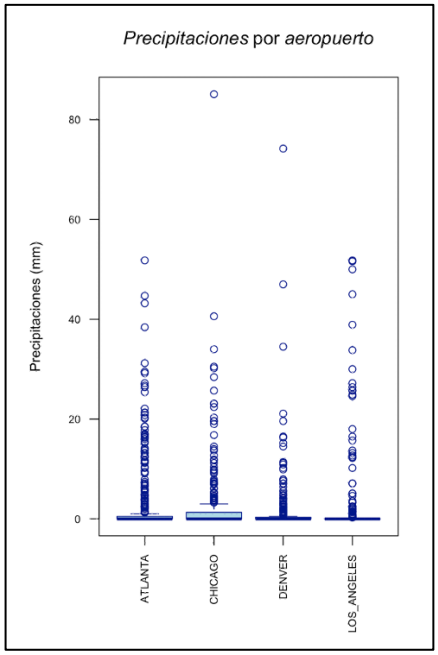
Atípicos en velocidad media diaria del viento			
Atlanta	Chicago	Denver	Los Ángeles
12	3	11	18

Atípicos por <u>estación</u> y <u>velocidad media diaria del viento</u>



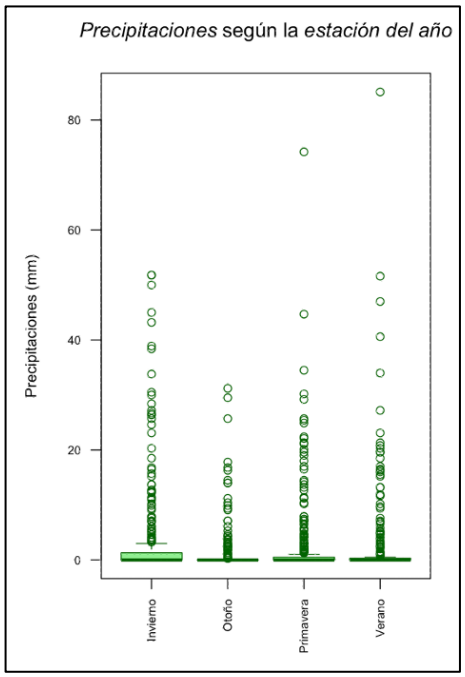
Atípicos en velocidad media diaria del viento			
Invierno	Otoño	Primavera	Verano
7	7	14	11

Atípicos por aeropuerto y precipitaciones



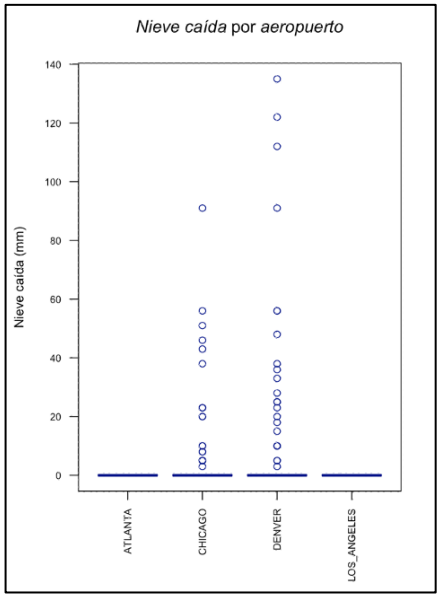
Atípicos en las precipitaciones			
Atlanta	Chicago	Denver	Los Ángeles
83	63	70	52

Atípicos por estación y precipitaciones



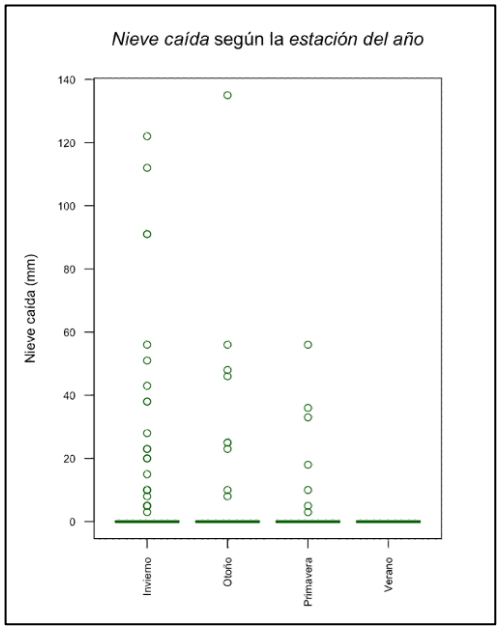
Atípicos en las precipitaciones			
Invierno	Otoño	Primavera	Verano
69	61	81	76

Atípicos por aeropuerto y nieve



Atípicos en la nieve caída			
Atlanta	Chicago	Denver	Los Ángeles
0	22	24	0

Atípicos por estación y nieve



Atípicos en la nieve caída			
Invierno	Otoño	Primavera	Verano
30	9	7	0

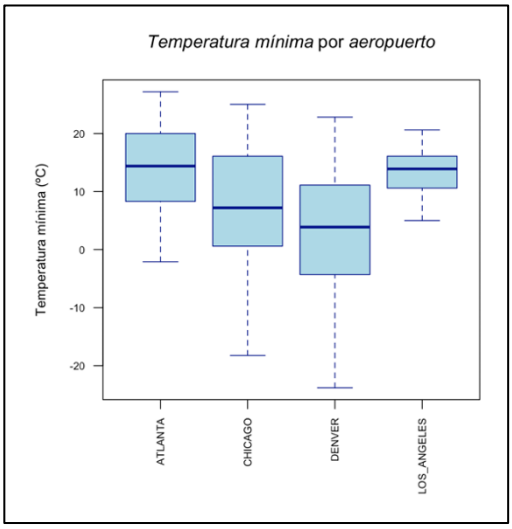
Atípicos por aeropuerto y temperatura media

