COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales
ICADE

Análisis y predicción de retrasos en la salida de vuelos a partir de condiciones climáticas en el aeropuerto de origen

Autor: 201905532
Director: Carlos Miguel Vallez Fernández

MADRID | Junio 2024

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Marta Ros Arroyo, estudiante de E2 + Business Analytics, de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "[Título del trabajo]", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Interpretador de código:** Para realizar análisis de datos preliminares.
2. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
3. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 9/6/2024

Firma: Marta Ros Arroyo

Resumen

El estudio se centra en determinar los factores meteorológicos que provocan retrasos en las salidas de vuelos en el Aeropuerto Internacional Chicago-O'Hare. Se trata de un problema crucial debido a la importancia operativa y económica en el ámbito de la aviación. El estudio busca mejorar la predicción de estos retrasos en los vuelos mediante un enfoque basado en el análisis de datos y la elaboración de modelos de aprendizaje automático. El objetivo final es aumentar la eficacia operativa, mejorando así la satisfacción del cliente. Los datos proceden de diversas fuentes, como Kaggle, el Departamento de Transporte de EE.UU. y Visual Crossing. Se combinan y modifican mediante un método ETL (Extraer, Transformar y Cargar) clásico. Para este proceso se utilizan herramientas como Python y RStudio, para seleccionar vuelos concretos en el aeropuerto internacional O'Hare, eliminar variables sin importancia y crear nuevas variables que tengan significado para el modelo. Se desarrollan y comparan dos modelos predictivos: una Regresión Logística y un *Random Forest*. El modelo de Regresión Logística ofrece un buen rendimiento con un AUC (*Area Under the Curve*) de 0,853 y una precisión de 0,787, destacando su especificidad. El modelo *Random Forest*, con un AUC de 0,803 y una precisión de hasta 0,894, muestra una sensibilidad y una precisión general superiores. Los resultados muestran que, aunque ambos métodos son buenos para predecir los vuelos retrasados por condiciones meteorológicas extremas, el modelo *Random Forest* tiene una precisión general superior. Para futuras investigaciones, se sugiere incorporar datos de múltiples años y aeropuertos. Además, explorar otros modelos y tipos de retrasos podría aumentar la precisión y utilidad del sistema.

Palabras clave: aprendizaje automático, regresión logística, bosque aleatorio, validación cruzada, balanceo de datos, sensibilidad, especificidad, precisión, área bajo la curva.

Abstract

The study focuses on determining the meteorological factors that cause delays in flight departures at Chicago-O'Hare International Airport. This is a critical problem due to its operational and economic importance in the aviation field. The study aims to improve the prediction of these flight delays through an approach based on data analytics and machine learning modelling. The goal is to increase operational efficiency, thereby improving customer satisfaction. The data comes from a variety of sources, including Kaggle, the US Department of Transportation and Visual Crossing. It is combined and modified using a classic ETL (Extract, Transform and Load) method. For this process, tools such as Python and RStudio are used to select specific flights at O'Hare International Airport, eliminate unimportant variables and create new variables that have meaning for the model. Two predictive models are developed and compared: a Logistic Regression and a Random Forest. The Logistic Regression model performs well with an AUC (Area Under the Curve) of 0,853 and an accuracy of 0,787, highlighting its specificity. The Random Forest model, with an AUC of 0,803 and an accuracy of up to 0,894, shows superior sensitivity and overall accuracy. The results show that, although both methods are good at predicting flights delayed due to extreme weather conditions, the Random Forest model has a superior general accuracy. For future research, it is suggested to incorporate data from multiple years and airports. In addition, exploring other models and types of delays could increase the accuracy and usefulness of the system.

Keywords: machine learning, logistic regression, random forest, cross-validation, data balancing, sensitivity, specificity, precision, area under the curve.

Agradecimientos

A Carlos Miguel Vallez Fernández por su dedicación, paciencia y esfuerzo

A mi familia y amigos por su preocupación y apoyo

Índice de Contenido

Índice de Figuras

Índice de Tablas

1 INTRODUCCIÓN

1.1 Objetivos

El trabajo tiene tres objetivos principales. En primer lugar, busca analizar y determinar cuáles son las variables climáticas que tienen un mayor impacto en los retrasos en la salida de vuelos causados por condiciones climáticas adversas. En segundo lugar, tiene como objetivo desarrollar modelos de *Machine Learning* que puedan predecir con precisión los retrasos en la salida de vuelos desde el Aeropuerto Internacional de Chicago-O'Hare (ORD) causados por condiciones climáticas y compararlos. Finalmente, el trabajo pretende proponer áreas de investigación futura y determinar las limitaciones del modelo seleccionado.

El primer objetivo se aborda de forma cuantitativa, analizando la correlación de las variables y la importancia de cada una de ellas en el modelo. El segundo objetivo se aborda de forma cualitativa, llevando a cabo una revisión de la literatura de los modelos actualmente usados en este tipo de problemáticas, así como de forma cuantitativa, al desarrollar el modelo con RStudio. El último objetivo se aborda de forma cualitativa.

1.2 Motivación

La motivación subyacente a esta investigación surge de la necesidad creciente de abordar los desafíos que plantean los retrasos en los vuelos, tanto desde una perspectiva económica, logística como social. En un contexto donde la demanda de viajes continúa en aumento, los aeropuertos y las aerolíneas enfrentan una presión operativa considerable, lo que subraya la importancia de contar con herramientas predictivas que mejoren la eficiencia de las operaciones aéreas.

Una revisión de la literatura revela que son pocos los estudios que incorporan datos climatológicos en sus predicciones, a pesar de que estos factores representan, en promedio, el 38,7% del total de minutos de retraso desde 2003 hasta 2021 (U.S. Department of Transportation, 2022). Se considera que estas variables climáticas pueden ser de gran utilidad, dado que son fácilmente accesibles y podrían enriquecer significativamente los modelos predictivos. Esta consideración es especialmente relevante para los aeropuertos ubicados en regiones con condiciones climáticas extremas,

donde el clima puede ser determinante para los retrasos de vuelo y anticipar dichos retrasos podría beneficiar tanto a pasajeros como a aerolíneas.

El enfoque analítico adoptado en este estudio, que implica el análisis de datos y el desarrollo de modelos predictivos, tiene como objetivo proporcionar a los pasajeros predicciones de retraso de vuelos que tengan en cuenta las condiciones climáticas del momento, permitiéndoles una mejor planificación con anticipación. Al mismo tiempo, se busca mejorar la asignación de recursos y la eficiencia operativa de las aerolíneas, lo que puede traducirse en una mejora significativa de la experiencia del usuario y en un ahorro de costes para las compañías aéreas.

Desde un punto de vista práctico, la aplicación de este trabajo puede incluir la integración del modelo predictivo en las aplicaciones móviles de las aerolíneas, lo que permitiría informar a los pasajeros en tiempo real sobre posibles retrasos en sus vuelos. Además, abre la posibilidad de comercializar seguros con tarifas dinámicas basadas en la probabilidad de retraso de vuelos en días o temporadas específicas. Este enfoque innovador podría tener un impacto significativo en la gestión de riesgos y en la satisfacción del cliente en la industria de la aviación.

1.3 Contexto

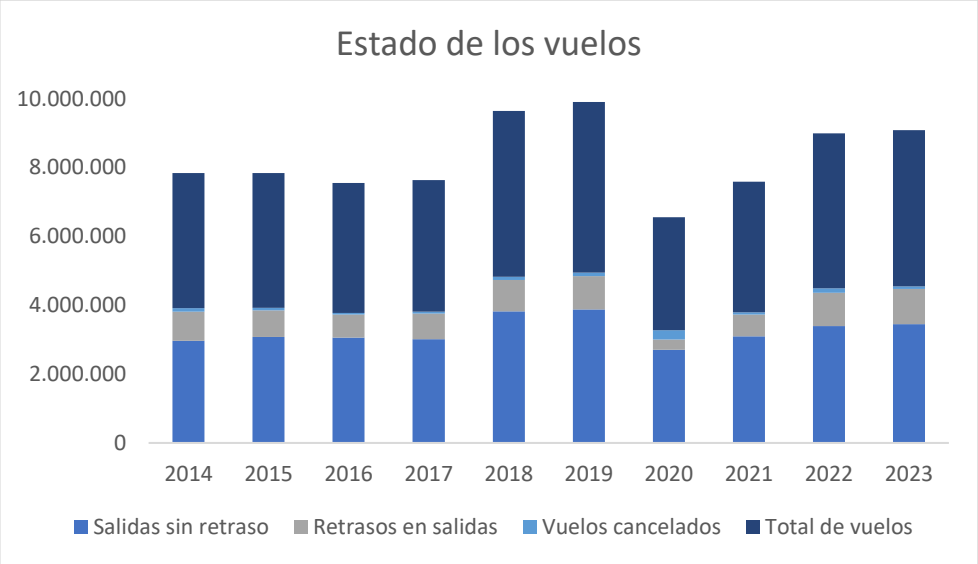
La industria de la aviación es un componente clave para la conectividad a nivel mundial, el transporte aéreo facilita vínculos entre personas y empresas, haciendo posible el comercio, la inversión y el turismo a nivel mundial. La creciente demanda en el tráfico de pasajeros, aumento un 40,5% interanual en el segundo trimestre de 2023, esto señala una fuerte demanda sostenida para el sector del tráfico aéreo (IATA, 2022)

Sin embargo, esta creciente demanda, supera la capacidad de las aerolíneas, que no están preparadas para aumentar sus operaciones al ritmo necesario para adaptarse al aumento de la demanda. A medida que intenten aumentar el número de operaciones para satisfacer la demanda, también aumentaran los retrasos (Federal Aviation Administration, 2023).

Los retrasos y la puntualidad en los vuelos siguen siendo un gran desafío en la industria de la aviación. En Estados Unidos, hasta noviembre de 2023, el 22,33% de los vuelos de salida sufrieron retrasos, lo que supone un 3,76% más de retrasos que el año anterior. Además, los datos históricos muestran como los retrasos en vuelos son factores

recurrentes (U.S. Department of Transportation, n.d.) lo cual podemos observar en la siguiente figura.

Figura 1. *Estado de los vuelos, 2014-2023.*



Fuente: Elaboración propia mediante datos de U.S. Department of Transportation (n.d.).

Estos retrasos en los vuelos además de ser simples inconvenientes tienen un impacto económico y social. Económicamente, los retrasos afectan a la eficiencia operativa y llevan asociados unos costes adicionales para las aerolíneas estadounidenses de 101,18 dólares por minuto, siendo los costes más representativos el combustible y la tripulación. Los costes están basados en los datos del formulario DOT 41 de las líneas aéreas de pasajeros (Airlines for America, 2023). No solo esto, si no que a su vez tienen un impacto social, afectando directamente a la satisfacción del cliente y la conectividad a nivel mundial. Los datos no solo ilustran la gravedad de la situación, sino que también resaltan la importancia de afrontar este desafío mediante métodos innovadores y predictivos.

El Departamento de Transporte de los Estados Unidos (2022) ha clasificado las causas de retraso de vuelos en cinco categorías distintas: aerolínea, condiciones meteorológicas extremas, Sistema Nacional de Aviación (NAS), retraso del avión y seguridad.

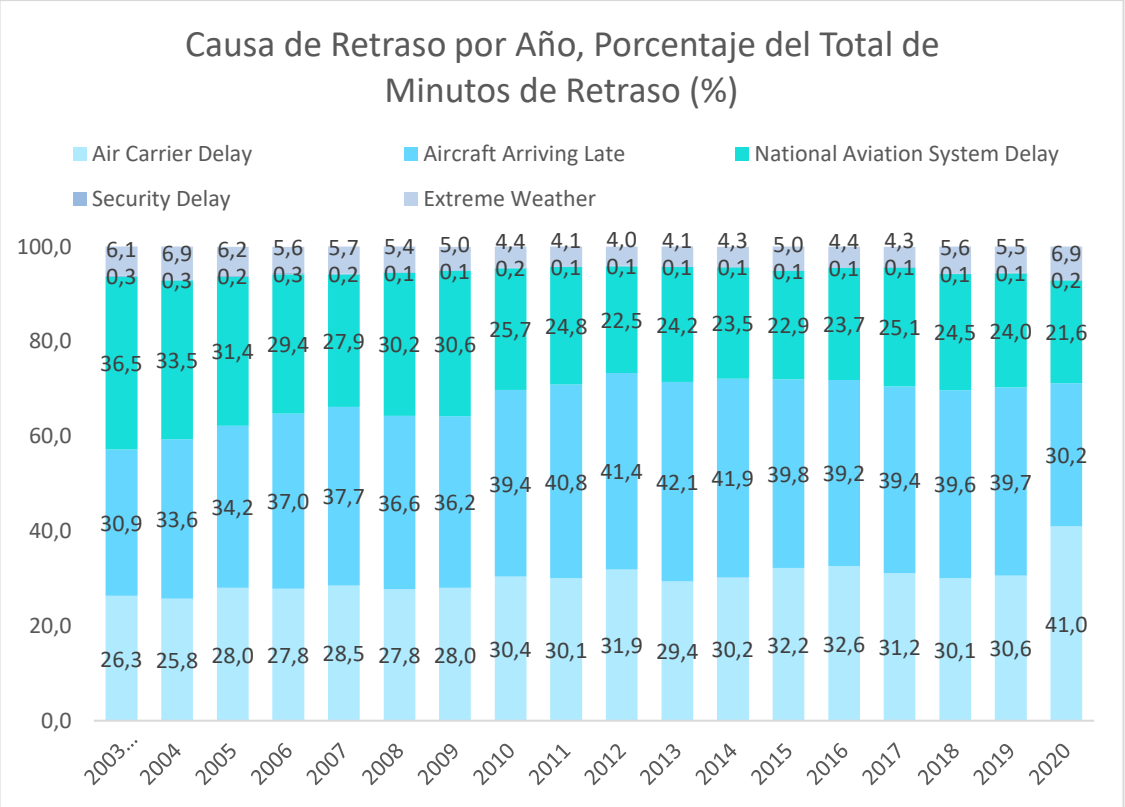
Los retrasos atribuidos a la aerolínea se deben a circunstancias que están bajo su control, como la disponibilidad de la tripulación, la limpieza de la aeronave, la carga de equipaje o el suministro de combustible. Por otro lado, los retrasos relacionados con condiciones meteorológicas extremas son aquellos causados por fenómenos climáticos, como tornados, nevadas o huracanes, que según la aerolínea retrasan o impiden la operación de vuelos.

Los retrasos clasificados como NAS son aquellos atribuibles al Sistema Nacional de Aviación y engloban una amplia variedad de situaciones, como operaciones en aeropuertos y control del tráfico aéreo. Los retrasos del avión se producen cuando una aeronave experimenta demoras previas que afectan el vuelo actual.

Finalmente, los retrasos por motivos de seguridad incluyen la evacuación de terminales o salas de espera, el reembarque de una aeronave debido a problemas de seguridad, fallas en los equipos de control y largas filas de espera de más de 29 minutos en los puntos de control de seguridad (U.S. Department of Transportation, 2022).

La Figura 2 ilustra estas cinco categorías de retraso y muestra cuáles han sido las principales causantes de minutos de retraso a lo largo del tiempo.

Figura 2. *Causa de retraso por año (2003-2020).*



Fuente: Elaboración propia mediante datos de U.S. Department of Transportation (n.d.).

En la industria de la aviación, los retrasos en los vuelos constituyen un desafío persistente que impacta tanto en las operaciones de las aerolíneas como en la experiencia de los pasajeros. Ante esta problemática, se hace patente la necesidad de adoptar enfoques innovadores y basados en el análisis de datos para comprender a fondo las causas subyacentes e implementar medidas preventivas. En este contexto, el análisis de datos emerge como una herramienta fundamental para abordar los retrasos en los vuelos. La recopilación, el procesamiento y el análisis exhaustivo de los datos operativos, meteorológicos, de tráfico aéreo y otros factores pertinentes pueden arrojar nuevos conocimientos valiosos sobre los patrones de retraso y sus principales impulsores. La capacidad de obtener nuevos conocimientos significativos a partir de estos datos puede conducir a la adopción de medidas proactivas que mejoren tanto la eficiencia operativa como la experiencia del cliente. En este sentido, este trabajo se sitúa en la intersección

entre la necesidad de aplicar enfoques analíticos y el gran potencial de los datos en el sector de la aviación, con el propósito último de contribuir a su mejora continua.

1.4 Estado del Arte

Predecir retrasos en vuelos representa un desafío importante en la industria de la aviación. Se han propuesto distintas técnicas y modelos para pronosticar con precisión estos retrasos, tanto a corto como a largo plazo. El siguiente resumen resume los enfoques y modelos utilizados en estudios recientes.

El estudio realizado por Nigam y Govinda (2017) se enfoca en la predicción de retrasos en vuelos mediante la utilización de datos sobre tráfico aéreo, cambios climáticos y retrasos anteriores. Emplean un modelo de regresión logística para determinar la puntualidad de los vuelos, utilizando información de 70 de los aeropuertos más transitados de Estados Unidos entre abril y octubre de 2013. A pesar de alcanzar una precisión del 80,6%, los autores reconocen limitaciones derivadas de la diversidad de causas de retraso que podrían no estar completamente representadas en los datos utilizados.

Por otro lado, Liu et al. (2020) proponen un método para prever retrasos en vuelos mediante el uso de *Gradient Boosting Decision Trees* (GBDT) y árboles CART. Este enfoque se basa en datos del sistema ADS-B para monitorear vuelos, incorporando información sobre condiciones climáticas, flujo de tráfico y horarios de vuelo como características clave. Las clasificaciones se dividen en dos categorías: retrasados y no retrasados, con distintos niveles de retraso. Para abordar el desafío de conjuntos de datos desequilibrados, emplean estrategias de submuestreo aleatorio. Los resultados exhiben una precisión del 87,72% en la clasificación binaria de retrasos, y del 79,45% y 67,36% para clasificaciones con tres y cuatro categorías, respectivamente.

Manna et al. (2017), por su parte, utilizan GBDT con un enfoque de regresión para predecir retrasos en vuelos, examinando tanto salidas como llegadas. Utilizan datos de vuelos proporcionados por el Departamento de Transporte de Estados Unidos o *Bureau of Transportation Statistics* (BTS), incluyendo todos los vuelos entrantes y salientes de los 70 aeropuertos más concurridos del país entre abril y octubre de 2013. Para mitigar retrasos inusuales, restringen los tiempos de retraso para cada día de la semana y

normalizan las características. Además, los modelos muestran una alta correlación, con valores de R cuadrado de 0,923185 y 0,948523 para llegadas y salidas, respectivamente.

Finalmente, Qu et al. (2020) proponen la utilización de *Deep Convolutional Neural Network* (DCNN) con retropropagación para anticipar retrasos en vuelos, integrando datos meteorológicos en su análisis. El modelo se entrena con aproximadamente 4 millones de registros y emplea datos de los años 2016 y 2017, suministrados por el Centro Nacional de Datos Climáticos y BTS. Se comparan dos modelos, un DCNN y una *Convolutional Neural Network* (CNN). El modelo DCNN exhibe un rendimiento superior al modelo CNN, logrando una precisión del 92,26%. Esto subraya la importancia de incorporar datos meteorológicos para mejorar la predicción de retrasos en vuelos, lo que resulta en una significativa mejora en la precisión de predicción al considerar información climática.

Otros autores como Monje Solá (2015) y Martínez Domenech (2016) han realizado estudios enfocados en uno o varios aeropuertos. El primero de ellos se centra específicamente en el aeropuerto de Seattle-Tacoma, analizando exclusivamente el retraso por motivos de la aeronave, mientras que el segundo se enfoca en los aeropuertos de Arizona, investigando los retrasos en la llegada de vuelos. Ambos comparten la utilización de la técnica conocida como *Gradient Boosting Machine*, la cual ha demostrado proporcionar los mejores resultados en términos de AUC y RMSE.

Asimismo, estudios como el de (Esperón Cespón, 2018) han explorado el análisis de datos de vuelos en múltiples aeropuertos de origen, considerando variables como datos climáticos, tamaño de los aeropuertos y días festivos. En esta investigación también se han evaluado una gama de modelos, destacando nuevamente la efectividad del *Gradient Boosting Machine* como uno de los más óptimos para abordar esta problemática.

Este repaso de la literatura proporciona una visión general de los enfoques y modelos recientes utilizados para predecir retrasos en vuelos, resaltando la diversidad de técnicas y la importancia de considerar múltiples factores, incluido el clima, para mejorar la precisión de la predicción.

Tabla 1. *Resumen de revisión de la literatura*

.

Titulo	Autor	Estudio	Datos	Periodo	Variable objetivo	Modelos	Medida de rendimiento
Análisis y Predicción de los Retrasos de Vuelo	Raul Monje Sola	Retrasos de vuelo en el aeropuerto de Seattle-Tacoma, mediante el impacto de los retrasos en el sistema de transporte aéreo	Bureau of Transportation Statistics (BTS)	enero 2014	Aircraft delay	Regresión logística *Gradient Boosting Machine* (GBM)	AUC: 0,77 AUC: 0,93
Construcción de un modelo de predicción para la puntualidad de vuelos comerciales	Iván Esperón Cespón	Predicción sobre los retrasos de los vuelos comerciales para 6 aeropuertos de Estados Unidos (Atlanta, LA, Chicago, Dallas, NY, Denver)	BTS National Oceanic and Atmospheric Administration Informacion de los aeropuertos (ourairports.com) Dias festivos (github)	2016	Hora de llegada real - hora de llegada programada	Regresión lineal Redes neuronales Arboles de decisión *Random forest* GBM *Ensamble*	 R^2: 0,34 (el mejor)
Predicción y Análisis de los Retrasos en los Vuelos	Nerea Martínez Domenech	Estudio de los retrasos en la hora de llegada en los aeropuertos de Arizona	BTS	enero – septiembre 2015	Hora de llegada real - hora de llegada programada	*Random forest* Gradiente de Árboles *Boosting*(GB)	RMSE: 1,86 RMSE: 0,80
Cloud Based Flight Delay Prediction using Logistic Regression	Rahul Nigam y K Govinda	Estudio de los retrasos a la hora de llegada en 70 de los aeropuertos más concurridos de Estados Unidos	Conjunto de datos de aerolíneas, clima y aeropuerto (no especificado)	abril – octubre 2013	Retraso en la hora de llegada	Regresión logística de dos clases	*Accuracy:* 80,6% *Precision* y *recall* > 50%
Generalized Flight Delay Prediction Method Using Gradient Boosting Decision Tree	Liu et al.	El estudio utiliza datos del sistema ADS-B para rastrear vuelos en China	Datos extraídos del sistema ADS-B (clima y flujo de tráfico) y Ctrip (Información de aeropuertos y horarios de vuelos)	diciembre 2018 – mayo 2019	Retraso en la hora de llegada	GBDT y árboles CART 2 clases 3 clases 4 clases	*Accuracy:* 87,72% 79,45% 67,36%
A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree	Manna et al.	Se estudian por separado los retrasos en llegadas y salidas de los 70 aeropuertos más concurridos de Estados Unidos	BTS	abril – octubre 2013	Retraso en hora de salida y llegada	GBDT	R^2: 0,94
Flight Delay Prediction Using Deep Convolutional Neural Network Based on Fusion of Meteorological Data	Qu et al.	Los resultados mejoran un 1% en precisión cuando se añaden variables climatológicas al modelo	BTS Quality Controlled Local Climatological Data (QCLCD)	2016 - 2017	Retraso del vuelo	DCNN	*Accuracy:* 92,26%

Fuente: Elaboración propia

2 MODELOS DE PREDICCIÓN

2.1 La importancia del Machine Learning y las predicciones

El aprendizaje automático, también conocido como *Machine Learning*, es una técnica en la que los algoritmos permiten a los ordenadores aprender y reconocer patrones automáticamente en grandes volúmenes de datos, para luego hacer predicciones basadas en estos patrones. La importancia del *Machine Learning* reside en su capacidad para procesar y analizar enormes conjuntos de datos, lo que es esencial para aprovechar la información recopilada, almacenada y gestionada. Hoy en día, esta técnica influye significativamente en nuestra vida diaria y es crucial en aplicaciones como la predicción del clima, la supervisión ambiental, la optimización de costes de inventario, el diseño de estrategias de venta y el desarrollo de vehículos autónomos (Zhou, 2021).

El *Machine Learning* ofrece numerosas ventajas para enfrentar el desafío de predecir retrasos en la salida de vuelos, un problema que implica múltiples variables. Una de sus principales fortalezas es la capacidad de manejar y analizar grandes volúmenes de datos provenientes de diversas fuentes, como registros históricos de vuelos y datos meteorológicos. Además, puede detectar patrones y relaciones complejas entre variables que no son evidentes a simple vista pero que influyen en los retrasos de los vuelos.

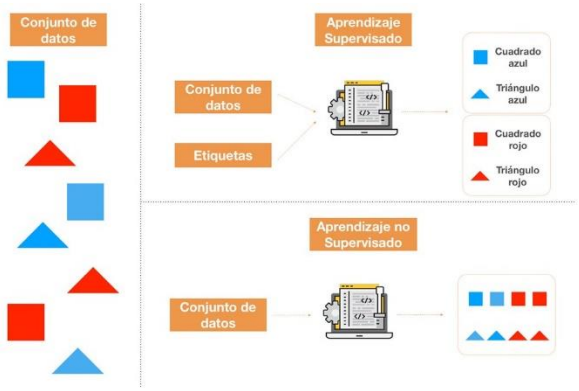
El uso de *Machine Learning* en este contexto es fundamental, ya que permite hacer predicciones de retrasos con mayor precisión. Los modelos de *Machine Learning* son adaptables, lo que es útil para ajustarse a los cambios en los datos con el tiempo. Además, ayudan a las aerolíneas a optimizar la programación de vuelos y mejorar la experiencia del cliente.

Finalmente, décadas de investigación en aprendizaje automático nos han proporcionado una amplia gama de opciones en cuanto a algoritmos y modelos (Jiang, 2021). En el próximo apartado, se presentarán y explicarán algunos de estos algoritmos y modelos.

2.2 Tipos de algoritmos (Supervisado/No supervisado)

En el ámbito del aprendizaje automático, se distinguen principalmente tres métodos: aprendizaje supervisado, no supervisado y semi-supervisado.

Figura 3. *Diferencia entre aprendizaje supervisado y no supervisado*



Fuente: Gonzalez (2018)

A continuación, se presenta una introducción a cada uno de ellos, teniendo en cuenta que el alcance del trabajo es el aprendizaje supervisado y que existen variedad de métodos de aprendizaje automático como el aprendizaje no supervisado, semisupervisado y el aprendizaje de refuerzo los cuáles están fuera del alcance de este trabajo.

