Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Ciencia de Datos

Trabajo Fin de Máster

Predicción de la Distancia de Separación de Umbral entre Aeronaves mediante Aprendizaje Automático

Autor: Andrei Saavedra Rivera

Tutor: Antonio Jiménez Martín

Madrid, junio del 2023

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Máster
Máster Universitario en Ciencia de Datos
Título: Predicción de la Distancia de Separación Umbral entre Aeronaves mediante Aprendizaje Automático
Junio, 2023

Autor(a): Andrei Saavedra Rivera

Tutor	*Co-Tutor*
Antonio Jiménez Martín	Juan A. Fernández del Pozo
ETSI Informáticos Departamento de Inteligencia Artificial Universidad Politécnica de Madrid	ETSI Informáticos Departamento de Inteligencia Artificial Universidad Politécnica de Madrid

Resumen

Los aeropuertos desempeñan un papel vital en el transporte aéreo, asegurando la puntualidad, eficiencia y seguridad de los vuelos. La gestión del tráfico aéreo es crucial debido al alto volumen de aeronaves por lo que es fundamental adherirse a los protocolos establecidos en los aeropuertos para todas las operaciones, como los aterrizajes o despegues. Estos protocolos tienen una importancia fundamental en términos de garantizar la seguridad y la eficiencia del flujo de las operaciones. Por lo tanto, es imperativo analizar los factores que pueden influir en cada operación y determinar las condiciones necesarias en diferentes escenarios, adaptándose así a diversas situaciones.

Este Trabajo de Fin de Máster se centra en el análisis de la separación de umbral durante el aterrizaje, con el objetivo de mejorar la seguridad y la eficiencia operativa del aeropuerto. No obstante, determinar el rango óptimo de separación presenta un desafío debido a la complejidad de los factores involucrados. Aunque se tienen en cuenta variables aeronáuticas como la estela, altitud y velocidad, el impacto de las condiciones meteorológicas aún no ha sido ampliamente investigado, a pesar de su posible relevancia en la determinación de la separación requerida. Por lo tanto, se recopilarán datos del Aeropuerto Adolfo Suárez Madrid-Barajas, incluyendo registros aeronáuticos y meteorológicos obtenidos de las bases de datos del Centro de Investigación, Desarrollo e Innovación en el ámbito de la Gestión del Tráfico Aéreo en España (CRIDA).

El aprendizaje automático, una subdisciplina de la inteligencia artificial, ofrece diversas técnicas y enfoques que permiten la capacidad de predicción. Su aplicación en los procesos aeroportuarios puede mejorar la seguridad, la velocidad de respuesta y la toma de decisiones, previniendo situaciones peligrosas.

Por lo tanto, el objetivo de este estudio es desarrollar un modelo de aprendizaje automático que prediga la separación de umbral entre aeronaves, considerando tanto las variables aeronáuticas conocidas como las variables meteorológicas no estudiadas. El objetivo final es promover una mayor seguridad y mejorar el tráfico aéreo, brindando apoyo a los controladores aéreos en la toma de decisiones.

Abstract

Airports play a vital role in air transport, ensuring the punctuality, efficiency and safety of flights. Air traffic management is crucial due to the high volume of aircraft so it is essential to adhere to the protocols established at airports for all operations, such as landings or takeoffs. These protocols are of fundamental importance in terms of ensuring the safe and efficient flow of operations. Therefore, it is imperative to analyze the factors that may influence each operation and determine the necessary conditions in different scenarios, thus adapting to various situations.

This Master's Thesis focuses on the analysis of threshold separation during landing, with the objective of improving the safety and operational efficiency of the airport. However, determining the optimal range of separation presents a challenge due to the complexity of the factors involved. Although aeronautical variables such as wake, altitude and speed are taken into account, the impact of meteorological conditions has not yet been extensively investigated, despite their possible relevance in determining the required separation. Therefore, data will be collected from Adolfo Suárez Madrid-Barajas Airport, including aeronautical and meteorological records obtained from the databases of the Center for Research, Development and Innovation in the field of Air Traffic Management in Spain (CRIDA).

Machine learning, a sub-discipline of artificial intelligence, offers various techniques and approaches that enable predictive capabilities. Its application in airport processes can improve safety, speed of response and decision making, preventing dangerous situations.

Therefore, the objective of this study is to develop a machine learning model that predicts threshold separation between aircraft, considering both known aeronautical variables and unstudied meteorological variables. The ultimate goal is to promote greater safety and improve air traffic, providing support to air traffic controllers in decision making.

Dedico este trabajo a CRIDA, por su continua labor en el ámbito de la aeronáutica y su impacto positivo en la sociedad.

Agradezco especialmente a Adrián Alfaro por su trabajo de fin de grado, el cual fue fundamental para el desarrollo de este estudio.

Agradezco a Tino por su apoyo constante durante todo el proceso de este trabajo de fin de máster.

Quiero expresar mi más profundo agradecimiento a Antonio Jiménez y Juan Antonio Fernández por su total implicación en este proyecto, sus invaluables consejos y por compartir su profundo conocimiento en el campo del aprendizaje automático.

A mi madre, quien ha sido la mayor fuente de motivación en mi vida, le dedico este trabajo con todo mi cariño y gratitud.

También quiero agradecer a mis compañeros Nataly, Delia y Erick, por su valiosa colaboración y compañerismo en cada etapa de este Máster.

A todos ellos, ¡muchas gracias!

Este TFM ha sido desarrollado en el marco del proyecto "Sistema de Ayuda a la Decisión basado en Aprendizaje Estadístico y Optimización en Redes. Aplicaciones a la Propagación de Pandemias a través del Transporte Aéreo", PID2021-122209OB-C31, Proyectos de Generación de Conocimiento 2021. Modalidad: Investigación Orientada Tipo Coordinado, subvencionado por el Ministerio de Ciencia e Innovación.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción y objetivos

En los últimos años, el transporte aéreo ha experimentado un crecimiento constante, representando aproximadamente el 10% de los viajes internacionales, lo que equivale a más de 4 mil millones de pasajeros al año a nivel mundial [1]. Esta creciente demanda ha generado una presión considerable en los aeropuertos para garantizar la seguridad y eficiencia de las operaciones.

El Aeropuerto Adolfo Suárez Madrid-Barajas, uno de los aeropuertos más importantes de Europa, maneja un tráfico aéreo significativo, registrando en 2021 más de 47 millones de pasajeros y aproximadamente 388,000 vuelos comerciales. Con un promedio de más de 1,000 vuelos diarios, la gestión eficiente del tráfico aéreo se convierte en un desafío crítico para garantizar la seguridad de los pasajeros y la puntualidad de las operaciones. El aeropuerto cuenta con cuatro pistas físicas ubicadas en dos grupos paralelos: el primer grupo comprende las pistas 18L/36R - 18R/36L, y el segundo grupo abarca las pistas 14L/32R - 14R/32L (véase la Figura 1). Estas pistas están diseñadas y numeradas según su dirección y posición para facilitar una organización eficiente del flujo de tráfico aéreo [2].

Figura 1: Distribución de pistas del aeropuerto Madrid-Barajas

Durante períodos de fuertes vientos cruzados en el aeropuerto de Barajas, se requiere ajustar la separación de umbral para compensar las condiciones adversas, según datos recopilados. La dirección y fuerza del viento son factores críticos a considerar en la gestión del tráfico aéreo, ya que influyen en la elección de las pistas de aterrizaje y despegue, así como en los procedimientos de separación. Es importante mencionar que la configuración preferente de las pistas puede cambiar según las condiciones del viento, evaluando la posibilidad

de cambio cuando el viento supera los 10 nudos para viento de cola y/o los 20 nudos para viento cruzado/lateral.

La separación adecuada entre las aeronaves es esencial para evitar colisiones y garantizar un flujo fluido del tráfico. La separación de umbral se refiere a la distancia entre aeronaves durante las operaciones de aterrizaje, como se muestra en la Figura 2. Actualmente, los controladores aéreos se encargan de gestionar la aproximación y dar instrucciones para garantizar la separación mínima entre llegadas sucesivas. El controlador asignado a la tarea determina la distancia de separación basándose en las aeronaves implicadas en la operación y su criterio, teniendo en cuenta factores como las condiciones meteorológicas y la visibilidad [3].

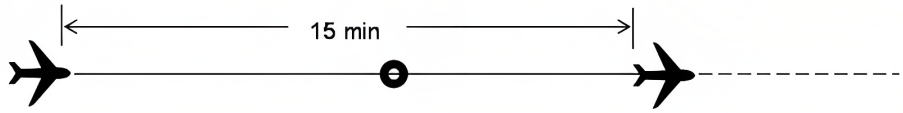

Figura 2: Distancia de separación de umbral entre aeronaves

Una de las variables críticas al establecer la separación de umbral es la estela de una aeronave. La estela es el rastro dejado por una aeronave en su camino y está compuesta por vórtices de aire que giran sobre sí mismos. Estos vórtices son generados por las hélices del avión y los vértices de la punta de las alas. La presencia de estas estelas puede tener un efecto significativo en la aeronave que sigue inmediatamente detrás, ya que las turbulencias creadas por estas pueden causar perturbaciones en su trayectoria. Es importante destacar que el tamaño de la aeronave está directamente relacionado con el tamaño de las estelas generadas [4]. Por lo tanto, cuanto mayor sea el tamaño del avión, mayor será el peligro de las estelas, como se muestra en la Figura 3.

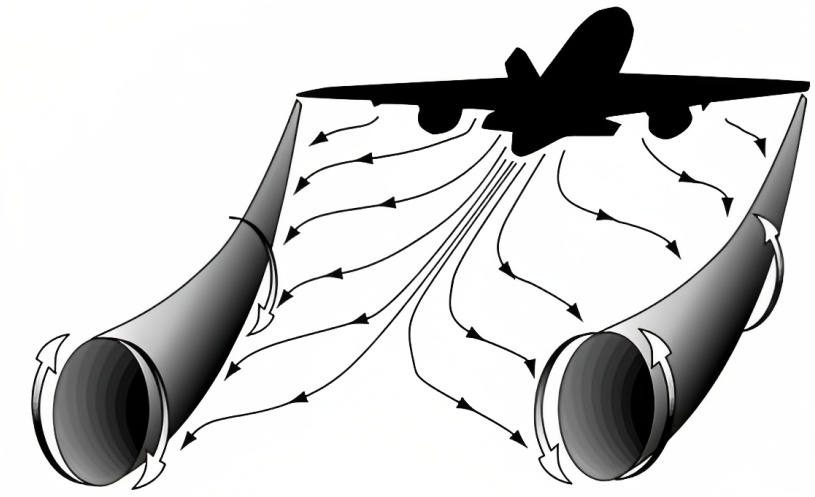

Figura 3: Estela turbulenta de aeronave

Durante las fases de aterrizaje y despegue de un vuelo, la estela turbulenta representa una amenaza significativa para la seguridad de la aeronave y sus pasajeros. Durante estas fases, el avión se mueve lentamente y con un ángulo de inclinación pronunciado, lo que aumenta la formación de vórtices en la punta de las alas. Es importante destacar que los aviones están mucho más cerca unos de otros durante estas fases del vuelo, especialmente en aeropuertos con alto tráfico aéreo, lo que aumenta la probabilidad de que un avión quede atrapado en la estela generada por la aeronave precedente. Además, debido a que el avión está cerca del suelo durante estas fases del vuelo, cualquier percance que surja podría resultar en un desastre catastrófico. Con el fin de minimizar el riesgo de incidentes, se ha establecido un sistema de clasificación de las estelas generadas por las aeronaves, basado en la masa máxima de despegue del avión.

Por lo tanto, se proporciona una clasificación de las aeronaves según su peso, con una descripción para cada categoría, dividida en cuatro grupos principales, como se muestra en la Tabla 1.

Tabla 1: Clasificación de la aeronave en base a su masa

Clasificación	***Descripción***
Super (J)	Aeronaves con una masa máxima muy elevada que se encuentran especificadas individualmente. Actualmente, algunas de estas aeronaves rondan los 560000 kg de masa.
Heavy (H)	Aeronave con un peso de más de 136000 kg exceptuando los J.
Medium (M)	Aeronaves con un peso de entre 7000 kg y136000 kg.
Light (L)	Aeronaves con un peso de 7000 kg o menos.

Con el objetivo de garantizar la seguridad de los vuelos comerciales, se establecen separaciones mínimas entre aeronaves durante las maniobras de aterrizaje como muestra la Tabla 2. Estas separaciones se basan en varios factores, como el tipo de aeronave involucrada, la velocidad de aproximación y la distancia entre la aeronave precedente y la que sigue.

Tabla 2: Separación mínima entre dos aeronaves

Aeronave anterior	***Aeronave posterior***	***Separación mínima***
SUPER	HEAVY	5.0 MN
SUPER	MEDIUM	7.0 MN
SUPER	LIGHT	8.0 MN
HEAVY	HEAVY	4.0 MN
HEAVY	MEDIUM	5.0 MN
HEAVY	LIGHT	6.0 MN
MEDIUM	LIGHT	5.0 MN

Es esencial que los pilotos sigan cuidadosamente las pautas de separación establecidas durante las maniobras de aterrizaje para evitar posibles conflictos y garantizar la seguridad de todos a bordo. En caso de detectar una posible infracción de las separaciones mínimas, los pilotos deben tomar medidas inmediatas para evitar un posible choque. Además, es importante destacar que las autoridades aeronáuticas pueden modificar las pautas de separación en cualquier momento en función de los datos y análisis de seguridad más recientes. En última instancia, la seguridad de los vuelos comerciales es responsabilidad de todos los involucrados, incluyendo pilotos, tripulación y autoridades aeronáuticas.

Además, es necesario considerar las condiciones meteorológicas, las cuales tienen un impacto significativo en todas las operaciones realizadas en los aeropuertos, lo que hace crucial comprender estas condiciones y sus consecuencias para predecir y tomar decisiones informadas en el futuro [5]. Uno de los procedimientos afectados por las condiciones meteorológicas es la separación de umbrales, que es un aspecto crítico de las operaciones de aterrizaje, convirtiéndolo en un tema excelente para un estudio en profundidad. Los datos recopilados del aeropuerto de Barajas indican que durante períodos de fuertes vientos cruzados, la separación de umbrales debe ajustarse para compensar las condiciones adversas. Esto resalta la necesidad de considerar variables meteorológicas en la predicción de la separación de umbrales y mejorar la toma de decisiones en tiempo real.

Para abordar este desafío, el presente Trabajo de Fin de Máster se propone desarrollar un modelo de aprendizaje automático capaz de predecir la separación de umbrales entre aeronaves en el aeropuerto de Barajas, teniendo en cuenta tanto las variables aeronáuticas como las variables meteorológicas. El objetivo es respaldar la toma de decisiones informadas y precisas, promoviendo una mayor seguridad, agilizando la respuesta a situaciones peligrosas y mejorando la eficiencia operativa en el aeropuerto de Barajas.

1.1 Objetivos del proyecto

El objetivo general es desarrollar un sistema de predicción de la distancia de separación de umbral entre aeronaves utilizando algoritmos de aprendizaje automático. El propósito es optimizar la capacidad del aeropuerto y garantizar la seguridad durante el aterrizaje, al tiempo que proporciona una herramienta eficiente para los controladores aéreos en la toma de decisiones. Los objetivos específicos se detallan a continuación:

- Analizar y comprender el funcionamiento y la metodología seguida en los procedimientos de aproximación a un aeropuerto.

- Identificar y seleccionar las variables más relevantes para la predicción de la separación en umbral, utilizando técnicas como conditional mutual information e information gain, adecuadas para variables discretas.

- Implementar técnicas de discretización de variables mediante clustering por observaciones, con el fin de mejorar el manejo y la representación de los datos.

- Evaluar el rendimiento de los modelos de clasificación utilizando técnicas como hold-out validation, nested cross-validation y stratified sampling, considerando la problemática de clases imbalanceadas.

- Aplicación de técnicas para abordar el desbalance de clases y mejorar la capacidad de generalización de los modelos.

- Implementar y comparar diferentes algoritmos de clasificación teniendo en cuenta la naturaleza del problema de clasificación multi clase con clases imbalanceadas

1.2 Estructura del documento

La presente memoria se estructura en seis capítulos. A continuación, se describen brevemente los diferentes capítulos que conforman este trabajo.

En el Capítulo 2 se aborda la obtención, análisis y depuración de datos utilizados en el estudio. Se detalla la procedencia de los datos, así como se realiza un análisis estadístico exhaustivo que incluye la identificación y manejo de valores atípicos y datos faltantes.

El Capítulo 3 se centra en la selección de variables, donde se describe el proceso de discretización y selección de variables para cada una de las pistas de aterrizaje en el aeropuerto de Barajas. Se presentan los métodos teóricos utilizados y se discuten los resultados obtenidos, definiendo así el conjunto de variables a utilizar en cada pista.

El Capítulo 4 se dedica a los modelos de clasificación utilizados en el estudio. Se presentan y describen los modelos a nivel teórico, discutiendo sus características y aplicabilidad en el contexto de la separación de umbral entre aeronaves. Además, se explican las medidas de rendimiento utilizadas para evaluar los modelos.

En el Capítulo 5 se realiza la aplicación práctica de los modelos de clasificación. Se detallan los pasos seguidos para implementar los modelos utilizando el lenguaje de programación R y las librerías pertinentes. Además, se presentan los resultados obtenidos para cada una de las pistas de aterrizaje evaluadas.

El Capítulo 6 se enfoca en las conclusiones y líneas de trabajo derivadas del estudio realizado. Se resumen los hallazgos más relevantes y se discuten las implicaciones de los resultados. También se sugieren posibles líneas de trabajo futuro para mejorar y ampliar la investigación en este campo.

La estructura del trabajo se ha diseñado de manera secuencial y lógica, desde la obtención y análisis de datos hasta la presentación de resultados y conclusiones. Cada capítulo aborda un aspecto específico del problema de predicción de la separación de umbral entre aeronaves, proporcionando una visión completa y detallada del estudio realizado.

Capítulo 2

Adquisición y preparación de datos

2.1 Obtención de datos

En este estudio, se utilizarán datos correspondientes al año 2019, considerado el último año con un tráfico aéreo regular antes de la aparición de la pandemia. Todos los datos relacionados con la aeronáutica, incluyendo la separación de umbral y las estelas generadas por los aviones, fueron obtenidos de las bases de datos de CRIDA.

Para obtener los datos meteorológicos necesarios, se recurrió a los informes METAR (***Meteorological Aerodrome Report***) del mismo año. Son informes meteorológicos rutinarios de los aeródromos, que se emiten en intervalos de media hora. En caso de producirse cambios bruscos en las condiciones meteorológicas, se puede enviar un informe no programado conocido como SPECI (***Special Weather Report***).

Cada informe METAR contiene información meteorológica detallada que afecta a un nivel de vuelo bajo y se refiere específicamente a un aeropuerto en particular. Estos informes pueden ser generados automáticamente por sistemas automatizados o emitidos por meteorólogos. Posteriormente, la información se envía a los pilotos para que puedan conocer las condiciones meteorológicas del aeropuerto de destino y actuar en consecuencia [6]. Los informes incluyen datos como los tipos, velocidades y direcciones de los vientos, la visibilidad, las características de las nubes (tipo, cantidad y altura), los fenómenos meteorológicos (tipo e intensidad), la temperatura, la presión y la información específica de la pista (cizalladura y depósitos en la pista).

El uso de estos conjuntos de datos combinados provenientes de las bases de datos del CRIDA y los informes METAR proporciona una fuente completa y confiable de información para el análisis de la separación de umbral en el aeropuerto de Barajas en Madrid durante el año 2019.

Es importante destacar que el proceso de recopilación, limpieza, transformación, selección e integración de los datos fue realizado por Adrián Alfaro en su Trabajo de Fin de Grado, quien también estudió la separación de umbral. De esta manera, se cuenta con un conjunto de datos íntegro que incluye variables aeronáuticas y meteorológicas relevantes para el proceso de toma de decisiones en la determinación de la separación de umbral [7].

Estas variables serán procesadas con el objetivo de obtener la máxima precisión posible al momento de predecir la separación de umbral, con el fin de garantizar la seguridad de los vuelos y mejorar la eficiencia del aeropuerto. A continuación, se muestra la distribución de las observaciones registradas para las pistas 32 y 18.

Tabla 3: Número de observaciones por pista

Pistas	*Cantidad*
32	55184
18	22659

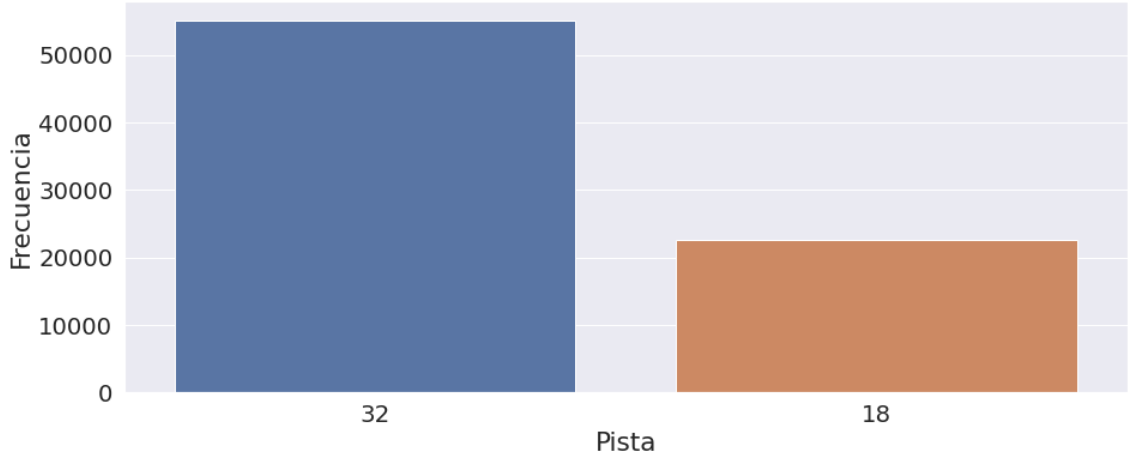

Figura 4: Distribución de observaciones por cada pista

2.2 Análisis estadístico de variables

En este capítulo, se llevará a cabo un análisis estadístico exhaustivo de las variables incluidas en el conjunto de datos. El objetivo principal es comprender en profundidad cada una de estas variables y su relación con la separación de umbral, con el fin de seleccionar las más relevantes para la predicción de esta medida.

El proceso de análisis se centrará en diversas etapas, que incluyen la identificación y manejo de datos ausentes, la detección y tratamiento de valores atípicos, y la verificación de la distribución de las variables mediante el uso de histogramas y diagramas de caja. Además, se registrará el número de observaciones por cada clase en el caso de variables categóricas, y se obtendrán estadísticas descriptivas como la media, desviación estándar y cuartiles para las variables numéricas.

El objetivo fundamental de este capítulo es obtener un conocimiento profundo de las variables presentes en el conjunto de datos, lo que nos permitirá seleccionar aquellas que sean más relevantes y contribuyan de manera significativa a la predicción de la separación de umbral. Al comprender a fondo estas variables, podremos mejorar la precisión y efectividad de nuestros algoritmos de aprendizaje automático, lo que a su vez mejorará la seguridad y eficiencia del aeropuerto.

A continuación, se presentarán los resultados obtenidos en este análisis estadístico, que nos proporcionarán información valiosa sobre la naturaleza de las variables y nos ayudarán en la selección de las características más influyentes para el modelo de predicción de la separación de umbral.

2.2.1 Variables aeronaúticas

Estas son las variables que se sabe que afectan la separación de umbral en la actualidad, y aunque no están relacionadas con factores meteorológicos, deben ser consideradas al analizar las variaciones en esta medida. Los datos han sido extraídos de la base de datos de CRIDA.

2.2.1.1 Estela de la aeronave previa

La variable representa el rastro que deja la aeronave previa y que tiene una gran influencia en la separación de umbral que se establece. Es decir, puede determinar si dos aeronaves pueden estar lo suficientemente separadas para evitar colisiones o si necesitan mantener una mayor distancia entre sí. En general, se considera que cuanto más grande sea la aeronave, más peligroso será el efecto del rastro, lo que significa que se debe tener una mayor precaución al establecer la separación entre aeronaves grandes.

Se trata de una variable categórica ordinal que puede tener cuatro valores: L (Bajo), M (Medio), H (Alto) o J (Jumbo). Estos valores representan el tamaño de la estela, ordenados de menor a mayor respectivamente. Para su uso posterior, se transformará en valores numéricos en función de su orden (L = 0, M = 1, H = 2 y J = 3).