Atípicos en la temperatura media			
Atlanta	Chicago	Denver	Los Ángeles
0	0	0	0

Atípicos por estación y temperatura media

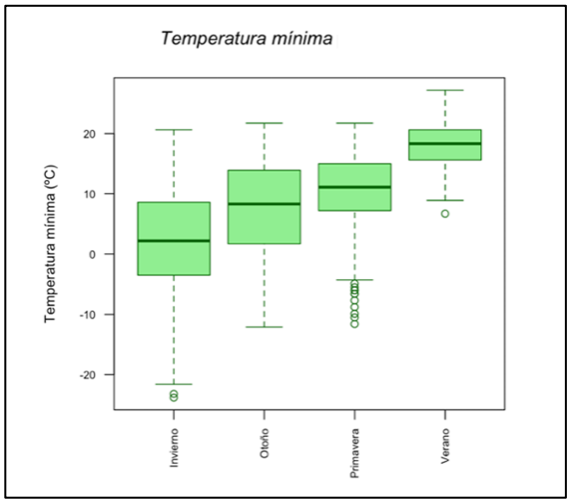
Atípicos en la temperatura media			
Invierno	Otoño	Primavera	Verano
2	1	17	0

Atípicos por aeropuerto y temperatura mínima



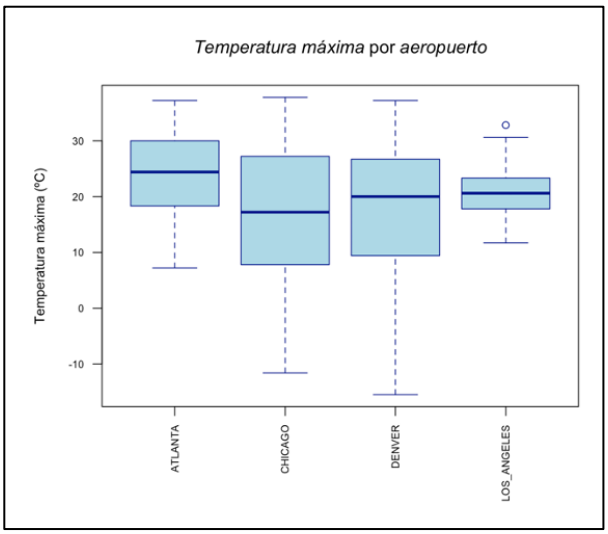
Atípicos en la temperatura mínima			
Atlanta	Chicago	Denver	Los Ángeles
0	0	0	0

Atípicos por estación y temperatura mínima



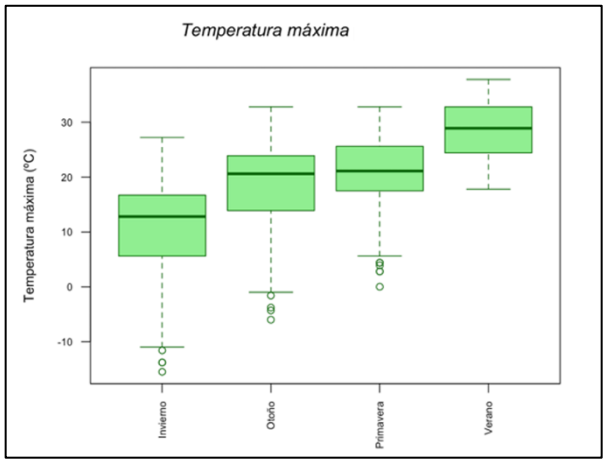
Atípicos en la temperatura mínima			
Invierno	Otoño	Primavera	Verano
2	0	11	1

Atípicos por aeropuerto y temperatura máxima



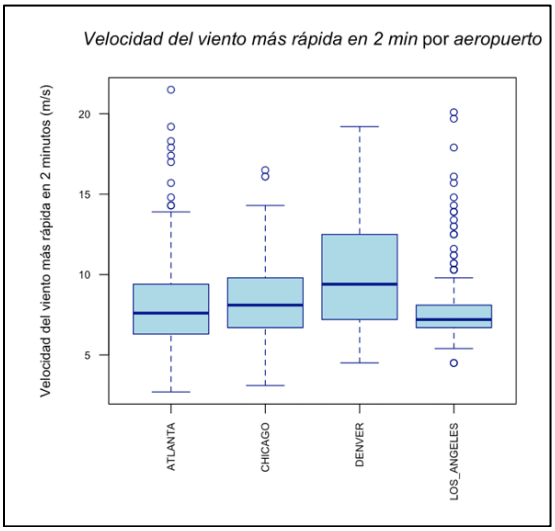
Atípicos en la temperatura máxima por aeropuerto			
Atlanta	Chicago	Denver	Los Ángeles
0	0	0	1

Atípicos por estación y temperatura máxima



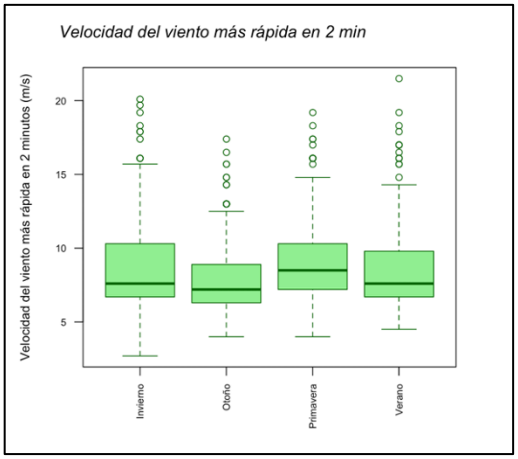
Atípicos en la temperatura máxima por aeropuerto			
Invierno	Otoño	Primavera	Verano
10	16	12	0