El aprendizaje supervisado implica el entrenamiento de modelos para asignar datos de entrada a salidas deseadas, utilizando conjuntos de datos previamente etiquetados. Cada ejemplo en estos conjuntos está claramente asociado con una salida o clasificación conocida. Este proceso permite a modelos como las redes neuronales y árboles de decisión aprender a identificar patrones y relaciones que pueden predecir el comportamiento de datos nuevos (Ayodele, 2010). El objetivo es desarrollar un modelo capaz de generalizar evitando el sobreajuste, que ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento y pierde la capacidad de predecir con precisión para datos nuevos. Este enfoque tiene aplicaciones como la detección de spam, el análisis de sentimientos o la clasificación de imágenes (Bonaccorso, 2017).

La clasificación es un tipo de aprendizaje supervisado, donde los algoritmos organizan los datos en categorías predefinidas basadas en características distintivas. Esta

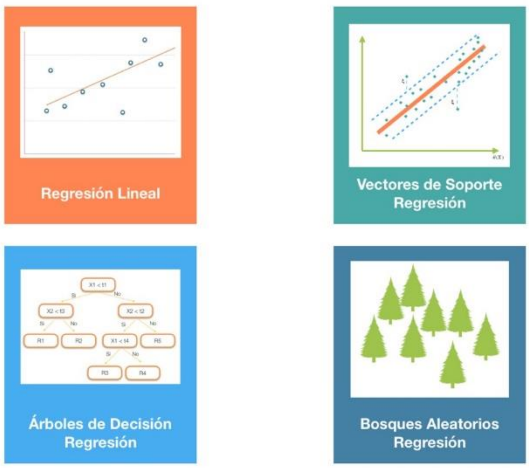
técnica es especialmente útil cuando las categorías correctas son conocidas, permitiendo al modelo clasificar nuevos datos basándose en lo aprendido en el conjunto de entrenamiento (Ayodele, 2010). Algunos algoritmos de clasificación conocidos son el *Naive Bayes*, regresión logística, *K-nearest neighbours* (KNN), *Support Vector Machine* (SVM) y varios otros incluyendo árboles de decisión y *Gradient Boosting*.

Figura 4. *Algoritmos de clasificación*

Fuente: Gonzalez (2019)

La regresión, se centra en predecir variables continuas, a diferencia de la clasificación que se orienta a etiquetas categóricas. Este método se emplea en campos como la previsión financiera y la modelización de respuestas a medicamentos, utilizando diversas formas de regresión para ajustar los modelos a los datos específicos, como pueden ser la regresión lineal, lasso o ridge Sarker, I.H. (2021)

Figura 5. *Algoritmos de regresión*

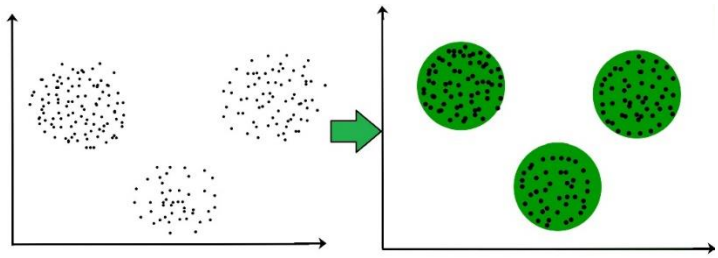


Fuente: Gonzalez (2019b)

El aprendizaje no supervisado permite a los algoritmos descubrir patrones y estructuras en datos sin etiquetar, facilitando la exploración de relaciones desconocidas previamente (Mahesh, 2018). Este método es útil para agrupar datos basados en sus similitudes o una métrica de distancia, utilizando técnicas como el *clustering*. A su vez también sirve para la reducción de la dimensionalidad (Bonaccorso, 2017).

El *clustering* es una técnica de aprendizaje no supervisado que organiza los datos en grupos o *clusters* basados en su proximidad o similitud. Se emplea frecuentemente para descubrir tendencias o patrones. Aunque es una muy buena técnica para el análisis de datos es importante evaluar críticamente los resultados para asegurar que los *clusters* identificados representen patrones verdaderos ya que el algoritmo siempre encontrará *clusters,* aunque no existan. Sarker, I.H. (2021)

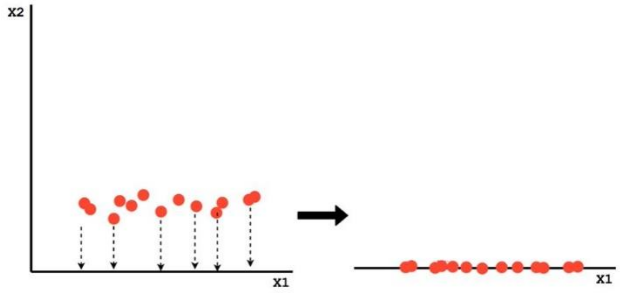
Figura 6. *Clustering*



Fuente: Priy (2024)

La reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA), se utiliza para reducir el número de dimensiones limitando la perdida de información de forma que los nuevos datos sigan siendo útiles. Esto se lleva a cabo transformando variables correlacionadas en componentes principales no correlacionadas (Mahesh, 2018). Esto facilita la visualización y sirve a su vez como método de preprocesamiento (Gatto, n.d.)

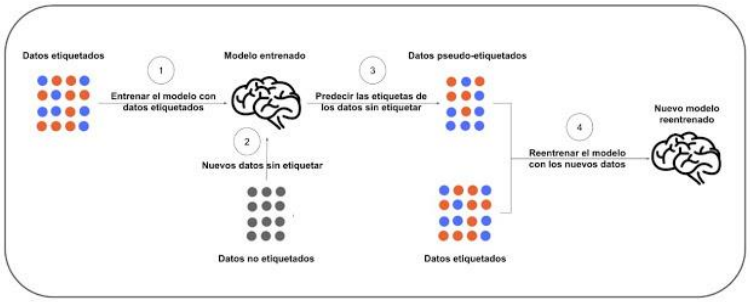
Figura 7. *Reducción de la dimensionalidad*



Fuente: Gil (n.d.)

Finalmente, el aprendizaje semi-supervisado es una combinación del aprendizaje supervisado y no supervisado. Utiliza tanto datos etiquetados como no etiquetados y se aplica en técnicas avanzadas como el aprendizaje por refuerzo y los modelos generativos.

Figura 8. *Aprendizaje semi-supervisado*



Fuente: Ibáñez Martín (2019)

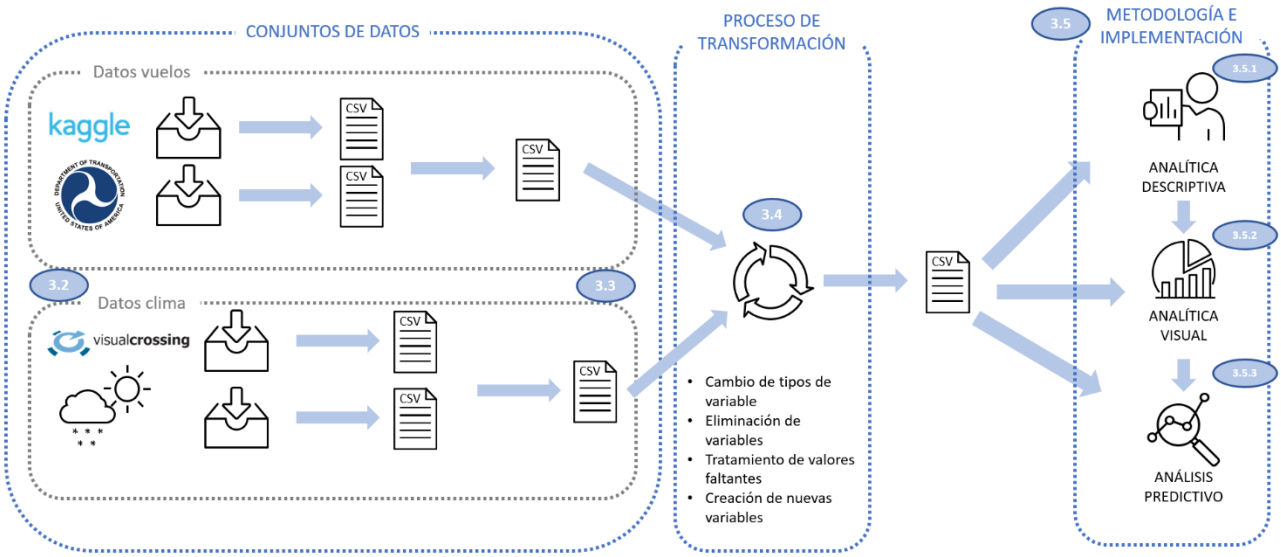
3 METODOLOGÍA

La construcción de un sistema de aprendizaje automático generalmente implica tres pasos fundamentales: primero, la recopilación de una cantidad suficiente de datos, a partir de los cuales las máquinas pueden aprender. Segundo, la aplicación de procedimientos para extraer características relevantes de los datos y finalmente, la selección de un algoritmo de aprendizaje para construir modelos basados en las características extraídas de los datos de entrenamiento (Jiang, 2021). En este capítulo, se llevarán a cabo todos estos pasos con el objetivo de desarrollar un modelo predictivo utilizando los datos recopilados inicialmente.

Se va a seguir una metodología ETL clásica la cual consiste en extraer, transformar y cargar los datos.

- Extracción: En esta fase, se recopilaron datos de diversas fuentes para su posterior análisis. Para este estudio, se utilizaron bases de datos provenientes de Kaggle, el Departamento de Transporte de Estados Unidos y Visual Crossing.
- Combinación: Se unificaron varios conjuntos de datos en uno solo. Primero, se integraron los datos de vuelos y los datos climáticos de 2015 y 2016. Luego, estos conjuntos de datos se combinaron en un archivo único llamado "data_combined.csv", utilizando claves comunes y aplicando las transformaciones necesarias para asegurar la coherencia.
- Transformación: Los datos fueron modificados para prepararlos para el análisis. Se seleccionaron específicamente los vuelos del Aeropuerto Internacional O'Hare, se eliminaron las variables irrelevantes, se ajustaron los tipos de datos, se trataron los valores faltantes y se crearon nuevas variables relevantes para el estudio.

Figura 9. *Metodología*



Fuente: Elaboración propia

3.1 Alcance del proyecto

El estudio se enfoca en el aeropuerto de Chicago y el año 2015 para establecer un marco específico inicial y dada la disponibilidad de los datos. Sin embargo, el diseño modular del estudio permite una fácil adaptación a otros aeropuertos y períodos temporales con mínimas modificaciones.

Para el objeto de este estudio se han eliminado algunos o todos los registros de las siguientes variables ya que no se consideran útiles para la predicción o están fuera del alcance del estudio:

- *CANCELLED*: Se eliminaron los registros en los que el vuelo fue cancelado por razones distintas al clima, ya que estos registros no son relevantes para la predicción.
- *DIVERTED*: Se eliminaron los registros en los que el vuelo fue desviado, ya que están fuera del alcance del estudio.
- *AIR_SYSTEM_DELAY*: Los retrasos debidos al sistema nacional de aviación se refieren a un amplio conjunto de condiciones, incluyendo "condiciones climáticas no extremas, operaciones del aeropuerto, gran volumen de tráfico y control del tráfico aéreo" (U.S. Department of Transportation, 2024). Dado que estos retrasos

se refieren a condiciones climáticas no extremas y que pueden ser causados por una combinación de factores, se considera que están fuera del alcance del estudio.

3.2 Extracción de datos de origen

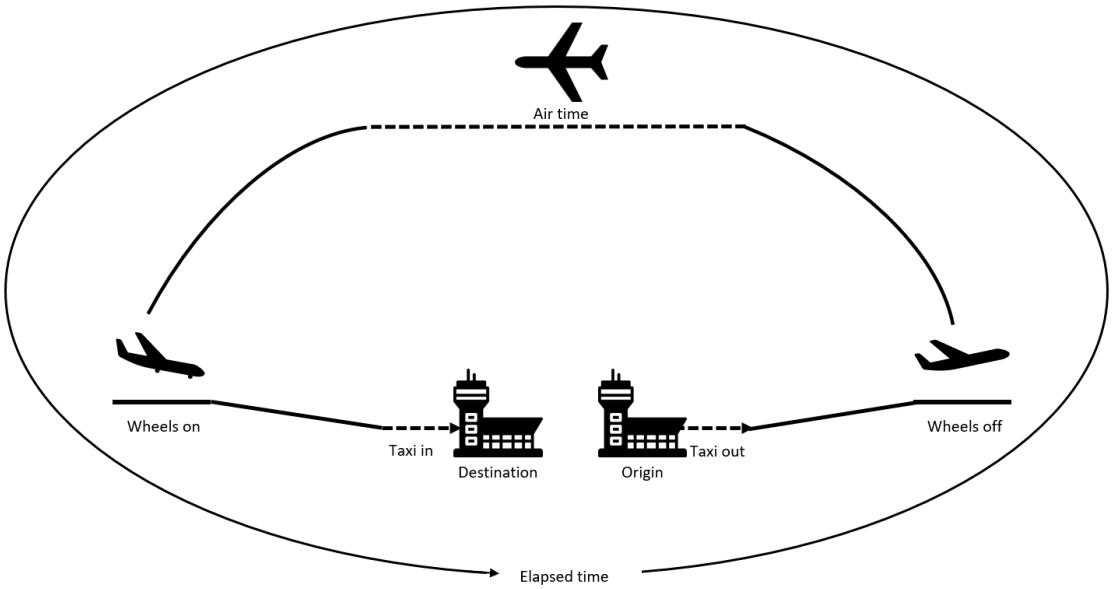
Para este estudio se han utilizado dos bases de datos, las cuales se han extraído para su posterior combinación y transformación.

Por un lado, se ha utilizado una base de datos con información sobre los vuelos la cual ha sido obtenida a través de Kaggle (Department of Transportation, 2017) y extraída en formato CSV, de ahora en adelante "flights.csv". El origen de estos datos es la base de datos "On-Time: Reporting Carrier On-Time Performance (1987-present)" desarrollada por el Departamento de Transporte de Estados Unidos. En ella se realiza el seguimiento de la puntualidad de los vuelos nacionales operados por grandes compañías aéreas. Los datos presentan información para vuelos de todos los meses del año 2015.

Puesto que en los datos obtenidos a través de Kaggle se identificó que faltaban los datos de octubre de 2015, estos se han obtenido directamente a través de la base de datos "On-Time: Reporting Carrier On-Time Performance (1987-present)" (United States Department of Transportation, 2024) y han sido extraídos en formato CSV, de ahora en adelante "T_ONTIME_REPORTING.csv". Incluye las mismas variables que los datos "flights.csv" pero con distintos nombres y formato. Por lo que se han tenido que tratar más adelante en la transformación de datos previamente a ser combinados.

Antes de proceder con un análisis detallado de cada una de las variables en los conjuntos de datos, se presenta el siguiente gráfico como herramienta de apoyo. Este gráfico facilita la comprensión del funcionamiento de las operaciones aéreas y la relación de cada variable con las diferentes etapas del proceso.

Figura 10. *Etapas de un vuelo*



Fuente: Elaboración propia.

Las variables de "flights.csv" con sus descripciones, tipo de variable y unidad de medida se muestran en la tabla a continuación:

Tabla 2. *Variables del conjunto de datos "flights.csv"*

Nombre	Descripción	Tipo de variable	Unidad de medida	Ejemplo de registro
YEAR	Año	Entero		2015
MONTH	Mes	Entero		1
DAY	Día del mes	Entero		1
DAY_OF_WEEK	Día de la semana	Entero		4
AIRLINE	Código asignado por la International Air Transport Association (IATA) y utilizado habitualmente para identificar a una compañía aérea.	Carácter		AS
FLIGHT_NUMBER	Número de vuelo	Entero		98
TAIL_NUMBER	Número de cola de la aeronave	Carácter		N407AS
ORIGIN_AIRPORT	Aeropuerto de origen	Carácter		LAX
DESTINATION_AIRPORT	Aeropuerto de destino	Carácter		ATL
SCHEDULED_DEPARTURE	Hora de salida prevista	Entero	Hora local: hhmm	5
DEPARTURE_TIME	Hora de salida real	Entero	Hora local: hhmm	2354
DEPARTURE_DELAY	Retraso en la salida. Diferencia entre la hora de salida programada y la real. Las salidas antes de tiempo aparecen con números negativos	Entero	minutos	-11
TAXI_OUT	Fase en la que la aeronave se mueve desde su lugar de estacionamiento hacia la pista en preparación para despegar	Entero	minutos	21

WHEELS_OFF	Cuando las ruedas o tren de aterrizaje de una aeronave están completamente despegadas del suelo durante la fase de despegue	Entero	Hora local: hhmm	15
SCHEDULED_TIME	Tiempo de vuelo programado	Entero	minutos	205
ELAPSED_TIME	Tiempo de vuelo. Computa desde la hora de salida de la puerta hasta la hora de llegada a la puerta del aeropuerto de destino	Entero	minutos	194
AIR_TIME	Flight Time. Desde el despegue (wheels off) en el aeropuerto de origen hasta el aterrizaje (wheels on) en el aeropuerto de destino	Entero	minutos	169
DISTANCE	Distancia entre los aeropuertos	Entero	millas	1448
WHEELS_ON	Cuando las ruedas o el tren de aterrizaje de una aeronave entran en contacto con la pista durante la fase de aterrizaje de un vuelo.	Entero	Hora local: hhmm	404
TAXI_IN	Fase en la que la aeronave se desplaza desde la pista hasta su lugar de estacionamiento o puerta	Entero	minutos	4
SCHEDULED_ARRIVAL	Hora de llegada prevista	Entero	Hora local: hhmm	430
ARRIVAL_TIME	Hora de llegada real	Entero	Hora local: hhmm	408
ARRIVAL_DELAY	Retraso en la llegada. Diferencia entre la hora de llegada prevista y la real. Las llegadas antes de tiempo aparecen con números negativos	Entero	minutos	-22
DIVERTED	Un vuelo que debe aterrizar en un destino distinto del previsto originalmente por razones ajenas al control del piloto/compañía. Vuelo desviado (1=Sí)	Entero	0/1	0
CANCELLED	Vuelo cancelado (1=Sí)	Entero	0/1	0
CANCELLATION_REASON	A: Aerolínea B: Clima C: Sistema Aéreo Nacional D: Seguridad	Carácter	A/B/C/D	A
AIR_SYSTEM_DELAY	Retraso debido al Sistema Aéreo Nacional	Entero	minutos	43
SECURITY_DELAY	Retraso debido a la seguridad	Entero	minutos	0
AIRLINE_DELAY	Retraso debido a la aerolínea	Entero	minutos	72
LATE_AIRCRAFT_DELAY	Retraso debido a el avión	Entero	minutos	0
WEATHER_DELAY	Retraso debido al clima	Entero	minutos	102

Fuente: Elaboración propia

Las variables de "T_ONTIME_REPORTING.csv" con sus descripciones, tipo de variable y unidad de medida se muestran en la tabla a continuación:

Tabla 3. *Variables del conjunto de datos "T_ONTIME_REPORTING.csv"*

Nombre	Descripción	Tipo de variable	Unidad de medida	Ejemplo de registro
YEAR	Año	Entero		2015
MONTH	Mes	Entero		10
DAY	Día del mes	Entero		1
DAY_OF_WEEK	Día de la semana	Entero		4

OP_UNIQUE_CARRIER	Código asignado por la International Air Transport Association (IATA) y utilizado habitualmente para identificar a una compañía aérea.	Carácter		AA
OP_CARRIER_FL_NUM	Número de vuelo	Entero		1044
TAIL_NUM	Número de cola de la aeronave	Carácter		N014AA
ORIGIN	Aeropuerto de origen	Carácter		IAH
DEST	Aeropuerto de destino	Carácter		MIA
CRS_DEP_TIME	Hora de salida prevista	Entero	Hora local: hhmm	500
DEP_TIME	Hora de salida real	Entero	Hora local: hhmm	508
DEP_DELAY	Retraso en la salida. Diferencia entre la hora de salida programada y la real. Las salidas antes de tiempo aparecen con números negativos	Número	minutos	8
TAXI_OUT	Fase en la que la aeronave se mueve desde su lugar de estacionamiento hacia la pista en preparación para despegar	Número	minutos	28
WHEELS_OFF	Cuando las ruedas o tren de aterrizaje de una aeronave están completamente despegadas del suelo durante la fase de despegue	Entero	Hora local: hhmm	536
CRS_ELAPSED_TIME	Tiempo de vuelo programado	Número	minutos	143
ACTUAL_ELAPSED_TIME	Tiempo de vuelo. Computa desde la hora de salida de la puerta hasta la hora de llegada a la puerta del aeropuerto de destino	Número	minutos	169
AIR_TIME	Flight Time. Desde el despegue (wheels off) en el aeropuerto de origen hasta el aterrizaje (wheels on) en el aeropuerto de destino	Entero	minutos	132
DISTANCE	Distancia entre los aeropuertos	Número	millas	964
WHEELS_ON	Cuando las ruedas o el tren de aterrizaje de una aeronave entran en contacto con la pista durante la fase de aterrizaje de un vuelo.	Entero	Hora local: hhmm	848
TAXI_IN	Fase en la que la aeronave se desplaza desde la pista hasta su lugar de estacionamiento o puerta	Número	minutos	9
CRS_ARR_TIME	Hora de llegada prevista	Entero	Hora local: hhmm	823
ARR_TIME	Hora de llegada real	Entero	Hora local: hhmm	857
ARR_DELAY	Retraso en la llegada. Diferencia entre la hora de llegada prevista y la real. Las llegadas antes de tiempo aparecen con números negativos	Número	minutos	34
DIVERTED	Un vuelo que debe aterrizar en un destino distinto del previsto originalmente por razones ajenas al control del piloto/compañía. Vuelo desviado (1=Sí)	Número	0/1	0
CANCELLED	Vuelo cancelado (1=Sí)	Número	0/1	0
CANCELLATION_CODE	A: Aerolínea B: Clima C: Sistema Aéreo Nacional D: Seguridad	Carácter	A/B/C/D	
NAS_DELAY	Retraso debido al Sistema Aéreo Nacional	Número	minutos	26
SECURITY_DELAY	Retraso debido a la seguridad	Número	minutos	0
CARRIER_DELAY	Retraso debido a la aerolínea	Número	minutos	8
LATE_AIRCRAFT_DELAY	Retraso debido a el avión	Número	minutos	0
WEATHER_DELAY	Retraso debido al clima	Número	minutos	0

Fuente: Elaboración propia

Por otro lado, se ha utilizado una base de datos con información climática, la cual ha sido obtenida a través de la web Visual Crossing (Visual Crossing, n.d.) y extraída en formato CSV. Puesto que hay vuelos que salen en la madrugada del 31 de diciembre de 2015, se necesitan datos climáticos para el primer día de enero de 2016. Se han extraído por separado, un CSV con todos los meses del año 2015, de ahora en adelante "CHICAGO O'HARE INTERNATIO... 2015-01-01 to 2015-12-31" y un CSV con los datos de enero de 2016, de ahora en adelante "CHICAGO O'HARE INTERNATIO... 2016-01-01 to 2016-01-31".

Las variables de ambos conjuntos de datos con sus descripciones, tipo de variable y unidad de medida se muestran en la tabla a continuación:

Tabla 4. *Variables de los conjuntos de datos "CHICAGO O'HARE INTERNATIO... 2015-01-01 to 2015-12-31" y "CHICAGO O'HARE INTERNATIO... 2016-01-01 to 2016-01-31".*

Nombre	Descripción	Tipo de variable	Unidad de medida	Ejemplo de registro
name	Nombre de la estación meteorológica	Carácter		CHICAGO O'HARE INTERNATIONAL AIRPORT STATION
datetime	Fecha y hora del registro	Carácter	Año/mes/día Hora:minutos:segundos	2015-01-01T00:00:00
temp	Temperatura o temperatura media	Número	°C	-1.1
feelslike	Sensación térmica: - si el índice de calor es alto, la sensación térmica se regirá por el índice de calor - si el viento helado es bajo, la sensación térmica se medirá por el viento helado - entre estos dos extremos, la sensación térmica se medirá por la temperatura real	Número	°C	-4.8
dew	Punto de rocío	Número	°C	-3.3

humidity	Humedad relativa: cantidad de vapor de agua presente en el aire comparada con la cantidad máxima posible para una temperatura dada, expresada como porcentaje medio - Los niveles de confort humano se encuentran normalmente entre el 30-70% - Los valores superiores al 70% se consideran húmedos - Los valores inferiores al 30% se consideran secos	Número	%	85.26
precip	La cantidad de precipitación que cayó o se predice que caerá en el período de tiempo especificado	Número	mm	0.461
precipprob	Probabilidad de precipitación expresada como un porcentaje de 0 a 100%	Entero	%	100
preciptype	Tipo de precipitación: lluvia, nieve, lluvia congelada y hielo	Carácter		rain, snow
snow	La cantidad de nueva nieve que ha caído en el período de tiempo	Número	cm	0.03
snowdepth	El promedio de la cantidad de nieve actualmente en el suelo durante el período de tiempo. La profundidad de la nieve aumentará con la caída de nieve adicional y disminuirá con el derretimiento y la compactación	Número	cm	0.55
windgust	La velocidad máxima del viento se mide durante un corto período de tiempo (normalmente menos de 20 segundos). Una ráfaga de viento requiere que la velocidad del viento a corto plazo medida sea significativamente mayor que la velocidad promedio del viento. Típicamente, la velocidad del viento debe ser 10 nudos más (11 mph o 18 km/h). Cuando la ráfaga de viento no cumple con estos criterios, se devuelve un valor nulo/vacío.	Número	km/h	N/A
windspeed	Promedio de la velocidad durante los dos minutos previos a la medición registrada	Número	km/h	10.7

winddir	Promedio de la dirección durante los dos minutos previos a las mediciones registradas. La dirección del viento indica la dirección desde la cual sopla el viento. Las unidades de la dirección del viento son grados desde el norte. El valor oscila entre 0 grados (desde el norte) hasta 90 grados (desde el este), 180 grados (desde el sur), 270 grados (desde el oeste) de vuelta a 360 grados	Número	Grados	130
sealevelpressure	La presión atmosférica al nivel del mar	Número	Milibares	1020.1
cloudcover	La cantidad de cielo cubierto por nubes expresado como un porcentaje	Número	%	100.0
visibility	La visibilidad es la distancia que se puede ver durante el día. Esto tiene en cuenta fenómenos meteorológicos como la neblina, la bruma, la niebla o el humo	Número	km	2.0
solarradiation	Radiación solar. Mide la potencia en el momento instantáneo de la observación	Número	W/m^2	0.0
solarenergy	Indica la energía total del sol que se acumula durante una hora o un día	Número	MJ/m^2	0.0
uvindex	Un valor entre 0 y 10 que indica el nivel de exposición a los rayos ultravioleta (UV) para esa hora o día. 10 representa un alto nivel de exposición, y 0 representa ninguna exposición. El índice UV se calcula en función de la cantidad de radiación solar de onda corta	Entero		0
severerisk	Riesgo climático grave	Lógica		N/A
conditions	Condiciones climáticas notables reportadas en una ubicación particular, como tormentas eléctricas, lluvias, etc	Carácter		Snow, Rain, Overcast
icon	Nieve: cantidad de nieve > 0 Lluvia: cantidad de lluvia > 0 Niebla: visibilidad es baja (<1 km) Viento: velocidad del viento es alta (>30 km/h) Nublado: cobertura de nubes es >90% Parcialmente nublado - día: cobertura de nubes es >20% durante el día. Parcialmente nublado - noche: cobertura de nubes es >20% durante la noche. Despejado - día: cobertura de nubes es <20% durante el día. Despejado - noche: cobertura de nubes es <20% durante la noche	Carácter		Snow
stations	Identificador único para la estación meteorológica	Carácter		72534014819,KORD ,72530094846,KMD

				W,74466504838,KPWK

Fuente: Elaboración propia

3.3 Combinación de conjuntos de datos

Inicialmente, se combinan los conjuntos de datos “flights.csv” y "T_ONTIME_REPORTING.csv", así como “CHICAGO O'HARE INTERNATIO... 2015-01-01 to 2015-12-31” y “CHICAGO O'HARE INTERNATIO... 2016-01-01 to 2016-01-31”. El objetivo es obtener un conjunto de datos consolidado para vuelos y otro para el clima, los cuales se integrarán posteriormente en un conjunto de datos completo.