Tabla 4: Distribución de observaciones de la estela por pista

Pistas	***M***	***H***	***L***	***J***
32	41377	5621	262	128
18	19962	2480	149	68

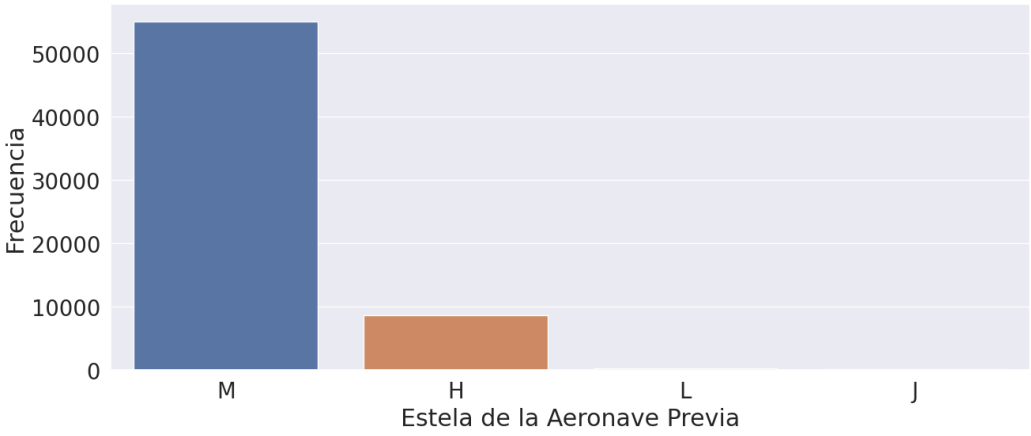

Figura 5: Distribución de la estela en la pista 32

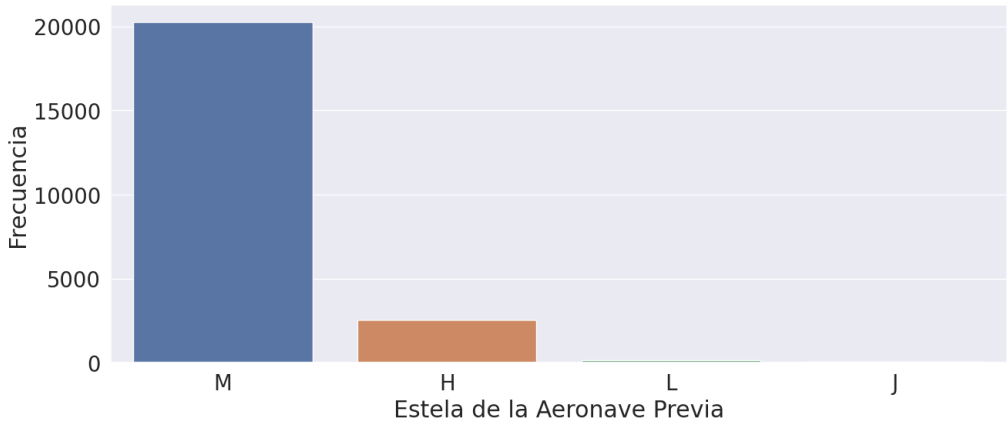

Figura 6: Distribución de la estela en la pista 18

2.2.1.2 Estela de la aeronave actual

La variable representa el rastro que deja la aeronave actual y que tiene una gran influencia en la separación de umbral que se establece. Es decir, puede determinar si dos aeronaves pueden estar lo suficientemente separadas para evitar colisiones o si necesitan mantener una mayor distancia entre sí. En general, se considera que cuanto más grande sea la aeronave actual, más peligroso será el efecto del rastro, lo que significa que se debe tener una mayor precaución al establecer la separación entre aeronaves grandes.

Se trata de una variable categórica ordinal que puede tener cuatro valores: L (Bajo), M (Medio), H (Alto) o J (Jumbo). Estos valores representan el tamaño de la estela, ordenados de menor a mayor respectivamente. Para su uso posterior, se transformará en valores numéricos en función de su orden (L = 0, M = 1, H = 2 y J = 3).

Tabla 5: Distribución de observaciones de la estela por pista

Pistas	***M***	***H***	***L***	***J***
32	43548	6066	251	135
18	20243	2542	147	64

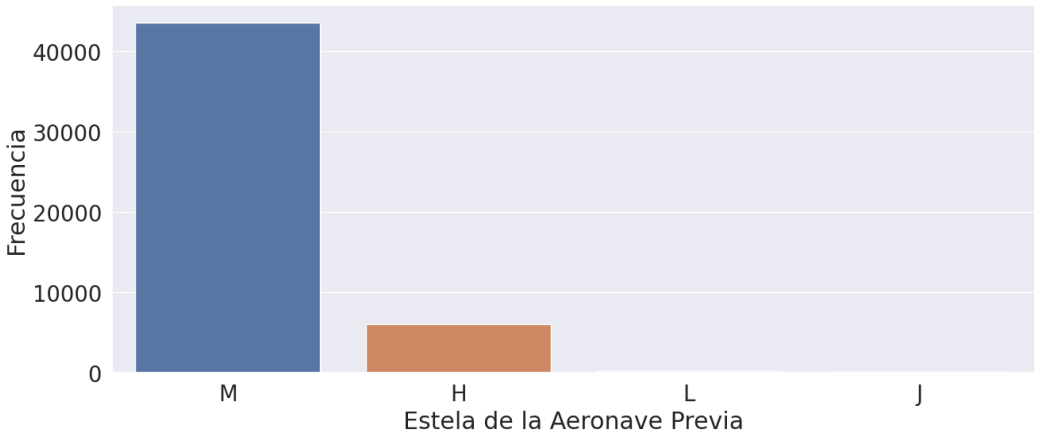

Figura 7: Distribución de la estela en la pista 32

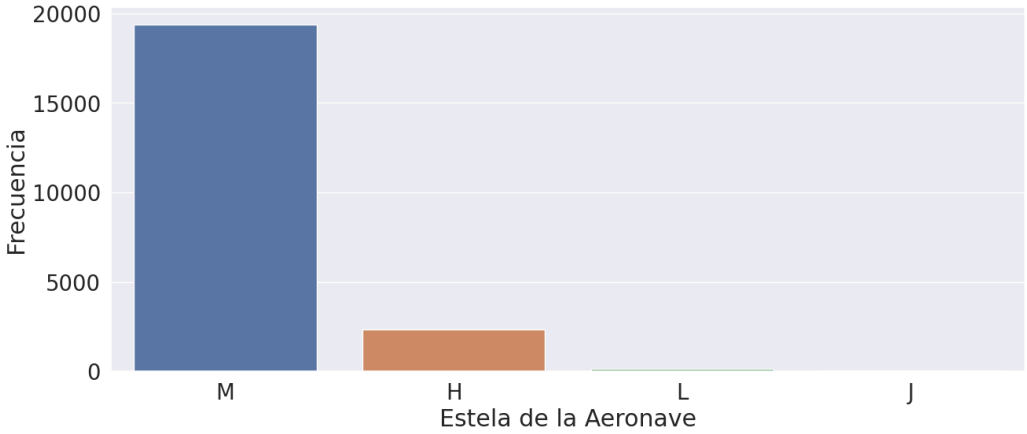

Figura 8: Distribución de la estela en la pista 18

2.2.2.2 Distancia diagonal

La distancia diagonal se refiere a la medida de distancia entre una aeronave en proceso de aterrizaje y otra aeronave que se encuentra en una pista paralela a la que ha sido asignada para el aterrizaje de la primera aeronave. Esta medida puede ser de gran importancia en la seguridad aérea, ya que garantiza que las aeronaves estén lo suficientemente separadas como para evitar colisiones durante el proceso de aterrizaje.

La variable se encuentra medida en millas náuticas. Es de tipo numérica continua. No existe un límite establecido para los valores que puede tomar esta variable, pero en los procedimientos se establece que debe existir una distancia diagonal mínima de al menos 2 mn.

Durante el análisis de la variable se observaron datos atípicos con el valor de –1. Los cuales, representan la ausencia de una aeronave en la pista paralela en el momento de aterrizaje. En la pista 32, estos valores representan el 7.33% del total de observaciones. Mientras que, en la pista 18 representan el 4.35%. Considerando tanto la opinión de los expertos como el resto de observaciones, se decidió reemplazar los valores -1 por el valor 19.999, que es el valor máximo que puede tomar esta variable. Lo que indica que no existe ninguna aeronave en la pista paralela a la pista de aterrizaje.

Tabla 6: Estadísticas descriptivas de distancia diagonal

Pistas	***Media***	***Desviación estándar***	***Mínimo***	***Percentil 25***	***Mediana***	***Percentil 75***	***Máximo***
32	5.77	5.26	0.92	2.65	3.39	6.62	19.99
18	4.99	4.47	0.65	2.46	3.31	5.28	19.99

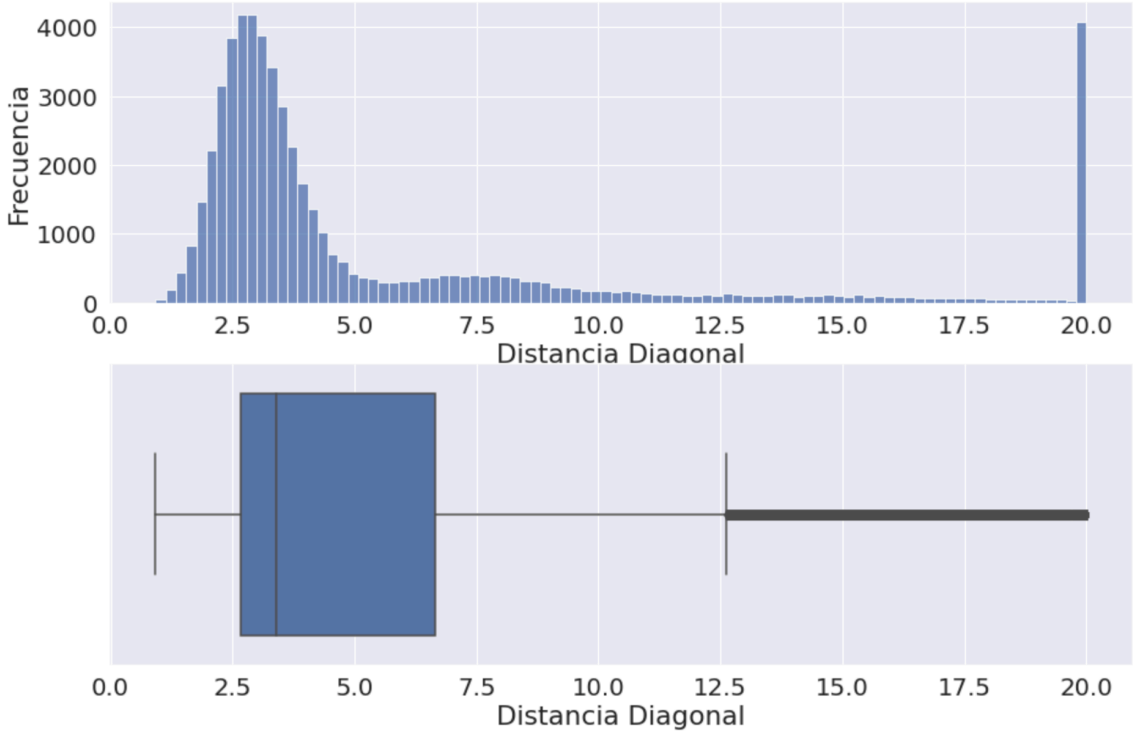

Figura 9: Distribución de distancia diagonal en la pista 32

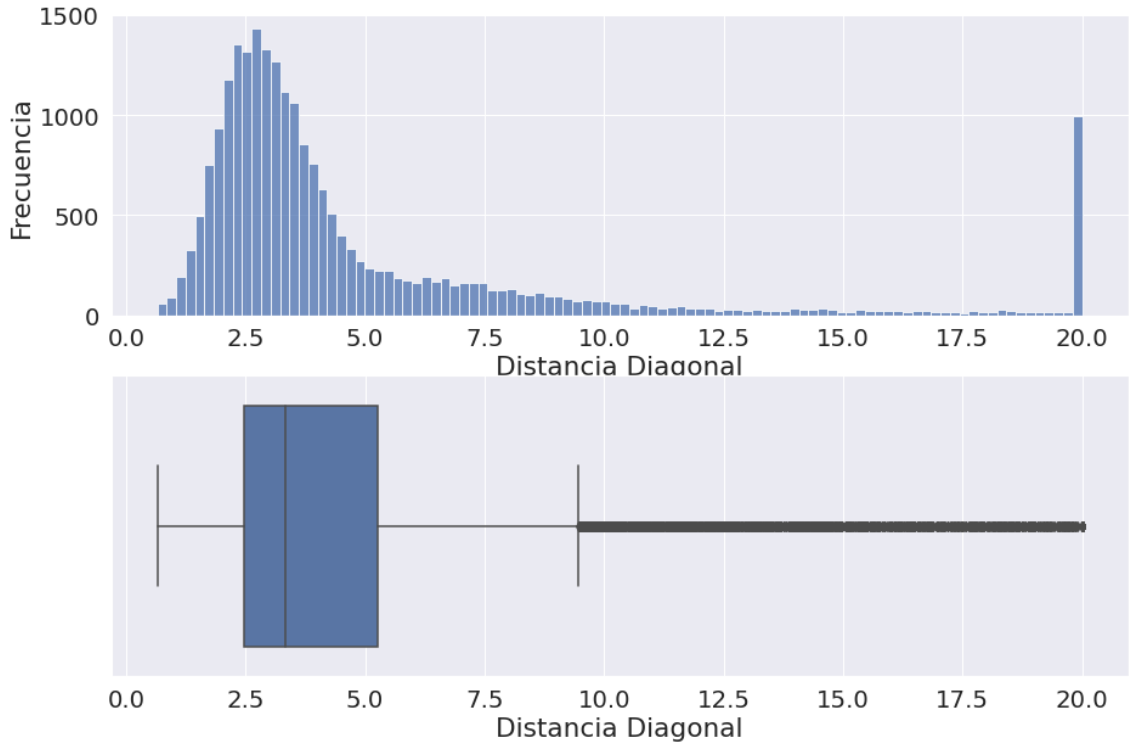

Figura 10: Distribución de distancia diagonal en la pista 18

2.2.2.3 Velocidad de la aeronave

La variable velocidad de la aeronave es de vital importancia en el control del tráfico aéreo y en la seguridad aérea, ya que permite determinar la distancia entre dos aeronaves en el momento del aterrizaje. De esta manera, se puede garantizar que las aeronaves estén lo suficientemente separadas como para evitar cualquier tipo de accidente o colisión en el momento del aterrizaje.

Es importante señalar que esta variable es continua y puede tomar cualquier valor numérico dentro de un determinado rango. El hecho de que esta variable se mida en nudos refleja la importancia de la velocidad aerodinámica en la aviación y destaca la necesidad de mediciones exactas y precisas para garantizar aterrizajes seguros y eficaces. Dicho esto, un nudo equivale a 1,852 km/h.

En base a la descripción de la variable, la velocidad máxima permitida está fijada en 225 nudos. Sin embargo, se encontraron registros que superaban este límite. Con el objetivo de prevalecer la integridad de los datos, los registros que contaban con velocidades mayores a lo permitido, fueron eliminados ya que ninguna aeronave en el Aeropuerto Adolfo Suárez Madrid-Barajas pueden llegar a obtener esa velocidad en el momento de aterrizaje. Este proceso fue validado con los expertos.

Tabla 7: Estadísticas descriptivas de la velocidad por pista

Pistas	***Media***	***Desviación estándar***	***Mínimo***	***Percentil 25***	***Mediana***	***Percentil 75***	***Máximo***
32	142.51	10.97	91.95	136.01	143.20	149.93	208.05
18	139.89	10.72	77.90	134.09	139.80	147.25	187.55

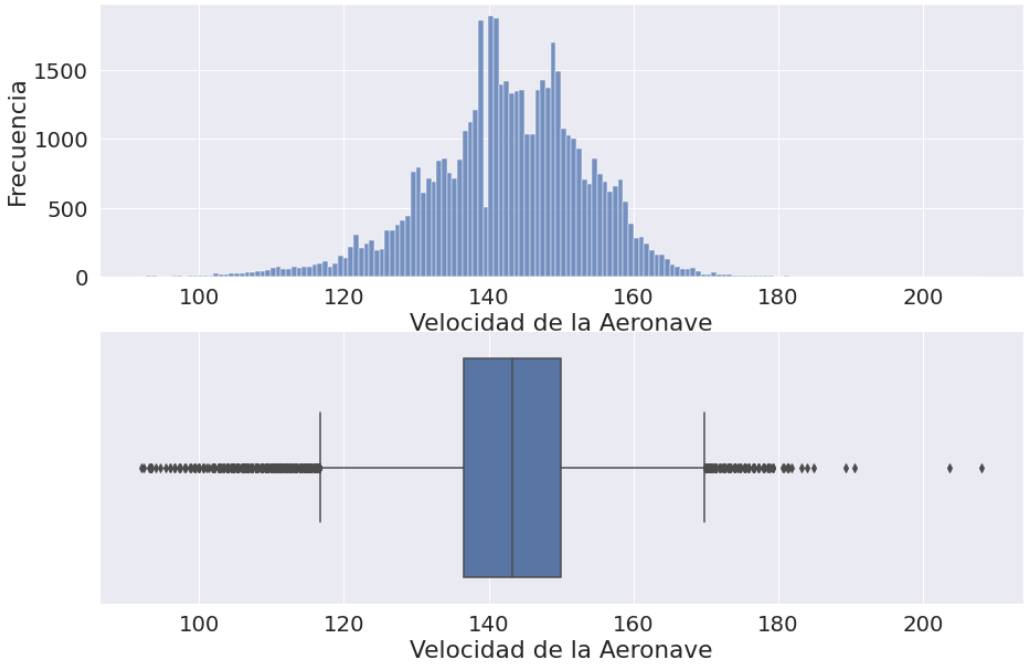

Figura 11: Distribución de la velocidad en la pista 32

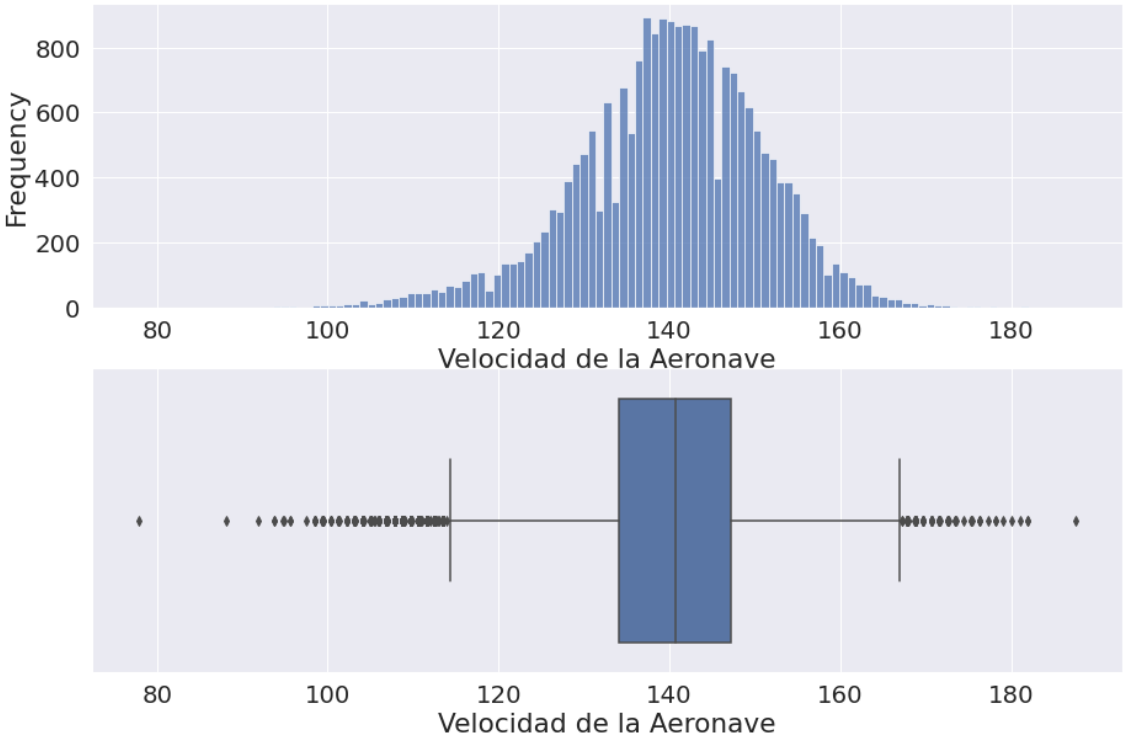

Figura 12: Distribución de la velocidad en la pista 18

2.2.2.4 Altura de la aeronave

La altitud de una aeronave durante el aterrizaje es un factor crítico para garantizar un aterrizaje seguro y eficaz. Esto se debe a que la altitud puede afectar a la velocidad de descenso y a la estabilidad general de la aeronave. Por lo tanto, las mediciones exactas y precisas de la altitud son esenciales para garantizar que la aeronave aterrice de forma segura y eficaz.

La variable que se describe en este contexto esta medida en pies, que equivale a 0.358 metros, y calculada a cuatro millas náuticas. Esta variable se considera una variable numérica discreta, lo que significa que sólo puede tomar valores numéricos específicos dentro de un rango determinado.

Durante el estudio de la variable, no se encontraron datos ausentes ni datos atípicos, los cuales son valores que se alejan significativamente de la tendencia general y pueden indicar observaciones inusuales o errores en la medición. Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Tabla 8: Distribución de observaciones de la altitud

Pistas	***20***	***21***	***23***	***22***	***24***	***19***	***25***	***26***	***18***	***27+***
32	6809	6592	6502	6428	6337	5828	5415	2645	491	341
18	3257	3062	3062	3024	3023	2818	1731	1593	336	33

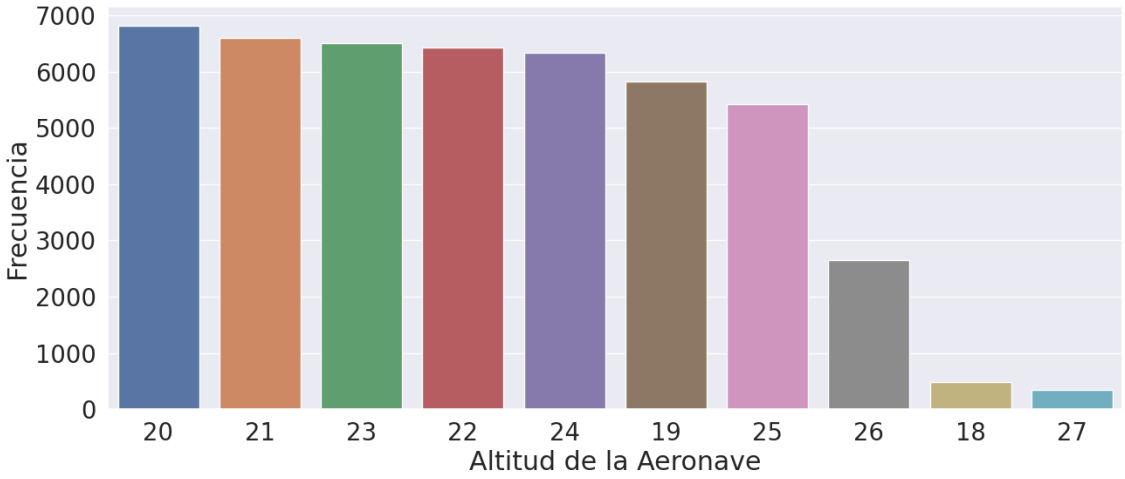

Figura 13: Distribución de altitud en la pista 32

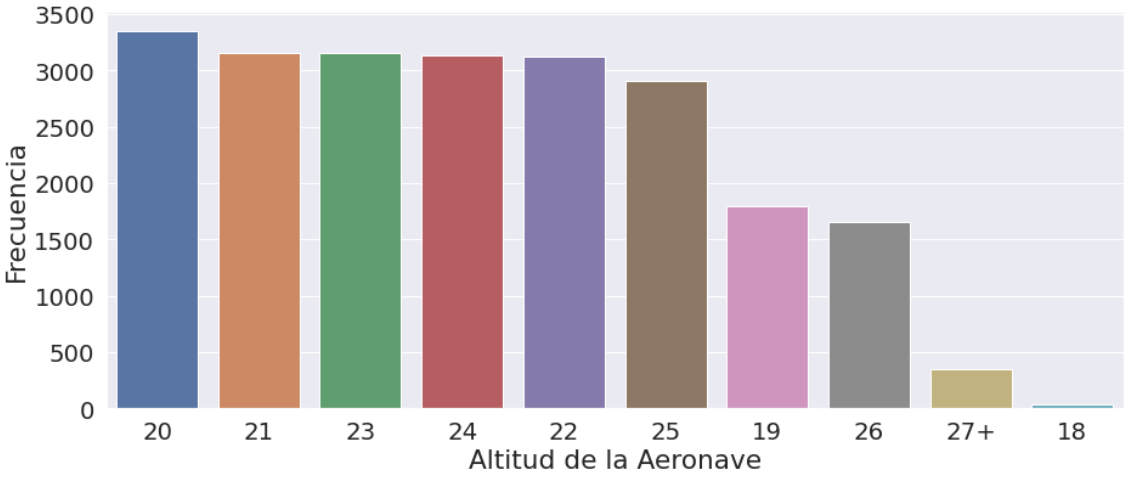

Figura 14: Distribución de altitud en la pista 18

2.2.2 Variables meteorológicas

La obtención de variables meteorológicas relevantes es fundamental en el análisis del METAR, el cual se extrae en intervalos de media hora. Dentro de este informe, se pueden identificar cuatro campos principales: datos misceláneos, nubes, fenómenos e información de la pista. En el apartado de datos misceláneos se incluyen las variables meteorológicas genéricas y obligatorias en el METAR. En el caso de las nubes, se han seleccionado las variables más representativas y las de mayor relevancia para los controladores aéreos al tomar decisiones relacionadas con la distancia de umbral.