Atípicos por aeropuerto y velocidad del viento más rápida en 2 minutos



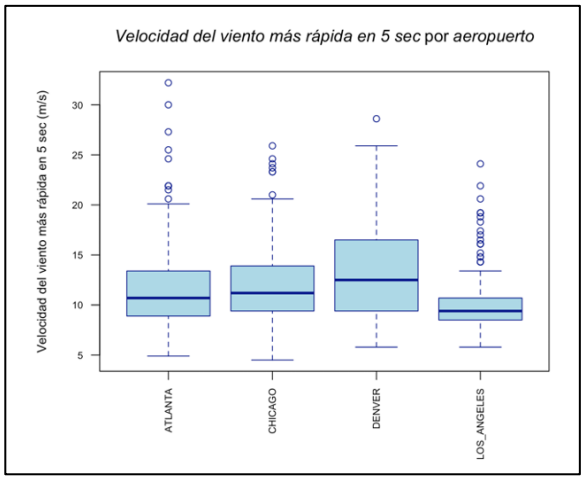
Atípicos en la velocidad del viento más rápida en 2 minutos			
Atlanta	Chicago	Denver	Los Ángeles
11	3	0	34

Atípicos por estación y velocidad del viento más rápida en 2 minutos



Atípicos en la velocidad del viento más rápida en 2 minutos			
Invierno	Otoño	Primavera	Verano
4	4	6	0

Atípicos por aeropuerto y velocidad del viento más rápida en 5 segundos



Atípicos por estación y velocidad del viento más rápida en 5 segundos

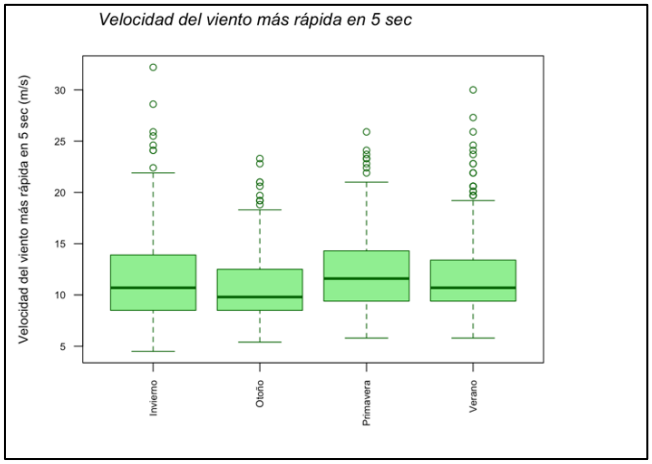


Gráfico de dispersión. Velocidad media del viento y velocidad del viento más rápida en dos minutos

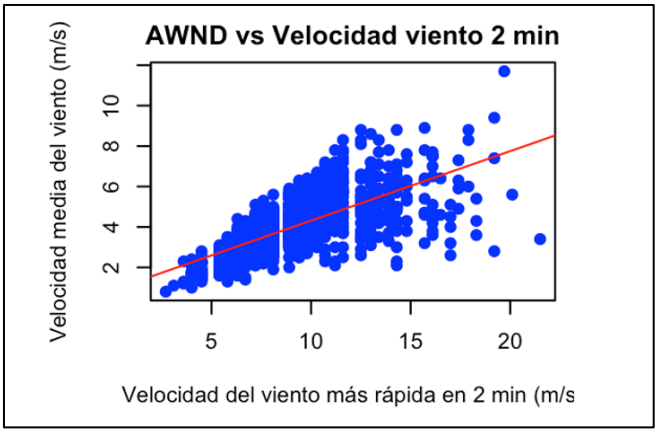


Gráfico de dispersión. Velocidad media del viento y velocidad del viento más rápida en 5 segundos.