Para combinar “flights.csv” y "T_ONTIME_REPORTING.csv", es necesario aplicar cambios en este último para alinearlo con el formato del primero y poder combinarlos a continuación. Esto incluye la modificación de nombres de variables, la reorganización de columnas y el ajuste de tipos de variables. Una vez ambos conjuntos de datos están en el mismo formato, se combinan para formar el conjunto “flights_combined”. Por otro lado, se han combinado los dos conjuntos de datos del clima, sin necesidad de hacer ninguna transformación pues tenían la misma estructura y variables. Se combinan para crear el conjunto de datos “weather”.

Posteriormente, es necesario realizar una serie de transformaciones en los conjuntos “weather” y “flights_combined” antes de combinarlos. Las variables utilizadas para la unión deben estar en un formato consistente. Tras estas transformaciones, ambos conjuntos de datos se exportan a Python para su combinación. En Python, se realizan ajustes en los tipos de variables, convirtiendo *DATE* del conjunto de vuelos y *date* del conjunto de clima al formato “datetime64”. La combinación de los dos conjuntos se lleva a cabo utilizando las columnas *ORIGIN_AIRPORT*, *DATE*, *TIME2* de los datos de vuelos y *airport*, *date*, *time* de los datos climáticos como claves de unión. Esta unión se realiza mediante una unión izquierda, conservando todas las filas del conjunto de datos de vuelos y añadiendo solo las correspondientes del conjunto de datos climáticos. Finalmente, el conjunto de datos combinado se exporta a un archivo CSV denominado “data_combined.csv”, el cual contiene 6305244 registros y se carga posteriormente en R Studio.

Tabla 5. *Resumen de transformaciones necesarias para la combinación de los conjuntos de datos*

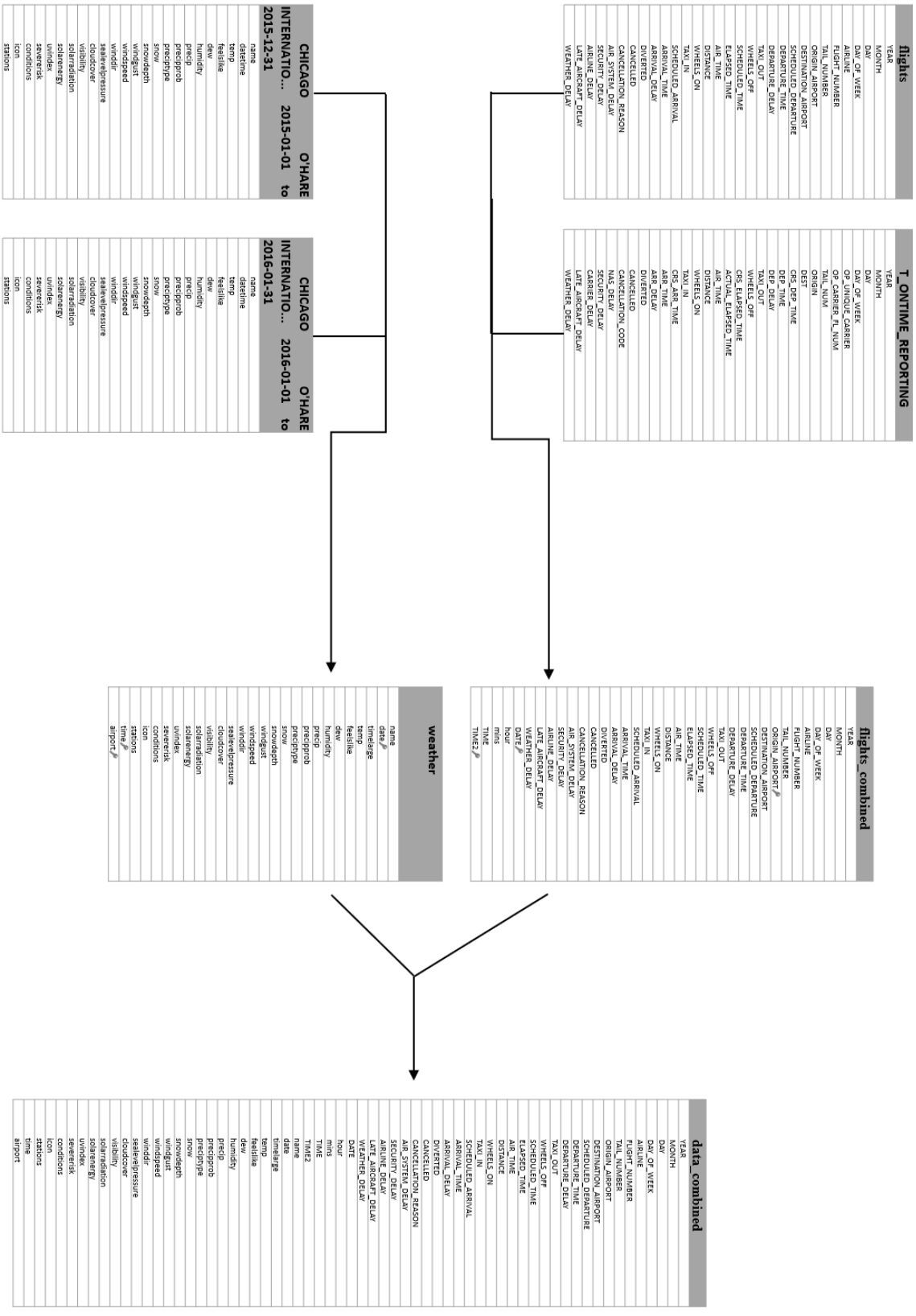
Tipo de transformación	Campos afectados	Fórmula	Conjunto de datos
Cambio nombre de variable	*DAY_OF_MONTH, OP_UNIQUE_CARRIER, OP_CARRIER_FL_NUM, TAIL_NUM, ORIGIN, DEST, CRS_DEP_TIME, DEP_TIME, DEP_DELAY, CRS_ARR_TIME, ARR_TIME, ARR_DELAY, CANCELLATION_CODE, CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME, CARRIER_DELAY, NAS_DELAY*	names(oct2015)[names(oct2015) == "DAY_OF_MONTH"] <- "DAY"	T_ONTIME_REPORTING.csv (denominado oct2015 en el código)
Reorganizar variables	*YEAR, MONTH, DAY_OF_MONTH, DAY_OF_WEEK, OP_UNIQUE_CARRIER, TAIL_NUM, OP_CARRIER_FL_NUM, ORIGIN, DEST, CRS_DEP_TIME, DEP_TIME, DEP_DELAY, TAXI_OUT, WHEELS_OFF, WHEELS_ON, TAXI_IN, CRS_ARR_TIME, ARR_TIME, ARR_DELAY, CANCELLED, CANCELLATION_CODE, DIVERTED, CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME, AIR_TIME, DISTANCE, CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY, LATE_AIRCRAFT_DELAY*	column_order_flights <- names(flights) oct2015 <- oct2015[, column_order_flights, drop = FALSE]	T_ONTIME_REPORTING.csv
Cambiar tipo de variable	*YEAR, MONTH, DAY, DAY_OF_WEEK, AIRLINE, FLIGHT_NUMBER,*	column_types_flights <- sapply(flights, class) oct2015 <- data.frame(Map(function(x, y) as(x, y), oct2015, column_types_flights))	T_ONTIME_REPORTING.csv

	TAIL_NUMBER, ORIGIN_AIRPORT, DESTINATION_AIRPORT, SCHEDULED_DEPARTURE, DEPARTURE_TIME, DEPARTURE_DELAY, TAXI_OUT, WHEELS_OFF, SCHEDULED_TIME, ELAPSED_TIME, AIR_TIME, DISTANCE, WHEELS_ON, TAXI_IN, SCHEDULED_ARRIVAL, ARRIVAL_TIME, ARRIVAL_DELAY, DIVERTED, CANCELLED, CANCELLATION_REASON, AIR_SYSTEM_DELAY, SECURITY_DELAY, AIRLINE_DELAY, LATE_AIRCRAFT_DELAY, WEATHER_DELAY		
Creación de variable *(DATE)*	*DATE*	flights_combined$DATE <- as.Date(paste(flights_combined$YEAR, flights_combined$MONTH, flights_combined$DAY, sep = "-"))	flights_combined
Cambiar tipo de variable	*SCHEDULED_DEPARTURE*	flights_combined$SCHEDULED_DEPARTURE <- sprintf("%04d", flights_combined$SCHEDULED_DEPARTURE)	flights_combined
Creación de variable *(hour)*	*SCHEDULED_DEPARTURE*	flights_combined$hour <- as.numeric(substr(flights_combined$SCHEDULED_DEPARTURE, 1, nchar(flights_combined$SCHEDULED_DEPARTURE) - 2))	flights_combined
Creación de variable *(mins)*	*SCHEDULED_DEPARTURE*	flights_combined$mins <- as.numeric(substr(flights_combined$SCHEDULED_DEPARTURE, nchar(flights_combined$SCHEDULED_DEPARTURE) - 1, nchar(flights_combined$SCHEDULED_DEPARTURE)))	flights_combined
Creación de variable *(TIME)*	*hour, mins*	flights_combined$TIME <- ifelse(flights_combined$mins > 30, flights_combined$hour + 1, flights_combined$hour)	flights_combined
Transformación de variable	*DATE*	flights_combined$DATE <- ifelse(flights_combined$TIME == 24, flights_combined$DATE + 1, flights_combined$DATE)	flights_combined
Cambiar tipo de variable	*DATE*	flights_combined$DATE<- as.Date(flights_combined$DATE, origin = "1970-01-01")	flights_combined

Creación de variable *(TIME2)*	*TIME*	flights_combined$TIME2 <- ifelse(flights_combined$TIME != 24, flights_combined$TIME, 0)	flights_combined
Creación de variable *(date, timelarge)*	*datetime*	weather <- separate(weather, datetime, into = c("date", "timelarge"), sep = "T")	weather
Creación de variable *(time)*	*timelarge*	weather$time <- as.numeric(sub("^0*([0-9]+).*", "\\1", weather$timelarge))	weather
Creación de variable *(airport)*	*airport*	weather$airport <- "ORD"	weather
Cambiar tipo de variable	*DATE*	data$DATE <- as.Date(data$DATE, format = "%Y-%m-%d")	data_combined.csv
Cambiar tipo de variable	*date*	data$date <- as.Date(data$date, format = "%Y-%m-%d")	data_combined.csv
Cambiar tipo de variable	*precipprob*	data$precipprob <- as.integer(data$precipprob)	data_combined.csv
Cambiar tipo de variable	*uvindex*	data$uvindex <- as.integer(data$uvindex)	data_combined.csv
Cambiar tipo de variable	*time*	data$time <- as.integer(data$time)	data_combined.csv

Fuente: Elaboración propia

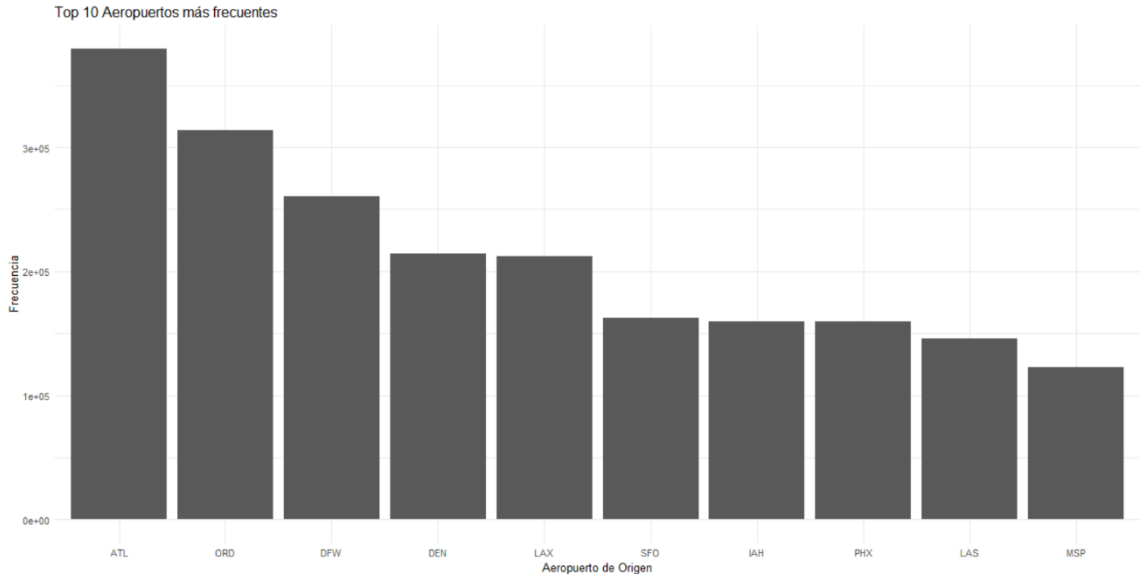
Figura 11. *Combinación de conjuntos de datos*



Fuente: Elaboración propia

3.4 Transformación de los datos

Figura 12. *10 aeropuertos de origen más frecuentes*



Fuente: Elaboración propia

El gráfico ilustra los diez aeropuertos de origen más frecuentes en el conjunto de datos, mostrando la frecuencia de vuelos para cada código de aeropuerto. La mayor frecuencia corresponde al Aeropuerto Internacional de Atlanta (ATL), seguido del Aeropuerto Internacional O'Hare de Chicago (ORD). Se selecciona el Aeropuerto Internacional O'Hare de Chicago para este estudio, ya que, además de ser el segundo en frecuencia de vuelos, presenta condiciones climáticas más extremas que otros aeropuertos de mayor tráfico, como ATL. Estas características lo hacen más útil para detectar retrasos debidos al clima, manteniendo un alto volumen de tráfico. Por lo tanto, se reduce el conjunto de datos a aquellos vuelos con el Aeropuerto Internacional O'Hare de Chicago (ORD) como aeropuerto de origen, resultando en 313536 registros.

Se eliminan las variables *TAIL_NUMBER, WHEELS_OFF, SCHEDULED_TIME, AIR_TIME, WHEELS_ON, TAXI_IN, SECURITY_DELAY, AIRLINE_DELAY, LATE_AIRCRAFT_DELAY, solarradiation, solarenergy* y *uvindex,* ya que no se consideran necesarias para la predicción de retrasos en la salida debidos al clima y están fuera del alcance del estudio. Además, se eliminan las variables creadas previamente para facilitar la combinación de los conjuntos de datos y que no son útiles para el análisis

posterior: *hour, mins, TIME, date, time* y *airport.* La variable *severerisk* también se elimina, ya que contiene únicamente valores vacíos.

Se cambian los tipos de algunas variables:

- *ORIGIN_AIRPORT* se cambia a tipo carácter, ya que incluye las siglas de los aeropuertos de origen.
- *DEPARTURE_TIME*, *ELAPSED_TIME*, *DEPARTURE_DELAY*, *TAXI_OUT*, *ARRIVAL_TIME*, *ARRIVAL_DELAY*, *AIR_SYSTEM_DELAY*, *WEATHER_DELAY* y *winddir* a tipo entero, ya que todas son números sin decimales.
- *DIVERTED*, *CANCELLED* y *precipprob* a tipo factor, ya que las dos primeras únicamente tienen valores de 0 o 1 y la ultima de 0 o 100.

Se tratan los valores faltantes, sustituyéndolos o eliminándolos

Variables numéricas

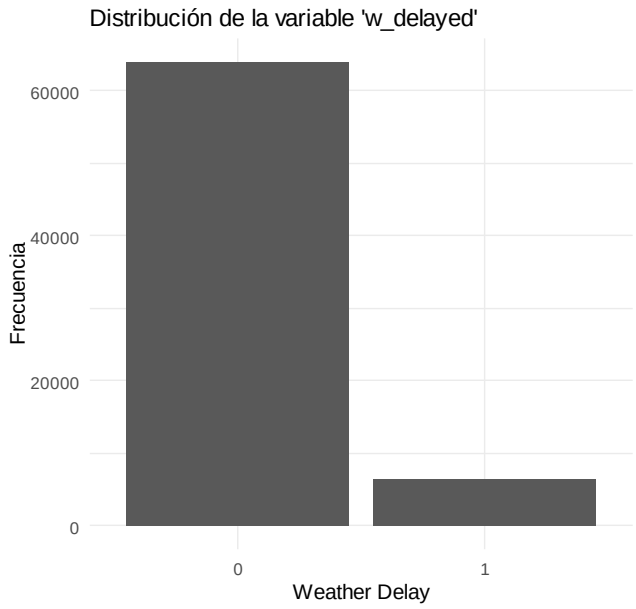
- *WEATHER_DELAY*: Los valores faltantes en esta variable ocurren cuando un vuelo no experimenta retraso, o cuando ha sido cancelado o desviado. Se sustituyen por 0 aquellos registros faltantes en los que el vuelo no ha sido cancelado, indicando así que no hubo retraso debido al clima. Por otro lado, a los registros vacíos en los que el vuelo fue cancelado debido al clima se les imputa un valor máximo de 180 minutos de *WEATHER_DELAY*. Este valor se asigna conforme a la normativa de retrasos en pista del Departamento de Transporte de Estados Unidos, la cual prohíbe que la mayoría de las aerolíneas permitan que un vuelo doméstico permanezca en la pista por más de tres horas, salvo por razones de seguridad u operativas. Según el Departamento de Transporte, "un retraso en la pista ocurre cuando un avión en tierra está esperando despegar o acaba de aterrizar y los pasajeros no tienen la oportunidad de bajarse del avión" (U.S. Department of Transportation, 2023).
- *Windgust*: los valores faltantes se dan cuando el criterio para una ráfaga de viento no se cumple y por lo tanto se da una ausencia de datos, en estos casos, se reemplaza el valor faltante por cero.

Variables categóricas

- *CANCELLATION_REASON*: se dan valores faltantes en aquellos casos en los que el vuelo no se cancela y por lo tanto no tiene razón de cancelación. Se sustituyen estos valores faltantes por *Not cancelled*.
- *preciptype*: se dan valores faltantes cuando no hay precipitación. Por ello cuando la precipitación sea cero se sustituyen estos valores faltantes por *No precip*.

Se crea la variable objetivo, denominada *w_delayed*, que representa la predicción de retrasos en la salida de los vuelos basándose en las condiciones climáticas en el aeropuerto de origen. A la variable *w_delayed* se le asigna un valor de 1 cuando el *WEATHER_DELAY*, es decir, los minutos de retraso atribuibles al clima, supera los 15 minutos. Si el *WEATHER_DELAY* es menor o igual a 15 minutos, se le asigna un valor de 0. Finalmente, la variable *w_delayed* se convierte en un factor con dos niveles: 0 y 1.

Figura 13. *Distribución de la variable objetivo*



Fuente: Elaboración propia

Se hace la proporción de cada nivel de la variable y revela que solo el 9,1% de las observaciones están en la clase positiva, la que representa el retraso en la salida de los

vuelos. Por lo tanto, se trata de un conjunto de datos imbalanceado, lo cual se deberá tener en cuenta más adelante a la hora de realizar los modelos.

Además de la variable objetivo, se crean nuevas variables adicionales con el objetivo de mejorar la explicación del modelo:

Tabla 6. *Nuevas variables creadas*

Tipo de transformación	Nueva variable	Variables incluidas	Formula
Creación de variable	*adverse_climate_condition*	humidity, precip, snow, snowdepth, windgust, windspeed, cloudcover, WEATHER_DELAY	dataORD2$adverse_climate_condition <- rowSums(dataORD2[, variables_interes] * pesos)
	season	MONTH	dataORD2$season <- ifelse(dataORD2$MONTH %in% c(12, 1, 2), "Winter", ifelse(dataORD2$MONTH %in% c(3, 4, 5), "Spring", ifelse(dataORD2$MONTH %in% c(6, 7, 8), "Summer", "Fall")))
	avg_wdelay_prev_flights	WEATHER_DELAY, FLIGHT NUMBER	dataORD2$avg_wdelay_prev_flights <- ave(dataORD2$WEATHER_DELAY, dataORD2$FLIGHT NUMBER, FUN = function(x) mean(x, na.rm = TRUE))
	departure_time_category	SCHEDULED_DEPARTURE	dataORD2$departure_time_category <- cut(dataORD2$SCHEDULED_DEPARTURE, breaks = c(-Inf, 600, 1200, 1900, Inf), labels = c("Madrugada", "Mañana", "Tarde", "Noche"), include.lowest = TRUE)
	wdelay_per_distance	WEATHER_DELAY, DISTANCE	dataORD2$wdelay_per_distance <- dataORD2$WEATHER_DELAY / dataORD2$DISTANCE
	wdelay_per_elapsedtime	WEATHER_DELAY, ELAPSED_TIME	dataORD2$wdelay_per_elapsedtime <- dataORD2$WEATHER_DELAY / dataORD2$ELAPSED_TIME

Fuente: Elaboración propia

Condición climática adversa: se crea calculando primero la correlación entre las distintas variables relacionadas con el clima y la variable objetivo, se selecciona únicamente aquellas que tienen una correlación positiva con ella y ya que no tienen correlaciones muy altas, se escogen todas para crear la nueva variable. La nueva variable se crea asignando pesos a cada una de estas variables, basándose en la correlación con la variable objetivo, esto se hace dividiendo cada correlación por la suma del valor absoluto de todas las correlaciones. Finalmente, crea una nueva variable llamada

adverse_climate_condition que representa una combinación ponderada de las variables climáticas seleccionadas, multiplicando cada valor de las variables climáticas por su respectivo peso y sumando los productos. Se crea esta variable ya que los retrasos debidos al clima pueden no solo deberse a una sola condición climática, si no que puede ser la combinación de distintas condiciones climáticas las que provoquen este retraso. De esta forma, en lugar de considerar cada condición climática por separado, tener una variable que represente las condiciones climáticas adversas de manera general puede ayudar a simplificar el modelo y a hacer que sea más fácil de interpretar.

Estación del año: se crea una variable que indica la estación del año, en función del mes. Si el mes es diciembre, enero o febrero se le asigna el valor *Winter*, si es marzo, abril o mayo se le asigna el valor *Spring*, si es junio, julio o agosto, se le asigna *Summer*, y de ser otro, se asigna *Fall*. Finalmente se convierte la variable en tipo factor con cuatro niveles. Se crea esta variable ya que las condiciones climáticas varían según estación y pueden afectar a la probabilidad de retrasos. Por ejemplo, es posible que los vuelos tengan más retrasos durante el invierno, cuando hay condiciones climáticas adversas como el hielo o la nieve.

Retraso promedio del clima en vuelos anteriores: se calcula la media de retraso por clima para cada grupo de vuelos con el mismo número de vuelo. Se crea esta nueva variable con el objetivo de identificar si hay patrones de retraso recurrentes para vuelos específicos y ayudar al modelo a predecir el retraso potencial en futuros vuelos con el mismo número.

Categoría de hora programada de salida: la variable agrupa las horas del día en cuatro categorías: madrugada, mañana, tarde y noche según la hora programada de salida. Se crea con el fin de observar si el momento del día tiene correlación con los retrasos y por lo tanto identificar si ciertos retrasos tienden a ocurrir con más frecuencia en ciertos momentos del día.

Retraso debido al clima por unidad de distancia: se dividen los minutos de retraso debidos al clima por la distancia del vuelo. Se hace con el objetivo de identificar patrones en la relación entre los retrasos debidos al clima y la distancia del vuelo. Por ejemplo, puede ser que los vuelos más largos experimenten más retrasos climáticos.

Retraso debido al clima por unidad de tiempo transcurrido: se dividen los minutos de retraso debidos al clima por el tiempo de vuelo transcurrido. La variable en un principio identificaría patrones parecidos a los de la variable anterior, ya que la distancia y el tiempo transcurrido son dos formas distintas pero equivalentes de medir la duración de un vuelo. Sin embargo, se crea esta variable con el fin de observar la correlación de ambas con la variable objetivo.

3.5 Metodología e implementación

3.5.1 Analítica descriptiva

Limpieza y transformación

Se crea un conjunto de datos y una serie de variables a partir de las ya existentes en el conjunto de datos "data_combined.csv" con el fin de representar gráficamente los datos de los que se parte. En la siguiente se tabla se muestran las transformaciones necesarias y que han sido realizadas.