Los fenómenos meteorológicos se describen mediante la indicación de múltiples tipos, cada uno con su nivel de intensidad asociado. Para facilitar la representación de estas diferentes condiciones meteorológicas, se ha establecido una clasificación ordenada. Se consideran variables categóricas ordinales con cuatro valores posibles, los cuales se detallan en la Tabla 9. Esta clasificación permitirá un análisis más preciso y comprensible de los fenómenos.

Tabla 9: Categorización de fenómenos

Categoría	***Valor***
Ausencia de fenómenos	0
Fenómeno de intensidad baja	1
Fenómeno con intensidad media	2
Fenómeno con intensidad alta	3

A lo largo de este capítulo, se examinará en detalle la extracción y categorización de estas variables, proporcionando una visión clara y sistemática de la información meteorológica contenida en el METAR.

2.2.2.1 Visibilidad

La visibilidad hace referencia al alcance visual en la aviación. Este es esencial para la operación segura y eficiente de las aeronaves. La capacidad de ver y navegar por el espacio aéreo depende en gran medida de la visibilidad, y la medición precisa del alcance visual es necesaria para que los pilotos tomen decisiones informadas y ajusten su aproximación en consecuencia. La comprensión de la variable de alcance visual es crucial para identificar áreas potenciales de mejora en la seguridad y eficiencia de la aviación, y puede ayudar a informar los procesos de toma de decisiones para pilotos y profesionales de la aviación.

La presente es una variable numérica que se mide en metros y puede tomar los valores en un rango de 0 a 9999. Un valor de 9999 representa una visibilidad perfecta, que corresponde a una distancia de 10 kilómetros o más, mientras que un valor de 0 indica una visibilidad nula.

Durante su estudio, no se encontraron datos ausentes ni datos atípicos que caen fuera del rango intercuartílico. Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Tabla 10: Estadísticas descriptivas de la visibilidad

Pistas	***Media***	***Desviación estándar***	***Mínimo***	***Mediana***	***Máximo***
32	9846.3	864.3	200.0	9999.0	9999.0
18	9886.0	679.5	2400.0	9999.0	9999.0

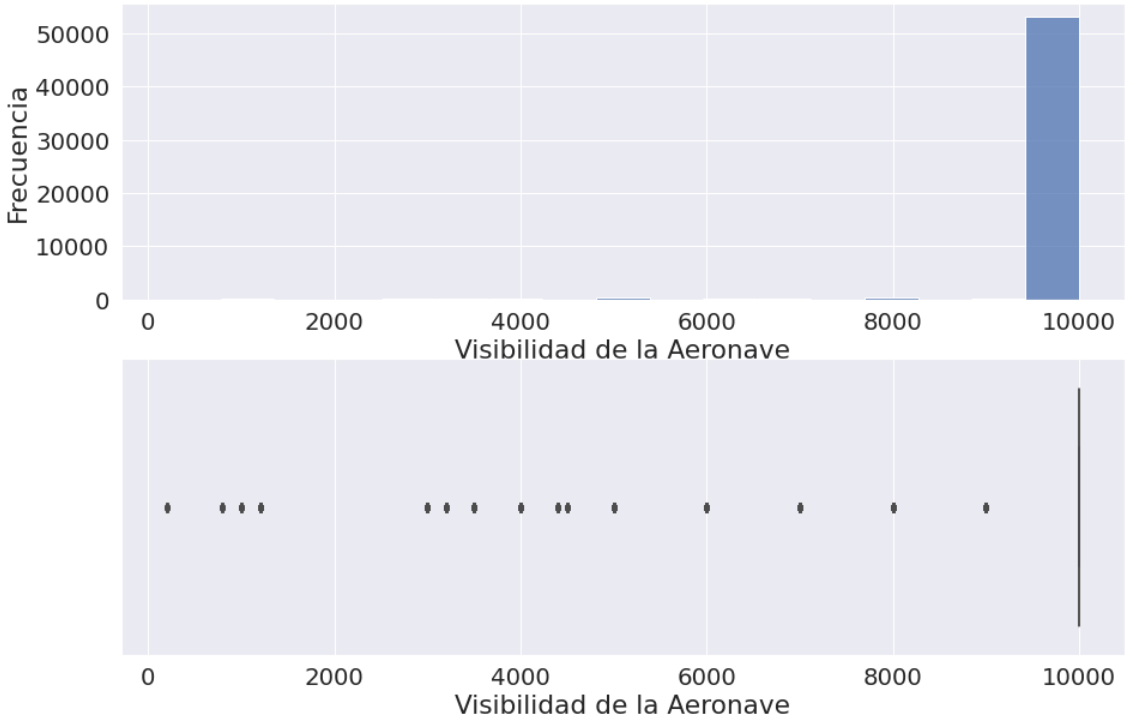

Figura 15: Distribución de visibilidad en la pista 32

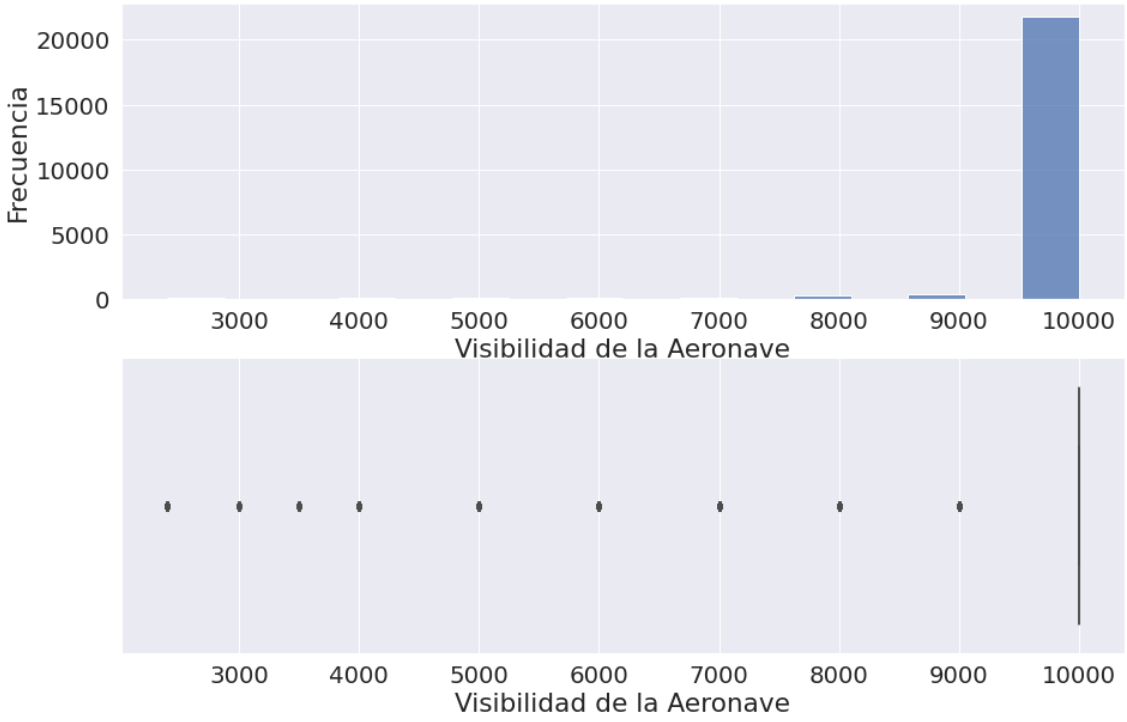

Figura 16: Distribución de visibilidad en la pista 18

2.2.2.2 CAVOK

La presencia de condiciones CAVOK (***Ceiling and Visibility OK***) es un factor importante para la seguridad y eficiencia de la aviación. Esta indica que no hay limitaciones significativas en cuanto a la visibilidad, el techo nuboso, la temperatura y la humedad, lo que proporciona a los pilotos información crítica sobre las condiciones meteorológicas actuales. Esta información les permite tomar decisiones informadas sobre sus planes de vuelo, asegurando un entorno más seguro para las operaciones aéreas.

La naturaleza binaria de esta variable significa que sólo puede tomar dos valores posibles, ya sea indicando la presencia o la ausencia de condiciones meteorológicas perfectas.

Durante su estudio, no se encontraron datos ausentes ni datos atípicos (valores diferentes a 0 y 1). Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Tabla 11: Distribución de observaciones CAVOK por pista

Pistas	***1***	***0***
32	30544	16844
18	11349	10590

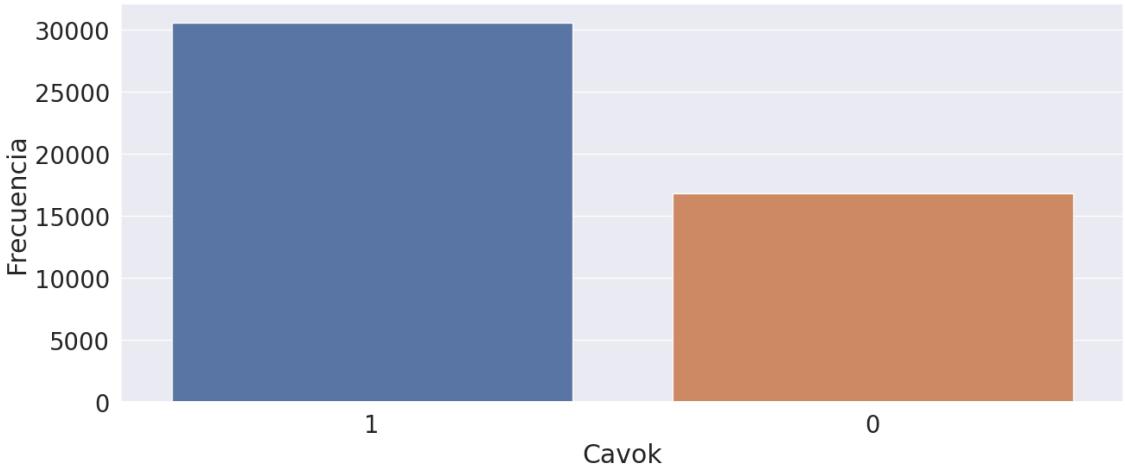

Figura 17: Distribución de CAVOK en la pista 32

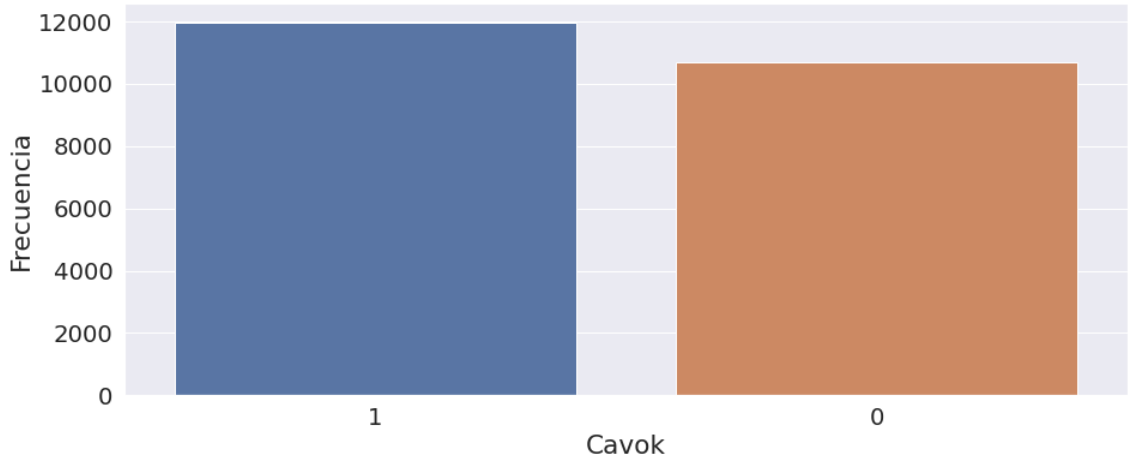

Figura 18: Distribución de CAVOK en la pista 18

2.2.2.3 Temperatura

La temperatura es una variable importante en la seguridad aérea, sobre todo cuando se trata de mantener una distancia segura entre aeronaves. La temperatura afecta a la densidad del aire, que a su vez afecta a las fuerzas de sustentación y resistencia que actúan sobre una aeronave. Esto puede influir en la velocidad y la altitud a las que puede operar una aeronave, así como en su capacidad para maniobrar y mantener distancias seguras con otras aeronaves.

La temperatura se mide normalmente en grados Celsius y se considera una variable numérica continua. La medición precisa de la temperatura es fundamental para predecir el rendimiento de las aeronaves y garantizar la seguridad de las operaciones de vuelo.

Durante su estudio y la continua comunicación con los expertos se descubrió que, si bien la variable es registra en el METAR y es enviada a un controlador aéreo, este no toma en cuenta la temperatura al momento de establecer la separación de umbral entre dos aeronaves. Por lo cual, se optó por eliminar la variable del conjunto de datos.

Tabla 12: Estadísticas de la temperatura

Pistas	***Media***	***Desviación estándar***	***Mínimo***	***Percentil 25***	***Mediana***	***Percentil 75***	***Máximo***
32	16.59	8.51	-5.00	10.00	16.00	23.00	39.00
18	21.41	9.56	3.00	12.00	21.00	30.00	41.00

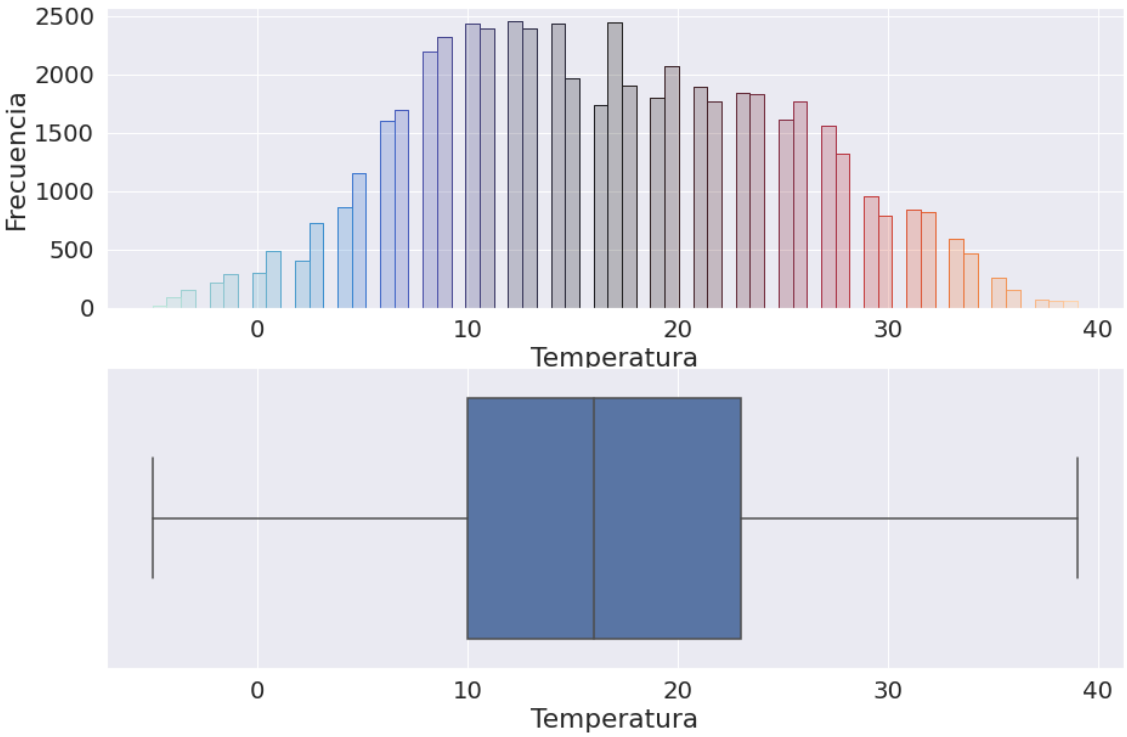

Figura 19: Distribución de temperatura en la pista 32

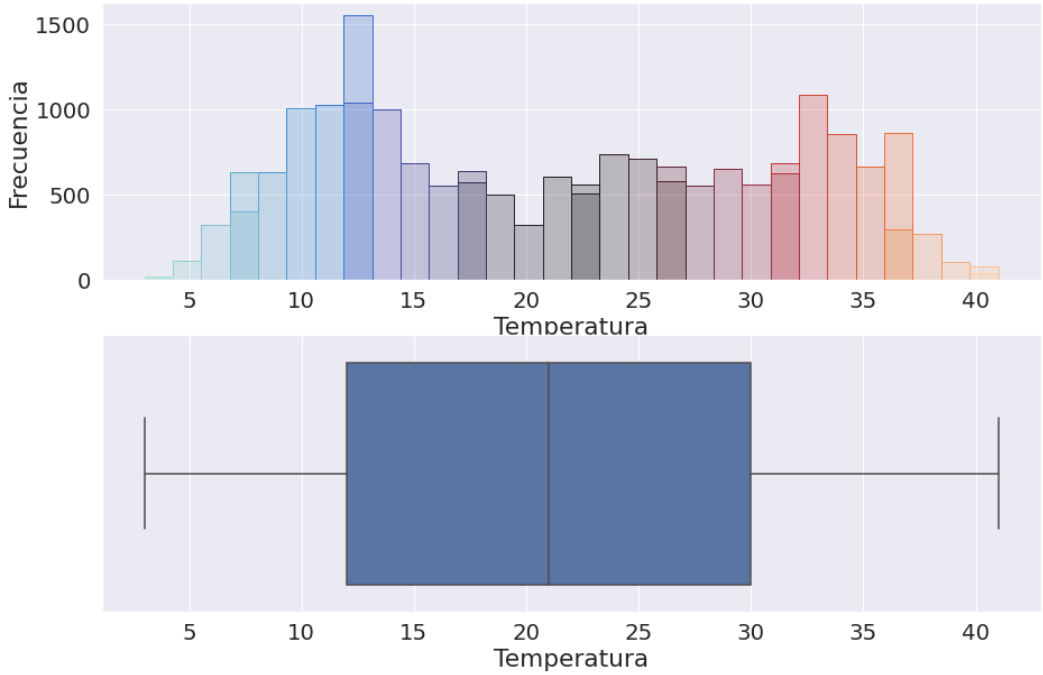

Figura 20: Distribución de temperatura en la pista 18

2.2.2.4 Presión

El QNH es un sistema que utiliza la presión para determinar la altitud, y se mide en hectopascales. Se trata de una variable importante en aviación y meteorología, ya que permite a pilotos y meteorólogos determinar la presión atmosférica en un lugar concreto, que a su vez afecta a factores como la densidad del aire, la temperatura y los patrones meteorológicos.

La variable es de tipo numérica continua, lo que significa que puede tomar cualquier valor dentro de un cierto rango de valores, y es una herramienta importante para garantizar la seguridad y la eficiencia de los viajes aéreos. Además, se utiliza en el cálculo de otras variables como la altitud y el nivel de vuelo, y los controladores aéreos y las autoridades de aviación lo vigilan de cerca.

Durante su estudio y la continua comunicación con los expertos se descubrió que, si bien la variable es registra en el METAR y es enviada a un controlador aéreo, este no toma en cuenta la presión al momento de establecer la separación de umbral entre dos aeronaves. Por lo cual, se optó por eliminar la variable del conjunto de datos.

Tabla 13: Estadísticas descriptivas de la presión por pista

Pistas	***Media***	***Desviación estándar***	***Mínimo***	***Percentil 25***	***Mediana***	***Percentil 75***	***Máximo***
32	1018.5	5.93	995.0	1015.0	1018.0	1022.0	1034.0
18	1014.3	6.77	991.0	1012.0	1015.0	1018.0	1032.0

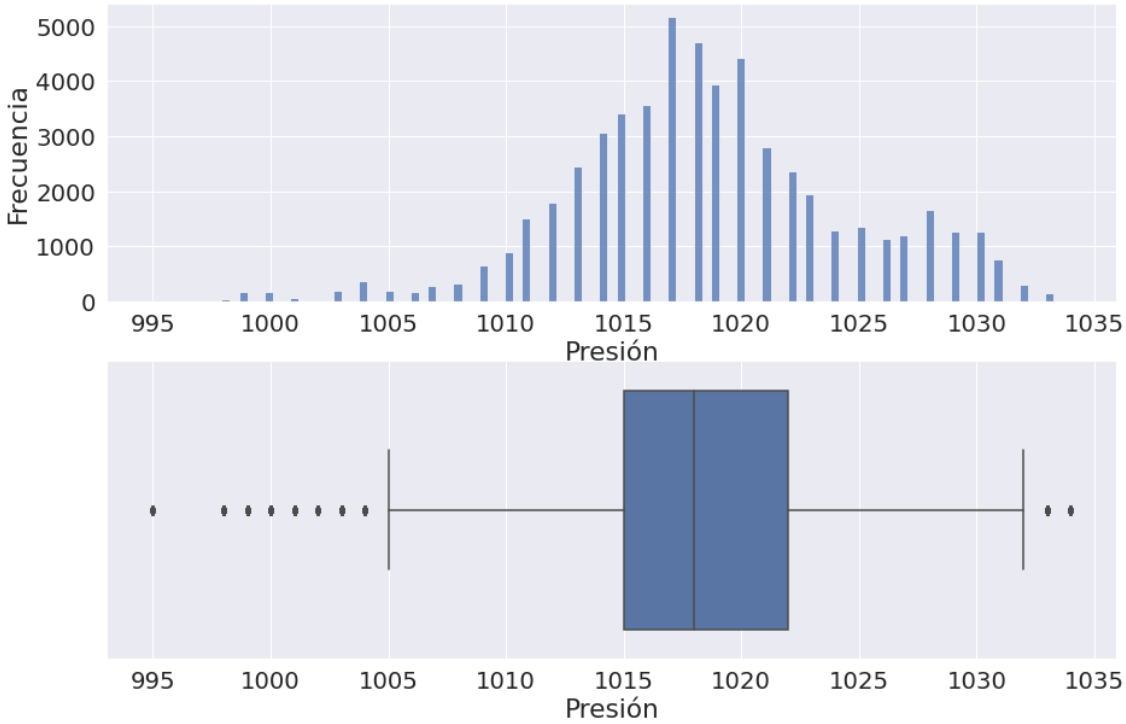

Figura 21: Distribución de presión en la pista 32

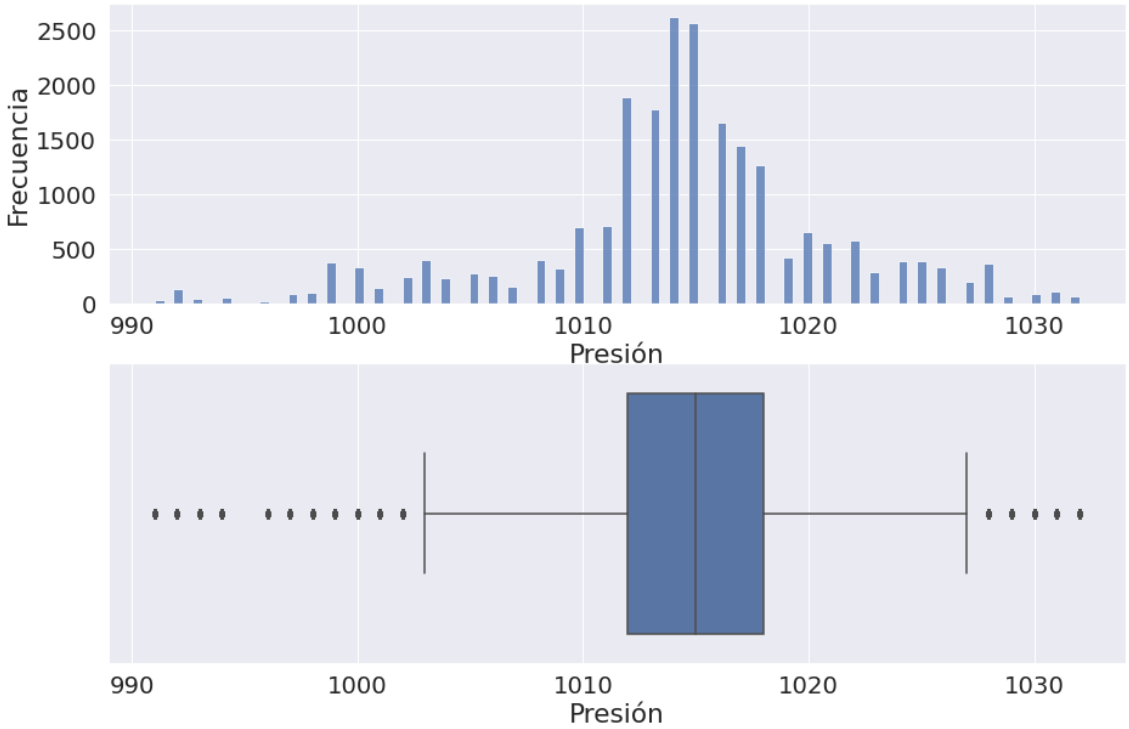

Figura 22: Distribución de presión en la pista 18

2.2.2.5 Viento variable

La variable es un indicador de la presencia de viento variable. Se refiere a la detección de la presencia de viento que no mantiene una dirección fija, sino que fluctúa dentro de un rango de direcciones diferentes.