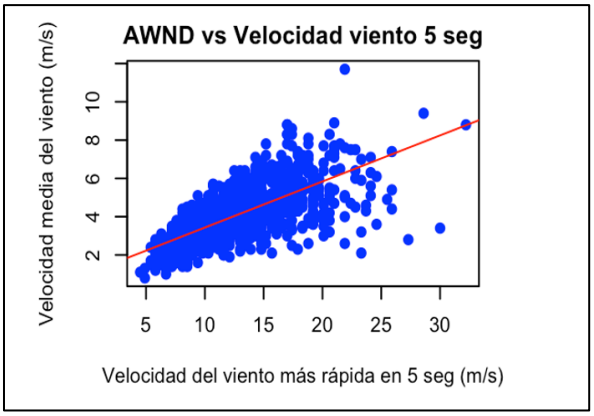


Gráfico de dispersión. Temperatura media y temperatura máxima

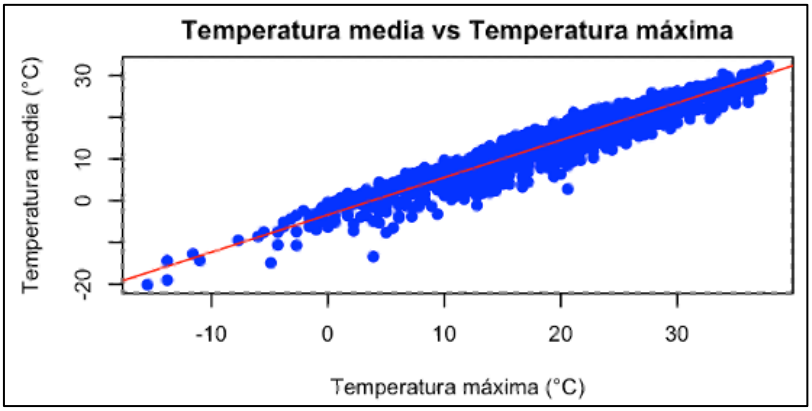


Gráfico de dispersión. Temperatura media y temperatura mínima

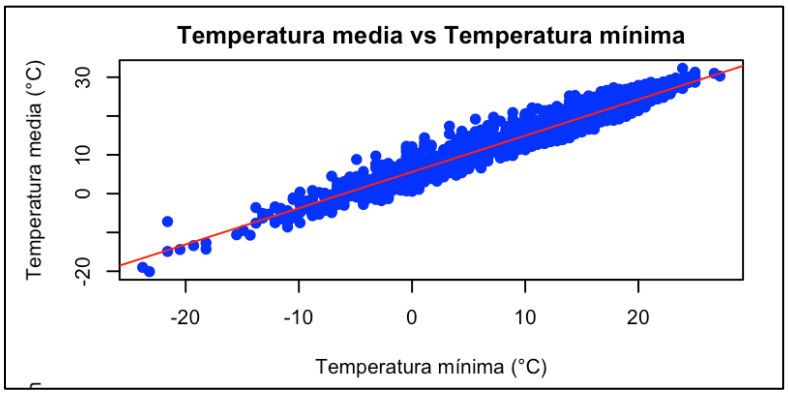


Gráfico de dispersión. Temperatura máxima y temperatura mínima

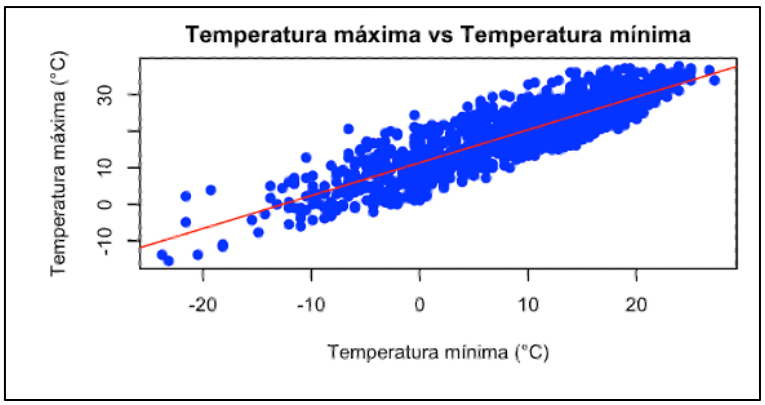


Gráfico de dispersión. Velocidad del viento más rápida en 2 minutos y 5 segundos

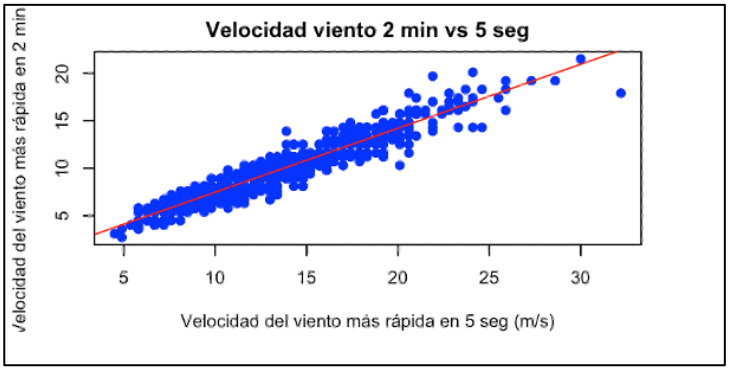


Gráfico de dispersión. Velocidad media del viento y temperatura media

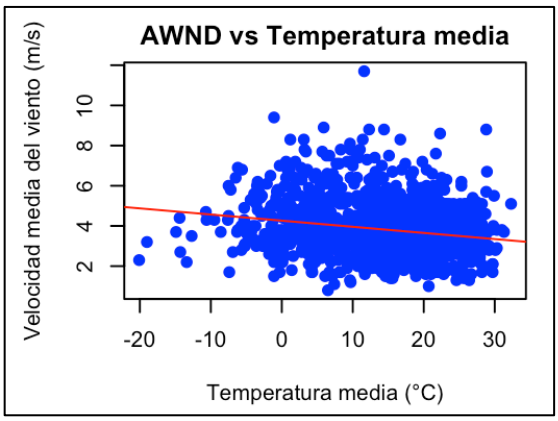


Gráfico de dispersión. Velocidad media del viento y temperatura media

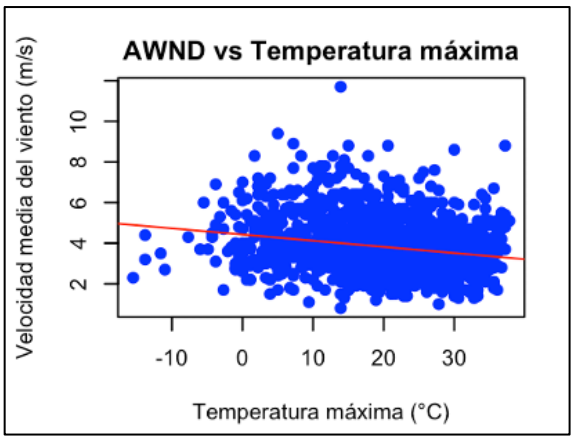