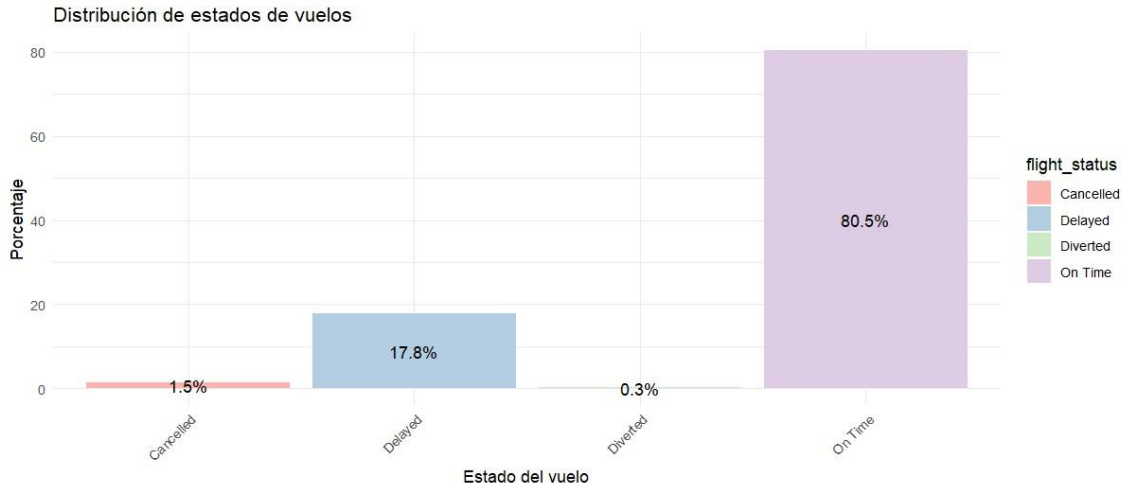
Tabla 7. *Transformaciones para analítica descriptiva*

Tipo de transformación	Campos afectados	Fórmula	Conjunto de datos
Creación de variable (flight_status)	*DIVERTED, CANCELLED, ARRIVAL_DELAY*	data$flight_status <- ifelse(data$DIVERTED == 1, "Diverted", ifelse(data$CANCELLED == 1, "Cancelled", ifelse(data$ARRIVAL_DELAY >= 15, "Delayed", "On Time")))	data_combined.csv
Creación de variable (delay_total_mins)	*AIR_SYSTEM_DELAY, SECURITY_DELAY, AIRLINE_DELAY, LATE_AIRCRAFT_DELAY, WEATHER_DELAY*	delay_total_mins <- colSums(data[, c("AIR_SYSTEM_DELAY", "SECURITY_DELAY", "AIRLINE_DELAY", "LATE_AIRCRAFT_DELAY", "WEATHER_DELAY")], na.rm = TRUE)	data_combined.csv
Creación de conjunto de datos	*delay_total_mins*	delay_percentage_data <- data.frame(delay_type = names(delay_total_mins), percentage = delay_total_mins / sum(delay_total_mins) * 100)	delay_percentage_data
Creación de variable	*ORIGIN_AIRPORT*	top10_airports <- names(sort(table(data$ORIGIN_AIRPORT), decreasing = TRUE)[1:10])	data_combined.csv

Fuente: Elaboración propia

Desarrollo

Figura 14. *Distribución de estados de vuelos de todos los aeropuertos*

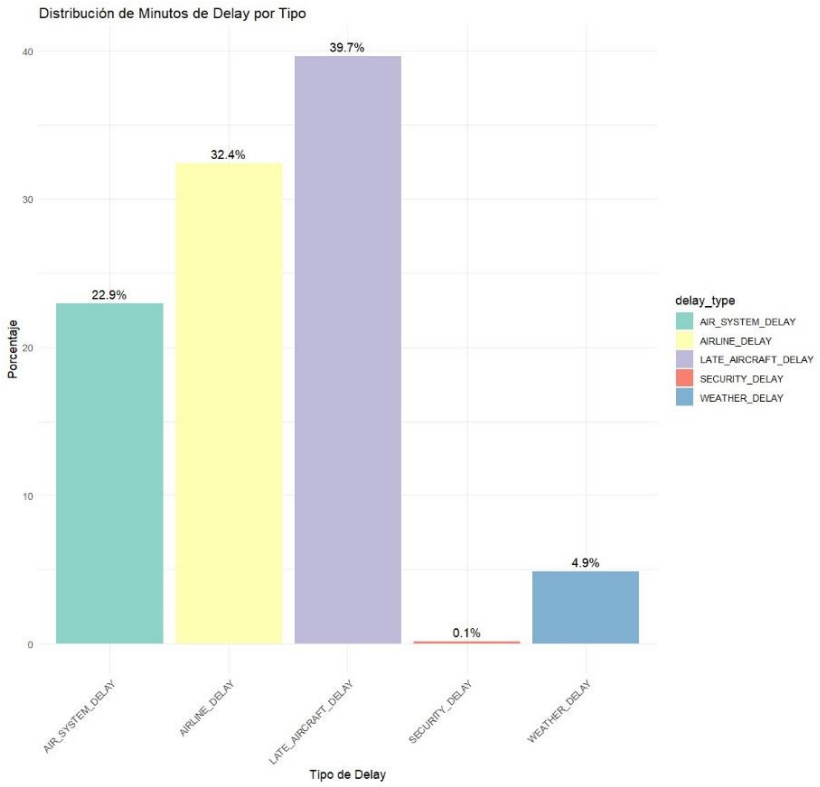


Fuente: Elaboración propia

Figura 15. *Distribución de estados de vuelos del aeropuerto de Chicago*

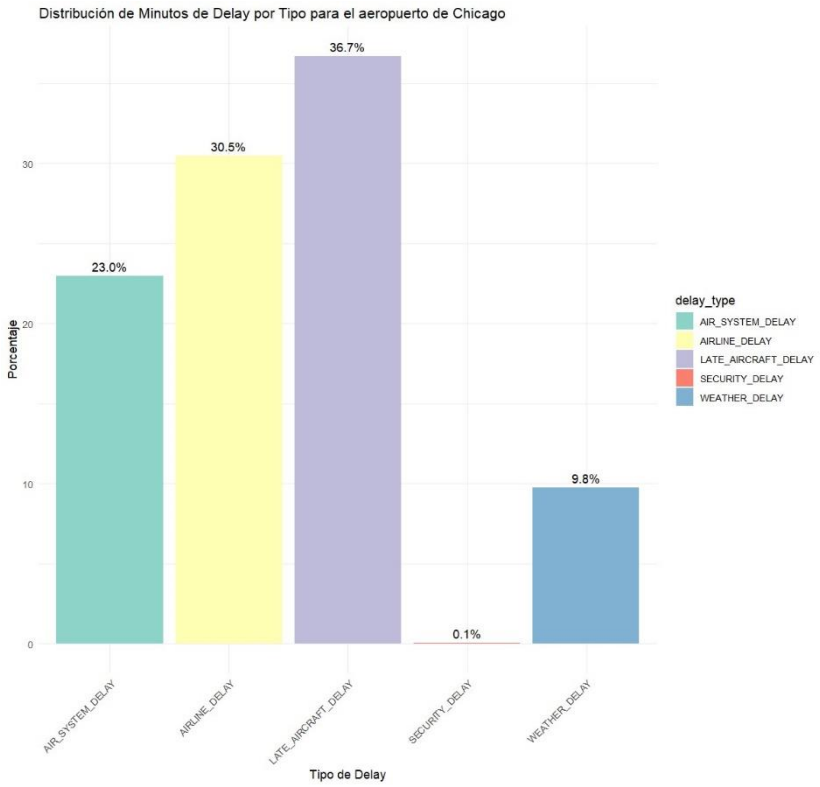
Fuente: Elaboración propia

Figura 16. *Distribución de minutos de retraso por tipo*



Fuente: Elaboración propia

Figura 17. *Distribución de minutos de retraso por tipo (Chicago ORD)*



Fuente: Elaboración propia

Resultado

El primer gráfico muestra la distribución del estado de los vuelos, teniendo en cuenta todos los datos de vuelos nacionales en Estados Unidos en el año 2015. La mayoría de los vuelos, un 80,5%, salieron a tiempo. Un 17,8% de los vuelos se retrasaron, lo cual es un porcentaje significativo y verifica que es un área de interés donde realizar mejoras operativas y estudios sobre las causas subyacentes de los retrasos. Por otro lado, solo un 1,5% de los vuelos se cancelaron y un 0,3% fueron desviados, si bien hay que tener en cuenta que estos dos últimos casos suelen causar más inconveniencia a los pasajeros. Debe tenerse en cuenta que un vuelo se considera retrasado cuando hay un retraso en la llegada superior a 15 minutos, según el Departamento de Transporte de Estados Unidos (U.S. Department of Transportation, n.d.). Al comparar este primer gráfico con el siguiente, el cual muestra los mismos datos, pero únicamente para el aeropuerto de Chicago, se observa como el porcentaje de vuelos cancelados es casi el doble en este

aeropuerto y los vuelos desviados muestran el mismo porcentaje. Por otro lado, un 22,4% de vuelos fueron retrasados, cifra superior a la de todos los aeropuertos y alrededor de un 6% menos de vuelos salieron a tiempo. Por lo tanto, podría decirse que el aeropuerto de Chicago presenta un panorama algo más desfavorable en cuanto a el estado de los vuelos, en comparación con los datos de todos los aeropuertos.

En el tercer gráfico se observa la distribución de los minutos de retraso por tipo. La mayoría de los retrasos se deben al sistema de control del tráfico aéreo y a problemas de la aerolínea. La llegada retrasada de las aeronaves también constituye una parte considerable de los retrasos, lo cual podría estar relacionado con los retrasos propagados a lo largo del día. Aunque menos frecuentes, los retrasos meteorológicos representan un 4,9% y presentan un desafío único debido a su naturaleza impredecible. Sin embargo, este estudio opta por centrarse en estos retrasos climáticos, aprovechando la disponibilidad de datos meteorológicos detallados. Al identificar qué factores climáticos ejercen mayor influencia en los retrasos y anticipándolos en las previsiones, las aerolíneas pueden tomar medidas preventivas como reprogramar vuelos y comunicar cambios a tiempo, lo que resulta en una mejora de la experiencia del cliente, seguridad, optimización de costes y una gestión más eficiente de sus operaciones. En el siguiente gráfico se observan los minutos de retraso por tipo, específicamente para el aeropuerto de Chicago. Al compararse, salta a la vista como el tipo de retraso objeto de este estudio, *WEATHER_DELAY*, duplica su porcentaje al analizarlo sobre este aeropuerto en concreto, siendo en este caso un 9,8% y un 4,9% si se tienen en cuenta todos los aeropuertos. Estas cifras cobran sentido al tener en cuenta el tipo de clima extremo que se da en la ciudad de Chicago, protagonizado por la humedad, nieve, viento y nubes.

3.5.2 Analítica visual

Limpieza y transformación

La siguiente tabla muestra las diversas transformaciones realizadas para limpiar y preparar el conjunto de datos para las visualizaciones. Las transformaciones más frecuentes incluyen la limitación de variables, la creación de nuevas variables y el cambio de tipo de variable.

Tabla 8. *Transformaciones realizadas para la analítica visual*

Tipo de transformación	Campos afectados	Fórmula	Conjunto de datos
Seleccionar registros sin valores faltantes	*DEPARTURE_DELAY, TAXI_OUT, ARRIVAL_DELAY, AIR_SYSTEM_DELAY*	dataORD2_cleaned <- dataORD2[complete.cases(dataORD2[, selected_vars]),]	dataORD2
Limitar la variable	*WEATHER_DELAY*	dataORD2$WEATHER_DELAY <- pmin(dataORD2$WEATHER_DELAY, 180)	dataORD2
Cambio tipo de variable	*cloudcover*	intervalos <- c(0, 25, 50, 75, 100) etiquetas <- c("Sin nubes", "Pocas nubes", "Nublado", "Muy nublado") dataORD2$cloudcover_categorico <- cut(dataORD2$cloudcover, breaks = intervalos, labels = etiquetas, include.lowest = TRUE)	dataORD2
Limitar la variable	*snow, snowdepth*	data_filtered <- dataORD2[dataORD2$snow > 0 & dataORD2$snowdepth > 0,]	dataORD2
Reemplazar valores faltantes	*windgust*	data_filtered$windgust <- ifelse(is.na(data_filtered$windgust), 0, data_filtered$windgust)	data_filtered
Limitar la variable	*windgust, windspeed*	data_filtered <- data_filtered[data_filtered$windgust > 0 & data_filtered$windspeed > 0,]	data_filtered
Limitar la variable	*precip*	data_filtered <- data_filtered[data_filtered$precip > 0]	data_filtered
Cambiar tipo de variable	*visibility*	dataORD2$categoria_visibility <- cut(dataORD2$visibility, breaks = c(0, 4, 8, 12, 16, Inf), labels = c("0-4", "4-8", "8-12", "12-16", ">16"))	dataORD2
Cálculo sobre variable	*cloudcover, visibility*	data_summary <- dataORD2 %>% group_by(YEAR, MONTH, week_in_month) %>% summarise(mean_cloudcover = mean(cloudcover), mean_visibility = mean(visibility, na.rm = TRUE))	dataORD2
Creación de variable	*WEATHER_DELAY*	delayed_flights <- dataORD2 %>% filter(WEATHER_DELAY > 15) %>% group_by(MONTH) %>% summarize(num_delayed_flights = n())	dataORD2
Creación de variable	*WEATHER_DELAY*	media_retrasos <- dataORD2 %>% group_by(AIRLINE) %>%	dataORD2

		summarize(Media_Retrasos = mean(WEATHER_DELAY > 15, na.rm = TRUE))	
Creación de variable	*conditions*	monthly_summary <- data_filtered %>% group_by(MONTH, conditions) %>% summarise(Num_Retrasos = n()) %>% ungroup()	data_filtered
Creación de variable	*DESTINATION_AIRPORT*	media_salidas <- dataORD2 %>% group_by(DESTINATION_AIRPORT) %>% summarize(Media_Retrasos_Salida = mean(DEPARTURE_DELAY, na.rm = TRUE)) %>% arrange(desc(Media_Retrasos_Salida)) %>% slice(1:10)	dataORD2
Creación de variable	*DESTINATION_AIRPORT*	media_llegadas <- dataORD2 %>% group_by(DESTINATION_AIRPORT) %>% summarize(Media_Retrasos_Llegada = mean(ARRIVAL_DELAY, na.rm = TRUE)) %>% arrange(desc(Media_Retrasos_Llegada)) %>% slice(1:10)	dataORD2
Creación de variable	*DESTINATION_AIRPORT*	media_retrasos <- inner_join(media_salidas, media_llegadas, by = "DESTINATION_AIRPORT")	dataORD2

Fuente: Elaboración propia

Desarrollo, resultado e interpretación

Las visualizaciones a continuación representan el conjunto de datos de 2015 para los vuelos con salida desde el Aeropuerto Internacional de Chicago O'Hare. Gracias a la estructura de programación modular y a la aplicación de un tema unificado para todas las visualizaciones, estas pueden ser fácilmente adaptadas para otros aeropuertos, diferentes años o para su integración en una empresa concreta.

Tabla 9. *Estadísticos básicos*

	Mínimo	Máximo	Media	Mediana	Cuartil 1	Cuartil 3	Moda	Desviación estándar	Rango
SCHEDULED_DEPARTURE	1.00	2357.000	1.466077e+03	1515.00	1105.0	1839.00	1800.00	442.75	2356.000
DEPARTURE_TIME	1.00	2400.000	1.537619e+03	1604.00	1200.0	1932.00	1357.00	473.89	2399.000
DEPARTURE_DELAY	-21.00	1258.000	5.775603e+01	41.00	20.0	76.00	0.00	62.65	1279.000
TAXI_OUT	1.00	162.000	2.580341e+01	21.00	15.0	31.00	16.00	16.51	161.000
ELAPSED_TIME	28.00	633.000	1.406188e+02	124.00	91.0	171.00	113.00	70.69	605.000
DISTANCE	67.00	4243.000	7.517139e+02	622.00	334.0	925.00	733.00	529.87	4176.000
SCHEDULED_ARRIVAL	1.00	2359.000	1.653038e+03	1720.00	1310.0	2059.00	1900.00	491.50	2358.000
ARRIVAL_TIME	1.00	2400.000	1.568772e+03	1715.00	1229.0	2057.00	2230.00	632.94	2399.000
ARRIVAL_DELAY	15.00	1265.000	6.183562e+01	41.00	24.0	77.00	15.00	60.47	1250.000
AIR_SYSTEM_DELAY	0.00	614.000	1.419649e+01	2.00	0.0	19.00	0.00	26.84	614.000
WEATHER_DELAY	0.00	991.000	6.035498e+00	0.00	0.0	0.00	0.00	27.10	991.000
Temp	-22.10	33.400	1.030307e+01	12.90	-0.6	21.80	-1.10	13.25	55.500
Feelslike	-33.10	37.000	8.036157e+00	12.90	-5.3	21.80	26.20	16.32	70.100
Dew	-27.80	24.900	4.130590e+00	5.60	-4.5	15.00	19.30	12.65	52.700
Humidity	13.61	100.000	6.810326e+01	68.76	55.3	81.87	92.66	17.38	86.390
Precip	0.00	20.176	2.315358e-01	0.00	0.0	0.00	0.00	1.06	20.176
Snow	0.00	7.890	2.037689e-02	0.00	0.0	0.00	0.00	0.24	7.890
Snowdepth	0.00	35.500	2.565166e+00	0.00	0.0	1.67	0.00	5.88	35.500
Windgust	0.00	81.300	1.453397e+01	0.00	0.0	35.40	0.00	20.62	81.300
Windspeed	0.00	53.000	1.784189e+01	16.70	11.7	23.30	11.30	8.61	53.000
Winddir	0.00	360.000	1.829454e+02	202.00	70.0	268.00	209.00	106.37	360.000
Sealevelpressure	986.20	1043.900	1.016929e+03	1016.20	1011.9	1022.30	1014.30	8.59	57.700
Cloudcover	0.00	100.000	7.486505e+01	88.70	48.2	100.00	100.00	30.93	100.000
visibility	0.00	16.000	1.382047e+01	16.00	14.3	16.00	16.00	4.31	16.000

Fuente: Elaboración propia

La siguiente tabla muestra un resumen de las variables que se representan en el análisis de datos exploratorio y el tipo de gráfico que se visualiza.

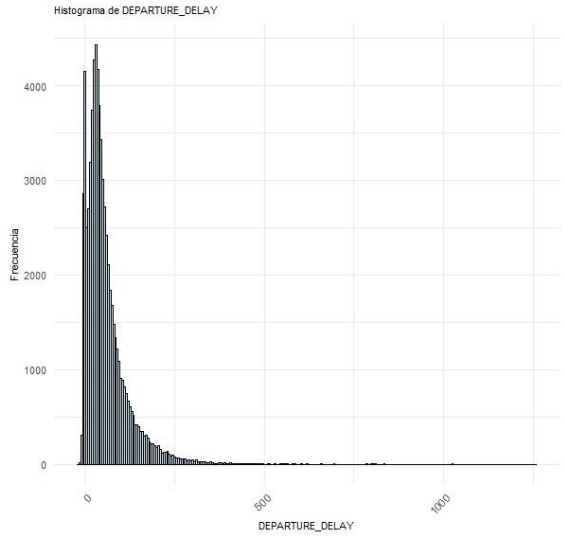
Tabla 10. *Resumen de variables en las visualizaciones siguientes*

Variable	Tipo de gráfico
DEPARTURE_DELAY	histograma
TAXI_OUT	histograma
ARRIVAL_DELAY	histograma
AIR_SYSTEM_DELAY	histograma
WEATHER_DELAY	histograma
temp	histograma, serie temporal
feelslike	histograma, serie temporal
dew	histograma, serie temporal
humidity	histograma, serie temporal
cloudcover	histograma, serie temporal
snow	histograma, serie temporal
snowdepth	histograma, serie temporal
windgust	histograma, serie temporal
windspeed	histograma, serie temporal
precip	histograma, serie temporal
visibility	histograma, gráfico de tarta, serie temporal
conditions	histograma
preciptype	histograma
icon	histograma
adverse_climate_condition	histograma
season	histograma
avg_wdelay_prev_flights	histograma
departure_time_category	histograma
wdelay_per_distance	histograma
wdelay_per_elapsedtime	histograma, gráfico de dispersión

Fuente: Elaboración propia

Análisis de la variable *DEPARTURE_DELAY*

Figura 18. *Histograma de la variable DEPARTURE_DELAY*



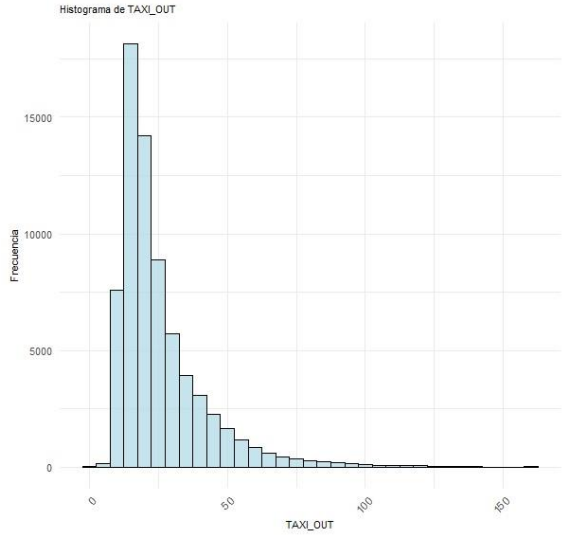
Fuente: Elaboración propia

El histograma representa la distribución de los retrasos en la salida de vuelos. El eje horizontal representa el tiempo de retraso en minutos y el vertical indica la frecuencia de vuelos que experimentaron esos retrasos. La mayoría de los vuelos tienen retrasos cercanos a cero minutos, con una alta frecuencia en esos valores. A medida que los minutos de retraso aumentan, la frecuencia de vuelos disminuye, mostrando una distribución sesgada hacia la derecha. Lo cual indica que, aunque hay algunos vuelos con retrasos muy largos, estos son menos frecuentes.

En conclusión, la mayoría de los vuelos salen a tiempo o con un retraso mínimo, lo cual es positivo para las aerolíneas y los pasajeros. Sin embargo, existe un pequeño número de vuelos que tienen retrasos muy largos. Abordar y mejorar estos retrasos largos sería un punto de mejora clave para las aerolíneas.

Análisis de la variable *TAXI_OUT*

Figura 19. *Histograma de la variable TAXI_OUT*



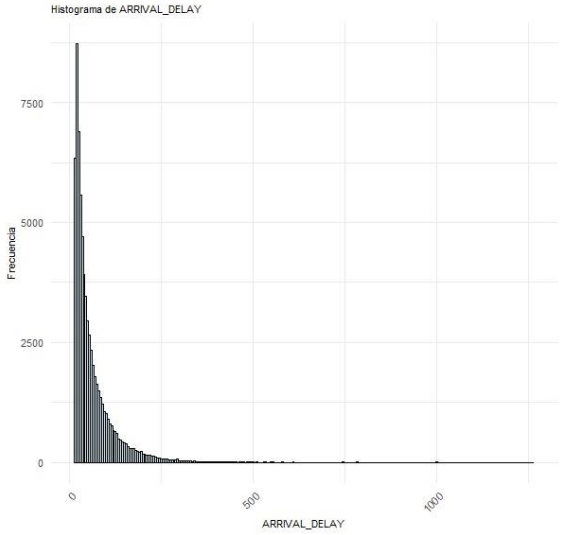
Fuente: Elaboración propia

El histograma representa la distribución de minutos que pasan rodando los aviones desde la puerta de embarque hasta la pista, expresado en el eje horizontal. Mientras que el eje vertical indica la frecuencia de los vuelos con cada uno de esos tiempos. Se observa un pico claro de frecuencia, en torno a los 15 minutos, que sugiere que este es el tiempo de *TAXI_OUT* más común en la mayoría de los vuelos. Además, los valores van disminuyendo a medida que aumenta el tiempo de *TAXI_OUT*, lo cual implica una distribución sesgada a la derecha, lo cual indica que, aunque hay algunos vuelos con tiempos de rodaje más largos, estos son menos comunes.

En conclusión, la mayoría de los tiempos de *TAXI_OUT* son relativamente cortos. Sin embargo, existen casos menos frecuentes donde los tiempos son más largos. Esto podría estar relacionado con la congestión del aeropuerto o con las condiciones meteorológicas.

Análisis de la variable *ARRIVAL_DELAY*

Figura 20. *Histograma de la variable ARRIVAL_DELAY*

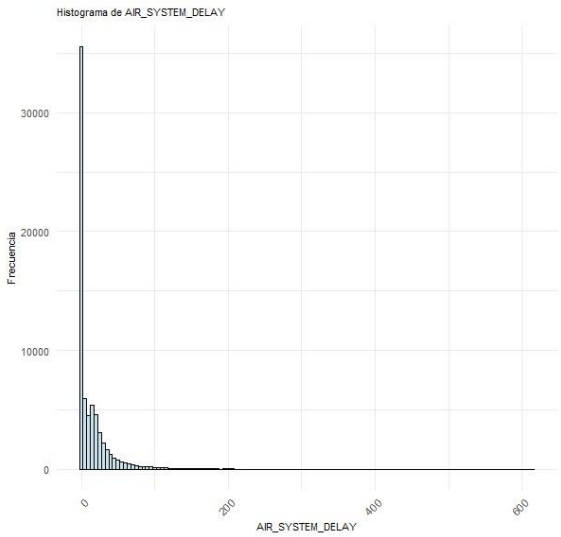


Fuente: Elaboración propia

El histograma representa la distribución de los retrasos en la llegada de los vuelos. Representado en minutos en el eje horizontal, mientras que el vertical indica la frecuencia de los vuelos con esos retrasos. Se observa como la mayoría de los vuelos tienen retrasos cercanos a cero minutos, con una alta frecuencia de vuelos en estos valores. Y a medida que los retrasos aumentan, la frecuencia disminuye rápidamente, indicando de nuevo que se trata de una distribución sesgada a la derecha. En conclusión, la mayoría de los vuelos llegan puntuales o con un pequeño retraso, lo cual es buena señal para las aerolíneas. Aun así, existen retrasos significativos que pueden llegar a superar los 1000 minutos y aunque estos casos son raros, deben abordarse las causas de estos retrasos para mejorar la puntualidad en todos los vuelos.

Análisis de la variable *AIR_SYSTEM_DELAY*

Figura 21. *Histograma de la variable AIR_SYSTEM_DELAY*

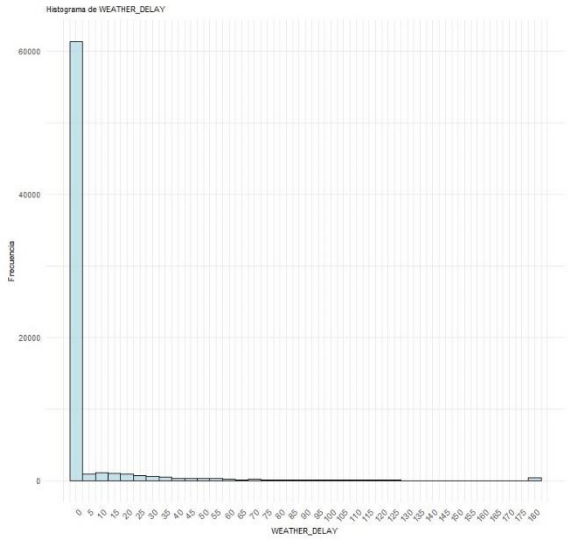


Fuente: Elaboración propia

El gráfico muestra la distribución de los retrasos debidos al sistema aéreo, representados en minutos en el eje horizontal. La gran mayoría de los vuelos tienen retrasos entorno a cero minutos, se observa por la frecuencia extremadamente alta en estos valores. La distribución esta sesgada a la derecha ya que la frecuencia va disminuyendo conforme el tiempo de retraso aumenta. Hay algunos casos raros en los que el tiempo de retraso puede llegar hasta los 600 minutos. En conclusión, la mayoría de los vuelos tienen retrasos debidos al sistema aéreo mínimos o nulos. Sin embargo, existen casos con retrasos muy grandes los cuales deben abordarse para mejorar los retrasos totales en los vuelos.