Es una variable de tipo binaria que registra la presencia (1) o no (0) de viento variable. El viento variable puede tener un impacto significativo en la separación de umbral entre dos aeronaves en el aeropuerto de Barajas, Madrid.

Durante el proceso de estudio, se observó que la información proporcionada por la variable "viento variable" era muy similar a la que se obtenía a través de otra variable denominada "variabilidad en la intensidad de viento". Esto generaba redundancia entre ambas variables, ya que una brindaba información más resumida mientras que la otra ofrecía detalles más específicos en términos de nudos. En consecuencia, se decidió utilizar la variable "viento variable" para validar los registros de la variable "variabilidad en la intensidad de viento". Una vez que se confirmó la validez de los valores obtenidos, se determinó que la variable "viento variable" no aportaba información adicional relevante para el estudio, por lo que se decidió eliminarla.

Tabla 14: Distribución de observaciones del viento variable por pista

Pistas	***Ausencia (0)***	***Presencia (1)***
32	47389	7795
18	21939	720

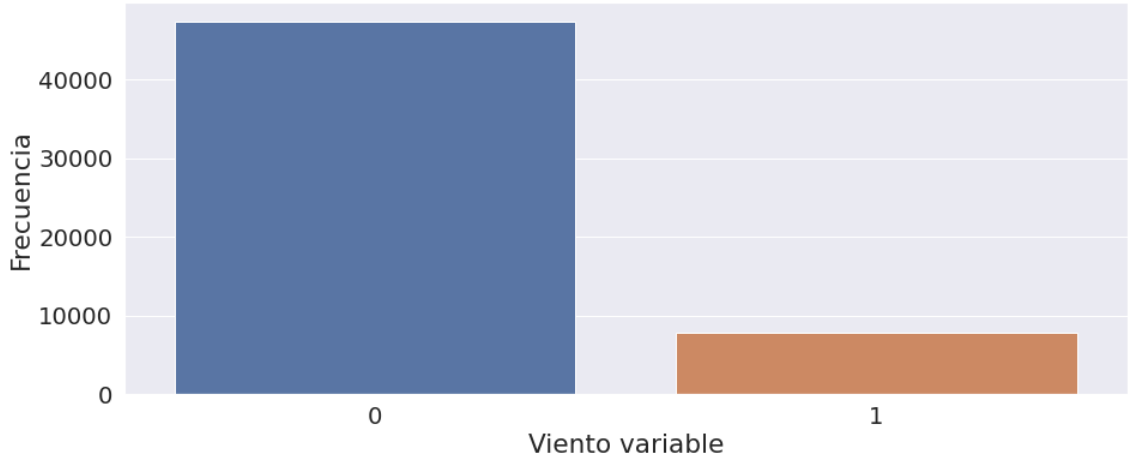

Figura 23: Distribución del viento variable en la pista 32

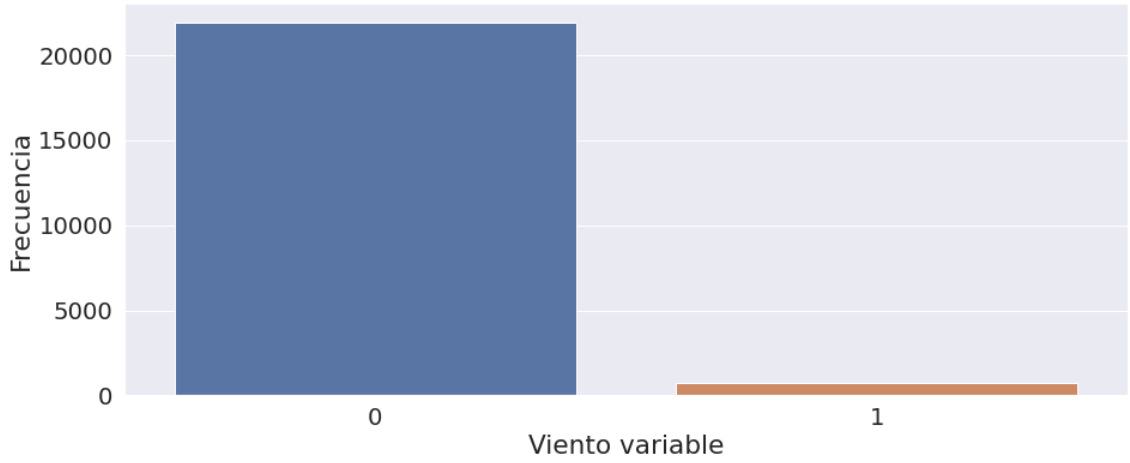

Figura 24: Distribución del viento variable en la pista 18

2.2.2.6 Variabilidad en la intensidad del viento

La medición precisa de la intensidad de la variabilidad del viento es esencial para predecir cómo afectará al rendimiento y la maniobrabilidad de las aeronaves, lo que puede afectar a la distancia entre aviones en vuelo. Al incorporar los datos de la fuerza del viento en modelos predictivos, los profesionales de la aviación pueden anticipar cambios en el comportamiento de las aeronaves y ajustar los planes de vuelo en consecuencia para mantener distancias seguras entre ellas.

Durante el análisis de las variables del presente estudio, se identificó tanto la variable "viento variable" y "variabilidad en la intensidad del viento" los cuales presentaban una información similar en sus registros. Mientras que la primera variable en cuestión representaba la presencia o no de viento variable, en esa condición, se registran los valores de la variabilidad en la intensidad del viento. De este modo, se decide solo considerar la variabilidad en la intensidad del viento, validando sus registros en base a los valores que se tenía del viento variable.

La presente variable solo se registra cuando dicha variación es de más diez nudos respecto de la intensidad media. Es por ello, que se tiene un valor atípico que representa una variación menor a diez nudos. Así, estos valores, tanto en la pista 32 como 18, se categorizaron en base a las siguientes clases:

- Variabilidad baja en la intensidad (L): Valores en un rango de 0 a 9
- Variabilidad media en la intensidad (M): Valores en un rango de 10 a 13
- Variabilidad alta en la intensidad (H): Valores en un rango de 14 a 16
- Variabilidad jumbo en la intensidad (J): Valores en un rango de 17 a 20

Tabla 15: Distribución de la variabilidad en la intensidad

Pistas	***Low***	***Medium***	***High***	***Jumbo***
32	45886	1467	35	0
18	17976	3330	497	136

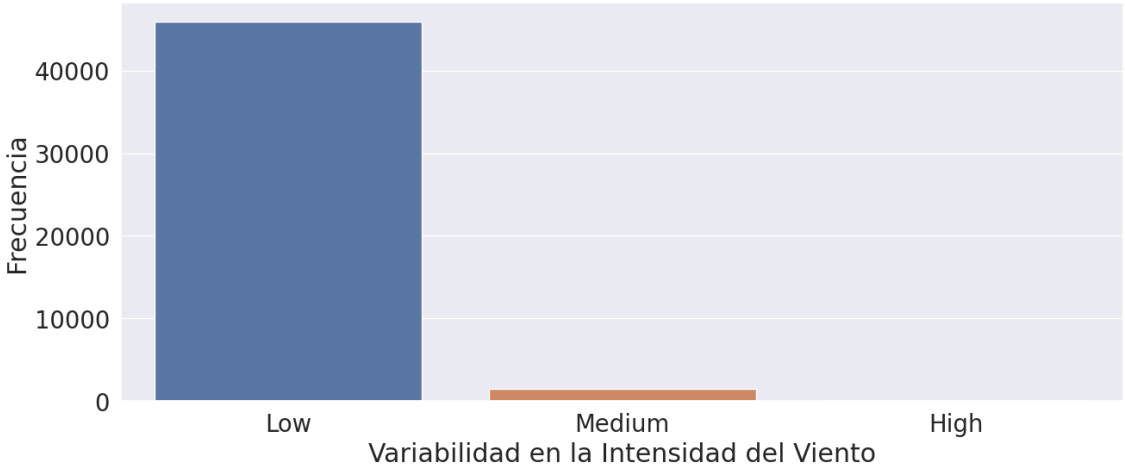

Figura 25: Distribución de variabilidad en intensidad en la pista 32

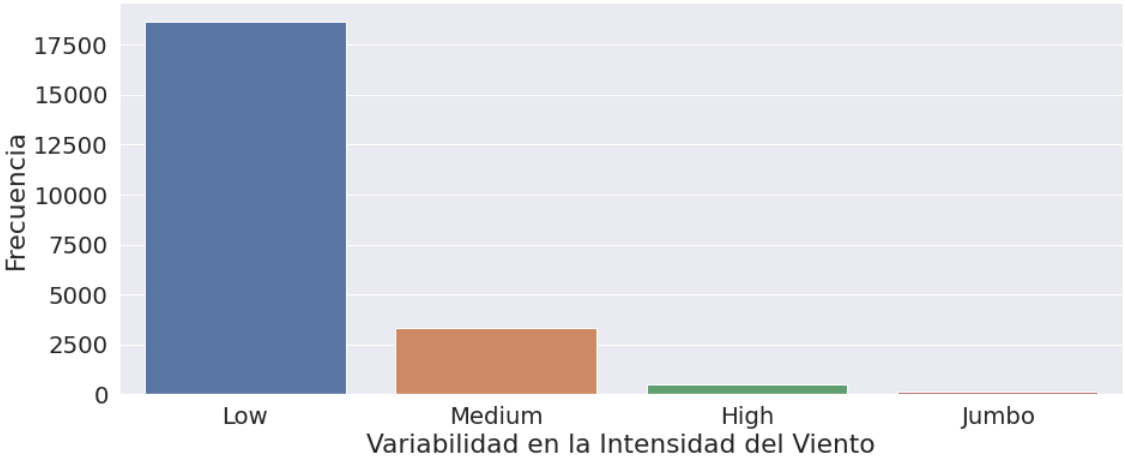

Figura 26: Distribución de variabilidad en intensidad en la pista 18

2.2.2.7 Dirección relativa del viento

La actual variable hace referencia a la dirección desde la que sopla el viento en relación con la pista en la que está aterrizando un avión. Puede tener valores dentro de un rango de 0 a 360 grados y es una variable categórica ordinal. Para comprender cómo afecta realmente el viento a la aeronave, se obtiene la dirección del viento en relación con la aeronave. Ésta puede tener cualquier valor dentro de un rango de 0 a 180 grados. Sin embargo, hay tres tipos principales de viento, ordenados de menor a mayor peligro: viento en contra (procedente de 0 grados), viento transversal (procedente de 90 grados) y viento de cola (procedente de 180 grados). El lado de la aeronave en el que sople el viento no importa, ya que afectará a la aeronave de la misma manera. Por lo tanto, los vientos dentro del rango (180, 360) se reflejan en el rango [0, 180]. El diagrama adjunto muestra las posibles direcciones del viento.

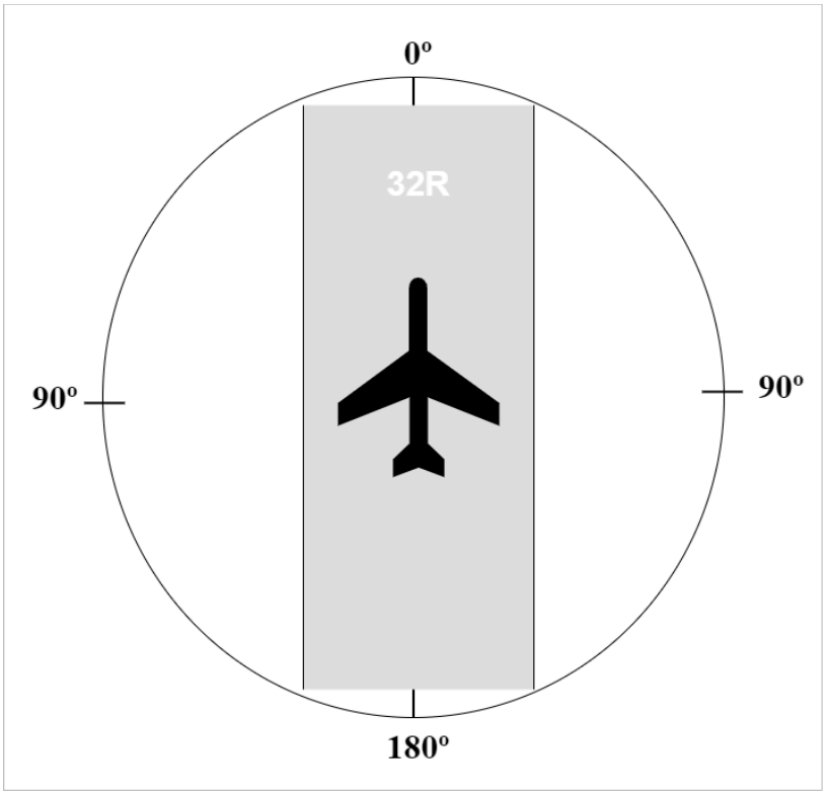

Figura 27: Orientación relativa del viento con respecto a la aeronave

Una vez se obtuvo la dirección relativa del viento, dentro del análisis de sus valores se registró un valor atípico (-1) que representa a la no existencia del viento o que no se puede determinar la dirección del viento en base a la constante variación de este. Dado que estos valores representan un 14.13% y un 3.18% en las pistas 32 y 18 respectivamente, se procedieron a eliminar los registros del conjunto de datos.

Tras obtener la dirección relativa del viento, los valores se han categorizado para reducir el número de valores o etiquetas posibles. Se han agrupado en función del nivel de peligrosidad del viento, como se ha mencionado anteriormente, dando lugar a cinco grupos diferentes: viento en contra (0 - 30 grados), viento

en contra diagonal (30 - 60 grados), viento en cola diagonal (120 - 150 grados), viento cruzado (60 - 120 grados) y viento en cola (150 - 180 grados). Están ordenados de menor a mayor peligro, respectivamente. Básicamente, esto significa que los valores de dirección del viento se han organizado en cinco grupos en función del peligro potencial que suponen para la aeronave.

La categorización de los valores se dio de la siguiente forma:

- Front = Viento de cara
- DFront = Viento de cara diagonal
- DTail = Viento de cola diagonal
- Lateral = Viento lateral
- Tail = Viento de cola

Tabla 16: Distribución de la dirección del viento por pista

Pistas	***DFront***	***Lateral***	***Front***	***Dtail***	***Tail***
32	15022	11838	10533	5376	4619
18	9462	1884	9457	266	870

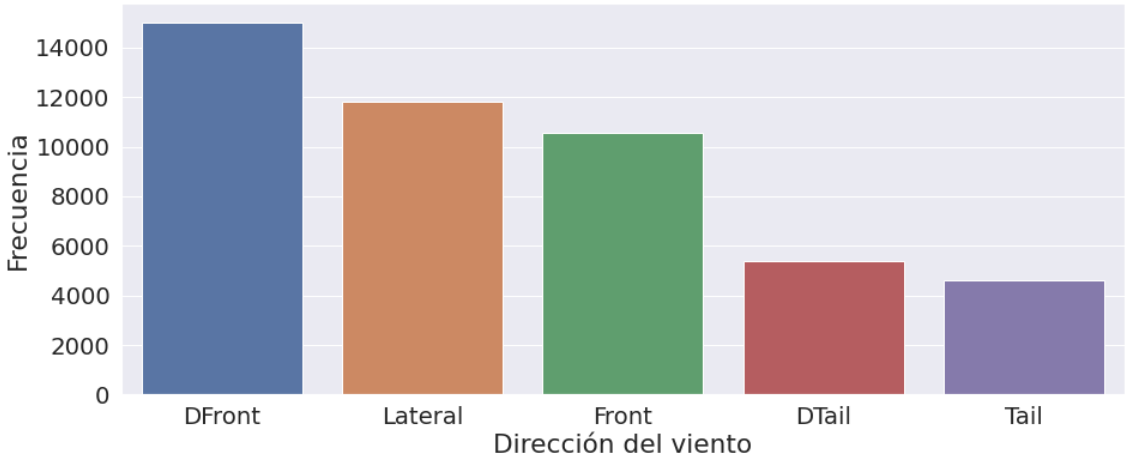

Figura 28: Distribución de dirección del viento en la pista 32

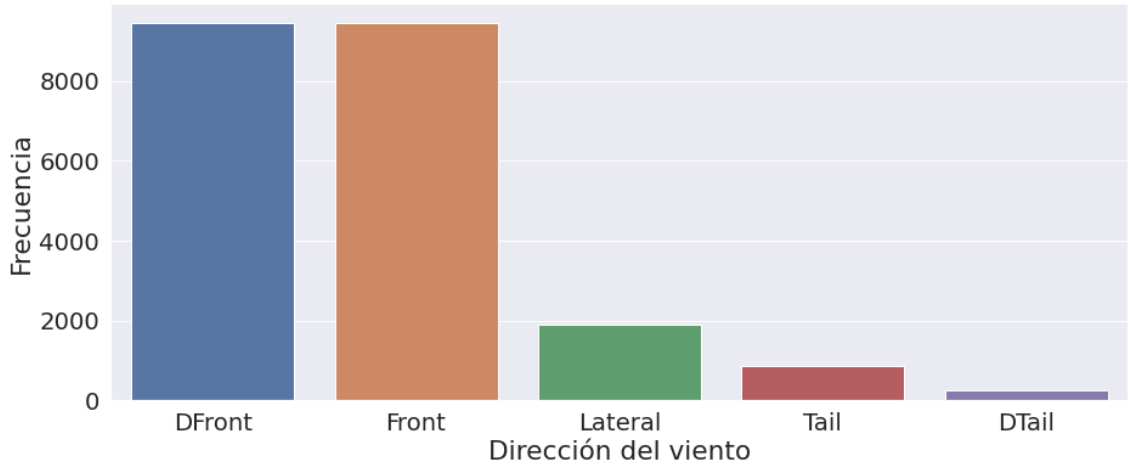

Figura 29: Distribución de dirección del viento en la pista 18

2.2.2.8 Intensidad del viento

En el ámbito de la aviación, el conocimiento de la velocidad del viento es de vital importancia, ya que puede tener un impacto significativo en el rendimiento de las aeronaves, particularmente durante las fases críticas de despegue y aterrizaje. Por lo tanto, al establecer distancias de operación seguras, es imperativo considerar detenidamente estos datos.

Se obtuvo a través los reportes METAR mediante la medición de la velocidad del viento en nudos durante un período de tiempo definido. A través de cálculos meticulosos, se determinó el valor promedio de la intensidad del viento, que desempeña un papel crucial en la determinación de la distancia adecuada entre las aeronaves para asegurar una operación segura. Cabe destacar que la intensidad del viento es una variable numérica continua, lo que implica que puede adoptar valores en un rango infinito de posibilidades.

Durante su estudio, no se encontraron datos ausentes ni datos atípicos. Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Tabla 17: Estadísticas de la intensidad del viento

Pistas	***Media***	***Desviación estándar***	***Mínimo***	***Percentil 25***	***Mediana***	***Percentil 75***	***Máximo***
32	5.05	4.05	0.00	2.00	4.00	7.00	31.00
18	9.30	4.91	0.00	6.00	9.00	12.00	30.00

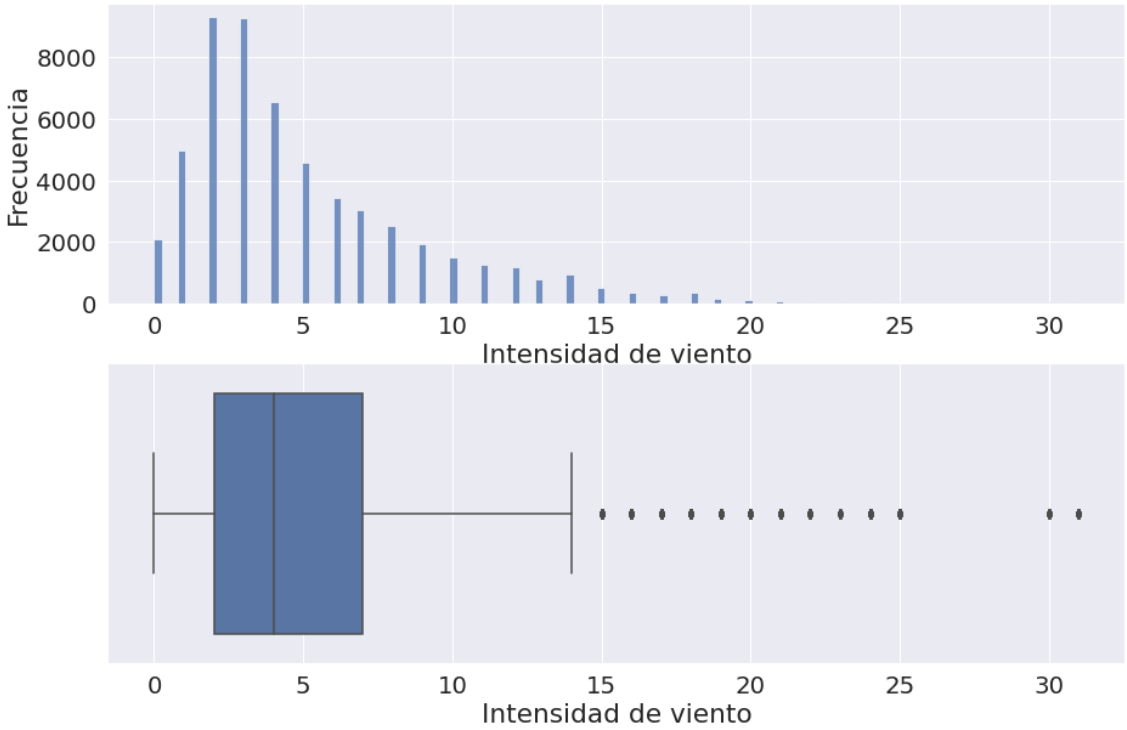

Figura 30: Distribución de la intensidad del viento en la pista 32

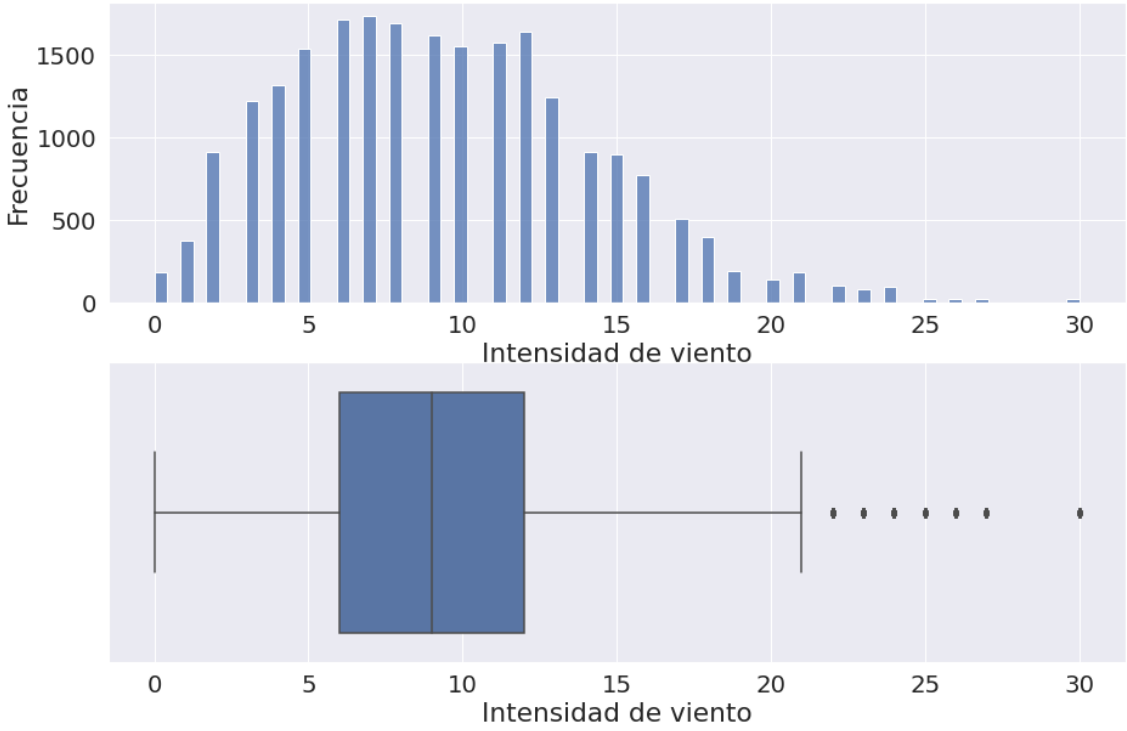

Figura 31: Distribución de la intensidad del viento en la pista 18

2.2.2.9 Existencia de nubes

El presente apartado aborda el indicador de la existencia de nubes. Para su creación, se ha llevado a cabo una combinación de dos variables que informaban sobre la ausencia de nubes. Una de estas variables se basaba en observaciones físicas realizadas por un controlador, mientras que la otra provenía de una fuente automática de datos. Como resultado, se obtuvo una variable binaria que proporciona una indicación clara sobre la presencia o ausencia de nubes.