Gráfico de dispersión. Velocidad media del viento y temperatura mínima

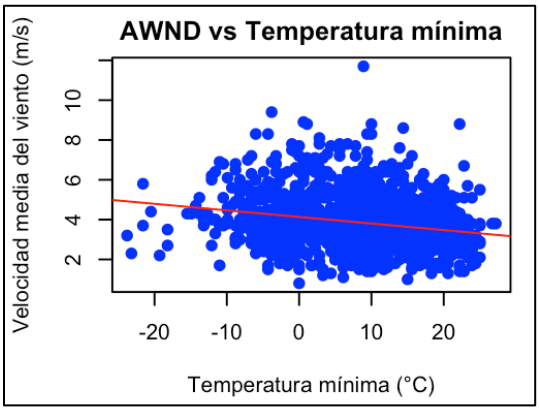


Gráfico de dispersión. Velocidad media del viento y precipitaciones

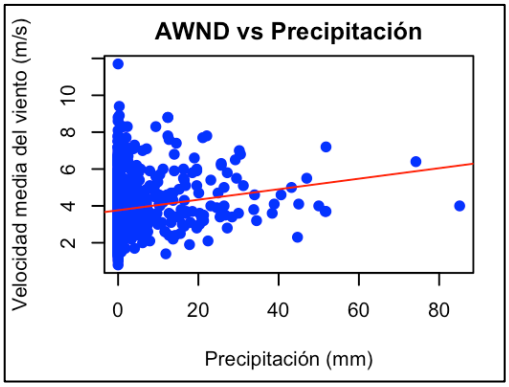


Gráfico de dispersión. Precipitaciones y temperatura media

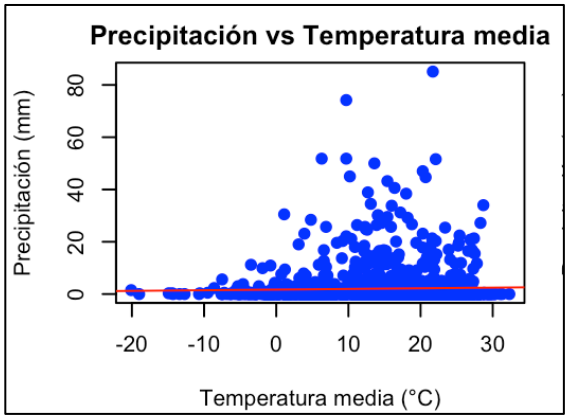


Gráfico de dispersión. Precipitaciones y temperatura máxima

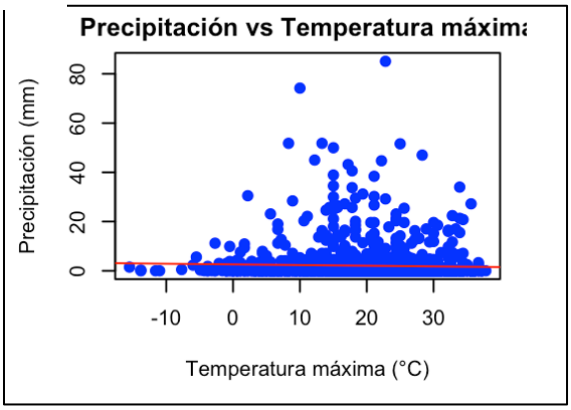


Gráfico de dispersión. Precipitaciones y temperatura mínima

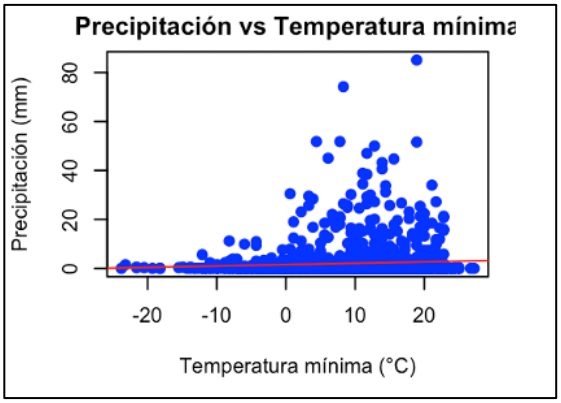


Gráfico de dispersión. Precipitaciones y velocidad del viento más rápida en 2 minutos

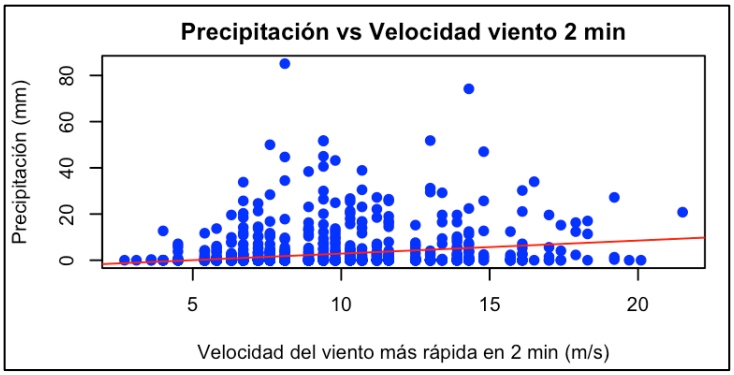
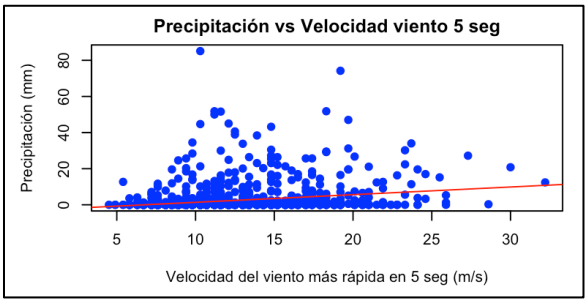