Análisis de la variable *WEATHER_DELAY*

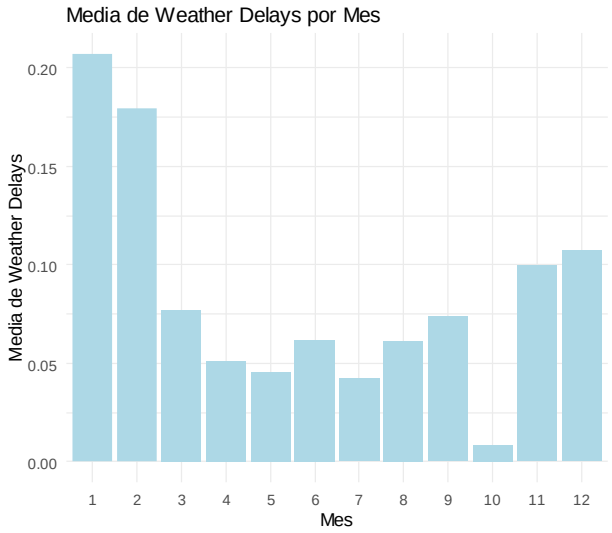
Figura 22. *Histograma de la variable WEATHER_DELAY*



Fuente: Elaboración propia

El histograma muestra la distribución de retrasos debidos al clima, presenta también una distribución de cola larga y sesgada a la derecha, en la que la mayoría de los retrasos climáticos son de corta duración y conforme aumenta su duración, disminuye la frecuencia de retrasos rápidamente y los retrasos largos son menos comunes. La mayoría de los retrasos climáticos están en el rango de 0 a 10 minutos. Aunque los retrasos largos de este tipo son menos comunes, observarlos en detalle también sería útil para ayudar a las aerolíneas a planificarse de cara a eventos climáticos extremos en los que pueden producirse retrasos de hasta 3 horas o más.

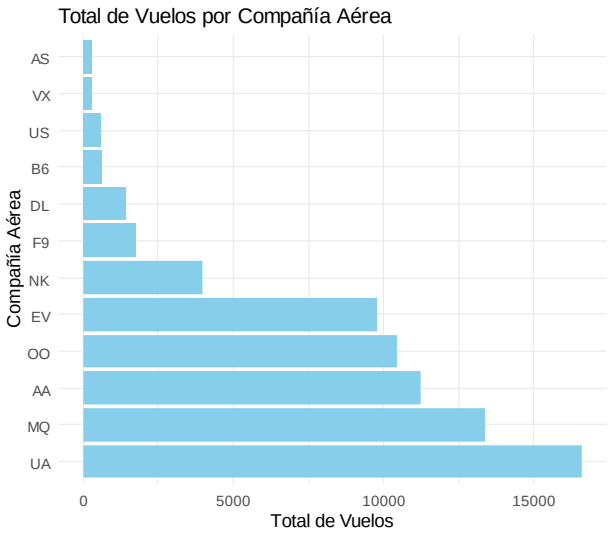
Figura 23. *Media de WEATHER_DELAY por mes*



Fuente: Elaboración propia

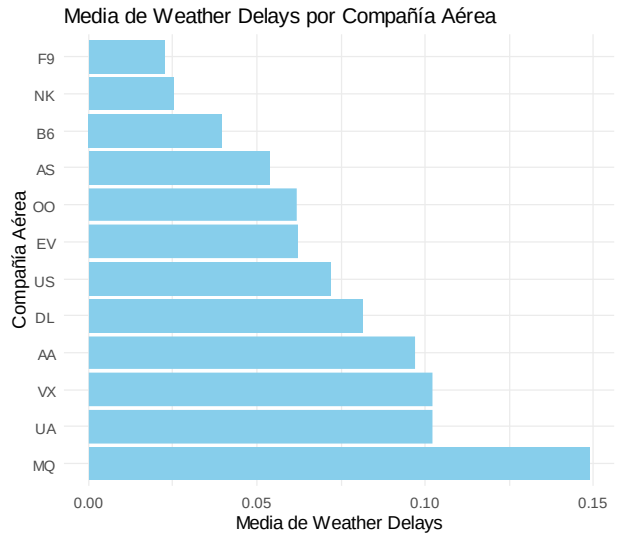
En este gráfico se observa la media mensual de vuelos retrasados por el clima, específicamente aquellos con retrasos mayores a 15 minutos en comparación con el número total de vuelos al mes. Se observa como la proporción es mayor en los primeros y últimos meses del año. Teniendo el mes de enero la mayor proporción con el 0,20 o 20% del total de vuelos, teniendo retrasos debidos al clima mayores a un cuarto de hora. Que estos valores sean mayores en estos meses puede ser debido a las condiciones climáticas durante los meses de invierno, las cuales pueden ser más severas.

Figura 24. *Total de vuelos por compañía aérea*



Fuente: Elaboración propia

Figura 25. *Media de WEATHER_DELAY por compañía aérea.*



Fuente: Elaboración propia

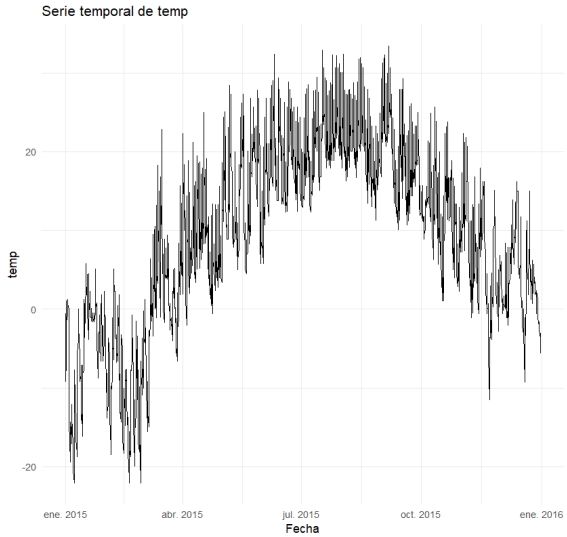
El primer gráfico muestra el total de vuelos por compañía aérea. Siendo la compañía con más vuelos UA (*United Airlines*), seguida por MQ (*American Eagle Airlines*) y AA (*American Airlines*). En el gráfico de la derecha representa la media de vuelos con retraso debido al clima mayor a 15 minutos, sobre el total de vuelos. Se observa cómo no la compañía con más vuelos tiene porque tener una mayor media de retrasos por clima. Siendo sin embargo MQ la que tiene una mayor media, en comparación a el resto, tiene

una media de casi 0,15, lo que significa que aproximadamente el 15% de sus vuelos tienen un retraso debido al clima mayor a un cuarto de hora. La siguen las aerolíneas UA y VX con un 10% de media de retrasos. Puede verse como aun siendo UA la compañía con más vuelos que salen del aeropuerto Internacional de Chicago, consigue no ser la que más media de retrasos tiene.

En conclusión, los gráficos muestran como los retrasos debidos al clima suelen ser cortos o nulos, sin embargo, también hay vuelos que sufren largos retrasos, de hasta 3 horas, habiendo puesto el límite para el gráfico en este valor ya que es el máximo legalmente permitido. Además, la distribución de los retrasos por clima varía a lo largo del año, habiendo un mayor porcentaje de vuelos retrasados en los meses de invierno, sugiriendo así una estacionalidad en los retrasos debidos a condiciones climáticas. Este tipo de retrasos también varía según la aerolínea, siendo American Eagle Airlines la que cuenta con más retrasos debidos al clima, ya que de media un 15% de sus vuelos tienen este tipo de retraso. En resumen, aunque este tipo de retrasos no son de los más comunes y generalmente son cortos, existen diferencias estacionales y entre aerolíneas a tener en cuenta para mejorar la puntualidad de los vuelos.

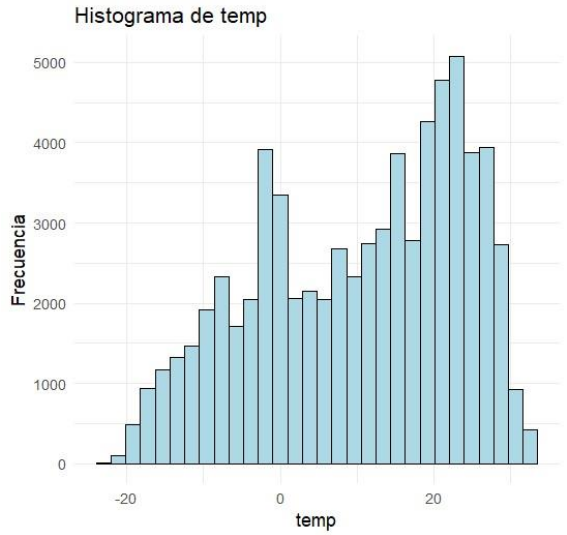
Análisis de la variable *temp*

Figura 26. *Serie temporal de la variable temp*



Fuente: Elaboración propia

Figura 27. *Histograma de la variable temp*



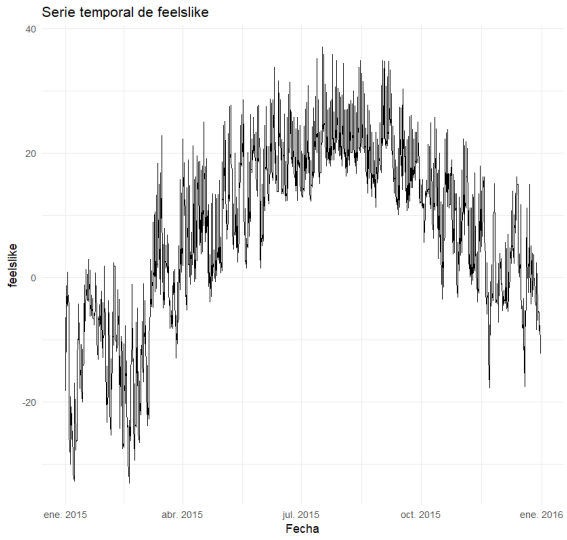
Fuente: Elaboración propia

Esta variable muestra claramente un patrón estacional con temperaturas bajas e incluso negativas a principios y finales de año y más altas en los meses centrales del año, alcanzando su punto máximo en julio y el mínimo en enero. Además, la temperatura

presenta cierta variabilidad diaria, lo cual concuerda con el clima en Chicago. Además, el histograma presenta la distribución de la temperatura, con valores de -20 a 30 grados, siendo los más frecuentes aquellos en torno a -5-0 grados y 20-25 grados. En conclusión, los gráficos representan un clima estacional que varía en función del momento del año.

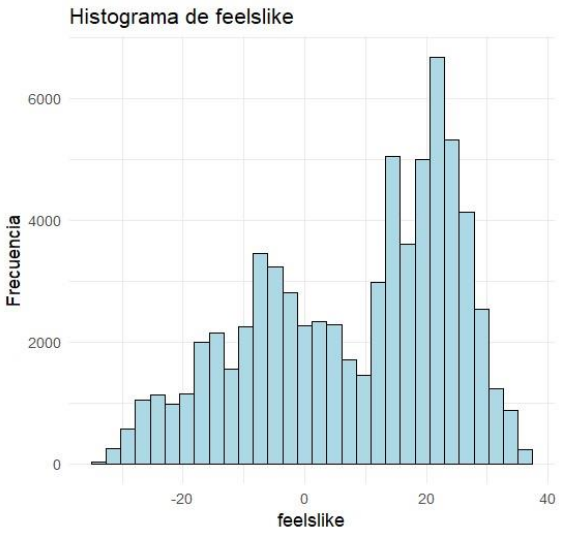
Análisis de la variable *feelslike*

Figura 28. *Serie temporal de la variable feelslike*



Fuente: Elaboración propia

Figura 29. *Histograma de la variable feelslike*

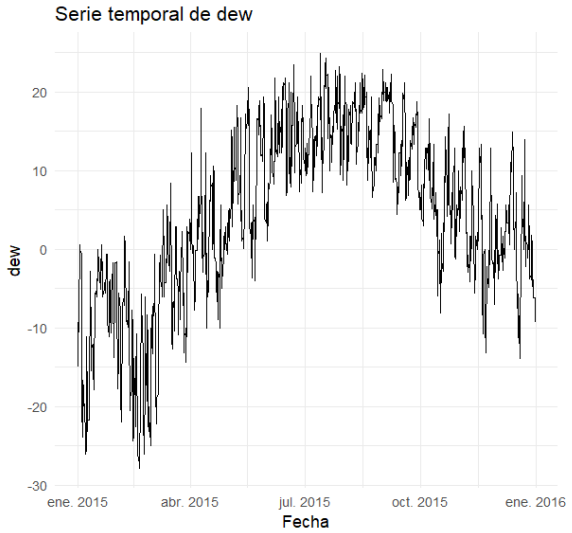


Fuente: Elaboración propia

La variable muestra un patrón estacional, al igual que la anterior y con una distribución similar, pues están relacionadas, pero con algunas diferencias pues la sensación térmica aparte de estar relacionada con la temperatura real, también lo está con el viento y la humedad. Se observa que las condiciones extremas de sensación térmica son poco comunes. También, que llega a haber sensación térmica por encima de los 30 grados, cosa que no se da en la temperatura y lo cual puede deberse a niveles altos de humedad o el índice de calor, ya que cuando no se trata de extremos, la sensación térmica se mide por la temperatura real. En los meses de invierno la sensación térmica es incluso menor a la temperatura, así como en los meses de verano donde la sensación térmica es mayor a la temperatura real. El histograma muestra como la temperatura más común o con mayor frecuencia es aquella entre los 20 y 25 grados. el histograma de la sensación térmica parece tener una distribución similar a la temperatura, En conclusión, los gráficos, siendo muy parecidos a los de la temperatura, explican un clima estacional.

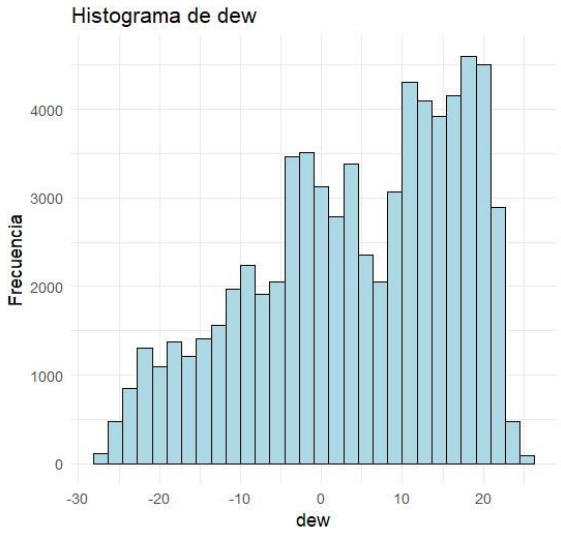
Análisis de la variable *dew*

Figura 30. *Serie temporal de la variable dew*



Fuente: Elaboración propia

Figura 31. *Histograma de la variable dew*

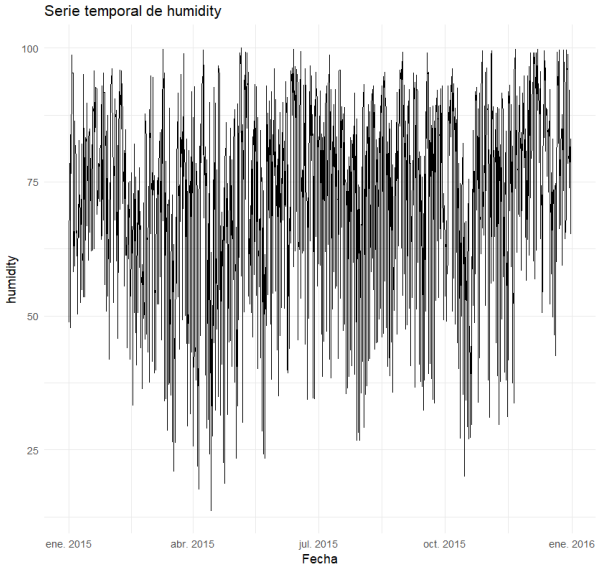


Fuente: Elaboración propia

La variable muestra un claro patrón estacional y es una medida que refleja a su vez la humedad, la cual suele ir relacionada con temperaturas altas, es por eso por lo que tiene sentido que haya valores más altos en verano y más bajos en invierno. Durante los meses de verano, el punto de rocío es más alto por la humedad, lo que a su vez contribuye a una mayor sensación térmica. La serie temporal tiene una forma parecida a la de la temperatura, pero con una mayor variabilidad.

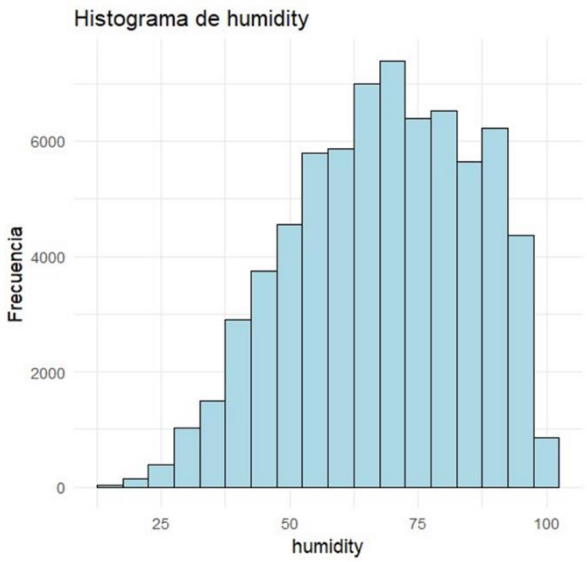
Análisis de la variable *humidity*

Figura 32. *Serie temporal de la variable humidity*



Fuente: Elaboración propia

Figura 33. *Histograma de la variable humidity*

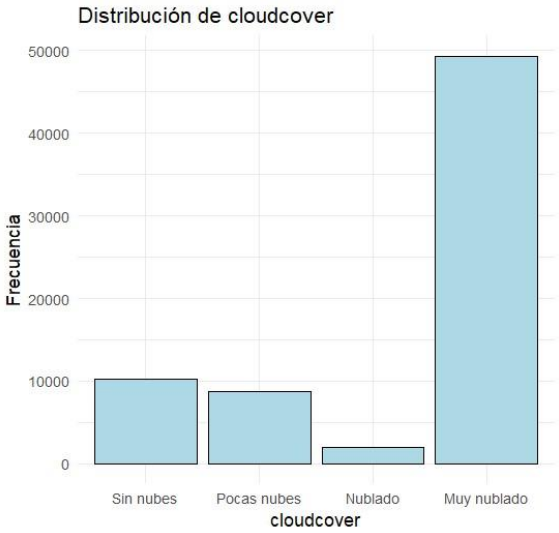


Fuente: Elaboración propia

Los gráficos presentan la humedad medida en porcentaje, el histograma de la derecha presenta una distribución ligeramente sesgada a la derecha con una media del 75%, por lo que puede considerarse un clima húmedo. La humedad está relacionada con la temperatura y el punto de rocío, de esta forma, la serie temporal de la humedad muestra una forma parecida a las anteriores, aumentando en los meses de verano, pero con más variabilidad.

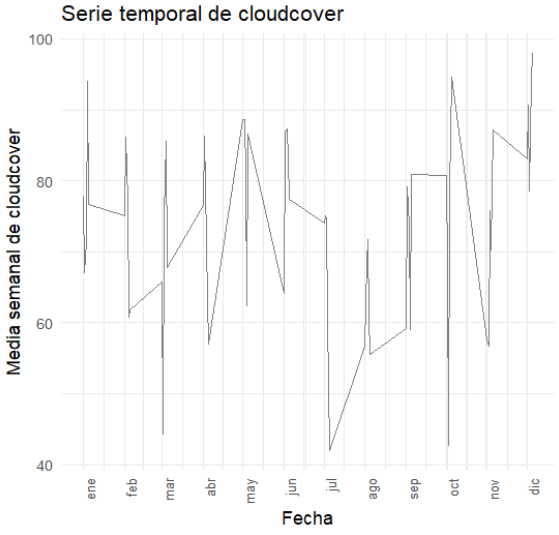
Análisis de la variable *cloudcover*

Figura 34. *Histograma de la variable cloudcover*



Fuente: Elaboración propia

Figura 35. *Serie temporal de la variable cloudcover*



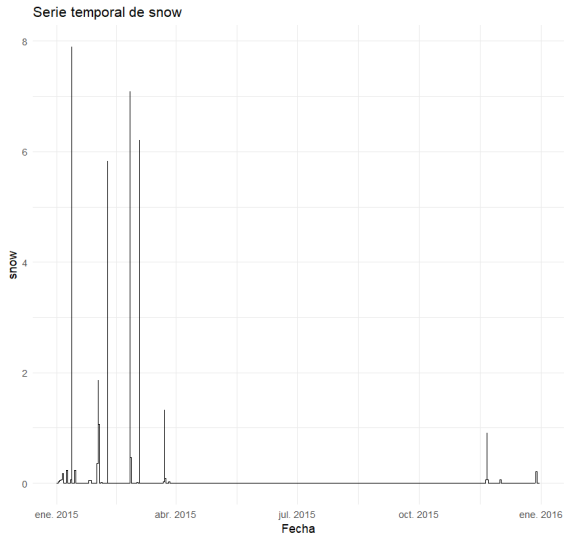
Fuente: Elaboración propia

El gráfico muestra como la mayoría de los datos tienen la condición “Muy nublado”, lo cual implica que el cielo está entre un 75% y 100% cubierto por nubes. El siguiente

gráfico representa la media por semana de la cobertura de las nubes a lo largo de todos los meses del año, mostrando una alta variabilidad ya que presenta un rango que va aproximadamente desde el 40% hasta casi el 100%. No hay datos que indiquen una cobertura inferior al 40%, lo cual sugiere que durante el año 2015 el cielo estuvo siempre parcial o completamente nublado. Además, se observa cierta estacionalidad, con menos nubes en los meses de verano y más en los meses de invierno.

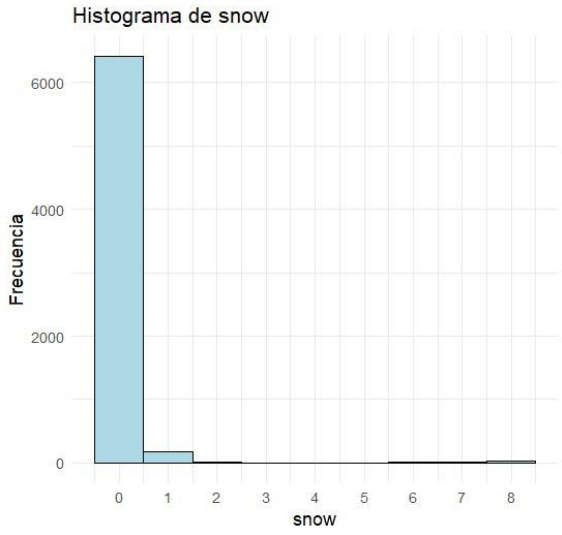
Análisis de la variable *snow*

Figura 36. *Serie temporal de la variable snow*



Fuente: Elaboración propia

Figura 37. *Serie temporal de la variable snow*

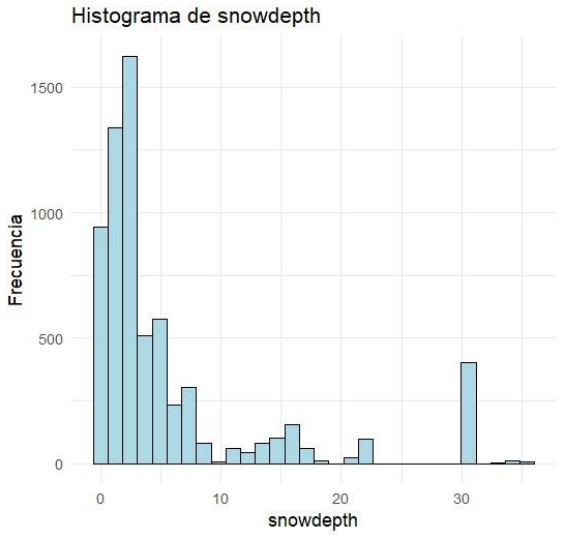


Fuente: Elaboración propia

El histograma presenta la cantidad de nieve que cae cuando nieva. La mayoría de las observaciones no llegan a un centímetro. Además, la serie temporal muestra como la nieve se da principalmente en los tres primeros meses del año, meses de invierno. Esto concuerda con el tipo de clima de Chicago, muy frío y ventoso en invierno.

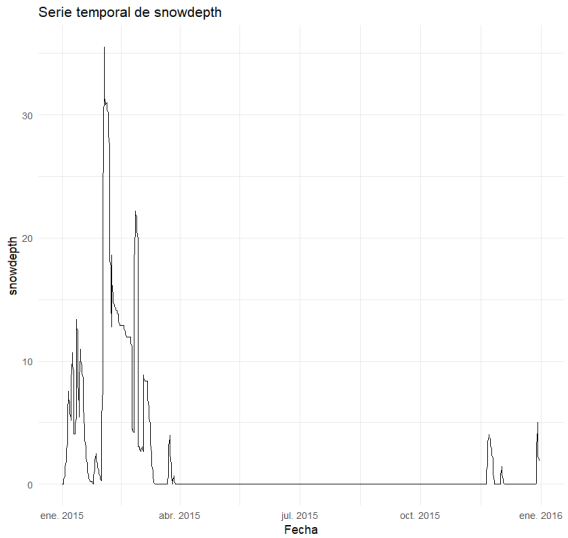
Análisis de la variable *snowdepth*

Figura 38. *Histograma de la variable snowdepth*



Fuente: Elaboración propia

Figura 39. *Serie temporal de la variable snowdepth*



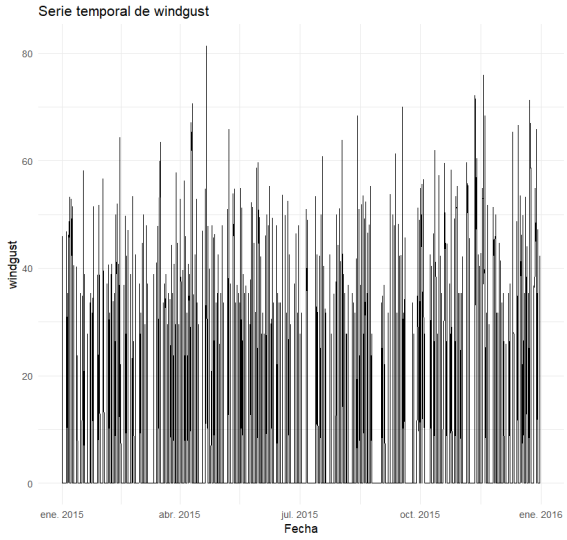
Fuente: Elaboración propia

Los gráficos muestran la profundidad de la nieve. En el histograma, la mayoría de las observaciones están cercanas al cero, lo cual indica que suele haber poca nieve en el

suelo, aunque sí que existen algunos picos esporádicos los cuales sugieren que ocasionalmente se registran profundidades de nieve más significativas. Es probable que esto ocurra en los meses más fríos como diciembre y enero. El siguiente gráfico muestra la serie temporal de la profundidad de la nieve, la cual verifica que los valores más altos se dan en los tres primeros meses del año. Siendo mayor en febrero, con valores por encima de 35 centímetros y seguida de marzo, esto es acorde con el gráfico visto anteriormente para la variable *snow*.