En términos más simples, este indicador se presenta como una herramienta fundamental para determinar si hay nubes presentes en un determinado momento. Su naturaleza binaria permite una clasificación precisa y directa en relación con la existencia de nubes.

Durante su estudio y la continua comunicación con los expertos se descubrió que, si bien la variable es registra en el METAR y es enviada a un controlador aéreo, este no toma en cuenta la variable "existencia de nubes" al momento de establecer la separación de umbral entre dos aeronaves. Por lo cual, se optó por eliminar la variable del conjunto de datos.

Tabla 18: Distribución de observaciones de nubes por pista

Pistas	***Ausencia (0)***	***Presencia (1)***
32	36568	18616
18	11973	10686

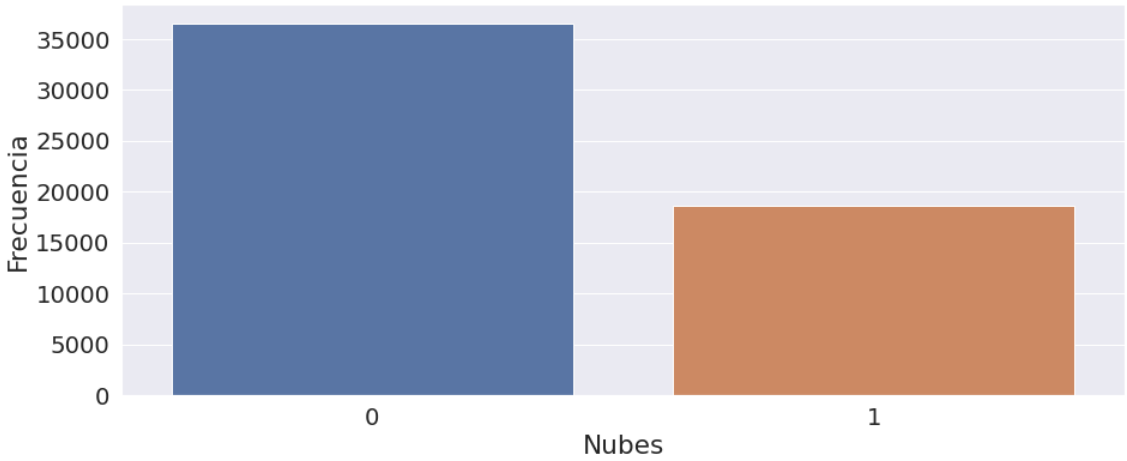

Figura 32: Distribución de presencia de nubes en la pista 32

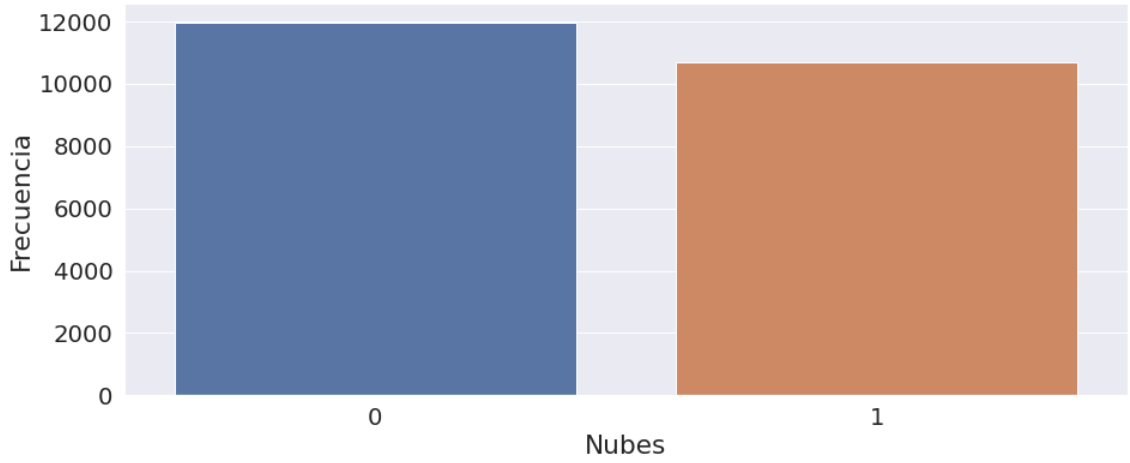

Figura 33: Distribución de presencia de nubes en la pista 18

2.2.2.10 Nubosidad baja

La presente variable establece la cantidad de nubes bajas presentes en la atmósfera. En este sentido, se llevó a cabo una observación y medición exhaustiva de las nubes situadas en la parte inferior de la troposfera, las cuales se ubican a una altitud de hasta 2.000 metros sobre la superficie terrestre. Cabe destacar que este tipo de nubes suele exhibir una mayor densidad y puede afectar la visibilidad y el comportamiento de las aeronaves durante las maniobras de aterrizaje y despegue. Por ende, resulta de vital importancia tener en consideración la presencia de nubes bajas al establecer la distancia de seguridad necesaria entre dos aeronaves, a fin de prevenir situaciones de riesgo y asegurar una operación aérea segura y eficiente.

La variable en cuestión ha sido clasificada en distintas clases, cada una con sus respectivos valores asociados, para un análisis más detallado:

Tabla 19: Categorización de valores de nubosidad baja

Categoría	***Valor***
Ausencia de nubes	0
Escasa presencia de nubes	1
Nubes dispersas	2
Cielo muy nuboso	3
Cielo cubierto	4

Durante su estudio, no se encontraron datos ausentes ni datos atípicos. Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Tabla 20: Distribución de la nubosidad baja por pista

Pistas	***0***	***1***	***2***	***3***	***4***
32	30883	10818	4275	1156	256
18	11349	6131	3193	1185	81

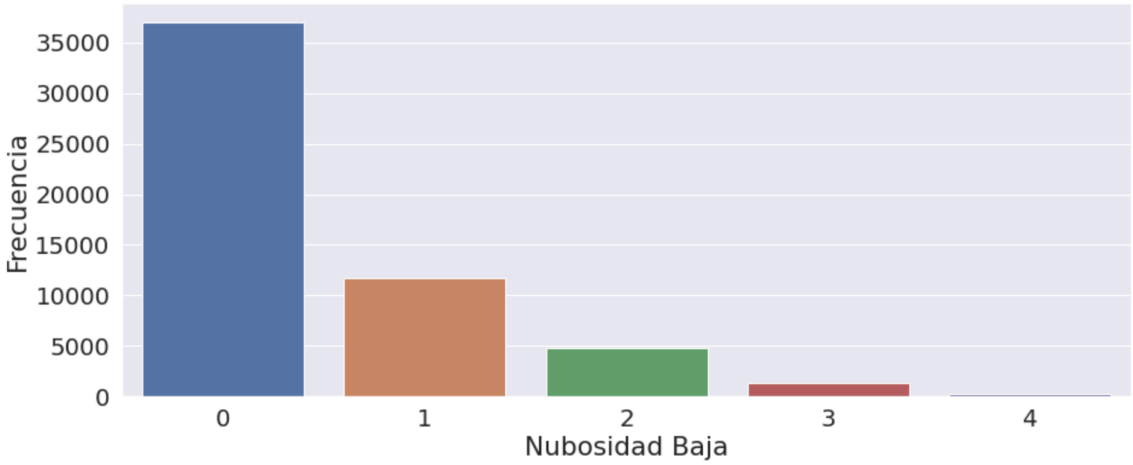

Figura 34: Distribución de nubosidad baja en la pista 32

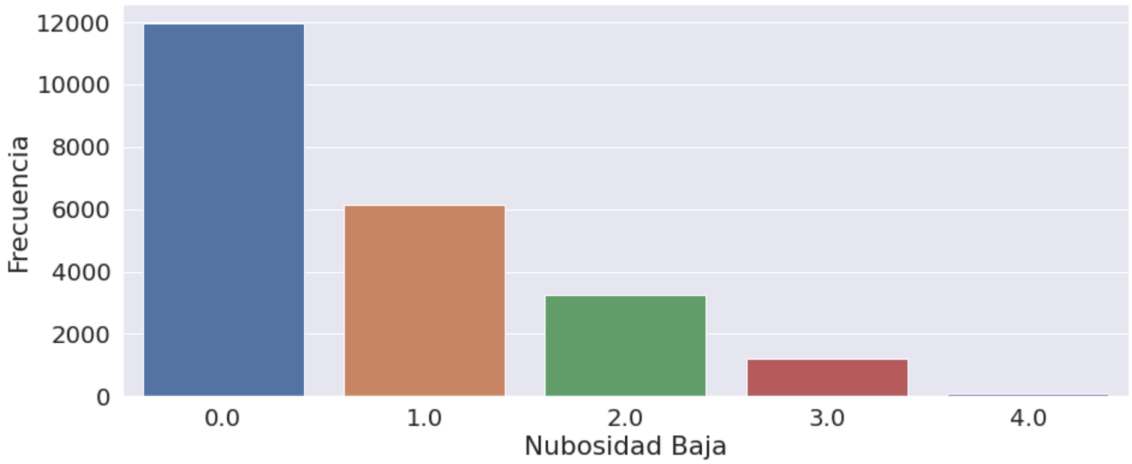

Figura 35: Distribución de nubosidad baja en la pista 18

2.2.2.11 Nubes peligrosas

La variable actual identifica la presencia de tipos de nubes altamente peligrosas en el contexto del estudio. Específicamente, se enfoca en las nubes Cumulonimbus (CB) y las nubes Cumulus Congestus (TCU) que se caracterizan por su considerable extensión vertical. Estas nubes representan un riesgo significativo en términos de seguridad aérea.

Cabe destacar que la variable en cuestión se clasifica como binaria, lo que implica que solo puede tener dos valores posibles.

Durante su estudio, no se encontraron datos ausentes ni datos atípicos. Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Tabla 21: Distribución de nubes peligrosas por pista

Pistas	***Ausencia (0)***	***Presencia (1)***
32	46210	1178
18	21291	648

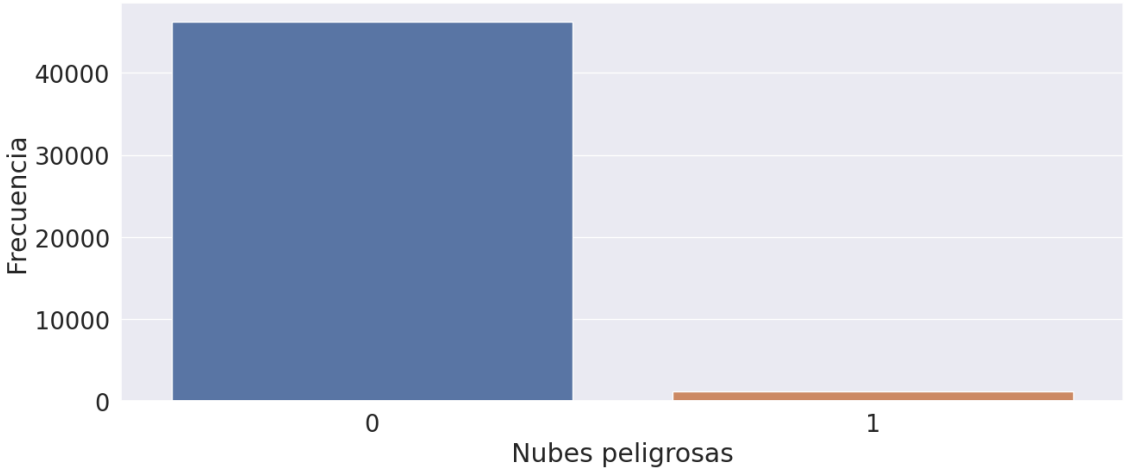

Figura 36: Distribución de nubes peligrosas en la pista 32

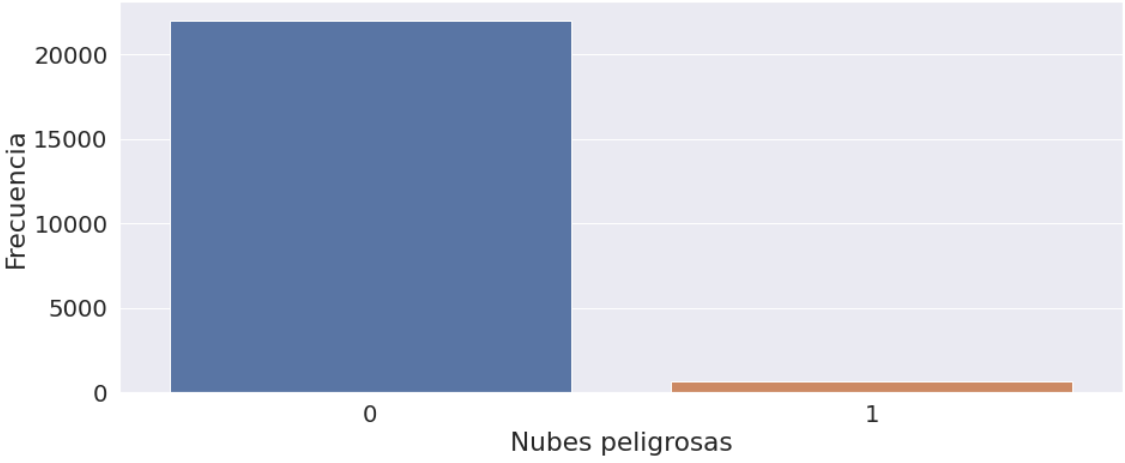

Figura 37: Distribución de nubes peligrosas en la pista 18

2.2.2.12 Lluvia

La lluvia suele medirse en milímetros o pulgadas, y su intensidad puede afectar a la visibilidad, así como al rendimiento general de una aeronave. Las precipitaciones intensas pueden provocar la acumulación de agua en las pistas, haciéndolas resbaladizas y peligrosas para el despegue y el aterrizaje. Además, las precipitaciones intensas también pueden reducir la visibilidad, provocando condiciones de baja visibilidad que pueden afectar al comportamiento de las aeronaves.

Durante su estudio, no se encontraron datos ausentes ni datos atípicos. Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Los números utilizados en este contexto se refieren a los siguientes valores relacionados con las condiciones de lluvia:

- No presencia de lluvias: 0
- Lluvia moderada: 1
- Lluvia media: 2
- Lluvia intensa: 3

Tabla 22: Distribución de observaciones de lluvia por pista

Pistas	***0***	***1***	***2***	***3***
32	47305	23	60	0
18	21865	0	74	0

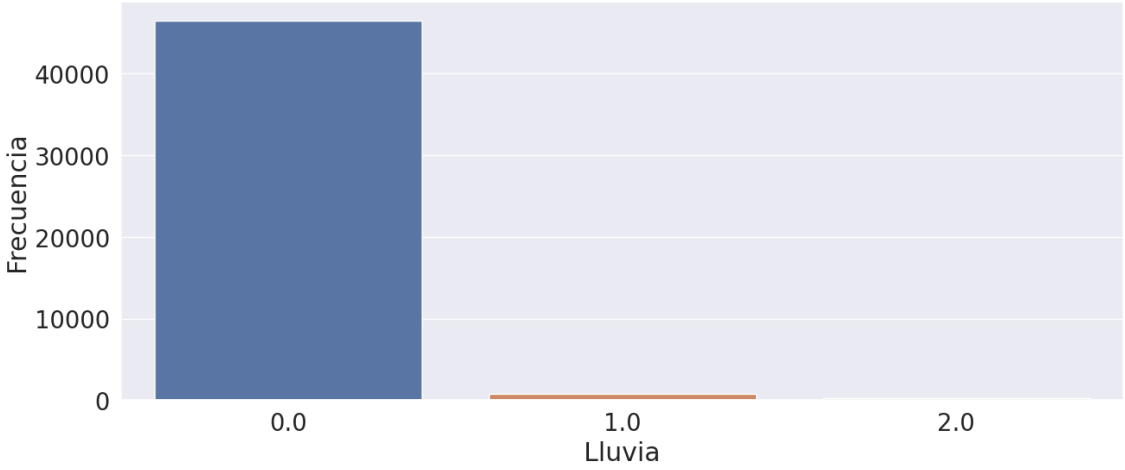

Figura 38: Distribución de presencia de lluvia en la pista 32

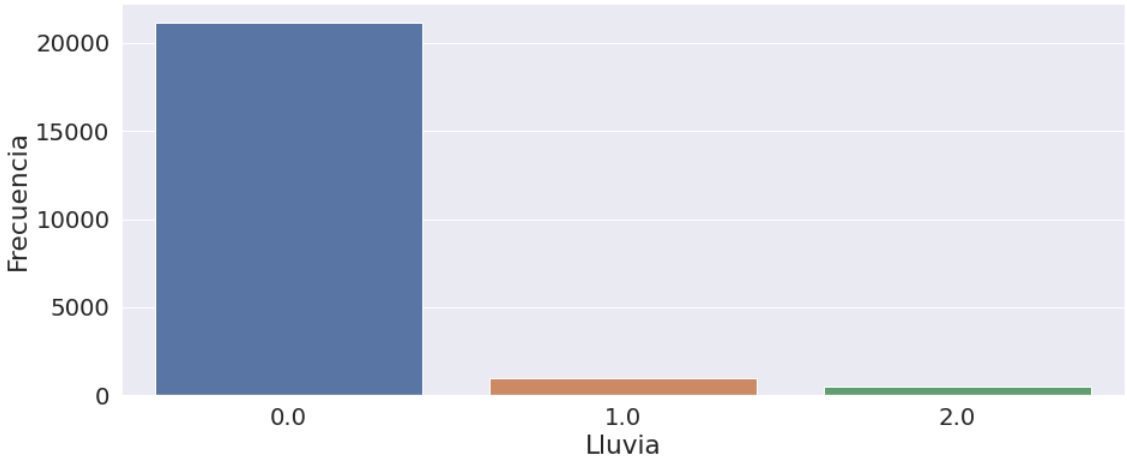

Figura 39: Distribución de presencia de lluvia en la pista 18

2.2.2.13 Niebla

La niebla es otra variable crítica que puede afectar a la seguridad de las aeronaves. La niebla se mide normalmente en metros o pies y está causada por pequeñas gotas de agua suspendidas en el aire. La niebla puede dar lugar a condiciones de baja visibilidad, lo que dificulta la visión y la navegación de los pilotos y, en última instancia, puede afectar al comportamiento de las aeronaves y a la distancia entre ellas.

Durante su estudio, no se encontraron datos ausentes ni datos atípicos. Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Los números utilizados en este contexto se refieren a los siguientes valores relacionados con las condiciones de niebla:

- No presencia de niebla: 0
- Niebla moderada: 1
- Niebla media: 2
- Niebla intensa: 3

Tabla 23: Distribución de observaciones de la niebla por pista

Pistas	***0***	***2***
32	46867	521
18	21883	56

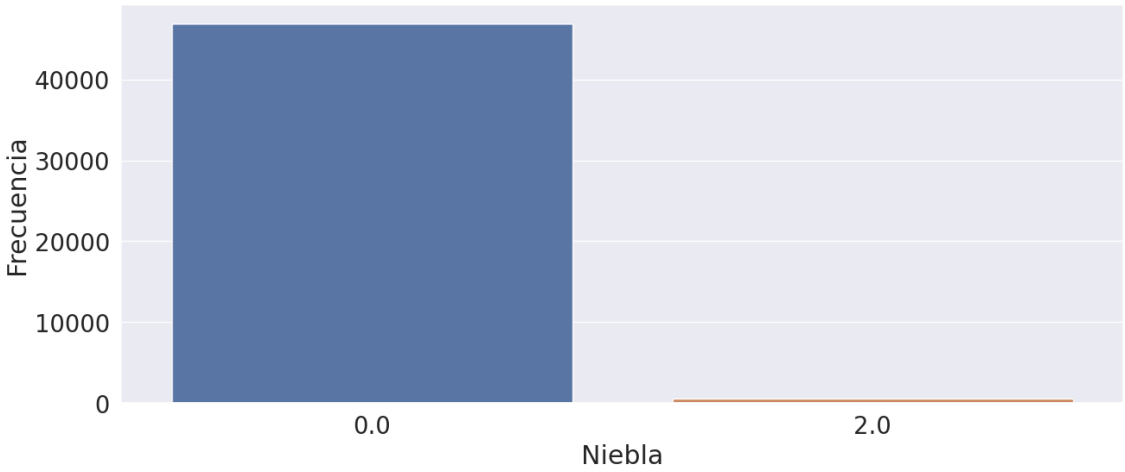

Figura 40: Distribución de niebla en la pista 32

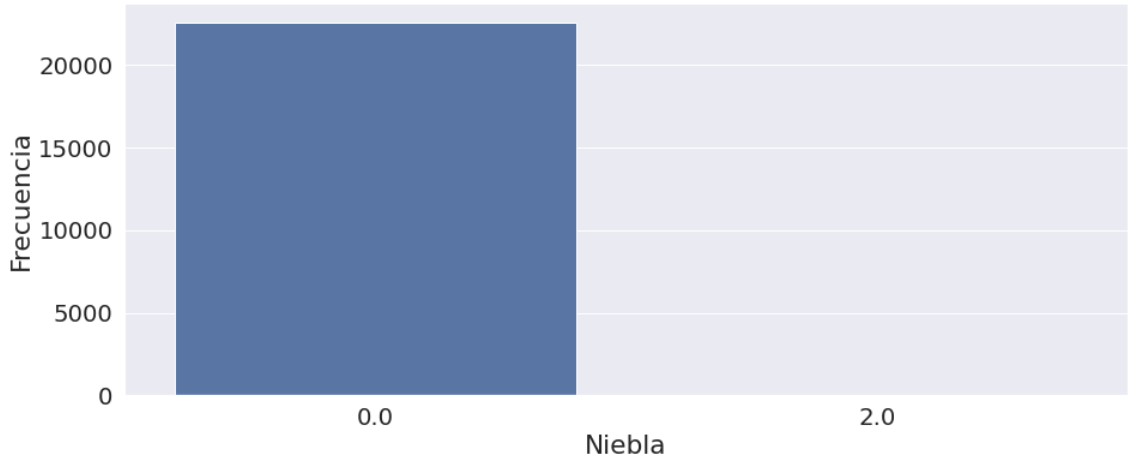

Figura 41: Distribución de niebla en la pista 18

2.2.2.14 Tormentas

Una tormenta es un fenómeno meteorológico que suele implicar fuertes vientos, precipitaciones intensas y relámpagos, y que también puede incluir truenos, granizo y tornados. Las tormentas pueden formarse rápidamente, lo que dificulta a los pilotos la navegación segura a través de ellas. Además, pueden crear condiciones de vuelo peligrosas, como turbulencias, cizalladura del viento y formación de hielo, que pueden afectar significativamente al rendimiento y la seguridad de las aeronaves.

Durante su estudio, no se encontraron datos ausentes ni datos atípicos. Se encontraba en un formato correcto, por lo que no fue necesario ningún procedimiento adicional.