Gráfico de dispersión. Precipitaciones y velocidad del viento más rápida en 5 segundos



C. Métodos de selección de variables

Variables escogidas por todos los métodos (7), 6 y 5 y 4

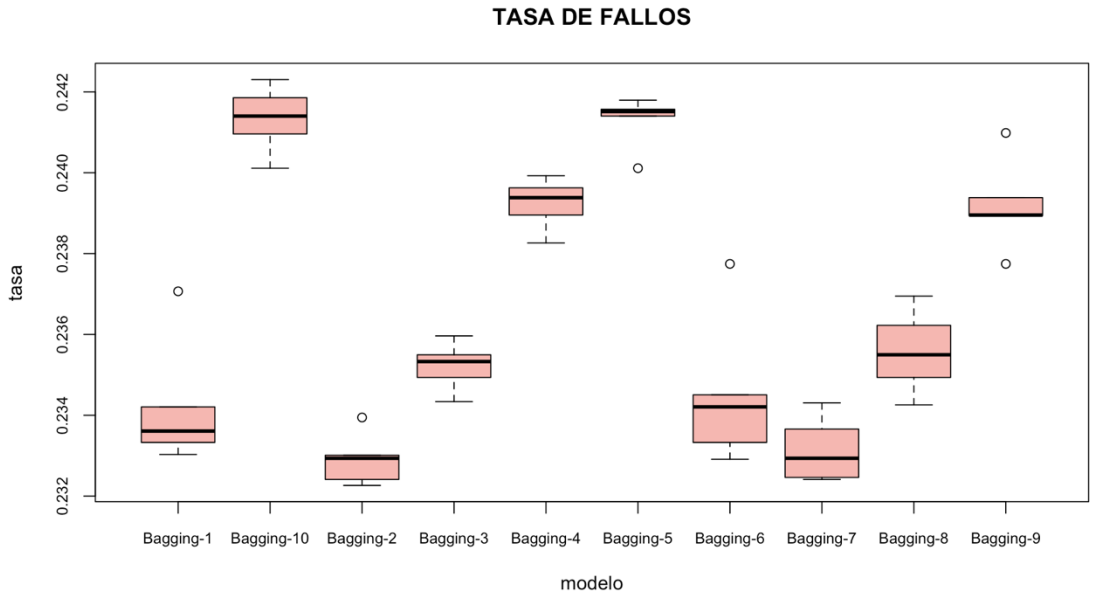
Métodos de selección de variables. Variables escogidas por todos los métodos, 6, 5, y 4.								
Variable	Stepwise AIC	Stepwise BIC	Boruta	Eliminación Recursiva con criterio Naive Bayes (NB)	Max-Min Parent Child (MMPC)	Subset Evaluation and Selection (SES)	Eliminación Recursiva con criterio Random Forest	Nº de veces que ha sido seleccionada la variable
Momento del día	Sí	Sí	Sí	Sí	Sí	Sí	Sí	7
Estación del año	Sí	Sí	Sí	Sí	Sí	Sí	Sí	7
Precipitaciones en el destino	Sí	Sí	Sí	Sí	Sí	Sí	Sí	7
Precipitaciones en el origen	Sí	Sí	Sí	Sí	Sí	Sí	Sí	7
Tormenta en el origen	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Nieve en el origen	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Compañía low cost	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Niebla en el origen	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Tormenta en el destino	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Aeropuerto de origen	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Día de la semana	Sí	Sí	Sí	No	Sí	Sí	Sí	6
Velocidad del viento más rápida en 5 sec en el origen	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Profundidad de la nieve en el origen	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Niebla en el destino	Sí	Sí	Sí	Sí	Sí	Sí	No	6
Nieve en el origen	Sí	Sí	Sí	No	Sí	Sí	No	5
Aeropuerto de destino	Sí	Sí	Sí	No	Sí	Sí	No	5
Escarcha o cencellada en el origen	Sí	Sí	Sí	No	Sí	Sí	No	5
Quincena del mes	Sí	No	Sí	Sí	Sí	Sí	No	5
Velocidad del viento más rápida en 2 minutos en el destino	Sí	Sí	Sí	Sí	No	No	No	4
Humo o neblina en el origen	Sí	Sí	Sí	Sí	No	No	No	4
Humo o neblina en el destino	Sí	Sí	Sí	Sí	No	No	No	4
Velocidad del viento más rápida en 2 minutos en el origen	Sí	Sí	Sí	Sí	No	No	No	4
Duración programada del vuelo	Sí	No	Sí	Sí	No	No	Sí	4
Escarcha o cencellada en el destino	Sí	Sí	No	No	Sí	Sí	No	4
Velocidad del viento más rápida en 2 minutos en el destino	No	No	Sí	Sí	Sí	Sí	No	4