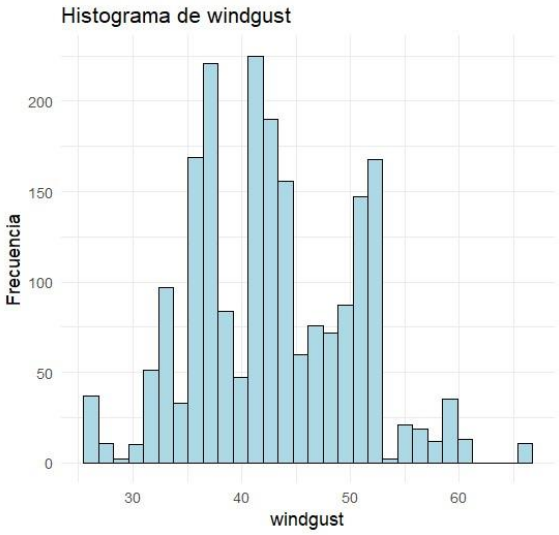
Análisis de la variable *windgust*

Figura 40. *Serie temporal de la variable windgust*



Fuente: Elaboración propia

Figura 41. *Histograma de la variable windgust*

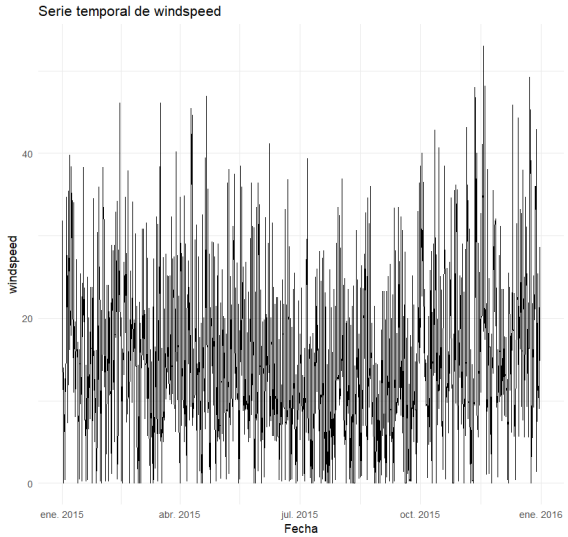


Fuente: Elaboración propia

Los gráficos muestran las ráfagas de viento, con valores por encima de 25 kilómetros por hora de la velocidad del viento, ya que para que se considere ráfaga la velocidad del viento tiene que estar típicamente 18 km/h por encima de la media. A su vez muestra múltiples picos o rangos en los que las ráfagas de viento son más frecuentes, esto tiene sentido ya que las ráfagas de viento suelen ser esporádicas e intermitentes. Por otro lado, el histograma de la velocidad del viento muestra también distintos picos, pero tiende a disminuir en frecuencia conforme aumenta la velocidad del viento, por lo que las velocidades de viento más altas son poco comunes. Por otro lado, la serie temporal muestra la variabilidad de las ráfagas de viento a lo largo del año, la cual toma valores desde 0 hasta 80 kilómetros por hora. Parece haber picos ocasionales los cuales representan ráfagas fuertes y podrían estar asociados con eventos climáticos como tormentas.

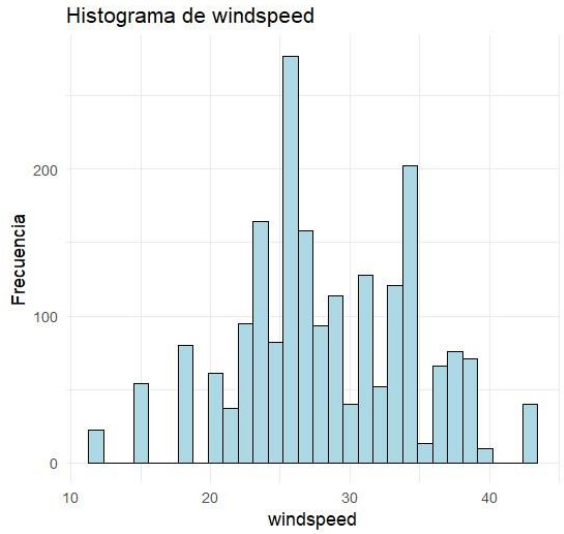
Análisis de la variable *windspeed*

Figura 42. *Serie temporal de la variable windspeed*



Fuente: Elaboración propia

Figura 43. *Histograma de la variable windspeed*

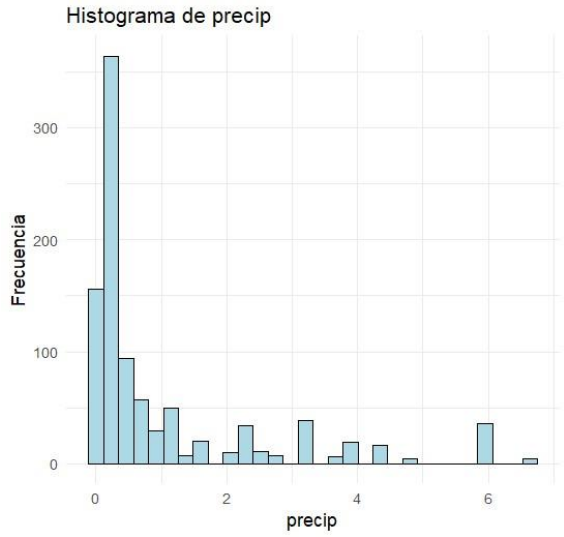


Fuente: Elaboración propia

La variable tiene una distribución uniforme, entre 10 y 40 km/h, con un pico sobre los 25 km/h lo cual indica que esta es la velocidad registrada de viento más frecuente. La serie temporal muestra una alta variabilidad de la velocidad del viento a lo largo del año, mostrando valores inferiores en los meses de verano y los más altos en los dos últimos meses del año. Estos valores explican porque se le llama a Chicago “la ciudad del viento”, ya que comparado con otras se la considera muy ventosa.

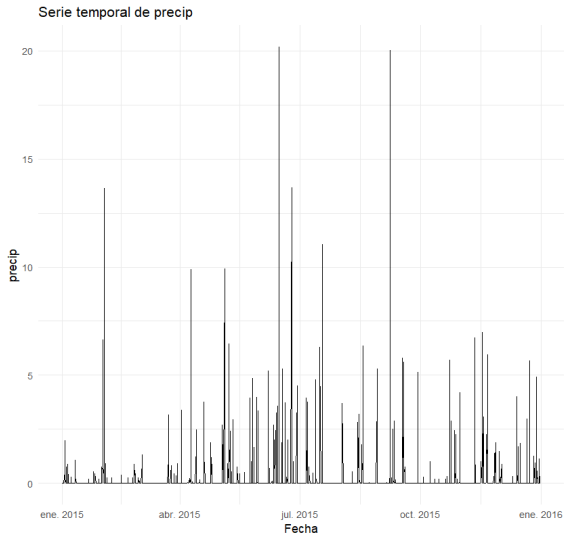
Análisis de la variable *precip*

Figura 44. *Histograma de la variable precip*



Fuente: Elaboración propia

Figura 45. *Serie temporal de la variable precip*

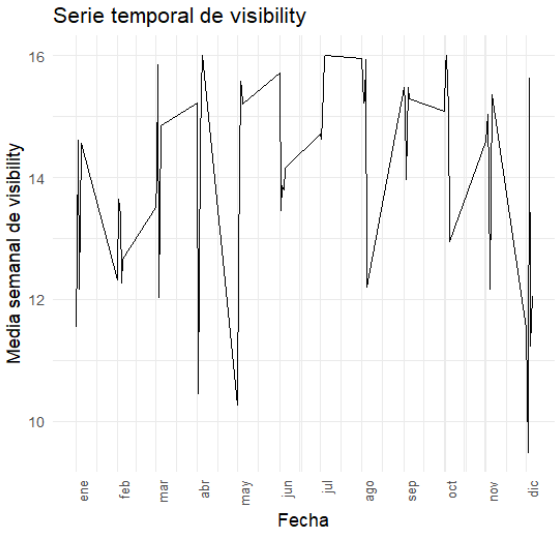


Fuente: Elaboración propia

La variable muestra la cantidad de precipitación en milímetros y como la mayoría de los días tienen poca o ninguna precipitación. Por lo tanto, son más frecuentes los eventos de lluvia ligera, siendo los eventos de gran precipitación menos comunes. Además, se observa una variabilidad diaria significativa, con algunos días en los que llueve mucho, pero son poco frecuentes. La variable no depende de la estación y sugiere que las lluvias son esporádicas e impredecibles, lo cual puede no ser bueno al intentar planificar según las condiciones meteorológicas.

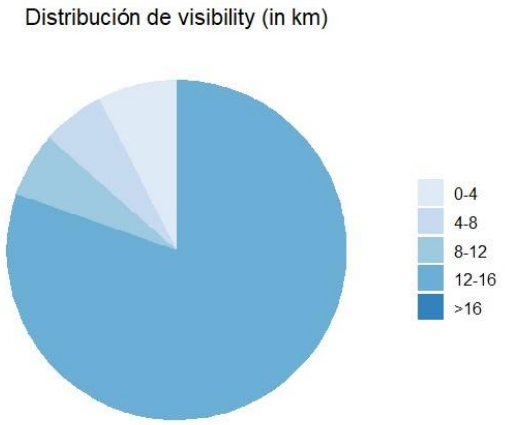
Análisis de la variable *visibility*

Figura 46. *Serie temporal de la variable visibility*



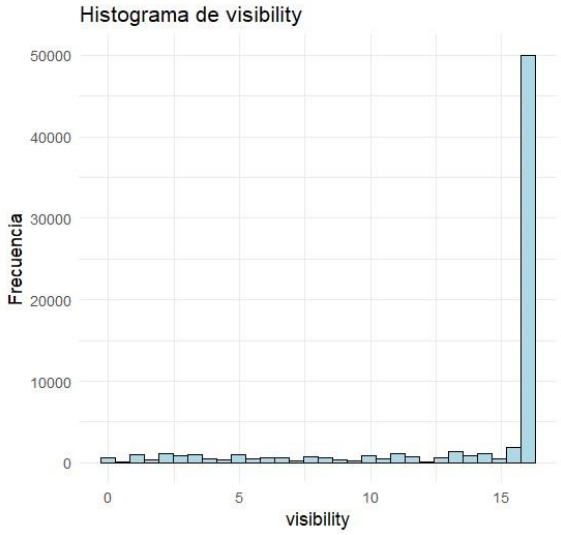
Fuente: Elaboración propia

Figura 47. *Gráfico de tarta de la variable visibility*



Fuente: Elaboración propia

Figura 48. *Histograma de la variable visibility*

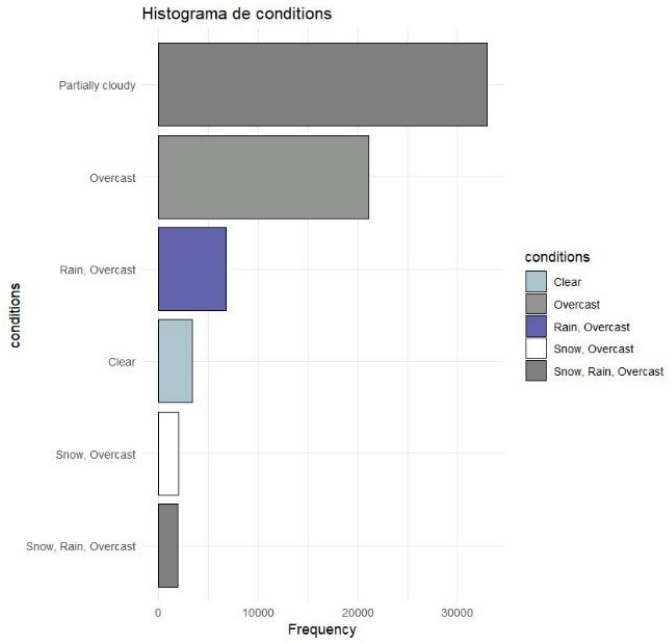


Fuente: Elaboración propia

La variable explica la visibilidad en kilómetros desde la estación climatológica del Aeropuerto Internacional de Chicago, e indica que la visibilidad máxima del histograma, en torno a 17 km, es la más común. Por lo que se entiende que la visibilidad en la mayoría de los casos es favorable. Además 16 kilómetros no se considera un valor extremo, ya que la máxima visibilidad desde el punto geográfico del aeropuerto es de 50 kilómetros. Calculada teniendo en cuenta una altitud del aeropuerto de 204 metros y el radio de la tierra, considerándola como una esfera perfecta para simplificar los cálculos y poder resolver el teorema de Pitágoras (University of Washington. Department of Mathematics, n.d.). El siguiente gráfico muestra como la mayoría de los datos presentan visibilidades entre 12 y 16 kilómetros. Además, la serie temporal muestra la media semanal de la visibilidad en kilómetros a lo largo del año, oscilando aproximadamente entre 10 y 16 kilómetros. Los meses con visibilidades más bajas para el año 2015, son abril, mayo y diciembre, pero por lo general los datos representan una buena visibilidad y no muestran estacionalidad, teniendo más del 75% de los vuelos una buena visibilidad en el despegue.

Análisis de la variable *conditions*

Figura 49. *Histograma de la variable conditions*

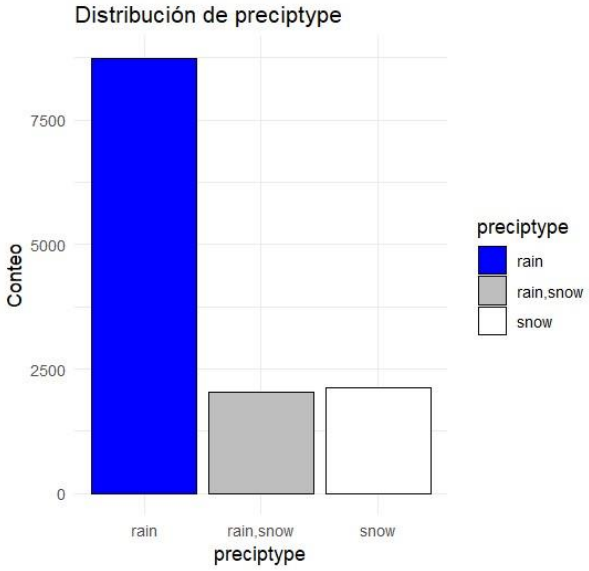


Fuente: Elaboración propia

La variable muestra las diferentes condiciones meteorológicas, siendo la más frecuente *Partially cloudy* o parcialmente nublado, seguida por *Overcast* o nublado. Sin embargo, la precipitación en forma de nieve es menos frecuente. Lo cual nos confirma el siguiente gráfico, que representa la frecuencia de los tipos de precipitación, donde se observa que la precipitación en forma de lluvia es la más frecuente. En conclusión, la variable muestra como el clima tiende a ser mayormente nublado, siendo los días claros y con clima severo más raros.

Análisis de la variable *preciptype*

Figura 50. *Histograma de la variable preciptype*

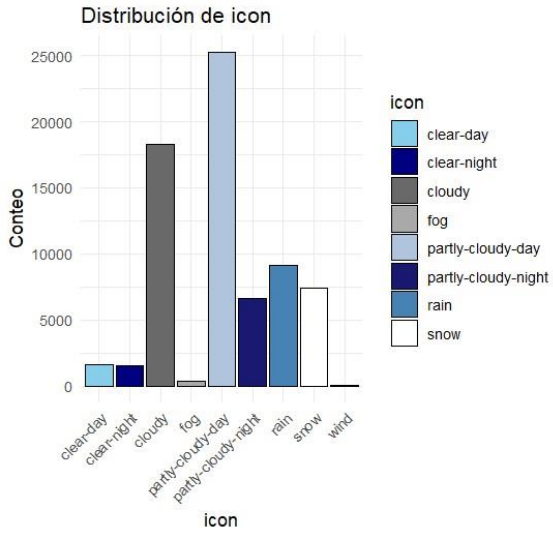


Fuente: Elaboración propia

La variable indica el tipo de precipitación, por lo que se observa que a la hora de salida de la mayoría de los vuelos hay lluvia, siendo esta el tipo de precipitación más común. Las condiciones de lluvia y nieve y la nieve sola son menos comunes, pero aun así son relevantes.

Análisis de la variable *icon*

Figura 51. *Histograma de la variable icon*

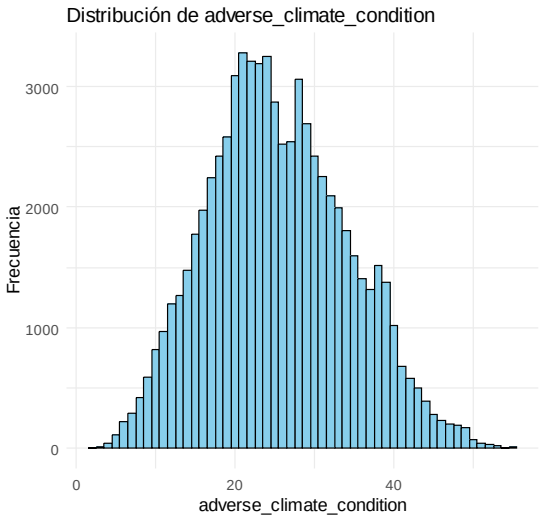


Fuente: Elaboración propia

El gráfico de barras muestra la frecuencia de distintas condiciones meteorológicas, más extensas que el gráfico previamente visualizado. Se observa que las condiciones más frecuentes son los días parcialmente nublados en los que la cobertura de las nubes es superior al 20% solo durante el día y días nublados en los que la cobertura de las nubes es superior a un 90%. Por otro lado, los eventos que menos se dan son la niebla y viento.

Análisis de la variable *adverse_climate_condition*

Figura 52. *Histograma de la variable adverse_climate_condition*

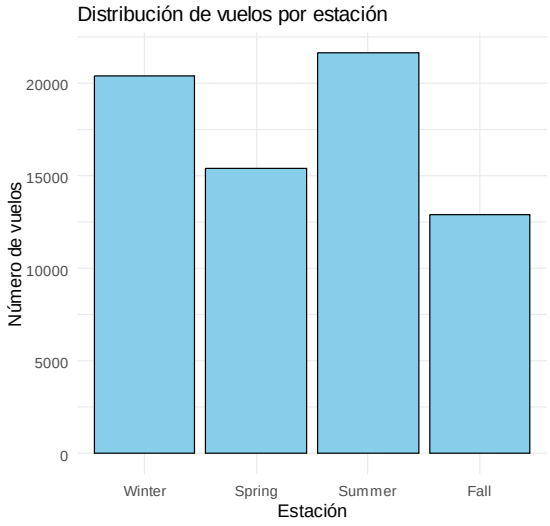


Fuente: Elaboración propia

La variable agrupa varias condiciones climáticas en una única medida ponderada y correlacionada con los retrasos por clima. Presenta una distribución normal con un pico entorno a 25, estando el máximo entorno a 50 esto sugiere que la mayoría de los retrasos por clima se deben a condiciones climáticas adversas de nivel moderado.

Análisis de la variable *season*

Figura 53. *Histograma de la variable season*

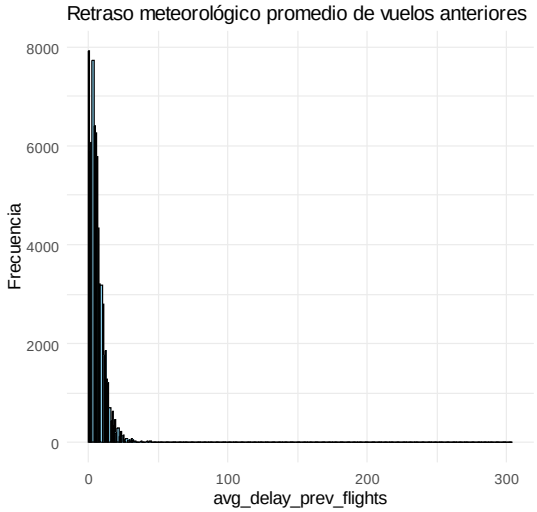


Fuente: Elaboración propia

La variable representa la estación del año en la cual se producen los vuelos. El gráfico muestra como el número de vuelos va variando por estación, siendo verano e invierno cuando más vuelos se producen, dadas las frecuencias más altas. Esto puede deberse a que se incremente el número de vuelos en esos meses por una mayor demanda.

Análisis de la variable *avg_wdelay_prev_flights*

Figura 54. *Histograma de la variable avg_delay_prev_flights*

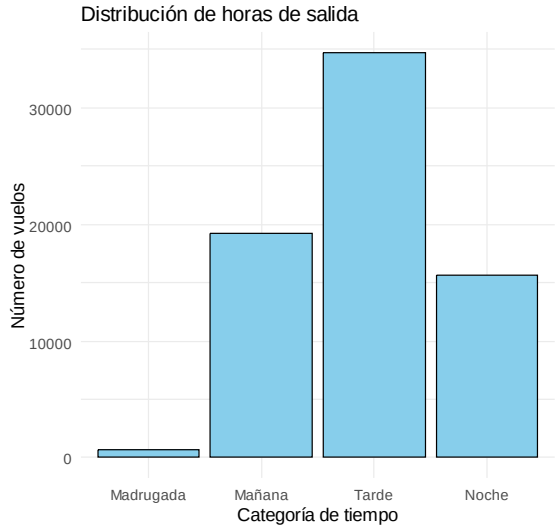


Fuente: Elaboracion propia

La variable muestra como la mayoría de los vuelos tienen un retraso promedio por clima bajo, con una mayor frecuencia de datos en el rango de 0 a 10 minutos de retraso. Aun así, sigue existiendo una pequeña cantidad de vuelos que tiene retrasos climáticos más significativos, acotados hasta 300 minutos. La distribución con la cola larga y sesgada a la derecha indica que los retrasos climáticos no son un gran problema para la mayoría de los vuelos, pero existen casos aislados con un gran impacto, por lo tanto, en esos casos, la variable puede ser útil para identificar patrones de retraso en vuelos con el mismo número y ayudar a predecir este retraso para el futuro.

Análisis de la variable *departure_time_category*

Figura 55. *Histograma de la variable departure_time_category*

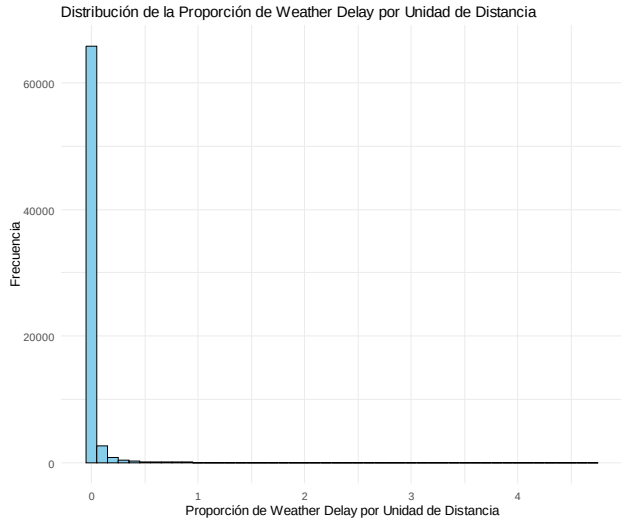


Fuente: Elaboración propia

La variable muestra el momento del día de la salida programada de los vuelos. Se observa como la mayoría de los vuelos están programados para salir por la tarde, seguidos de la mañana, noche y madrugada. Esta variable puede ayudar a entender al modelo como varían los retrasos por condiciones climáticas según el momento del día. Además, dependiendo de la condición climática podrá afectar más a un momento del día u otro.

Análisis de la variable *wdelay_per_distance*

Figura 56. *Histograma de la variable wdelay_per_distance*

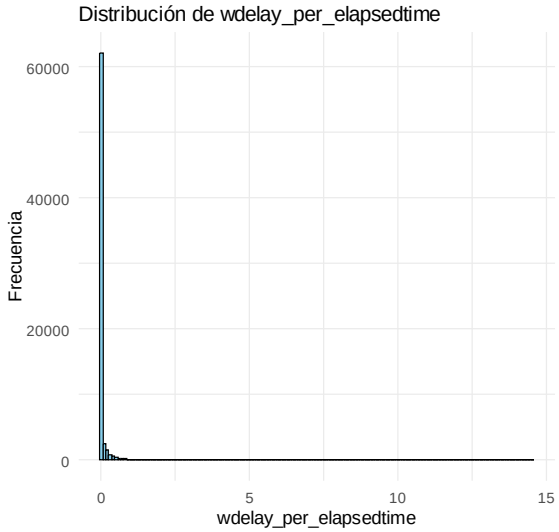


Fuente: Elaboración propia

La variable muestra como la mayoría de los vuelos tienen un retraso debido al clima por unidad de distancia muy bajo. No indica que vuelos de más distancia tengan porque tener un retraso debido al clima más alto. Aun así, hay casos extremos en los que el retraso por unidad de distancia es más alto y deben ser estudiados.

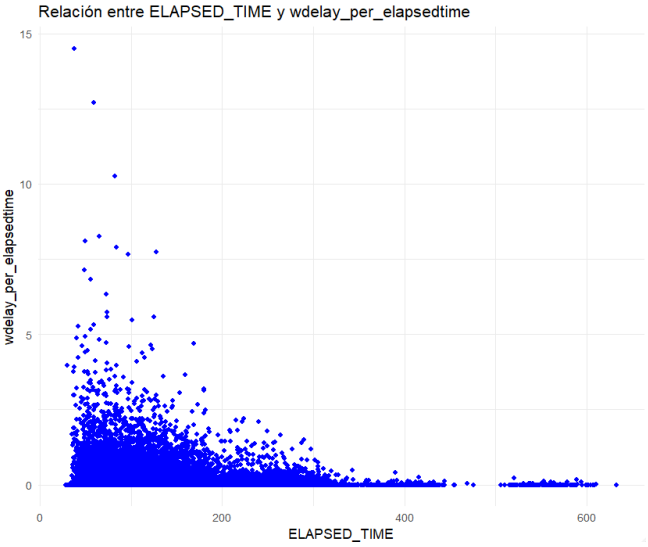
Análisis de la variable *wdelay_per_elapsedtime*

Figura 57. *Histograma de la variable wdelay_per_elapsedtime*



Fuente: Elaboración propia

Figura 58. *Gráfico de dispersión de la variable wdelay_per_elapsedtime*



Fuente: Elaboración propia

La variable muestra como generalmente, los retrasos por clima en relación con el tiempo total de vuelo son bajos. Además, con el gráfico de dispersión se observa una

relación inversa entre el tiempo de vuelo y la variable y se observa como los vuelos más cortos tienen más variabilidad en los retrasos debidos al clima por unidad de tiempo.