Los números utilizados en este contexto se refieren a los siguientes valores relacionados con las condiciones de tormentas:

- No presencia de tormenta: 0
- Tormenta moderada: 1
- Tormeta media: 2
- Tormenta intensa: 3

Tabla 24: Distribución de valores de tormentas por pista

Pistas	***0***	***1***	***2***
32	47305	23	60
18	21865	0	74

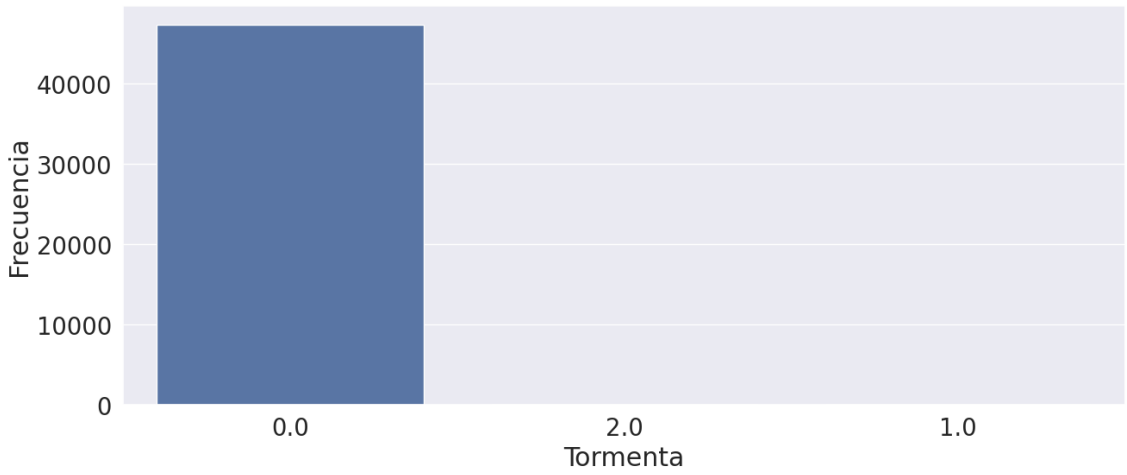

Figura 42: Distribución de presencia de tormentas en la pista 32

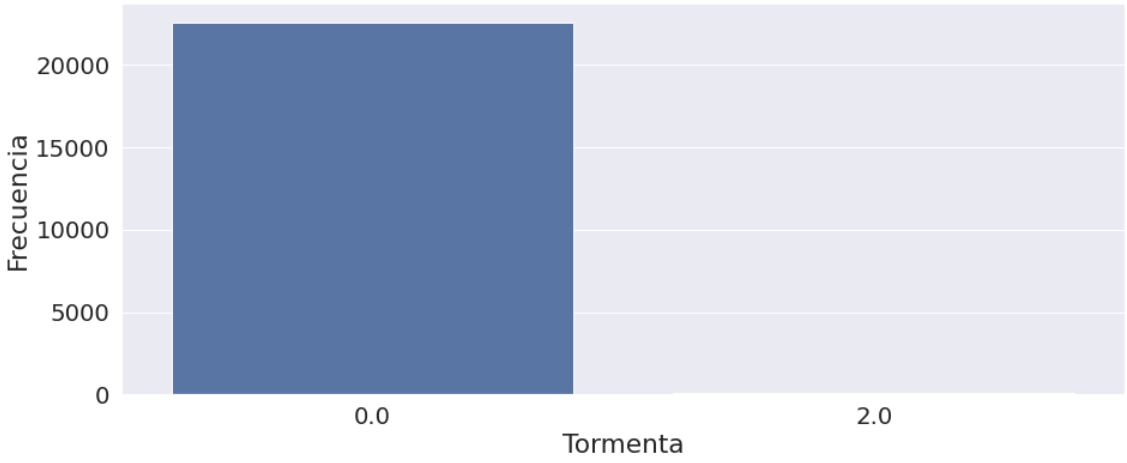

Figura 43: Distribución de presencia de tormentas en la pista 18

2.2.3 Separación de umbral

La separación de umbral, que representa la distancia entre dos aeronaves durante el aterrizaje, es el objetivo central de este estudio. Su predicción desempeña un papel fundamental en el mantenimiento de la seguridad en el aeropuerto de Barajas, y también tiene implicaciones importantes para la capacidad operativa del mismo. Al predecir la distancia necesaria entre las aeronaves, podemos optimizar la capacidad del aeropuerto en términos de aterrizajes, lo que a su vez impacta directamente en los despegues y mejora el flujo de tráfico general en el aeropuerto.

Inicialmente, la separación de umbral es una variable continua que puede variar en un rango de 1.36 a 19.2 millas náuticas. Sin embargo, debido a la naturaleza variable de los registros de separación de umbral reportados por los controladores aéreos, surge la necesidad de establecer un enfoque de clasificación para abordar este desafío. Actualmente, no existe un procedimiento o estándar establecido que regule las decisiones de los controladores aéreos en cuanto a la separación de umbral.

En este contexto, se ha tomado la decisión de enfocar este proyecto de investigación en la clasificación de la variable de separación de umbral a través de su discretización en rangos específicos. El objetivo de esta discretización es lograr una mayor precisión en las predicciones y recomendar un rango de separación de umbral que pueda mejorar el proceso de toma de decisiones de los controladores aéreos. Este enfoque ha sido validado con la ayuda de expertos en el campo de la aviación, asegurando la fiabilidad y relevancia de los resultados obtenidos. La discretización fue la siguiente.

Tabla 25: Categorización de la separación de umbral

Distancia (millas náuticas)	***Categoría***
0 – 3	Distancia límite (DL)
3 – 5	Distancia baja (DB)
5 – 8	Distancia media (DM)
8 – 15	Distancia alta (DA)
+15	Indiferente (DI)

Tabla 26: Distribución de la separación de umbral por pista

Pistas	***DM***	***DB***	***DA***	***DL***	***DI***
32	26678	14662	5680	331	37
18	11504	5563	4432	297	143

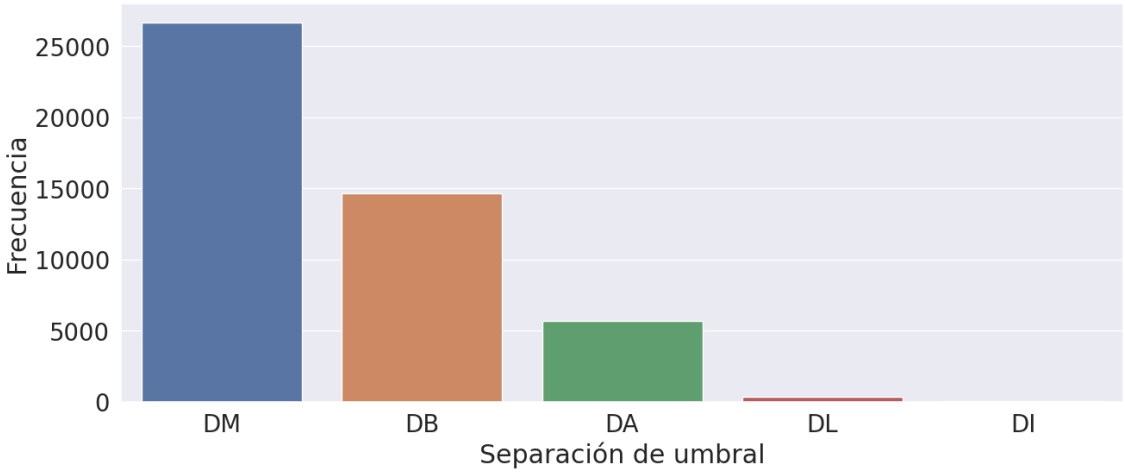

Figura 44: Distribución de separación de umbral en la pista 32

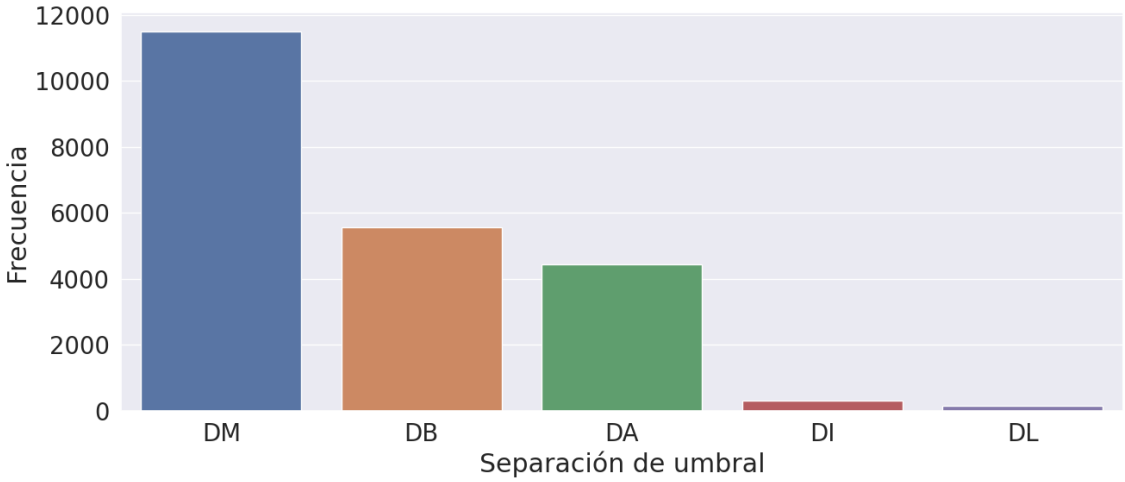

Figura 45: Distribución de separación de umbral en la pista 18

2.2.4 Vuelos en los siguientes 15 minutos

Durante el análisis exhaustivo de las variables previamente mencionadas, se identificó la necesidad de incorporar una nueva variable que complementara la información existente. Se observó que la distancia diagonal compartía una considerable cantidad de información con la velocidad de la aeronave. A través de una investigación preliminar, se determinó que sería beneficioso considerar el número de vuelos que se encontraban próximos a llegar en los 15 minutos siguientes al vuelo en cuestión. Esta variable adicional permite evaluar la densidad del tráfico aéreo y la eficiencia en términos de tiempo, proporcionando así una base sólida para la toma de decisiones óptimas.

Con el objetivo de capturar esta información relevante, se creó la variable "***flights_in_next_15mins***" a partir de la variable "***date***" en el conjunto de datos original. Esta representa el momento en que se registraron los datos de cada vuelo, y cada vuelo se identifica mediante un ID único (***Flight Key***).

Para llevar a cabo esta transformación, se convirtió la variable al formato ***DATETIME*** y se procedió a ordenar los registros de forma descendente, asegurando así un orden temporal coherente.

Utilizando la librería ***timedelta***, la cual ofrece funcionalidades para operaciones aritméticas con fechas y horas, se añadieron 15 minutos a la hora de cada vuelo registrado. Esto permitió obtener la variable "***flights_in_next_15mins***", la cual indica la cantidad estimada de aviones que se espera que aterricen en los próximos 15 minutos para cada vuelo considerado en el estudio.

La incorporación de esta nueva variable brinda una perspectiva adicional al modelo, al permitirle evaluar la situación del tráfico aéreo en un horizonte temporal cercano. Al considerar el número de vuelos que se aproximan, se logra capturar las condiciones actuales del aeropuerto y su potencial impacto en la separación de umbral entre las aeronaves. Esta información enriquece el conjunto de características y contribuye a una predicción más precisa y contextualizada.

- Los valores van desde 0 a 15 para ambas pistas.

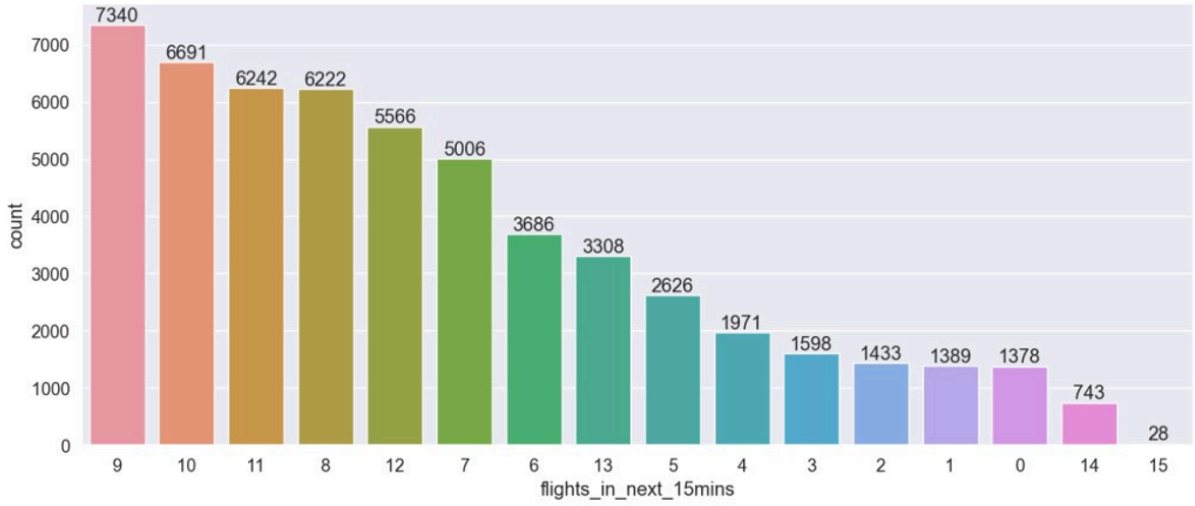

Figura 46: Distribución del número de vuelos en la pista 32

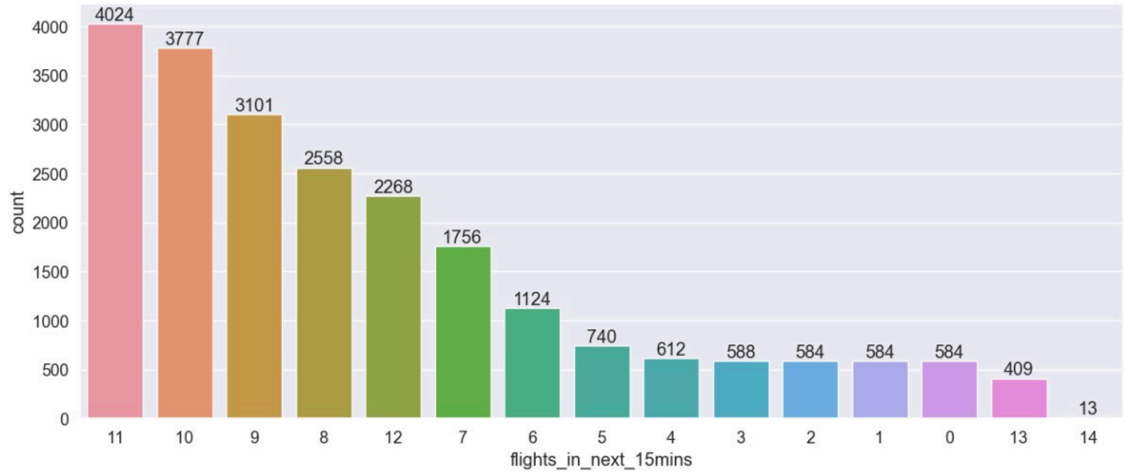

Figura 47: Distribución del número de vuelos en la pista 18

Capítulo 3

Selección de variables

Una vez se obtuvo el conjunto de datos totalmente analizado en cuanto a distribuciones, medidas estadísticas y valores, se procedió a analizar el comportamiento de cada una de las variables, así como su relación conjunta con otras variables predictoras y la variable objetivo. El objetivo era seleccionar las variables más importantes para este estudio. Para lograrlo, se implementó un plan de ejecución que incluyó la discretización de las variables. Se procedió a discretizar únicamente aquellas variables que requerían este proceso. Esto permitió analizar su comportamiento a través de diferentes métodos y técnicas, con el fin de identificar las variables de mayor impacto y que contribuyeran al objetivo de predecir la separación de umbral entre dos aeronaves

3.1 Discretización de variables: Métodos y técnicas

3.1.1 Métodos de discretización basados en intervalos

El método de discretización basado en intervalos es una técnica ampliamente utilizada en el ámbito de la ciencia de datos para transformar variables numéricas continuas en variables categóricas discretas mediante la creación de intervalos. Su objetivo principal es proporcionar una aproximación inicial del comportamiento de los datos. Este enfoque permite reducir la complejidad de los datos y obtener una visión general de la distribución y los patrones subyacentes de las variables, lo cual resulta útil en problemas de clasificación, como es el caso del estudio presente [8].

El proceso de discretización se lleva a cabo mediante la selección de diferentes rangos de valores para una variable en particular. A continuación, se generan puntos de corte en los bordes de estos rangos para dividir la variable en intervalos. Cada intervalo representa una categoría discreta y se contabiliza el número de observaciones que caen dentro de cada intervalo.

Es importante mencionar que existen enfoques específicos para la selección de rangos y puntos de corte en el método de discretización. En el caso de variables que siguen una distribución normal, se recomienda generar rangos con puntos de corte en los quartiles de la distribución. Estos quartiles dividen los datos en cuatro partes iguales, lo que proporciona una división equitativa de la variable.

Al utilizar el método de discretización basado en intervalos, se obtiene una representación discreta de las variables continuas, lo que facilita su uso en algoritmos de clasificación [9]. Sin embargo, es importante tener en cuenta que este proceso implica ciertas decisiones subjetivas, como la selección de los rangos y puntos de corte, que pueden afectar los resultados finales. Por lo tanto, es fundamental realizar un análisis exhaustivo de los datos y evaluar el impacto de la discretización en el rendimiento del algoritmo de aprendizaje automático utilizado.

Además, es importante destacar que este enfoque de discretización basado en intervalos no es el único disponible. Existen otros métodos, como el de discretización basado en frecuencia o el de discretización basado en clustering,

que pueden adaptarse mejor a determinados conjuntos de datos o problemas específicos.

3.1.2 Métodos de discretización basados en clústeres

El método de discretización basado en clústeres es una técnica utilizada para agrupar variables continuas en diferentes rangos, clases o categorías según la similitud que comparten los valores de los datos. Este enfoque resulta especialmente útil cuando los datos presentan distribuciones o estructuras de agrupación, o cuando se requiere una discretización precisa que siga el comportamiento de los valores contenidos en la variable y sea adaptativa a los patrones presentes en los datos [10].

Para lograr esto, se emplean diversos algoritmos de agrupación o clusterización, siendo uno de los más conocidos el algoritmo k-means. Este algoritmo se utiliza para dividir los valores de una variable continua en k grupos, donde k representa el número deseado de clústeres [11]. Sin embargo, antes de aplicar el algoritmo k-means, es necesario determinar el valor óptimo de k. Para ello, se pueden emplear técnicas conocidas como el método del codo o el silhouette score, las cuales permiten analizar diferentes valores de k y seleccionar el óptimo en base a criterios estadísticos.

Una vez obtenido el valor óptimo de k, se procede a aplicar el algoritmo k-means para realizar la agrupación de los datos. El objetivo es minimizar la distancia entre los puntos de datos y el centroide de cada clúster, de manera que se asignen los valores de la variable a los clústeres correspondientes. Una vez finalizado este proceso, cada centroide de la clusterización realizada representa un punto de corte para la discretización. De esta manera, se logra una mayor precisión en los puntos de corte y se identifican de mejor manera los rangos para la discretización.

Es importante tener en cuenta que la selección adecuada de k y la elección del algoritmo de agrupación más apropiado dependen del conjunto de datos y del problema específico. Además, es recomendable evaluar la calidad de la discretización obtenida y su impacto en la tarea de clasificación. Al igual que en otros métodos de discretización, es posible que exista una pérdida de información durante este proceso, por lo que es fundamental analizar los resultados y considerar su idoneidad para el problema en cuestión.

En resumen, el método de discretización basado en clústeres es una técnica que utiliza algoritmos de agrupación, como k-means, para dividir variables continuas en categorías o rangos basados en la similitud de los valores de los datos. Este enfoque permite una discretización precisa y adaptativa a los patrones presentes en los datos, mejorando la identificación de los puntos de corte. Sin embargo, es importante seleccionar adecuadamente el valor de k y evaluar la calidad de la discretización obtenida.

3.2 Selección de variables: Métodos y técnicas

3.2.1 Información condicional mutua

La información condicional mutua es una medida utilizada en el campo de la teoría de la información para cuantificar la dependencia entre dos variables aleatorias. Proporciona una medida de la cantidad de información que una variable proporciona sobre otra variable, dada cierta información adicional [12].

En el contexto de la ciencia de datos, la información condicional mutua es una herramienta valiosa para comprender las relaciones entre variables y su relevancia para un problema dado. Permite evaluar la dependencia entre variables y puede ser útil en tareas como selección de características, reducción de dimensionalidad y modelado de relaciones condicionales.

La información condicional mutua se define matemáticamente como la diferencia entre la entropía de una variable antes de conocer otra variable y la entropía de la variable después de conocer la segunda variable. Es decir, mide cuánta información se ha ganado sobre la primera variable al conocer la segunda [13].

Para calcular la información condicional mutua, se utiliza la entropía de las variables involucradas y la entropía conjunta. La entropía es una medida de incertidumbre y se calcula a partir de la distribución de probabilidad de una variable aleatoria.

En el caso de dos variables discretas, la fórmula para la información condicional mutua se define como la suma de las probabilidades conjuntas de las variables multiplicadas por el logaritmo del cociente de las probabilidades condicionales. Esta fórmula captura la dependencia entre las variables y proporciona una medida cuantitativa de la información compartida. Para su aplicación se utilizará la función ***mutinformation*** que pertence a la librería ***infotheo*** de R.

3.2.2 Ganancia de información

La ganancia de información es un concepto importante en la teoría de la información que se utiliza para medir la reducción de incertidumbre que se obtiene al conocer el valor de una variable. En el contexto de la ciencia de datos, la ganancia de información es una métrica útil para evaluar la relevancia de las características en un problema de clasificación.

La ganancia de información se basa en el principio de la entropía, que es una medida de la incertidumbre en una variable aleatoria. Cuanto mayor sea la incertidumbre, mayor será la entropía. Al calcular la ganancia de información, se evalúa cómo la información proporcionada por una característica específica reduce la incertidumbre en la variable objetivo [14].

El cálculo de la ganancia de información implica medir la diferencia entre la entropía de la variable objetivo antes y después de conocer el valor de una característica en particular. Cuanto mayor sea la diferencia, mayor será la ganancia de información y más relevante se considerará la característica para la clasificación.

En el caso de un problema de clasificación con variables discretas, la ganancia de información se puede calcular utilizando la fórmula de la entropía y las

probabilidades condicionales. Se evalúa la reducción en la entropía de la variable objetivo al conocer los diferentes valores de la característica. Para su aplicación se utilizará la función ***info.gain*** que pertence a la librería ***rpart*** de R.

3.3 Proceso de discretización y selección de variables

Una vez establecidos los métodos y técnicas a utilizar en este capítulo, se procedió al proceso de discretización de las variables. Esta elección se basó en la necesidad de obtener una visión inicial de cómo las variables interactúan, tanto de forma individual como en conjunto con otras variables del conjunto de datos. Inicialmente, las variables presentaban diferentes atributos: nominales, ordinales, continuas y discretas, lo que dificultaba su análisis y limitaba las opciones de algoritmos y técnicas aplicables. Aunque la discretización puede ocasionar pérdida de información al representar los datos en intervalos o categorías, lo cual podría afectar la precisión de los modelos y análisis subsiguientes, consideramos que era una etapa importante para explorar y comprender el comportamiento de los datos en este dominio [15].

El proceso de discretización de las variables en el conjunto de datos se llevó a cabo siguiendo una metodología coherente y consistente basada en el tipo de cada variable. El primer paso consistió en identificar correctamente los tipos de variables a partir de su descripción y registros, como se muestra en la Tabla 27.

Tabla 27: Tipos de variables

Categóricas	Ordinales	Estela de la aeronave
		Variabilidad de la intensidad
		Nubosidad baja
		Lluvia
		Niebla
		Tormentas
	Nominales	Dirección del viento
Numéricas	Continuas	Distancia diagonal
		Velocidad
		Visibilidad
		Intensidad del viento
	Discretas	Altura
		Nubes peligrosas
		CAVOK

- Las variables ordinales se identificaron dos situaciones: unas variables categóricas sin un orden fijo y otras variables discretas que tampoco seguían un orden. Para el primer caso, se utilizó la función ***factor***, que permite crear una variable categórica ordenada a partir de un vector numérico o de caracteres, y se empleó la función ***fct_relevel*** de la librería ***forcats*** para establecer un orden específico para las variables ordinales. Finalmente, se utilizó la función ***as.numeric***, función base de R, para representar los valores de la columna en formato numérico. Para el segundo caso, se estableció el orden para cada valor de las variables ordinales mediante la función ***ordered*** de R.
- Las variables nominales, no necesitaron discretización.
- Las variables continuas se discretizaron considerando una estrategia distinta dependiendo de su distribución. Si una variable mostraba una distribución normal, como muestra la Figura 48, se procedió a discretizarla en 5 intervalos utilizando los cuartiles como puntos de corte. Para esto, se utilizó la función cut de la librería base de R, que permite generar los 5 bins con un número igual de observaciones en cada uno.