Variables escogidas por todos 3,2,1 o ningún método

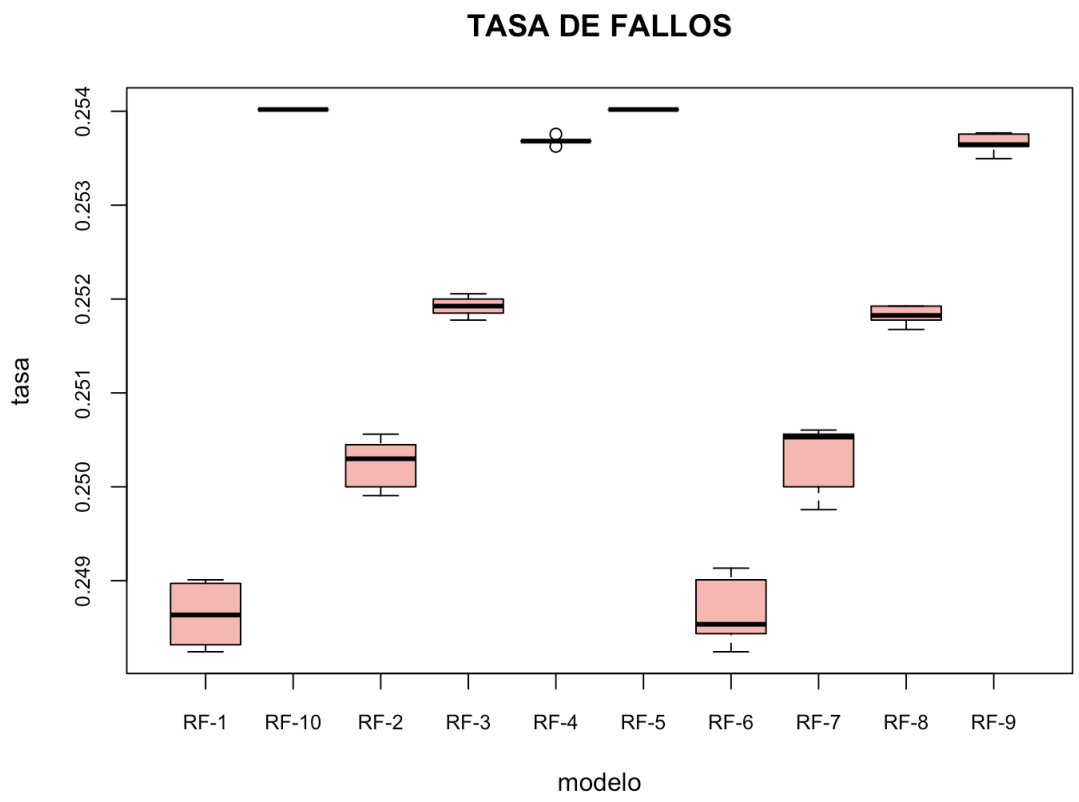
Métodos de selección de variables. Variables escogidas por 3, 2, 1 y ningún método								
Variable	Stepwise AIC	Stepwise BIC	Boruta	Eliminación Recursiva con criterio Naive Bayes (NB)	Max-Min Parent Child (MMPC)	Subset Evaluation and Selection (SES)	Eliminación Recursiva con criterio Random Forest	Nº de veces que ha sido seleccionada la variable
Dirección del viento más rapido en 2 min en el destino	Sí	Sí	Sí	No	No	No	No	3
Temperatura mín en el origen	Sí	No	Sí	Sí	No	No	No	3
Temperatura media en el origen	Sí	No	Sí	Sí	No	No	No	3
Temperatura máxima en el origen	No	No	Sí	Sí	No	No	Sí	3
Pellets de hielo, aguanieve, gránulos de nieve o granizo pequeño en el origen	No	No	Sí	No	Sí	Sí	No	3
Granizo en el destino	Sí	No	Sí	No	No	No	No	2
Profundidad de la nieve en el destino	Sí	No	Sí	No	No	No	No	2
Dirección del viento más rapido en 5 sec en el destino	Sí	No	Sí	No	No	No	No	2
Niebla densa o niebla helada en el origen	Sí	No	Sí	No	No	No	No	2
Velocidad media diaria del viento en el origen	Sí	No	Sí	No	No	No	No	2
Velocidad media diaria del viento en el destino	No	No	Sí	Sí	No	No	No	2
Temperatura media en el destino	No	No	Sí	Sí	No	No	No	2
Temperatura máxima en el destino	No	No	Sí	Sí	No	No	No	2
Temperatura mínima en el destino	No	No	Sí	Sí	No	No	No	2
Polvo, ceniza volcánica, polvo, arena o algún obstáculo soplado en el origen	Sí	No	No	No	No	No	No	1
Polvo, ceniza volcánica, polvo, arena o algún obstáculo soplado en el destino	Sí	No	No	No	No	No	No	1
Tornado, tromba marina o nube embudo en el origen	Sí	No	No	No	No	No	No	1
Niebla densa o niebla helada en el destino	No	No	Sí	No	No	No	No	1
Dirección del viento más rapido en 2 min en el origen	No	No	Sí	No	No	No	No	1
Dirección del viento más rapido en 5 sec en el origen	No	No	Sí	No	No	No	No	1
Granizo en el origen	No	No	Sí	No	No	No	No	0
Nieve soplada o arrastrada por el viento en el origen	No	No	Sí	No	No	No	No	0
Pellets de hielo, aguanieve, gránulos de nieve o granizo pequeño en el destino	No	No	No	No	No	No	No	0
Nieve soplada o arrastrada por el viento en el destino	No	No	No	No	No	No	No	0
Tornado, tromba marina o nube embudo en el destino	No	No	No	No	No	No	No	0