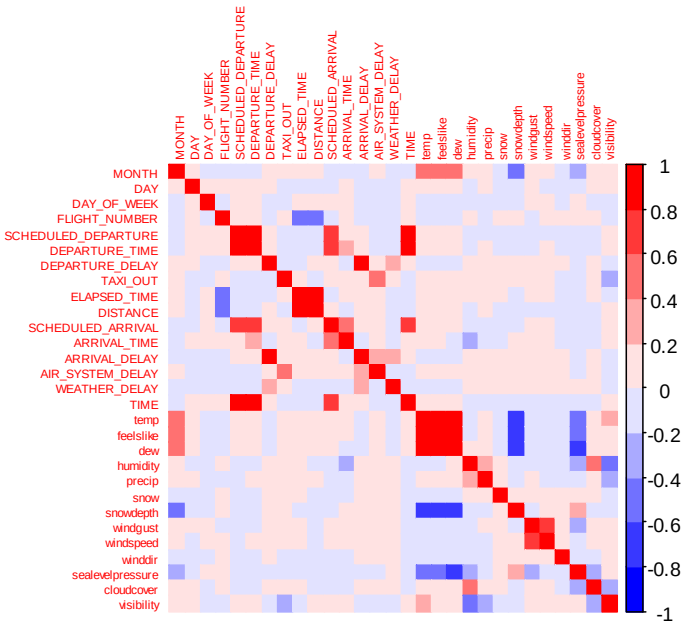
3.5.3 Análisis predictivo

Selección de variables

Para realizar la selección de variables primero se va a visualizar una matriz de correlación con el fin de identificar aquellas variables que están muy correlacionadas entre sí, y así eliminar variables redundantes que no aportan información adicional al modelo y evitar a su vez problemas de multicolinealidad. La multicolinealidad se da cuando dos o más variables están muy correlacionadas entre sí y por lo tanto dificulta la interpretación de los efectos de cada variable sobre la variable objetivo, afectando de manera negativa a la precisión del modelo.

Dado que el coeficiente de correlación necesita datos numéricos para calcularlo, en la matriz de correlación se incluyen únicamente variables numéricas. Para ello, se extraen del conjunto de datos "dataORD2" las variables numéricas, se centran y escalan, con el fin de igualar las escalas de medida y así interpretar mejor la matriz de correlaciones. Se calcula la matriz y se visualiza a continuación.

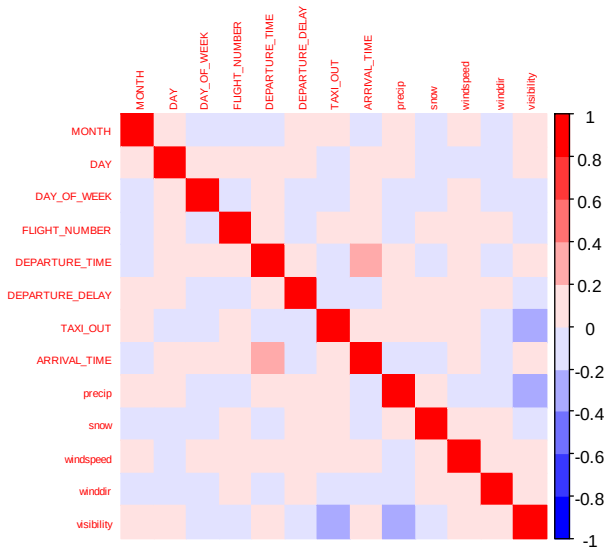
Figura 59. *Matriz de correlaciones*



Fuente: Elaboración propia

Se observa que en ningún caso hay correlaciones muy altas, por lo que se pone el umbral para considerar una correlación "alta" en 0,4 y por cada par de variables con correlación mayor a 0,4 se descarta una de ellas, para no tener información redundante. Se descartan por lo tanto las variables *SCHEDULED_ARRIVAL*, *TIME*, *ARRIVAL_DELAY*, *AIR_SYSTEM_DELAY*, *DISTANCE*, *ELAPSED_TIME*, *humidity*, *windgust*, *date*, *dew*, *snowdepth*, *sealevelpressure*, *temp*, *week_in_month*, *SCHEDULED_DEPARTURE*, *feelslike*, *wdelay_per_distance*, *wdelay_per_elapsedtime*, *cloudcover*, *precipprob*, *DIVERTED*, *CANCELLED*, *WEATHER_DELAY*. Y se visualiza la nueva matriz de correlaciones sin estas variables.

Figura 60. *Matriz de correlaciones*



Fuente: Elaboración propia

Se considera que no hace falta eliminar ninguna otra variable de las presentes en la matriz de correlación. Por lo tanto, se crea un nuevo conjunto de datos, "dataORD2_selecc" el cual es igual que "dataORD2", pero sin las variables eliminadas tras la primera visualización de las correlaciones y elimina también otras como *YEAR*, *name* y *stations* ya que de cada una de ellas solo hay un valor, todos los datos son del año 2015 y de la estación climática del Aeropuerto Internacional O'Hare, con el mismo

nombre de estación. Además, se elimina *FLIGHT_NUMBER,* pues no es seguro que siempre el mismo número de vuelo haga el mismo recorrido. Por lo tanto, el conjunto de datos "dataORD2_selecc" del cual se va a partir para hacer los modelos predictivos, tiene las variables que se muestran a continuación.

Tabla 11. *Variables y tipos del conjunto de datos "dataORD2_selecc"*

Variable	**Tipo**
MONTH	Entero
DAY	Entero
DAY_OF_WEEK	Entero
AIRLINE	Carácter
ORIGIN_AIRPORT	Carácter
DESTINATION_AIRPORT	Carácter
DEPARTURE_TIME	Entero
DEPARTURE_DELAY	Entero
TAXI_OUT	Entero
ARRIVAL_TIME	Entero
CANCELLATION_REASON	Carácter
DATE	Fecha
timelarge	Carácter
precip	Número
preciptype	Carácter
snow	Número
windspeed	Número
winddir	Entero
visibility	Número
conditions	Carácter
icon	Carácter
w_delayed	Factor
adverse_climate_condition	Número
season	Factor
avg_wdelay_prev_flights	Número
departure_time_category	Factor

Fuente: Elaboración propia

Más adelante se guardan estos datos en formato CSV y se escribe una sentencia para cargarlos fácilmente sin tener que ejecutar todo el código anterior. Tras cargarlos se realizan algunas transformaciones sobre los datos. Se convierten las variables *departure_time_category*, *season*, *w_delayed* a tipo factor y *DATE* a tipo fecha, ya que al volver a cargar los datos se han cambiado los tipos de algunas variables.

Modelos predictivos: Introducción

Teniendo en cuenta la revisión de la literatura, los distintos modelos predictivos usados en está y sus rendimientos, se decide elaborar dos modelos para el presente estudio. Una regresión logística y un *Random Forest*. Ya que ambos funcionan con una variable objetivo binaria y aceptan variables predictoras tanto numéricas como categóricas.

Regresión logística

Es un método de regresión utilizado para resolver problemas de clasificación. Ese utiliza cuando la variable dependiente u objetivo es dicotómica (0 o 1) y las variables independientes son numéricas y categóricas, en cuyo caso deben ser transformadas a variables de tipo *dummy* para que tome valores de 0 o 1. El modelo trata de predecir la probabilidad de una variable dependiente categórica, en este caso, un retraso debido al clima, la cual es una variable binaria con 1 si hay retraso y 0 si no lo hay. Además, para estimar esta probabilidad se basa en las variables independientes (Chitarroni, 2002). Además de ser un algoritmo fácil de interpretar, no abarca muchos recursos computacionales por lo que es fácil de ejecutar. La regresión logística utiliza la función logística para asignar estos valores de 0 y 1, si el resultado de esta función es mayor a 0,5 se clasificará como 1 y como 0 si ocurre, al contrario (Gonzalez, 2019a)

Random Forest

Es un modelo de predicción útil tanto para problemas de clasificación como de regresión. El método ejecuta varios árboles de decisión al mismo tiempo y combina sus resultados, lo hace mediante el algoritmo de agregación de *bootstrap (bagging)*. Para ello, extrae muchas muestras aleatorias con reemplazo de los datos, usa un subconjunto aleatorio de predictores para cada división y construye un árbol no podado para cada muestra, resultando en un bosque de múltiples árboles. Al seleccionar aleatoriamente los predictores se evita que algunas variables influyan demasiado en las primeras divisiones de cada árbol, logrando árboles más diversos y poco correlacionados. El número de predictores seleccionados es un hiperparámetro clave a optimizar. Por último, en clasificación, la predicción final es la más votada por los árboles, y en regresión es el promedio de sus resultados.

Random Forest puede manejar grandes cantidades de datos, acepta muchas variables independientes e identifica las más significativas. También estima datos faltantes y mantiene la precisión incluso con una gran proporción de datos ausentes. Sin embargo, una desventaja es que funciona como una "caja negra", siendo difícil de interpretar comparado con los árboles de decisión individuales (Gonzalez, 2018a).

Transformación del conjunto de datos a utilizar para los modelos predictivos

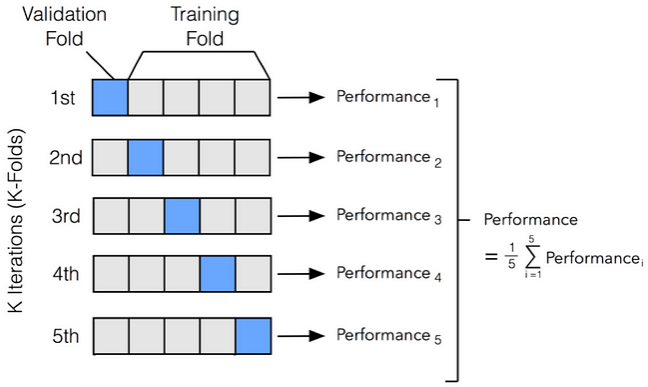
Es necesario realizar ciertas transformaciones en las variables antes de aplicar el modelo. Se convierten las variables *MONTH, DAY, DAY_OF_WEEK, AIRLINE, DESTINATION_AIRPORT, ORIGIN_AIRPORT, CANCELLATION_REASON, preciptype, conditions* e *icon* a tipo factor, ya que son variables con distintos niveles. A continuación, se eliminan las variables tipo factor con solo un nivel, como *ORIGIN_AIRPORT, CANCELLATION_REASON, DATE* y *timelarge*, ya que no proporcionan información adicional.

Además, se escalan las variables numéricas y se convierten en variables *dummy* aquellas variables categóricas o de tipo factor, excepto la variable objetivo. Por último, se asegura que todas las variables sean numéricas o factor con al menos dos niveles, ya que es lo que el modelo de regresión logística interpretará bien y se cambian los nombres de algunas variables para que no den lugar a error.

Validación cruzada

La validación cruzada es un método de remuestreo de datos que evalúa la capacidad de generalización de los modelos predictivos y evita el sobreajuste. Específicamente se utiliza la validación cruzada *k fold* por el cual se definen el número de particiones del conjunto de datos y en cada una de ellas se realiza entrenamiento y validación. Se van cogiendo distintas particiones y dejando una fuera, se entrena al modelo, calculando los parámetros y guardando el rendimiento obtenido con el conjunto de entrenamiento k-1 al evaluarlo sobre el conjunto de prueba. Una vez terminadas las iteraciones habrá k medidas de rendimiento para cada partición. El desempeño final del modelo será el promedio de todos estos rendimientos.

Figura 61. *Validación cruzada*



Fuente: Guerrero, 2021

Balanceo de datos con ROSE

Para ambos modelos se utiliza un método de balanceo de datos ya que hay una clase mayoritaria muy diferenciada y si no se balancea los resultados pueden estar sesgados. A la hora de hacer el balanceo se consideran dos opciones: sobremuestreo o submuestreo. El sobremuestreo aumenta el número de registros de la clase minoritaria, generando ejemplos sintéticos o duplicando registros existentes para aumentar los registros de la clase minoritaria hasta que el número de ambas clases sea parecido y el submuestreo reduce el número de registros de la clase mayoritaria eliminando aleatoriamente registros hasta que el número de registros de ambas clases sea aproximadamente igual.

Se decide utilizar el sobremuestreo ya que no elimina registros que podrían ser útiles y se considera mejor opción generar nuevos registros sintéticos. Dentro del sobremuestreo se analizan distintos tipos: ROSE, SMOTE y ADASYN, tres de los más conocidos. Tras un análisis de las capacidades de cada uno se elige ROSE ya que, aunque puede llegar a ser más lento en conjuntos de datos grandes como el aquí utilizado, es el único capaz de incluir variables tanto numéricas como categóricas sin tener que convertir estas últimas a numéricas previamente.

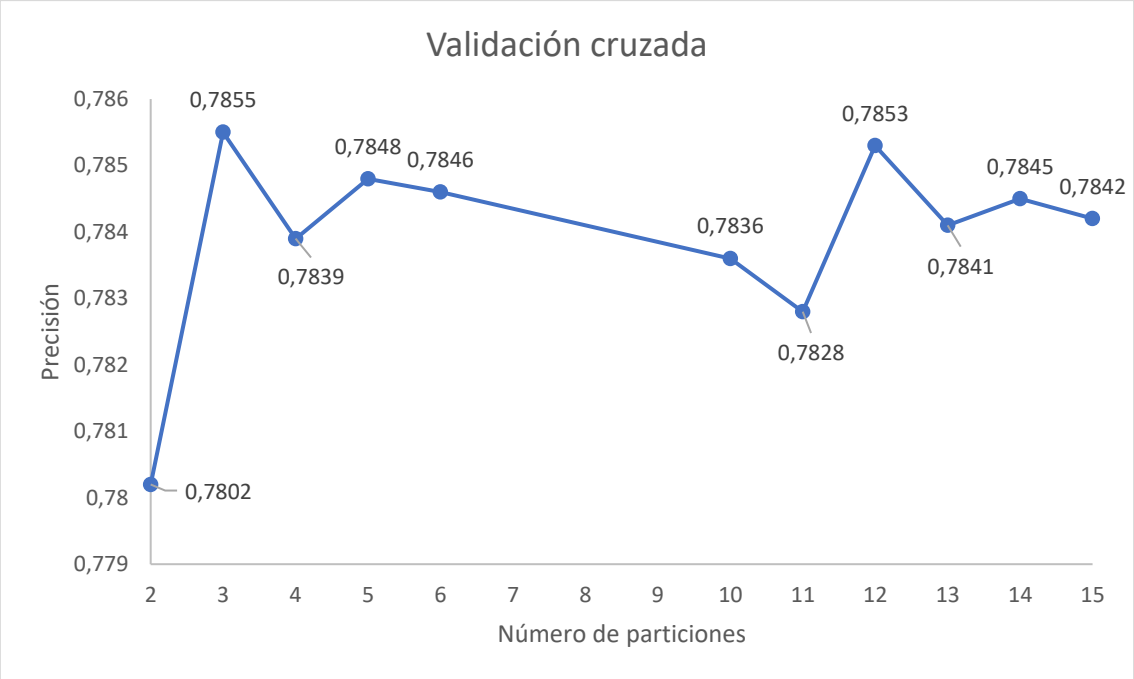
Modelos predictivos: Aplicación

Regresión logística

Para la regresión logística primero se definen los parámetros del control de entrenamiento, incluyendo ahí el método de validación cruzada, en este caso *k-fold cross validation*, el número de particiones y el método de balanceo de datos, en este caso ROSE, por el cual se balancearán los datos con este método durante la validación cruzada.

Típicamente se usan 5 o 10 particiones en la validación cruzada, se prueban ambas y aunque 5 particiones tienen una mejor precisión, se decide experimentar con distintas particiones cercanas a 5 y 10, aumentando o reduciendo de uno en uno, con el fin de observar cual ofrece una mejor precisión.

Figura 62. *Particiones y precisión en validación cruzada de regresión logística*



Fuente: Elaboración propia

Finalmente, se observa cómo utilizar 3 particiones proporciona una mejor precisión para un conjunto de datos grande como el tratado en este estudio, además de reducir el tiempo computacional. Lo siguiente será entrenar al modelo, mediante la siguiente sentencia:

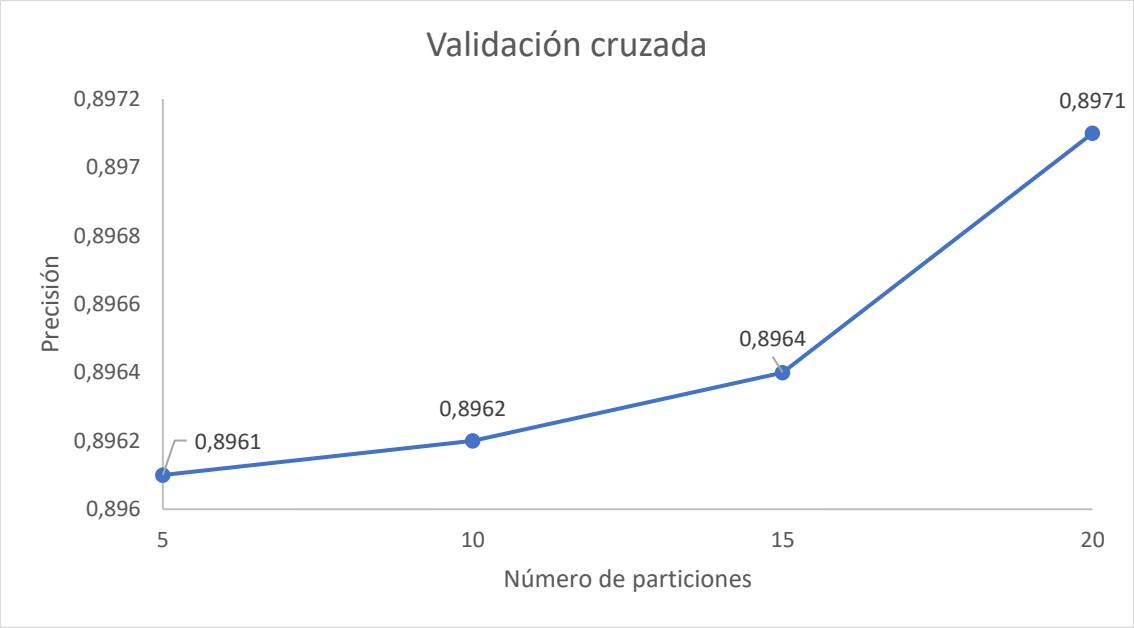
```
modelo <- train(
  w_delayed ~ .,
  data = dataORD2_selecc2,
  method = "glm",
  family = "binomial",
  trControl = control
)
```

En ella, se utiliza *w_delayed* como la variable dependiente y se usan todas las variables independientes. Se utiliza el conjunto de datos "dataORD2_selecc2", transformado previamente y se utiliza el método "glm" ya que se trata de una regresión logística y una distribución binomial, teniendo en cuenta que se trata de una regresión logística binaria. Por último, se añade la validación cruzada y el balanceo de datos definido anteriormente.

Random Forest

Al igual que en el modelo anterior, primero se definen los parámetros del control de entrenamiento, iguales a los del modelo anterior, pero ajustando el número de particiones para la validación cruzada según el desempeño de este modelo particular de bosques aleatorios. Al igual que en el anterior, se analiza la precisión con distinto número de particiones y se ajusta según esto, dejándolo en 20 particiones ya que es la que mejor desempeño muestra.

Figura 63. *Particiones y precisión en validación cruzada de Random Forest*



Fuente: Elaboración propia

Se entrena al modelo mediante la siguiente sentencia:

```
modelo_rf <- train(w_delayed ~ .,
                   data = dataORD2_rf,
                   method = "rf",
                   trControl = control)
```

Se utiliza la variable dependiente *w_delayed* y todas las variables independientes. Con respecto al conjunto de datos, inicialmente se prueba con "dataORD2_selecc2" el cual tiene variables numéricas y categóricas convertidas a tipo *dummy* pero resulta en un modelo lento y de mala precisión, por lo que pasa a usarse el conjunto de datos "dataORD2_rf" el cual contiene únicamente variables numéricas y la variable objetivo, y se entrena el modelo con este último. Como método se utiliza "rf" ya que se trata de un bosque aleatorio y por último se aplica la validación cruzada y el balanceo de datos definido previamente. En cada partición de la validación cruzada, se entrena un modelo de bosque aleatorio, utilizando todas las particiones excepto la actual y luego se evalúa

en la partición que se dejó a parte, también llamado conjunto de validación. Este proceso se repite para las 20 particiones en este caso.

Por otro lado, se ajusta el parámetro "mtry" el cual se refiere al número de predictores que se seleccionan de manera aleatoria en cada división de cada árbol de decisión del bosque. Es decir, el número de variables que se seleccionan para entrenar el modelo. El código presente prueba 3 valores diferentes de este parámetro: 2, 6 y 11. Para cada valor de "mtry" se realiza la validación cruzada con 20 particiones descrita anteriormente y se evalúa su rendimiento. Finalmente, tras haber probado los tres valores del parámetro, el modelo compara la precisión entre las particiones y selecciona el modelo con la precisión más alta

Comparación de modelos predictivos

La tabla a continuación muestra distintas combinaciones para los modelos de regresión logística y *Random Forest*. Muestra la precisión como medida de calidad de cada uno de los modelos. Los modelos varían según si están balanceados o no, si tienen validación cruzada o no y el número de particiones en estas. Todas las particiones de la validación cruzada mencionadas previamente en cada modelo se muestran a continuación en formato tabla, además de algún otro modelo.

Tabla 12. *Resumen de modelos predictivos*

Modelo	Balanceado	Validación cruzada	Número de particiones	Detalle	Medidas
Al azar	No	No			0,5007
Regresión logística	No	No		- Datos: dataORD2_selecc2 - 70% train 30% test; se entrena con train y se predice en test - Overfitting	*Accuracy:* 0,9166 Precisión: 0,6181 *Recall:* 0,2168
Regresión logística	No	Sí	10	- Datos: dataORD2_selecc2 - Overfitting	*Accuracy:* 0,9174 Kappa: 0,2888

Regresión logística	Sí	No		- Datos: dataORD2_selecc2 - Se balancea el train (70% de dataORD2_selecc2) y se predice sobre el test	
Regresión logística	Sí	Sí	2	Datos: dataORD2_selecc2	*Accuracy:* 0,7802 Kappa: 0,2955
Regresión logística	Sí	Sí	3	Datos: dataORD2_selecc2	*Accuracy:* 0,7855 Kappa: 0,2999
Regresión logística	Sí	Sí	4	Datos: dataORD2_selecc2	*Accuracy:* 0,7839 Kappa: 0,3002
Regresión logística	Sí	Sí	5	Datos: dataORD2_selecc2	*Accuracy:* 0,7848 Kappa: 0,2998
Regresión logística	Sí	Sí	6	Datos: dataORD2_selecc2	*Accuracy:* 0,7843 Kappa: 0,2992
Regresión logística	Sí	Sí	10	Datos: dataORD2_selecc2	*Accuracy:* 0,7836 Kappa: 0,2989
Regresión logística	Sí	Sí	11	Datos: dataORD2_selecc2	*Accuracy:* 0,7828 Kappa: 0,2981
Regresión logística	Sí	Sí	12	Datos: dataORD2_selecc2	*Accuracy:* 0,7853 Kappa: 0,3003
Regresión logística	Sí	Sí	13	Datos: dataORD2_selecc2	*Accuracy:* 0,7841 Kappa: 0,2995
Regresión logística	Sí	Sí	14	Datos: dataORD2_selecc2	*Accuracy:* 0,7845 Kappa: 0,3001
Regresión logística	Sí	Sí	15	Datos: dataORD2_selecc2	*Accuracy:* 0,7842

					Kappa: 0,2994
Random forest	Sí	Sí	2	Datos: dataORD2_selecc2	*Accuracy:* 0,0910 Kappa: 0
Random forest	Sí	Sí	5	Datos: dataORD2_rf	*Accuracy:* 0,8961 Kappa: 0,3492
Random forest	Sí	Sí	10	Datos: dataORD2_rf	*Accuracy:* 0,8962 Kappa: 0,3498
Random forest	Sí	Sí	15	Datos: dataORD2_rf	*Accuracy:* 0,8964 Kappa:
Random forest	Sí	Sí	20	Datos: dataORD2_rf	*Accuracy:* 0,8942 Kappa: 0,3480
Random forest	No	Sí	5	Datos: dataORD2_rf	*Accuracy:* 0,9307 Kappa: 0,4657

Fuente: Elaboración propia

Graficas comparativas

Tras haber realizado distintos tipos de modelos tanto de regresión logística como de *Random Forest*, balanceados o no, con validación cruzada o sin ella y cambiando el parámetro del número de particiones, todos ellos resumidos en la tabla anterior, se llega a la conclusión de que los dos mejores modelos, por presentar una precisión más alta son:

- Regresión logística: balanceado, con validación cruzada y 3 particiones
- *Random Forest*: balanceado, con validación cruzada y 20 particiones

Por ello se va a realizar un análisis más exhaustivo de ambos modelos, analizando, además de la precisión, otras medidas de calidad.

Regresión logística balanceado, con validación cruzada y 3 particiones

Se realiza un código más extenso con el fin de poder sacar más medidas de calidad del modelo predictivo y se muestra a continuación:

```
control <- trainControl(
  method = "cv",
  number = 3,
  sampling = "rose",
  verboseIter = TRUE,
  classProbs = TRUE,
  summaryFunction = twoClassSummary,
  savePredictions = "final"
)
modelo <- train(
  w_delayed ~ .,
  data = dataORD2_selecc2,
  method = "glm",
  family = "binomial",
  trControl = control,
  metric = "ROC"
)
```

El modelo realiza 3 particiones y agrega los resultados para sacar una media de las medidas de calidad de cada modelo entrenado en cada partición.