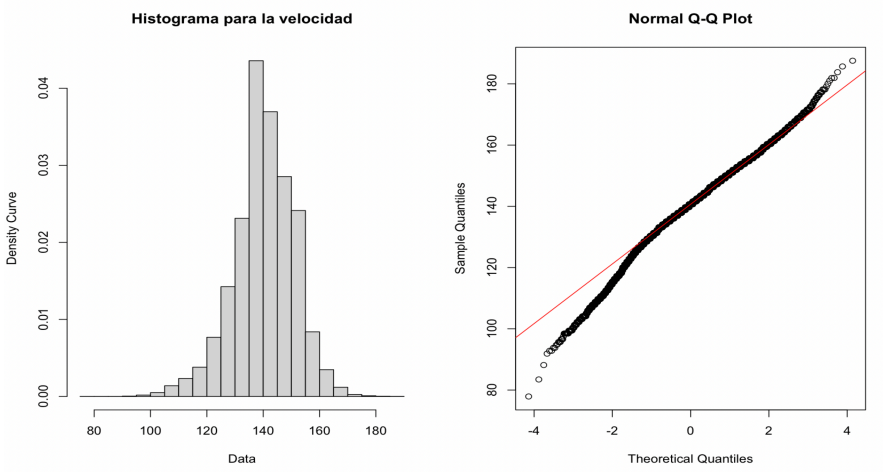

Figura 48: Distribución normal de variables

Si, después de aplicar una transformación logarítmica, la variable lograba la simetría, se continuó con la discretización utilizando los cuartiles como puntos de corte. Sin embargo, si la asimetría persistia incluso después de la transformación logarítmica, se optó por una discretización basada en técnicas de clustering. Para determinar el valor óptimo de k en los algoritmos de clustering, se utilizaron métodos como el método del codo (***elbow method***) y el coeficiente de silueta (***silhouette score***). Una vez identificado el mejor valor de k, se realizó el clustering y se utilizaron los centroides identificados como puntos de corte en la discretización, como se muestra en la Figura 49.

Método del codo ***Coeficiente de silueta***

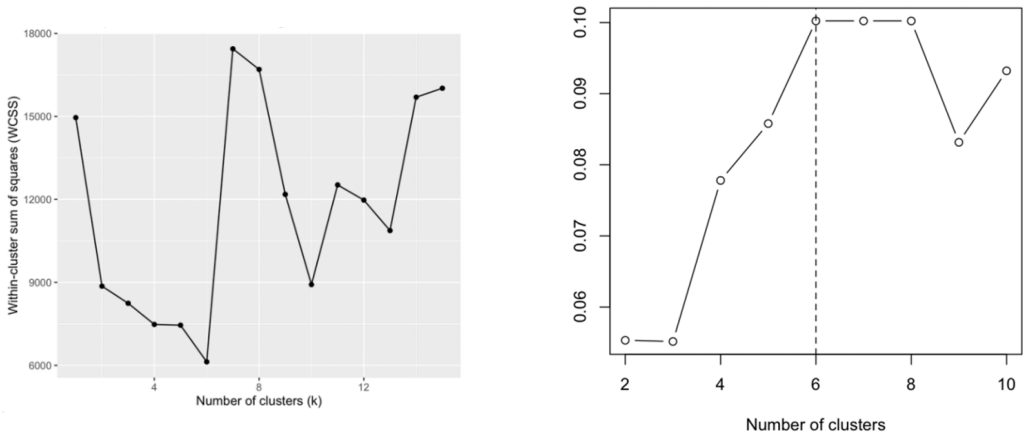

Figura 49: Métodos de evaluación para el valor de K

El análisis de clustering desempeñó un papel crucial en nuestro proyecto, ya que nos permitió comprender la estructura subyacente de las variables predictoras y su relación con la variable objetivo. Para validar la idoneidad del método, consideramos la opinión de expertos y utilizamos técnicas como el "silhouette score" para determinar el número óptimo de clusters. Sorprendentemente, los resultados de estas evaluaciones convergieron en un rango similar, indicando que un número de clusters de 5 o 6 podría ser adecuado para nuestro análisis.

Dado que la visualización es una parte esencial del proceso de clustering, inicialmente consideramos utilizar la técnica análisis de componentes principales (***PCA)*** para representar los resultados del clustering. Sin embargo, después de una cuidadosa reflexión, decidimos adoptar un enfoque alternativo. Evaluaremos cada clúster individualmente mediante diagramas de caja múltiples, que nos brindarán una representación visual más detallada de las características y distribuciones de las variables en cada grupo como se myestra en la Figura 50.

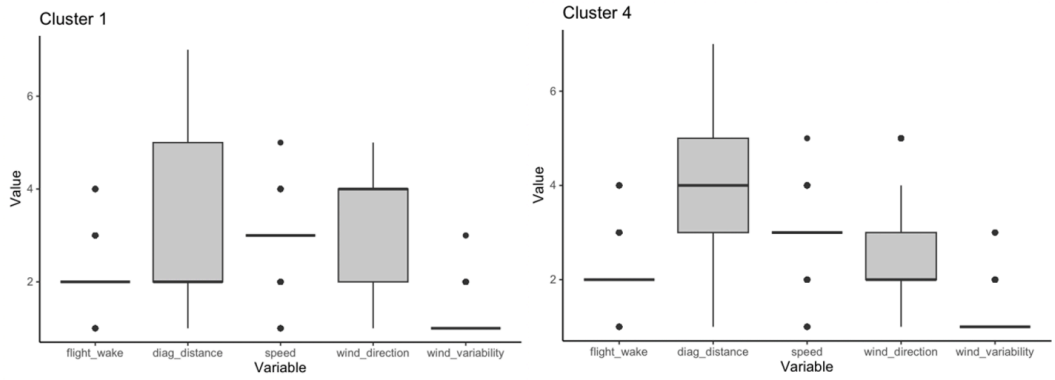

Figura 50: Representación del clustering en diagramas de cajas

Estos puntos de corte fueron validados tanto en términos de su influencia en las métricas de clasificación como a través de la consulta a expertos en el campo. Se siguió una metodología sistemática y rigurosa adaptada a cada tipo de variable presente en el conjunto de datos. Esta metodología garantizó una discretización coherente y bien fundamentada, contribuyendo a la confiabilidad y utilidad de los resultados obtenidos en el análisis de datos y en la toma de decisiones subsiguientes.

Una vez, se obtuvieron todas las variables discretizadas, se procedió a realizar una la selección de variables. Procedemos a evaluar la importancia y relevancia de cada una de las variables predictoras en el conjunto de datos. Para esto utilizaremos la técnica de información mutua condicional (ICM), la cual es adecuada para datos discretos y cuantifica la cantidad de información compartida entre dos variables teniendo en cuenta el conocimiento que tenemos sobre la variable objetivo.

La técnica ICM se utiliza para analizar las dependencias y relaciones entre las variables predictoras, con el objetivo de evaluar el proceso de selección de variables y asegurar que tenga un impacto positivo en las métricas de clasificación posteriores. Aplicaremos esta técnica a las variables predictoras y la variable objetivo. Para esto, se utilizó la función ***mutinformation*** que pertence a la librería ***infotheo***, la cual aplica diversas medidas de la teoría de la información basadas en varios estimadores de entropía. De esta forma, se obtuvieron los siguientes resultados para la diferentes pistas, Véase la Figura 51.

Pista 32

variable1	variable2	cmi_score
flights_in_next_15mins	diag_distance	0.163581175
diag_distance	flights_in_next_15mins	0.163581175
flight_wake	diag_distance	0.106940430
diag_distance	flight_wake	0.106940430
speed	diag_distance	0.105381277
diag_distance	speed	0.105381277
diag_distance	visibility	0.095630802
visibility	diag_distance	0.095630802
wind_intensity	diag_distance	0.094337846
diag_distance	wind_intensity	0.094337846
wind_direction	diag_distance	0.093451170

Pista 18

variable1	variable2	cmi_score
diag_distance	flights_in_next_15mins	0.202636144
flights_in_next_15mins	diag_distance	0.202636144
speed	diag_distance	0.127589874
diag_distance	speed	0.127589874
diag_distance	flight_wake	0.117655067
flight_wake	diag_distance	0.117655067
diag_distance	altitude	0.115831422
altitude	diag_distance	0.115831422
diag_distance	wind_intensity	0.115291649
wind_intensity	diag_distance	0.115291649
diag_distance	wind_direction	0.114801016

Figura 51: Información condicional mutua para ambas pistas

Asimismo, para complementar la información o los resultados brindados por la métrica de ***CMI***, se aplica la métrica de ganancia de información (***IG***). La técnica de ganancia de información cuantifica la reducción de incertidumbre al dividir los datos en función de cada variable predictora con respecto a la variable objetivo. Esto nos permite calcular el poder predictivo de cada variable en relación a la variable objetivo.

Al calcular la ganancia de información para todas las variables predictoras para la pista 32 y 18, obtuvimos los siguientes resultados, ordenados de mayor a menor, como se muestra en la Figura 52.

Pista 32

	variable	ig_score
1	diag_distance	8.567030e-03
2	speed	7.877737e-03
3	flight_wake	2.033744e-03
4	flights_in_next_15mins	1.365411e-03
5	wind_intensity	8.870255e-04
6	wind_direction	5.530192e-04
7	altitude	4.962280e-04
8	visibility	2.924993e-04
9	fog	2.916988e-04
10	shear	1.622641e-04
11	wind_variable	1.217558e-04

Pista 18

	variable	ig_score
1	speed	6.749030e-03
2	diag_distance	5.992132e-03
3	flights_in_next_15mins	1.675005e-03
4	flight_wake	1.517342e-03
5	wind_intensity	9.488692e-04
6	wind_direction	6.590189e-04
7	altitude	3.161209e-04
8	cloudiness_low	2.652503e-04
9	CAVOK	2.032881e-04
10	visibility	1.191043e-04
11	wind_variable	1.144883e-04

Figura 52: Ganancia de información entre variables predictoras

Estos valores reflejan la utilidad de cada variable para predecir la variable objetivo. Estos hallazgos nos ayudarán a tomar decisiones informadas en la selección final de variables para nuestro modelo de predicción del umbral de separación entre aeronaves.

3.4 Análisis de resultados y generación de conjuntos de datos

Analizar los resultados de la selección de variables es crucial para comprender qué variables son más relevantes y útiles en el modelado de la separación de umbral entre aeronaves. Para este propósito, se llevará a cabo un análisis individual para cada pista, tanto considerando el uso de la variable resumida "CAVOK" como el uso de las variables independientes relacionadas con los fenómenos atmosféricos, nubes y visibilidad.

El objetivo de este análisis es determinar las variables que funcionan mejor en la predicción de la separación de umbral entre aeronaves, evitando la redundancia y modelando de manera precisa la forma en que los controladores aéreos realizan dicha predicción.

En el caso de la pista 32, en base a los resultados de la métrica CMI, la variable ***flights_in_next_15mins*** muestra una dependencia significativa con la variable objetivo, ya que tiene el mayor valor de puntuación CMI (0.16358). Esto indica que la cantidad de vuelos en los próximos 15 minutos puede ser un factor importante en la separación de umbral en la pista 32. De la misma forma, las variables distancia diagonal, velocidad y la estela son importantes para predicción de la separación de umbral, según los resultados de la métrica IG.

Se puede añadir que las variables ***flights_in_next_15mins*** y distancia diagonal tienen la misma puntuación CMI, lo que indica una fuerte dependencia mutua. La distancia diagnóstica entre vuelos y la cantidad de vuelos en los próximos 15 minutos parecen estar correlacionados y pueden tener un impacto conjunto en la separación de umbral en la pista 32.

Por otro lado, se menciona que las variables wind_variability y wind_variable obtuvieron puntuaciones insignficantes por lo que podrian ser posibles variables a eliminar en la pista 32.

Para el caso específico de la pista 18, la variable distancia diagonal tiene el mayor puntaje de CMI con las variables predictoras ***flights_in_next_15mins*** y velocidad. Esto indica una fuerte dependencia entre estas variables y la variable objetivo de separación de umbral. Mientras que, si nos enfocamos en las variables menos importantes al momento de predecir la separación de umbral en base a los resultados de los métodos IG, podemos asegurar que las variables wind_variable, fog, dangerous_clouds, storm, rain, wind_direction y cloudiness_low podrían considerarse para su eliminación, ya que muestran una dependencia más débil con la variable objetivo y podrían tener menos influencia en la predicción de la separación de umbral.

En base al análisis de los resultados de la pista 32 y 18, se descubrieron dos variables que no tenian impacto al momento de predecir la variable objetivo, son las siguienes: variabilidad de la intensidad del viento y nubes peligrosas.

Una vez se seleccionaron las variables a eliminar, se generan dos conjuntos de datos diferentes. Uno de ellos contendrá la variable "CAVOK" como resumen de las condiciones meteorológicas, véase la Figura 53.

	flight_wake	diag_distance	speed	altitude	CAVOK	wind_variable	wind_direction	wind_intensity	threshold_separation	flights_in_next_15mins
1	2	4	3	25	1	0	1	2	2	9
2	2	3	4	22	1	0	1	2	3	9
3	2	2	4	20	1	0	1	2	4	9
4	3	4	4	20	1	0	1	2	2	10
5	2	2	4	19	1	0	1	2	2	10
6	2	1	3	21	1	0	1	2	2	11
7	2	1	4	26	1	0	1	2	3	11

Figura 53: Conjuntos de datos con la variable CAVOK

Mientras que el otro conjunto incluirá las variables independientes relacionadas con los fenómenos atmosféricos, nubes y visibilidad, véase la Figura 54.

	flight_wake	diag_distance	speed	altitude	visibility	wind_variable	wind_direction	wind_intensity
1	2	4	3	25	6	0	1	2
2	2	3	4	22	6	0	1	2
3	2	2	4	20	6	0	1	2
4	3	4	4	20	6	0	1	2
5	2	2	4	19	6	0	1	2

cloudiness_low	rain	fog	storm	threshold_separation	flights_in_next_15mins
0	0	0	0	2	9
0	0	0	0	3	9
0	0	0	0	4	9
0	0	0	0	2	10
0	0	0	0	2	10

Figura 54: Conjunto de datos con las variables independientes

El análisis exhaustivo de cada pista y la comparación entre el uso de la variable resumida y las variables independientes permitirá identificar cuáles son las variables que ofrecen un mejor rendimiento y contribuyen de manera más significativa a la predicción precisa de la separación de umbral entre aeronaves.

Este análisis de resultados y la generación de los conjuntos de datos son etapas fundamentales para obtener una comprensión completa del problema y sentar las bases para el desarrollo de un modelo efectivo de predicción de la separación de umbral en entornos aéreos controlados.

Capítulo 4

Modelos de clasificación

Durante esta sección, teniendo en cuenta todo los procesos previos realizados desde la exploración descriptiva mediante distribuciones, técnicas como el clustering y la selección de variables, y finalmente terminando con la discretización de las variables, se obtuvo un conjunto de datos fiable que permitirá la evaluación de distintos modelos de clasificación, los cuales serán evaluados en diversas secciones.

4.1 Algoritmos de clasificación

Para la selección de algoritmos de clasificación, primero se tuvo que tener en cuenta el conjunto final de datos en cuanto a su número de observaciones y el tipo de cada variable, asi se procedieron a seleccionar diversos modelos de clasificación que, según sus referencias, descripcion y formulación era buenos en el manejo de variables discretas. Para lo cual, la metodología que se siguió es seleccionar modelos iniciales como ***naïve bayes*** y ***tree augmented naïve bayes (TAN)*** que permitirán evaluar el comportamiento de los algoritmos de clasificación y tener una idea inicial de como se desenvuelven o que clases les cuesta clasificar. Posteriormente, se probaron con modelos más complejos que traten de identificar o manejar todas las relaciones complejas que se manejan dentro del conjunto de datos, los cuales son el algoritmo de decision trees que permitio mejorar el estudio y que finalmente se utilizo el algoritmo de random forest, el cual tiene mayor complejidad y el cual permitió obtener los mejores resultados con respecto a las métricas de performance.

4.1.1 Naïve Bayes

El algoritmo Naïve Bayes es un popular clasificador probabilístico utilizado en problemas de clasificación, especialmente cuando todas las variables predictoras son de naturaleza discreta. Se fundamenta en el teorema de Bayes y parte del supuesto de independencia condicional entre las variables predictoras, dado el valor de la variable objetivo, como muestra la Figura 55. Este enfoque simplificado permite un cálculo eficiente de las probabilidades y un rendimiento satisfactorio en muchos casos de clasificación [16].

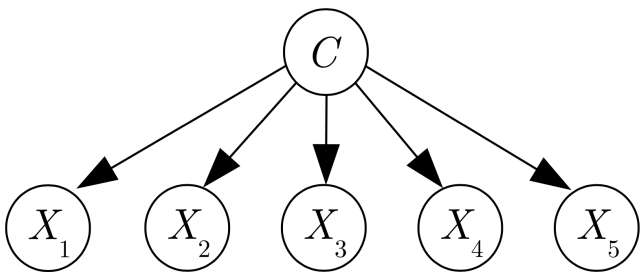

Figura 55: Estructura del algoritmo Naïve Bayes

Esta suposición simplificadora, conocida como "ingenuidad" o "naïve" en inglés, permite un cálculo más eficiente de las probabilidades condicionales y una mayor escalabilidad del algoritmo.

La formulación matemática del Naïve Bayes se basa en el teorema de Bayes:

$$P(C|X) = P(X|C) * P(C)/P(X) \quad (4.1)$$

Donde:

- $P(C|X)$ es la probabilidad condicional de la clase C dado el vector de características X.
- $P(X|C)$ es la probabilidad condicional del vector de características X dado la clase C.
- $P(C)$ es la probabilidad a priori de la clase C.
- $P(X)$ es la probabilidad marginal del vector de características X.

Para realizar la clasificación, Naïve Bayes calcula la probabilidad de pertenencia a cada clase para una instancia de entrada y asigna la clase con la probabilidad más alta. Esto se logra mediante la siguiente fórmula:

$$C = argmaxP(C) * \prod P(Xi|C) \quad (4.2)$$

Donde:

- C es la clase objetivo.
- Xi s el valor de la característica i en el vector de características X.

Una de las ventajas destacadas de Naïve Bayes en problemas de clasificación con variables discretas es su eficiencia computacional y su capacidad para lidiar con conjuntos de datos grandes. Dado que solo requiere estimar las probabilidades de las variables predictoras y la probabilidad a priori de cada clase, el tiempo de entrenamiento y predicción es generalmente rápido en comparación con otros algoritmos más complejos.

4.1.2 TAN

El algoritmo TAN, o ***Tree Augmented Naïve Bayes***, es una versión mejorada del algoritmo Naïve Bayes que permite capturar las relaciones de dependencia entre las variables predictoras. A diferencia del Naïve Bayes clásico, que asume independencia condicional entre las variables predictoras, el algoritmo TAN amplía esta capacidad al modelar las relaciones entre las variables. Esta mejora en la modelización de las relaciones entre las variables proporciona una mayor capacidad de representación y, en muchos casos, mejora el rendimiento predictivo en comparación con el Naïve Bayes tradicional [16].

En el algoritmo TAN, se construye un árbol de dependencia para modelar las relaciones entre las variables predictoras. Este árbol es una estructura acíclica dirigida que conecta las variables predictoras y la variable objetivo. A diferencia del Naïve Bayes, donde todas las variables predictoras son consideradas independientes entre sí dado el valor de la variable objetivo, en TAN, se establecen conexiones adicionales entre las variables predictoras basadas en su relación de dependencia condicional mutua. Véase la Figura 56.

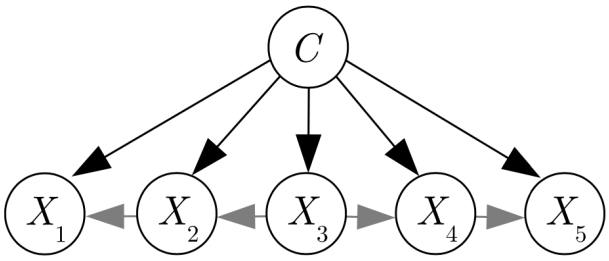

Figura 56: Estructura del algoritmo TAN

Una vez construido el árbol de dependencia, la clasificación se realiza utilizando la misma formulación matemática y enfoque probabilístico que Naïve Bayes.

El algoritmo TAN tiene varias ventajas sobre Naïve Bayes en problemas de clasificación. Al considerar las dependencias entre las variables predictoras, puede capturar relaciones más complejas y, por lo tanto, obtener un modelo más preciso. Además, TAN sigue siendo computacionalmente eficiente y escalable, ya que utiliza algoritmos de aprendizaje de estructuras de redes bayesianas que están diseñados para conjuntos de datos de gran tamaño.

En el contexto de nuestro problema de predicción de la distancia de umbral entre aeronaves, el uso de TAN resultó apropiado debido a su capacidad para modelar las relaciones de dependencia entre nuestras variables predictoras. Esto nos permitió considerar interacciones más complejas entre las características discretas, lo que podría ser relevante en la predicción precisa de la distancia de umbral. Al emplear TAN, esperamos obtener mejoras en la capacidad predictiva en comparación con el Naïve Bayes clásico.

4.1.3 Decision Trees

El método de aprendizaje automático conocido como árboles de decisión, o ***Decision Trees*** en inglés, es un algoritmo ampliamente empleado en problemas de clasificación y regresión. Su enfoque se basa en la construcción de un modelo en forma de árbol, en el cual cada nodo interno plantea una pregunta relacionada con una característica, y cada hoja representa una decisión o predicción. Esta técnica ha ganado popularidad debido a su capacidad para abordar diferentes tipos de problemas y su capacidad de interpretación [17].

La construcción de un árbol de decisión se realiza a través de un proceso de particionamiento recursivo. En cada nodo, se selecciona la característica más relevante para dividir el conjunto de datos en subconjuntos más puros o homogéneos en términos de la variable objetivo. Esta selección se basa en métricas como la ganancia de información o la impureza de Gini, que miden la capacidad predictiva de la característica. Véase la Figura 57.

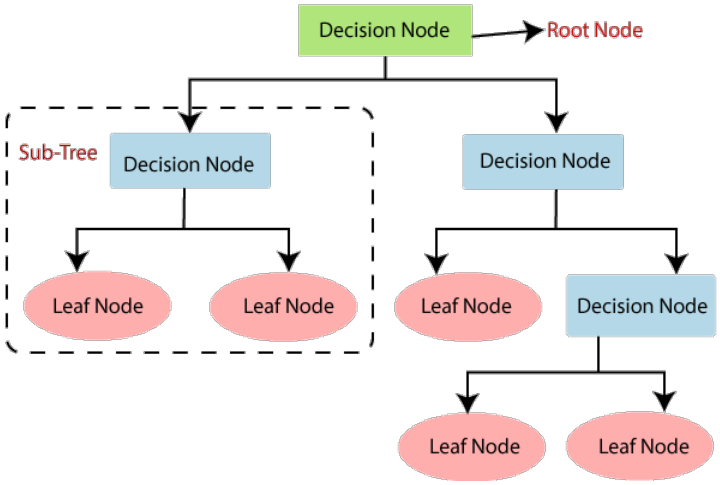

Figura 57: Algoritmo de árboles de decisiones

En el contexto de nuestro problema de predicción de la distancia de umbral entre aeronaves, el uso de Decision Trees resultó adecuado debido a la naturaleza de nuestras variables discretas y el deseo de evaluar relaciones más complejas en el conjunto de datos. Al emplear Decision Trees, pudimos explorar patrones y reglas más complejas que podrían estar presentes en la predicción de la distancia de umbral. La naturaleza discreta de los árboles de decisión y su capacidad de modelar relaciones no lineales fue una ventaja clave en nuestro enfoque.

4.1.4 Random Forest

El algoritmo conocido como Bosques Aleatorios, o ***Random Forest*** en inglés, es una técnica de aprendizaje automático que aprovecha la combinación de múltiples árboles de decisión para realizar predicciones. Este enfoque se destaca por su habilidad para abordar de manera efectiva problemas de clasificación y regresión, y su popularidad se debe a su rendimiento y versatilidad en diversas aplicaciones [18].

La construcción del algoritmo implica dos niveles de aleatoriedad: la selección de subconjuntos de datos y la selección de características para cada árbol individual. Estas técnicas de aleatorización ayudan a reducir el sobreajuste y mejorar la generalización del modelo. Asimismo, su predicción se realiza tomando la mayoría de votos (en el caso de clasificación) o promediando las predicciones (en el caso de regresión) de todos los árboles individuales. Esta agregación de múltiples predicciones reduce el sesgo y la varianza, mejorando así la precisión y robustez del modelo. Véase la Figura 58.