D. Modelización

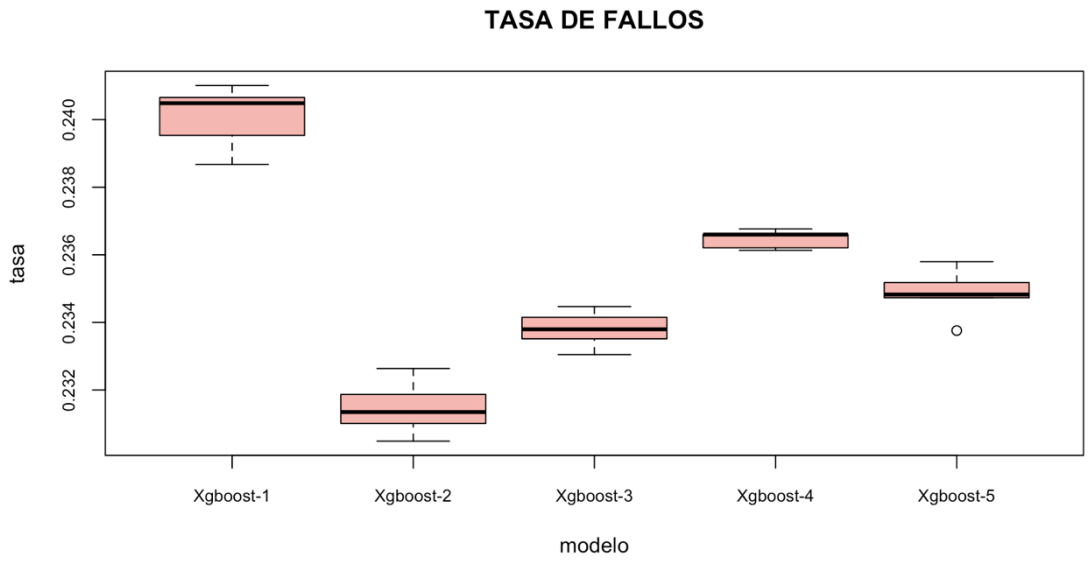
Tasa de fallos de los modelos *Bagging*



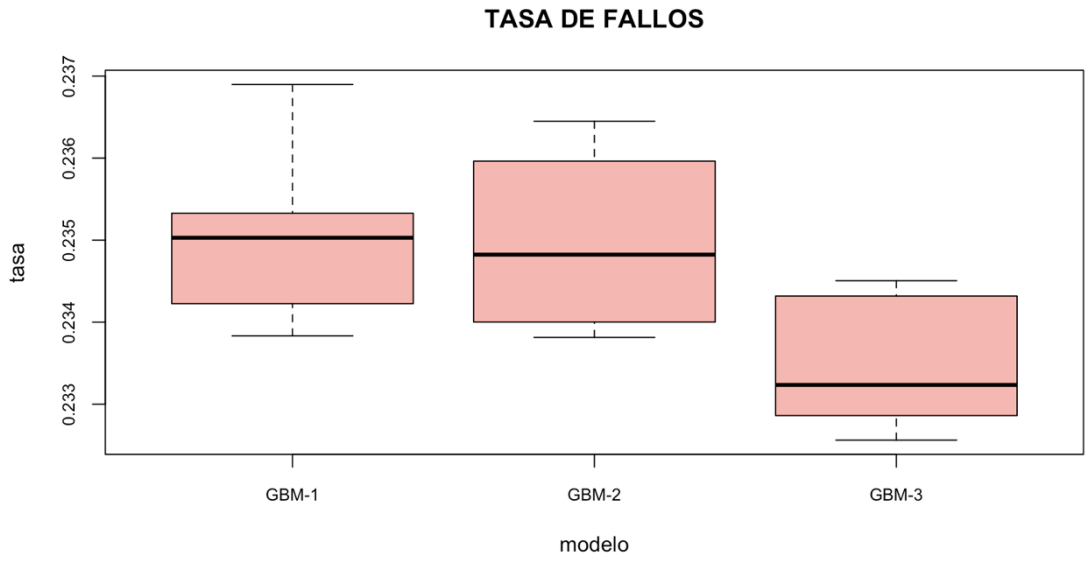
Tasa de fallos de los modelos *Random Forest*



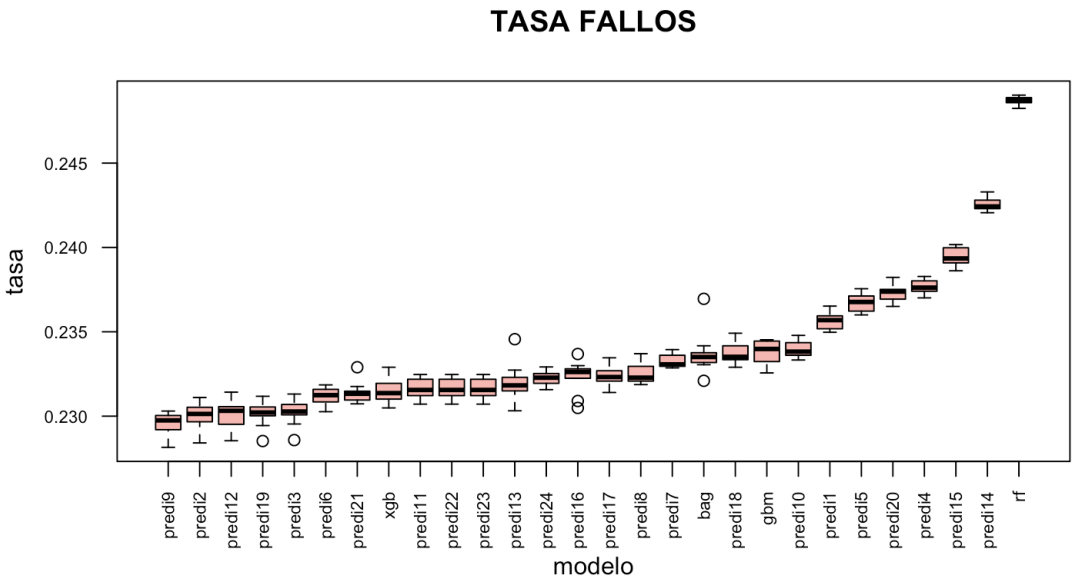
Tasa de fallos de los modelos *XGBoost*



Tasa de fallos de los modelos *Gradient Boosting*



Tasa de fallos de los modelos de *ensamblado*



Tasa de fallos de los mejores modelos de *ensamblado*