Tabla 13. *Medidas de rendimiento del modelo de Regresión Logística*

AUC *(Area Under the Curve)*	Sensibilidad	Especificidad	*Accuracy*	Valor predictivo positivo	Valor predictivo negativo
0,853	0,789	0,769	0,787	0,972	0,267

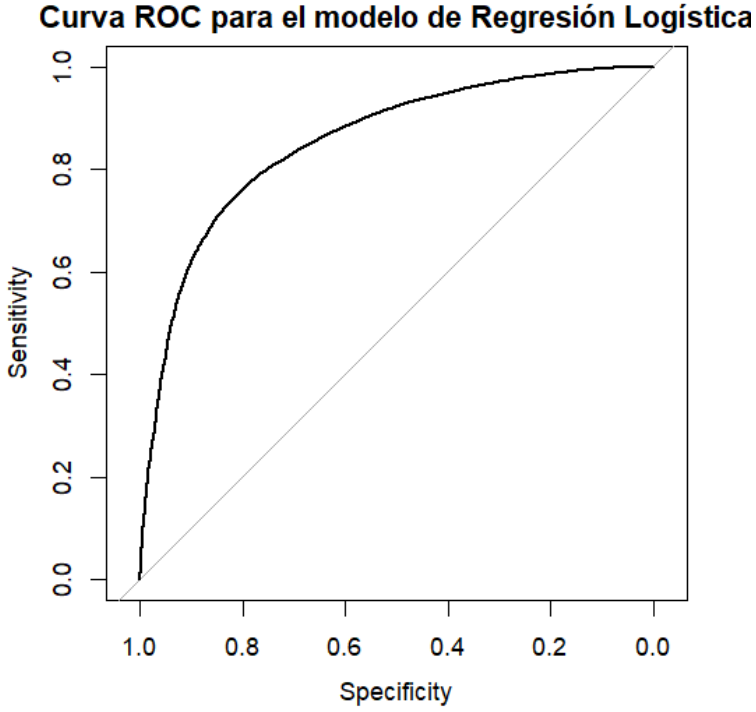
Fuente: Elaboración propia

Figura 64. *Matriz de confusión para el modelo de Regresión Logística*

```
          Reference
Prediction    X1    X0
        X1 50412  1479
        X0 13502  4920
```

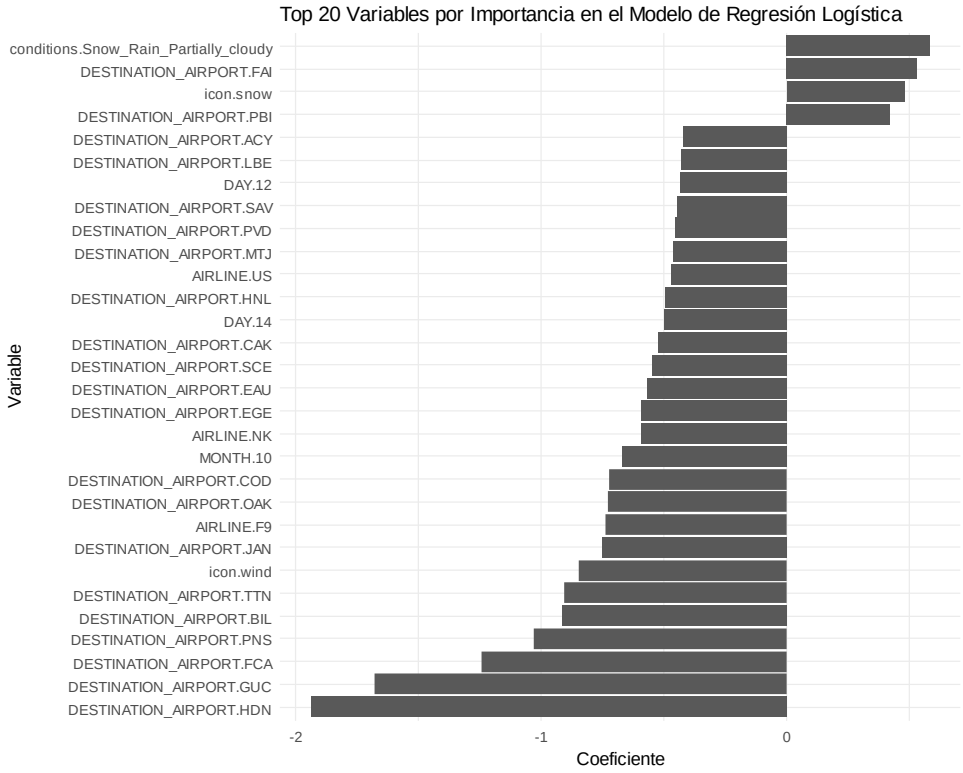
Fuente: Elaboración propia

Figura 65. *Curva ROC para el modelo de Regresión Logística*



Fuente: Elaboración propia

Figura 66. *Importancia de las variables en el modelo de Regresión Logística*



Fuente: Elaboración propia

Random Forest **balanceado, con validación cruzada y 20 particiones**

Se realiza un código más extenso con el fin de poder sacar más medidas de calidad del modelo predictivo y se muestra a continuación:

```
control <- trainControl(method = "cv",
                        number = 20,
                        sampling = "rose",
                        verboseIter = TRUE,
                        classProbs = TRUE,
                        summaryFunction = twoClassSummary,
                        savePredictions = "final")
modelo_rf <- train(w_delayed ~ .,
```

data = dataORD2_rf,

method = "rf",

trControl = control,

metric = "ROC")

El modelo realiza 20 particiones y evalúa el número óptimo de predictores, el cual deja finalmente en 2, siendo este el modelo óptimo según el valor de la curva ROC.

Tabla 14. *Medidas de rendimiento del modelo Random Forest*

mtry	AUC *(Area Under the Curve)*	Sensibilidad	Especificidad	*Accuracy*	Valor predictivo positivo	Valor predictivo negativo
2	0,803	0,943	0,404	0,894	0,940	0,416

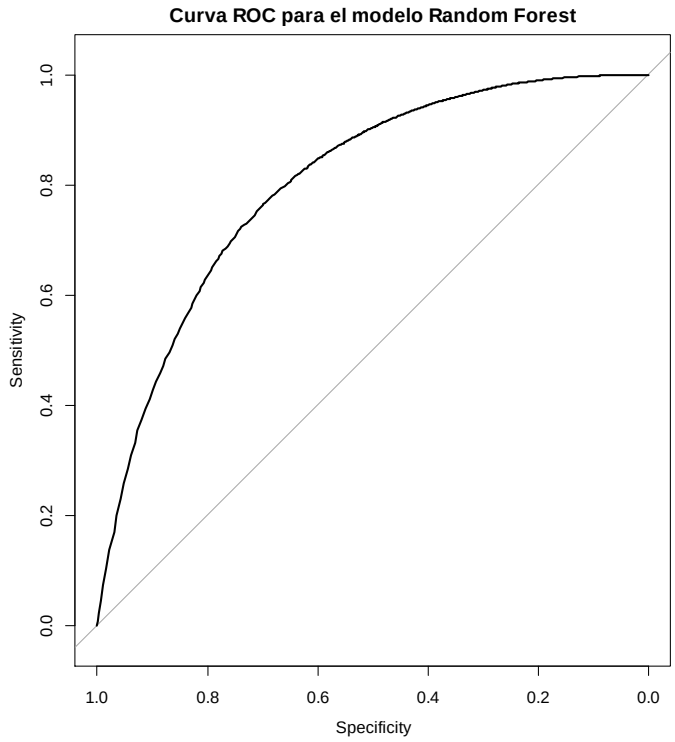
Fuente: Elaboración propia

Figura 67. *Matriz de confusión para el modelo de Random Forest*

```
          Reference
Prediction    X1    X0
        X1 60287  3811
        X0  3627  2588
```

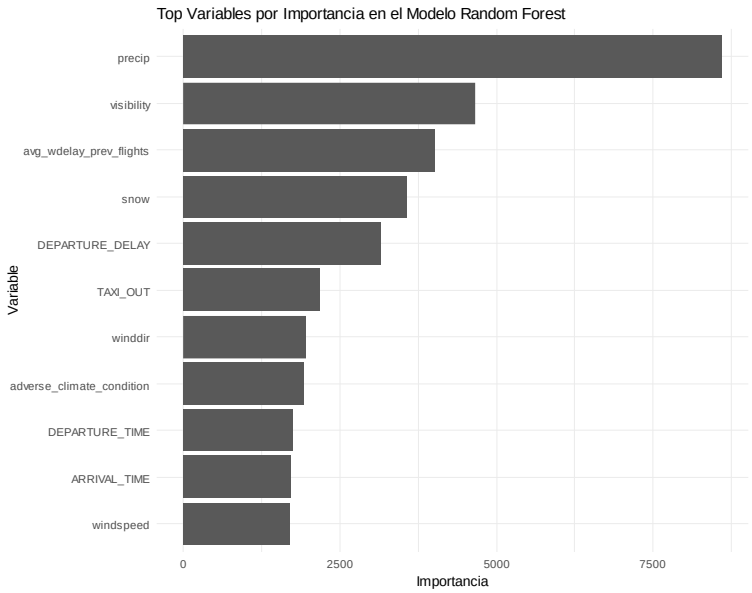
Fuente: Elaboración propia

Figura 68. *Curva ROC para el modelo de Random Forest*



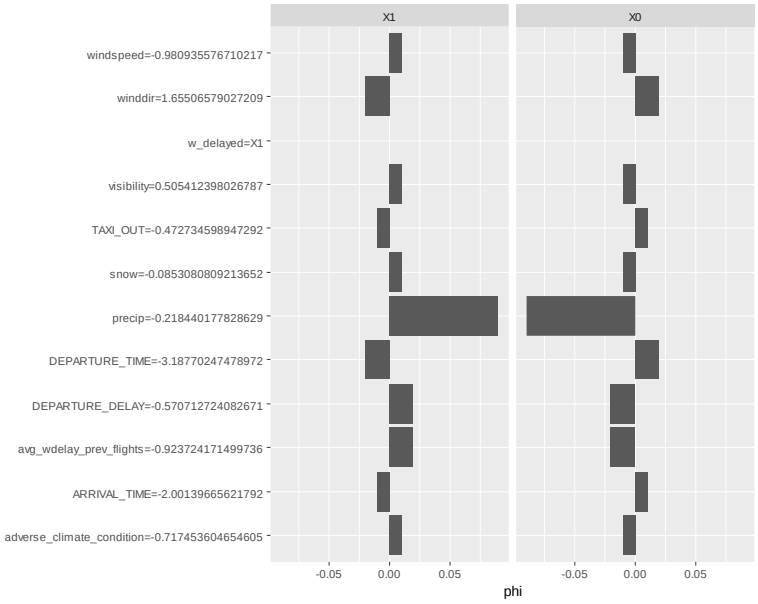
Fuente: Elaboración propia

Figura 69. *Importancia de las variables en el modelo de Random Forest*



Fuente: Elaboración propia

Figura 70. *Valores de Shapley en el modelo de Random Forest*



Fuente: Elaboración propia

4 CONCLUSIONES

4.1 Interpretar resultados

En esta sección, se interpretan los resultados del modelo de regresión logística, seguidos por los del modelo de *Random Forest*, y finalmente se realiza una comparación entre ambos.

El modelo de regresión logística presenta un área bajo la curva ROC (AUC) de 0,853. La curva ROC permite comparar distintos modelos de clasificación, indicando que un modelo es mejor cuanto mayor sea el área bajo la curva. Este valor sugiere un buen rendimiento del modelo de clasificación.

En la matriz de confusión se observan varias métricas clave. La sensibilidad del modelo o tasa de verdaderos positivos es de 0,789, lo que indica que el modelo predice correctamente el 78,9% de los vuelos retrasados debido a condiciones climáticas extremas. La especificidad, que mide la tasa de verdaderos negativos, es de 0,769, indicando que el modelo identifica correctamente el 76,9% de los vuelos que no sufren retrasos. El valor predictivo positivo es de 0,972, lo que significa que cuando el modelo predice un retraso, hay un 97,2% de probabilidad de que esta predicción sea correcta. En cambio, el valor predictivo negativo es de 0,267, indicando que cuando el modelo predice que no habrá un retraso, hay un 26,7% de probabilidad de que esta predicción sea correcta. Finalmente, la precisión del modelo es de 0,787, lo que significa que el 78,7% de las predicciones del modelo son correctas.

Considerando todas estas métricas, se puede observar que el modelo tiene una buena capacidad para identificar vuelos retrasados. Sin embargo, el número de falsos negativos, es decir casos en los que se predice que no se va a retrasar y en la realidad se acaba retrasando, es relativamente alto, lo que disminuye la sensibilidad del modelo. Este aspecto es crucial para el objetivo del trabajo, ya que para una aerolínea y para los pasajeros es más perjudicial predecir incorrectamente que un vuelo no se retrasará (falsos negativos), comparado con predecir un retraso cuando no ocurre (falsos positivos).

Por otro lado, en la Figura 66, obtenida a partir de los coeficientes del modelo, se muestran las 20 variables más influyentes y sus efectos positivos o negativos. Dos variables climáticas que aumentan la probabilidad de retraso son *conditions.Snow_Rain_Partially_cloudy* e *icon.snow*. Además, los aeropuertos de destino

FAI (*Fairbanks International Airport*) y PBI (*Palm Beach International Airport*) también tienen un impacto positivo significativo en la probabilidad de retraso.

Por otro lado, variables como *icon.wind*, ciertos días del mes (días 10 y 12) y el mes de octubre tienen coeficientes negativos, disminuyendo la probabilidad de retraso. Esto puede deberse a condiciones climáticas más favorables en estos períodos. Las aerolíneas F9 (*Frontier Airlines*), NK (*Spirit Airlines*) y US (*US Airways*) también muestran un impacto negativo, sugiriendo una menor probabilidad de retraso en sus vuelos. Asimismo, los aeropuertos de destino HDN (*Yampa Valley Airport*), GUC (*Gunnison-Crested Butte Regional Airport*) y FCA (*Glacier Park International Airport*) indican una menor probabilidad de retraso.

En resumen, el modelo de Regresión Logística muestra un buen rendimiento en la predicción de retrasos en vuelos debido a condiciones climáticas, con una AUC de 0,853 y una precisión de 0,787. La sensibilidad y especificidad son razonablemente altas, aunque hay margen de mejora en la reducción de falsos negativos. Las variables climáticas, específicamente la nieve y lluvia, y los aeropuertos de destino son los factores más influyentes en la predicción de retrasos.

El modelo de *Random Forest* presenta un AUC de 0,803, indicando también un buen rendimiento. De la matriz de confusión, se observa que la sensibilidad es de 0,943, lo que significa que el modelo identifica correctamente el 94,3% de los vuelos retrasados. Sin embargo, la especificidad es de 0,404, indicando que el modelo identifica correctamente solo el 40,4% de los vuelos no retrasados. El valor predictivo positivo es de 0,940, sugiriendo que cuando el modelo predice un retraso, hay un 94% de probabilidad de que esta predicción sea correcta. El valor predictivo negativo es de 0,416, indicando que cuando el modelo predice que no habrá un retraso, hay un 41,6% de probabilidad de que esta predicción sea correcta. La precisión del modelo es de 0,894, lo que significa que el 89,4% de las predicciones del modelo son correctas.

Las figuras 69 y 70 muestran la importancia de las variables y sus efectos sobre las predicciones. El primer gráfico muestra las variables importantes para el modelo, entre las que se incluyen variables climáticas: *precip, visibility, snow, windspeed, winddir, adverse_climate_condition* y variables relacionadas con las operaciones del vuelo como *DEPARTURE_TIME, DEPARTURE_DELAY, ARRIVAL_TIME, TAXI_OUT,*

avg_wdelay_prev_flights. La Figura 70 muestra los efectos positivos o negativos de cada variable sobre la predicción, según los valores de Shapley, los cuales muestran como cada variable contribuye a la predicción de una instancia especifica. La variable con un efecto más fuerte es *precip*, que tiene un efecto positivo sobre la predicción de retrasos, por lo que un aumento en la precipitación aumenta la probabilidad de que un vuelo se retrase. Otras variables con efectos positivos incluyen *adverse_climate_condition, avg_wdelay_prev_flights, DEPARTURE_DELAY, snow, visibility* y *windspeed.* Por otro lado, las variables con efectos negativos, es decir, que cuando aumentan, la probabilidad de retraso debido a condiciones climáticas en la salida disminuye, son *winddir, TAXI_OUT, DEPARTURE_TIME y ARRIVAL_TIME.*

En conclusión, el modelo de *Random Forest* muestra un buen rendimiento en la predicción de retrasos en vuelos debido a condiciones climáticas, con una AUC de 0,803 y una precisión de 0,894. La sensibilidad es muy alta, lo que indica que el modelo es eficaz en la identificación de vuelos retrasados. Sin embargo, la especificidad es baja, lo cual significa que el modelo tiene dificultades para identificar vuelos que no están retrasados. En general, las variables climáticas tienen un efecto positivo sobre la predicción de retrasos, mientras que las variables operacionales del vuelo tienen efectos negativos.

Tabla 15. *Medidas de rendimiento de ambos modelos*

Medida	Regresión Logística	*Random Forest*
AUC	0,853	0,803
Sensibilidad	0,789	0,943
Especificidad	0,769	0,404
Precisión *(Accuracy)*	0,787	0,894
Valor predictivo positivo	0,972	0,940
Valor predictivo negativo	0,267	0,416

Fuente: Elaboración propia

Al comparar ambos modelos, se observa que el modelo de Regresión Logística presenta una mayor área bajo la curva (AUC), lo que indica una mejor capacidad general para discriminar entre vuelos retrasados y no retrasados. En términos de sensibilidad, el modelo *Random Forest* destaca con un valor significativamente mayor, lo que sugiere

una superior capacidad para predecir correctamente los vuelos retrasados. Por otro lado, en cuanto a especificidad, el modelo de Regresión Logística es superior, lo que implica una mayor precisión al identificar vuelos que no experimentarán retrasos.

En cuanto a la precisión global, el modelo *Random Forest* es superior, indicando una mejor capacidad para predecir tanto vuelos retrasados como no retrasados. Ambos modelos presentan un valor predictivo positivo alto, siendo ligeramente mayor en el modelo de Regresión Logística. Esto significa que cuando este modelo predice un retraso, existe una mayor probabilidad de que dicha predicción sea correcta. Por otro lado, el valor predictivo negativo es mayor en el modelo *Random Forest*, lo que indica que cuando predice que no habrá retraso, la predicción es más probable de ser correcta en comparación con el modelo de Regresión Logística.

En relación con las variables importantes para cada modelo, ambos coinciden en que las variables climáticas relacionadas con la precipitación y la nieve tienen un impacto significativo en la predicción de retrasos, aumentándolos. El modelo de Regresión Logística captura días, meses y aeropuertos específicos, mientras que el modelo *Random Forest* incluye variables más genéricas y tiene la capacidad de capturar interacciones complejas entre ellas.

Desde el punto de vista de la aplicabilidad, la Regresión Logística, aunque tiene menor sensibilidad y precisión, ofrece una mejor especificidad. Por lo tanto, este modelo es útil cuando es crucial identificar correctamente los vuelos que no experimentarán retrasos. Por otro lado, el modelo *Random Forest*, con mayor sensibilidad y precisión, es más adecuado para identificar correctamente los vuelos retrasados. Además, tiene la capacidad de capturar interacciones no lineales y más complejas entre las variables. Sin embargo, su menor especificidad puede resultar en la predicción de retrasos cuando no los hay. A pesar de ello, este escenario es preferible al inverso, donde se predice que no habrá retraso cuando en realidad sí lo hay.

En conclusión, el modelo *Random Forest* es más adecuado cuando se busca una mayor precisión y sensibilidad, es decir, cuando es más importante identificar correctamente los vuelos retrasados.

4.2 Limitaciones del modelo y trabajos futuros

La principal limitación del modelo desarrollado es su aplicación exclusivamente al año 2015 y a el Aeropuerto Internacional de Chicago (ORD). Sin embargo, esto no representa una limitación significativa, ya que se ha diseñado un código modular que permite una fácil adaptación a otros años y aeropuertos. Ya que la intención es utilizar este modelo como un prototipo para una implementación más extensa, abarcando todos los aeropuertos de Estados Unidos e incorporando registros de vuelos de múltiples años. Esta expansión permitirá realizar predicciones más precisas para vuelos futuros y facilitar la comercialización de la aplicación tanto a aerolíneas, para la optimización de sus operaciones, como a empresas privadas que podrían ofrecer diversos servicios basados en estas predicciones.

Otra limitación significativa del modelo es su enfoque exclusivo en los retrasos causados por condiciones climáticas extremas (*WEATHER_DELAY*), excluyendo otros tipos de retrasos que pueden ser inducidos por condiciones climáticas no extremas u otras razones (*AIR_SYSTEM_DELAY*).

Para trabajos futuros, se sugiere no solo ampliar la base de datos para incluir más años y aeropuertos, sino también explorar la construcción de modelos avanzados como *Gradient Boosting Machine* y *Deep Convolutional Neural Network*. Según la literatura, estos modelos presentan un excelente desempeño en problemas similares. Además, sería beneficioso estudiar otros tipos de retrasos, no solo los retrasos en la salida causados por condiciones climáticas extremas, sino también aquellos en las llegadas y causados por otros motivos como la aerolínea, el Sistema Nacional de Aviación (NAS), retrasos del avión y cuestiones de seguridad. Asimismo, resulta de interés la predicción de otros eventos adversos como la cancelación o desvío de vuelos. Esto permitiría un análisis más completo de todas las eventualidades que pueden afectar a un vuelo.

5 BIBLIOGRAFÍA

Airlines for America. (24 de mayo de 2023). *U.S. Passenger Carrier Delay Costs*. Airlines for America. https://www.airlines.org/dataset/u-s-passenger-carrier-delay-costs/#:~:text=In%202022%2C%20the%20average%20cost

Ayodele, T. O. (2010). Types of Machine Learning Algorithms. *New Advances in Machine Learning* (pp. 19–49). InTech.

Bonaccorso, G. (2017). *Machine learning algorithms : reference guide for popular algorithms for data science and machine learning*. Packt.

Chitarroni, H. (2002). *La regresión logística*. https://racimo.usal.edu.ar/83/1/Chitarroni17.pdf

Department of Transportation. (2017). 2015 Flight Delays and Cancellations. *Www.kaggle.com*. https://www.kaggle.com/datasets/usdot/flight-delays

Esperón Cespón, I. (2018). *Construcción de un modelo de predicción para la puntualidad de vuelos comerciales.* https://docta.ucm.es/rest/api/core/bitstreams/a7179624-7425-43da-b3d1-18ac941582c5/content

Federal Aviation Administration. (2023). *FAA Aerospace Forecast Fiscal Years 2023–2043*. https://www.faa.gov/data_research/aviation/aerospace_forecasts

Gatto, L. (n.d.). UCLouvain-CBIO/WSBIM1322: Bioinformatics. *uclouvain-cbio.github.io*. https://github.com/UCLouvain-CBIO/WSBIM1322

Gil, E. (n.d.). Reducción de la dimensionalidad: Análisis de Componentes Principales (PCA). *profesorDATA.com*. Recuperado el 14 de mayo de 2024 de https://profesordata.com/2020/09/01/reduccion-de-la-dimensionalidad-analisis-de-componentes-principales-pca/

Gonzalez, L. (23 de marzo de 2018a). *Aprendizaje Supervisado: Random Forest Classification*. Aprende IA. https://aprendeia.com/aprendizaje-supervisado-random-forest-classification/

Gonzalez, L. (2018b). Diferencia entre aprendizaje supervisado y no supervisado. *Aprende IA*. https://aprendeia.com/diferencia-entre-aprendizaje-supervisado-y-no-supervisado/

Gonzalez, L. (28 de junio de 2019a). *Regresión Logística - Teoría*. Aprende IA. https://aprendeia.com/algoritmo-regresion-logistica-machine-learning-teoria/

Gonzalez, L. (2019b). Ventajas y Desventajas de los Algoritmos de Clasificación. *Aprende IA*. https://aprendeia.com/ventajas-y-desventajas-de-los-algoritmos-de-clasificacion-machine-learning/

Gonzalez, L. (2019c). Ventajas y Desventajas de los Algoritmos de Regresión. *Aprende IA*. https://aprendeia.com/ventajas-y-desventajas-de-los-algoritmos-de-regresion-machine-learning/

Guerrero, G. A. R. (2021). Cross-Validation. *Medium*. https://gladysandrea-rodriguez.medium.com/cross-validation-11e9ea688506

IATA. (2022). Quarterly Air Transport Chartbook IATA Economics Q3 2022. *Iata.org/economics*. IATA. https://www.iata.org/en/iata-repository/publications/economic-reports/quarterly-air-transport-chartbook-q3-2022/

Ibáñez Martín, A. (2019). *Semi-Supervised Learning…el gran desconocido*. https://telefonicatech.com/blog/semi-supervised-learningel-gran-desconocido

Jiang, H. (2021). *Machine learning fundamentals : a concise introduction*. Cambridge University Press.

Liu, F., Sun, J., Liu, M., Yang, J., & Gui, G. (2020). *Generalized Flight Delay Prediction Method Using Gradient Boosting Decision Tree*. https://doi.org/10.1109/vtc2020-spring48590.2020.9129110

Mahesh, B. (2018). Machine Learning Algorithms -A Review Machine Learning Algorithms -A Review. *International Journal of Science and Research (IJSR) ResearchGate Impact Factor*, *9*(1). https://doi.org/10.21275/ART20203995

Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., & Barman, S. (2017). A statistical approach to predict flight delay using gradient boosted decision tree. *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)*. https://doi.org/10.1109/iccids.2017.8272656

Martínez Domenech, N. (2016). *Predicción y Análisis de los Retrasos en los Vuelos Estudio del Aeropuerto de Arizona (E.E.U.U.)*.

Monje Solá, R. (2015). *Análisis y Predicción de los Retrasos de Vuelo Estudio del Aeropuerto de Seattle-Tacoma*. https://core.ac.uk/download/pdf/78534675.pdf

Nigam, R., & Govinda, K. (2017). *Cloud Based Flight Delay Prediction using Logistic Regression*.

Priy, S. (20 de marzo de 2024). *Clustering in Machine Learning - GeeksforGeeks*. GeeksforGeeks. https://www.geeksforgeeks.org/clustering-in-machine-learning/

Qu, J., Zhao, T., Ye, M., Li, J., & Liu, C. (2020). Flight Delay Prediction Using Deep Convolutional Neural Network Based on Fusion of Meteorological Data. *Neural Processing Letters*, *52*. https://doi.org/10.1007/s11063-020-10318-4

Ros, M. (2024). *TFGv1*. https://github.com/mrosarro/TFGv1

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(160), 1–21. Springer. https://doi.org/10.1007/s42979-021-00592-x

U.S. Department of Transportation. (n.d.). *On-Time Performance - Reporting Operating Carrier Flight Delays at a Glance*. Bts.gov. https://www.transtats.bts.gov/HomeDrillChart.asp

U.S. Department of Transportation. (6 de mayo de 2023). *Fly Rights*. Transportation.gov. https://www.transportation.gov/airconsumer/fly-rights#:~:text=DOT%20rules%20prohibit%20most%20U.S.

U.S. Department of Transportation. (15 de abril de 2024). *Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics*. Www.bts.gov. https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations

United States Department of Transportation. (2024). On-Time : Reporting Carrier On-Time Performance (1987-present). *Bts.gov*. https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr

University of Washington. Department of Mathematics. (n.d.). *How far away is the horizon?* https://sites.math.washington.edu/~conroy/m120-general/horizon.pdf

Visual Crossing. (n.d.). *Historical weather data for CHICAGO O'HARE INTERNATIONAL AIRPORT STATION*.

Zhou, Z.-H. (2021). *Machine learning*. Springer.

6 APÉNDICES

Apéndice 1. *Código fuente*

https://github.com/mrosarro/TFGv1

Fuente: (Ros, 2024)