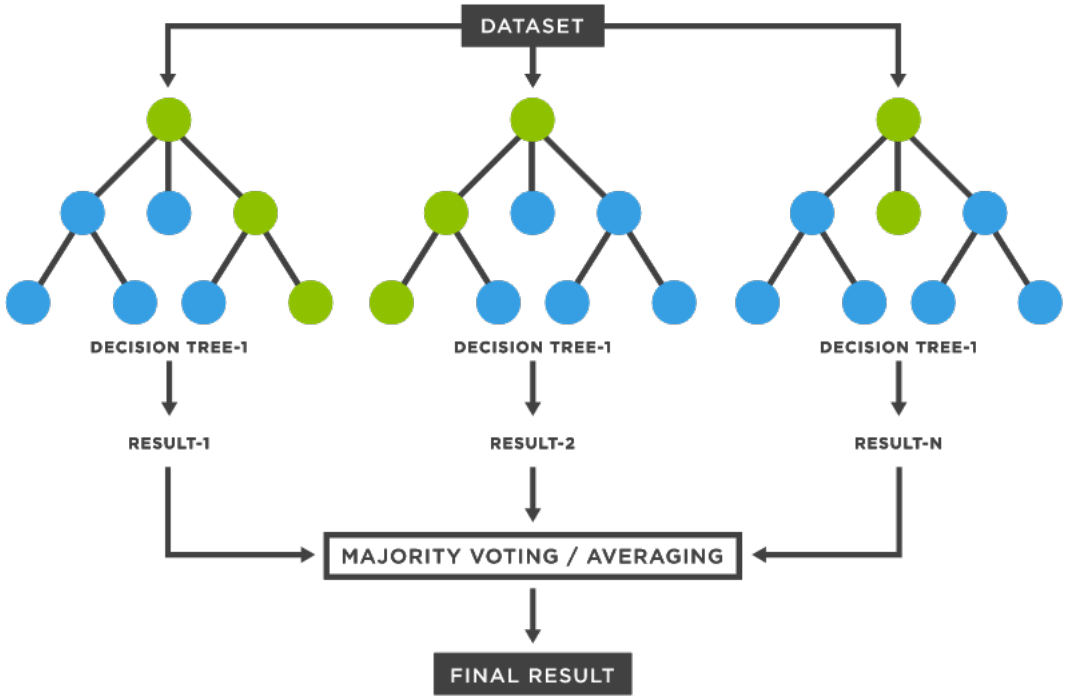

Figura 58: Algoritmo de bosques aleatorios

Una ventaja clave del algoritmo en problemas de clasificación con variables discretas es su capacidad para manejar relaciones complejas y no lineales. Al combinar múltiples árboles de decisión, el modelo puede capturar patrones más complejos y realizar predicciones más precisas.

Otra ventaja importante es su resistencia al sobreajuste. La aleatorización en la construcción del bosque y la agregación de múltiples árboles ayudan a reducir el sobreajuste y mejorar la capacidad de generalización del modelo. Esto es especialmente útil cuando se trabaja con conjuntos de datos grandes y variables discretas, donde el riesgo de sobreajuste puede ser mayor.

En el contexto de nuestro problema de predicción de la distancia de umbral entre aeronaves, el uso de Random Forest resultó beneficioso debido a su

capacidad para manejar relaciones complejas y no lineales en el conjunto de datos. Al emplear Random Forest, pudimos obtener predicciones más precisas y robustas en comparación con modelos más simples, como Decision Trees. Además, la capacidad de Random Forest para manejar variables discretas y lidiar con el riesgo de sobreajuste fue una ventaja clave en nuestro enfoque.

4.2 Evaluación y validación de modelos

4.2.1 Validación Hold-out

La técnica de ***hold out*** es una metodología comúnmente utilizada para evaluar modelos de clasificación cuando se dispone de conjuntos de datos de grandes dimensiones. Consiste en dividir el conjunto de datos en dos partes principales: un conjunto de entrenamiento y un conjunto de evaluación [19].

Considerando que el conjunto de datos se considera de grandes dimensiones, fue la primera opción como método de evaluación para los modelos de clasificación. Véase la Figura 59.

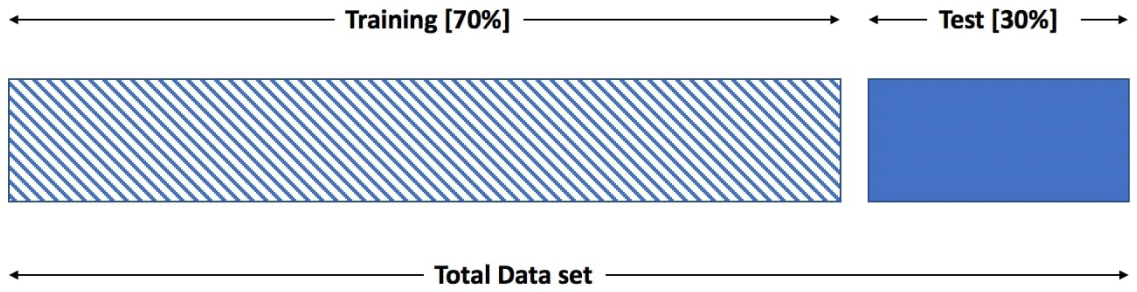

Figura 59: Técnica de validación Hold-out

Así, se propone tener una evaluación realista de los modelos de clasificación con las pruebas sobre datos nunca antes vistos y que proporcione las siguientes ventajas al proceso de evaluación:

- La eficiencia en cuanto a tiempos de evaluación de los modelos se considera una gran ventaja para el proyecto ya que se tiene que evaluar dos pistas de aterrizaje
- Esta técnica permitirá obtener el máximo impacto de la gran magnitud de datos con las que se cuenta ya que permitirá tener un conjunto de datos de entrenamiento mayor, lo cual terminará directamente por influir sobre la partición de evaluación.
- La flexibilidad será un tema interesante a evaluar para elegir el porcentaje correcto para cada partición que maximicen los resultados.

Para el estudio, finalmente se utilizó el 70 porciento para entrenamiento y el 30 porciento para la evaluación. Esta decisión esta validada en base a pruebas reiterativas sobre el modelo y teniendo en cuenta las observaciones sobre cada una de las particiones.

4.2.2 Nested cross-validation

La técnica de ***Nested cross-validation***, también conocida como validación cruzada anidada, es un enfoque más avanzado para evaluar modelos de clasificación, especialmente cuando se dispone de conjuntos de datos más pequeños o se busca obtener estimaciones más precisas del rendimiento del modelo [20].

La técnica fue utilizada con el objetivo de dividir nuestro conjunto de datos tanto para una partición de entrenamiento como de evaluación que nos permita primero entrenar al modelo, ajustar los hiperparámetros para lograr el mejor resultado posible para que posteriormente se puedan evaluar los resultados del entrenamiento, sobre un conjunto de datos nunca antes visto como el de evaluación. El porcentaje que finalmente se decidió para la division entre entrenamiento y evaluación es la misma. Cabe resaltar la importancia de la técnica nested cross-validation ya que al momento de dividir el conjunto de datos se aplicará la técnica de cross validation, lo cual permitirá dividir el conjunto de entrenamiento en subconjuntos que serán evaluados independimiento, como muestra la Figura 60.

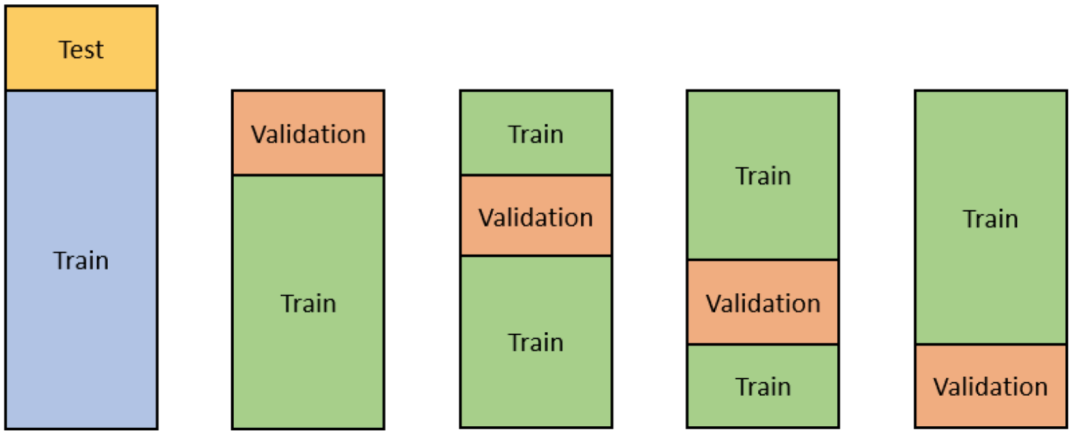

Figura 60: Técnica de validación nested cross-validation

Esto brindará todos los beneficios de la técnica cross validation a los modelos de clasificación como:

- Evaluación mas precisa del rendimiento del modelo al evaluarlo en múltiples particiones del conjunto de datos.
- Se maximizará la utilización de la gran de cantidad de datos que se poseen en el presente proyecto
- Ayudará en el proceso de detección y control del sobreajuste del modelo.
- Se logrará una mejor selección de hiperparámetros en algoritmos de aprendizaje automático al evaluar el modelo en diferentes configuraciones de hiperparámetros.
- Al repetir el proceso de validación cruzada en varias particiones y promediar los resultados, se obtiene una medida mas robusta del rendimiento del modelo.

Finamente, en base a diferentes pruebas reiterativas y teniendo en cuenta el impacto sobre las métricas de performance, se decidió como valor de k=5 el cual representa el k fold cross validation.

4.2.3 Muestreo probabilístico no aleatorio

El muestreo no probabilístico es una estrategia empleada para seleccionar muestras de conjuntos de datos, que se diferencia del muestreo probabilístico al no seguir principios estadísticos de probabilidad. En contraste con este último, donde cada elemento tiene una probabilidad definida de ser seleccionado, el muestreo no probabilístico se basa en criterios subjetivos o de conveniencia para elegir los elementos de la muestra [21].

Durante el desarrollo de la investigación y con las técnicas previamente mencionadas, y teniendo en cuenta el objetivo del proyecto y al tener un problema de clasificación multiclase con clases no balanceadas se propuso el uso del muestreo probabilistico no aleatorio, véase la Figura 61.

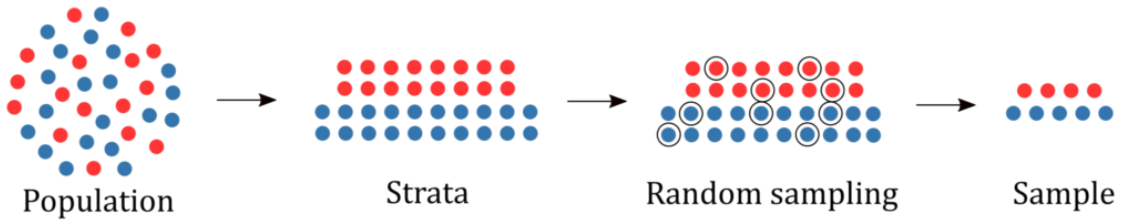

Figura 61: Técnica del muestreo probabilístico no aleatorio

Este consistio en de una manera propocional considerar que tanto el conjunto de entrenamiento y testeo tengan una proporcion significativa de observaciones en cada una de las clases. Asi se logró:

- Se logro obtener un equilibrio sobre todas las clases para evitar el sesgo sobre las clases mas dominantes
- Se tiene mas informacion para aprender sobre cada una de las clases minotarias, lo que puede llevar a una solución con una mayor generalización y capacidad de predicción.ss
- Se puede reducir el riesgo de sobreajuste de las clases mayoritarias debido a la cantida de observaciones.

Estas ventajas contribuyen a un mejor rendimiento y capacidad de generalización del modelo en escenarios donde el desequilibrio de clases es un desafío importante.

Capítulo 5

Aplicación y resultados

En base a todos los algoritmos de clasificación utilizados en este estudio de ciencia de datos, se empleará la matriz de confusión como una herramienta fundamental de evaluación. Además, se analizarán y se derivarán métricas clave a partir de la matriz de confusión, incluyendo el ***accuracy***, ***precision***, ***recall***, ***specificity*** y ***f1 score***. Estas métricas nos proporcionarán una visión detallada del rendimiento de los modelos de clasificación, permitiéndonos evaluar tanto la capacidad de predicción global como la precisión en la identificación de casos positivos y negativos [22].

A continuación se mostraran todos los modelos y sus resultados sobre cada una de las pistas de estudio. Finalmente, la sección de conclusiones describe los mejores resultados, el modelo y la configuración a nivel de parámetros que se obtuvo.

5.1 Ponderación de pesos para abordar desbalanceo de clases

En el ámbito de la predicción de la distancia de umbral entre aeronaves mediante algoritmos de aprendizaje automático, uno de los desafíos comunes es el desbalanceo de clases. El desbalanceo de clases se refiere a una situación en la que las instancias de diferentes clases en el conjunto de datos de entrenamiento están distribuidas de manera desproporcionada. En otras palabras, una o varias clases están representadas por un número significativamente menor de ejemplos en comparación con otras clases. Esta discrepancia puede tener un impacto negativo en el rendimiento de los algoritmos de aprendizaje automático, ya que pueden tener dificultades para reconocer adecuadamente las clases minoritarias [23].

Así, abordaremos el desbalanceo de clases mediante la técnica de ponderación de pesos. Esta técnica se utiliza para asignar diferentes pesos a las instancias de las clases minoritarias y mayoritarias, de manera que se tenga en cuenta la distribución desigual durante el proceso de entrenamiento. En nuestro estudio, utilizamos una variante específica de ponderación de pesos conocida como la técnica de la raíz cuadrada para calcular los pesos de cada clase.

La técnica de la raíz cuadrada asigna pesos a las clases en función de la inversa de la raíz cuadrada de su frecuencia relativa en el conjunto de datos de entrenamiento.

La fórmula para calcular el peso de cada clase utilizando la técnica de la raíz cuadrada es la siguiente:

$$w_i = \frac{1}{f_i} \tag{5.1}$$

Donde:

- w_i es el peso asignado a la clase ι
- f_i es la frecuencia relativa de la clase ii en el conjunto de datos de entrenamiento.

Al aplicar la ponderación de pesos utilizando la técnica de la raíz cuadrada, logramos mitigar los efectos del desbalanceo y mejorar el rendimiento de los algoritmos. Los pesos asignados a las clases minoritarias aumentaron significativamente, lo que permitió que los algoritmos prestaran mayor atención a estas clases durante el entrenamiento. Como resultado, observamos una mejora en la capacidad de los algoritmos para reconocer y clasificar correctamente las clases minoritarias en nuestras predicciones.

5.2 Matriz de confusión

La matriz de confusión y las métricas derivadas de ella son herramientas esenciales en la evaluación de modelos de clasificación, especialmente en problemas donde las clases no están balanceadas en términos de su distribución. En nuestro caso, al abordar la predicción de la distancia de umbral entre aeronaves, nos encontramos con un escenario donde las clases pueden tener una representación desigual [24].

La matriz de confusión es una representación tabular que muestra la relación entre las clases reales y las clases predichas por el modelo. Es una herramienta valiosa para comprender el rendimiento del modelo en diferentes clases y evaluar su capacidad para discriminar correctamente entre ellas. Véase la Figura 62.

	Predicted Category 1	**Predicted Category 2**	**Predicted Category 3**	**Predicted Category 4**	**Predicted Category 5**
Actual Category 1	True Positives (TP)	False Positives (FP)	False Positives (FP)	False Positives (FP)	False Positives (FP)
Actual Category 2	False Positives (FP)	True Positives (TP)	False Positives (FP)	False Positives (FP)	False Positives (FP)
Actual Category 3	False Positives (FP)	False Positives (FP)	True Positives (TP)	False Positives (FP)	False Positives (FP)
Actual Category 4	False Positives (FP)	False Positives (FP)	False Positives (FP)	True Positives (TP)	False Positives (FP)
Actual Category 5	False Positives (FP)	False Positives (FP)	False Positives (FP)	False Positives (FP)	True Positives (TP)

Figura 62: Matriz de confusión de 5 por 5

En el caso de clases no balanceadas, las métricas tradicionales como la exactitud (***accuracy***) pueden ser engañosas. Esto se debe a que un modelo puede lograr una alta exactitud al clasificar la clase mayoritaria correctamente, mientras que puede tener dificultades para clasificar correctamente las clases minoritarias. Esto lleva a una aparente buena precisión en la clasificación general, pero en realidad, el modelo puede no ser útil para identificar las clases minoritarias.

Es por eso que en este análisis, utilizaremos métricas más apropiadas para problemas de clasificación con clases no balanceadas, como la precisión, la sensibilidad y la medida f1. La precisión mide la proporción de observaciones clasificadas correctamente como positivas dentro de la clase positiva, mientras que el recall mide la capacidad del modelo para detectar correctamente las instancias positivas. El F1-score es una medida que combina precisión y recall, proporcionando una evaluación equilibrada del rendimiento del modelo. Para encontrar estas métricas se utilizará la función ***performance*** del paquete ***ROCR***, una herramienta especializada en la evaluación y visualización del rendimiento de modelos de clasificación.

Estas métricas nos permitirán evaluar el rendimiento de los modelos en la clasificación de todas las clases, incluidas las minoritarias. Además, nos ayudarán a identificar el equilibrio adecuado entre la precisión y el recall, según las necesidades específicas del problema. Esto es crucial para garantizar que el modelo pueda clasificar correctamente las instancias de todas las clases, incluidas las menos representadas.

5.3 Aplicación de los modelos de clasificación

La Pista 32, ubicada en la configuración Norte del Aeropuerto Adolfo Suárez Madrid-Barajas, es una de las pistas más importantes para los aterrizajes durante las operaciones normales debido a su amplio uso y relevancia en el flujo de tráfico aéreo. En este estudio, se recopilaron un total de 64,261 observaciones para la Pista 32, lo que proporciona una base de datos robusta y representativa. Por otro lado, la Pista 18, utilizada en casos de tráfico alto o cuando la Pista 32 no puede operar adecuadamente, cuenta con 28,651 observaciones, lo que representa una cantidad menor pero aún significativa de datos.

Para la aplicación de los modelos de clasificación, se procedió definiendo el tamaño del conjunto de entrenamiento y evaluación. Dado que algunas clases de la variable objetivo presentaban un número limitado de observaciones, se optó por asignar el 70% de las observaciones al conjunto de entrenamiento y el 30% al conjunto de evaluación. Esto garantiza que se cuente con suficientes datos para el entrenamiento de los modelos y una evaluación adecuada de su rendimiento. Además, se utilizó un muestreo probabilístico para obtener una muestra representativa en ambos conjuntos, asegurando así una distribución equilibrada de los datos. Este proceso de muestreo se llevó a cabo utilizando la librería ***caret***, que proporciona funcionalidades específicas para este propósito.

Una vez se obtuvo el número de observaciones en cada clase de la variable objetivo "separación de umbral", se aplicó la técnica de ponderación de pesos. Esta técnica es útil cuando existe un desequilibrio en la distribución de clases y se busca dar mayor importancia a las clases menos frecuentes. Se asignaron pesos inversamente proporcionales a la raíz cuadrada del recuento de observaciones de cada clase, de manera que las clases menos frecuentes recibieran mayores pesos y, por lo tanto, tuvieran una influencia más significativa en el proceso de clasificación.

Es importante destacar que este proceso de ponderación de pesos se realizó tanto para la Pista 32 como para la Pista 18, teniendo en cuenta tanto la variable "CAVOK" como las variables generales. Esto permite una comparación justa y adecuada del rendimiento de los modelos en ambos contextos. A continuación,

se presentan los resultados del rendimiento de los cuatro modelos evaluados en este estudio, lo que proporciona una visión general de su desempeño y su capacidad para predecir la separación de umbral entre aeronaves. Estos resultados serán clave para evaluar la eficacia de los modelos y realizar conclusiones significativas acerca de su utilidad en la práctica de la gestión del tráfico aéreo.

5.3.1 Pista 32

5.3.1.1 Modelo con la variable CAVOK

Para este modelo, al tener un menor número de variables, la ejecución y evaluación de cada uno de los modelos aplicados requieren menos tiempo.

A partir de los resultados de la Tabla 28, se puede observar que las clases 1 y 5 presentan desafíos importantes en términos de clasificaciones incorrectas. Clasificar incorrectamente una instancia de clase 5 como clase 1 puede tener un impacto significativo en la seguridad de las operaciones aeroportuarias. Por lo tanto, es crucial prestar especial atención a estos casos extremos y mejorar la precisión en la predicción de la distancia de separación en el umbral para estas clases.

Tabla 28: Matriz de confusión, modelo cavok

	Clase 1	***Clase 2***	***Clase 3***	***Clase 4***	***Clase 5***
Clase 1	56	67	20	57	0
Clase 2	195	2693	1555	603	8
Clase 3	227	1714	5991	1431	20
Clase 4	165	657	1541	2233	10
Clase 5	0	4	11	17	0

La Tabla 29 muestra que el clasificador Random Forest (RF) tiene el valor más alto de recall con 0.52 y un F1-score de 0.52. Estas métricas indican que el clasificador RF tiene una buena capacidad para identificar correctamente las instancias positivas, lo cual es especialmente importante en este contexto donde se busca maximizar el recall. Sin embargo, es importante tener en cuenta que el accuracy por sí solo no proporciona una imagen completa del rendimiento del modelo.

Tabla 29: Indicadores de rendimiento, CAVOK en pista 32

Clasificador	***accuracy***	***precision***	***recall***	***specificity***	***f1***
NB	0.55	0.52	0.42	0.88	0.46
TAN	0.53	0.53	0.42	0.88	0.47
DT	0.53	0.56	0.39	0.90	0.46
RF	0.57	0.51	0.52	0.85	0.52

5.3.1.2 Modelo con las variables generales

Dentro de esta pista, se pudo evaluar que las variables generales no ofrecían un mejor rendimiento en términos de tiempos de ejecución y métricas de rendimiento en comparación con el modelo que utiliza únicamente la variable ***CAVOK***.

En la siguiente Tabla 30, la clase 4 también presenta una cantidad significativa de clasificaciones incorrectas, lo que resalta la importancia de mejorar la precisión en la predicción para garantizar una clasificación precisa y salvaguardar la seguridad de las operaciones aéreas.

Tabla 30. Matriz de confusión, modelo general

	Clase 1	***Clase 2***	***Clase 3***	***Clase 4***	***Clase 5***
Clase 1	62	68	19	51	0
Clase 2	223	2691	1545	588	7
Clase 3	247	1682	6047	1364	16
Clase 4	174	605	1565	2253	9
Clase 5	0	4	14	14	0

La Tabla 31 muestra que el clasificador Random Forest (RF) tiene el valor más alto de precision con 0.52. Esto indica que el clasificador tiene una mayor proporción de verdaderos positivos en comparación con los falsos positivos.

Tabla 31: Indicadores de rendimiento, generales en pista 32

Clasificador	***accuracy***	***precision***	***recall***	***specificity***	***f1***
NB	0.29	0.44	0.42	0.65	0.35
TAN	0.28	0.43	0.41	0.63	0.34
DT	0.31	0.45	0.41	0.68	0.38
RF	0.55	0.52	0.52	0.85	0.52

5.3.2 Pista 18

5.3.2.1 Modelo con la variable CAVOK

Para este modelo, es importante tener en cuenta que la pista 18 cuenta con un menor número de observaciones en comparación con la pista 32, pero se mantiene el mismo número de variables. Sin embargo, a pesar de esta diferencia, el tiempo de ejecución y evaluación es aún mayor en este modelo.

A partir de estos resultados, véase la Tabla 32, se puede observar que las clases 2 y 3 presentan desafíos importantes en términos de clasificaciones incorrectas, lo que indica la necesidad de mejorar la precisión en la predicción de la distancia de separación en el umbral.

Tabla 32: Matriz de confusión, modelo cavok

	Clase 1	***Clase 2***	***Clase 3***	***Clase 4***	***Clase 5***
Clase 1	8	29	14	34	7
Clase 2	33	993	421	220	43
Clase 3	81	784	1816	770	127
Clase 4	109	457	800	1443	140
Clase 5	9	34	69	139	13

La Tabla 33 muestra que el clasificador Random Forest (RF) tiene el valor más alto de recall con 0.58. Esto indica que el clasificador tiene la capacidad de identificar correctamente el 58% de las instancias positivas.

Tabla 33: Indicadores de rendimiento, CAVOK en pista 18

Clasificador	***accuracy***	***precision***	***recall***	***specificity***	***f1***
NB	0.56	0.50	0.46	0.88	0.48
TAN	0.57	0.50	0.47	0.88	0.49
DT	0.55	0.46	0.50	0.85	0.48
RF	0.49	0.43	0.58	0.81	0.49

5.3.2.2 Modelo con las variables generales

En este contexto, se llevó a cabo una evaluación de las variables generales dentro de esta pista, y se determinó que no ofrecían un rendimiento superior al modelo anterior.

De la Tabla 34 se puede mencionar Es crucial prestar especial atención a estos casos extremos y mejorar la precisión en la predicción de la distancia de separación en el umbral para la clase 1

Tabla 34: Matriz de confusión, modelo generales

	Clase 1	***Clase 2***	***Clase 3***	***Clase 4***	***Clase 5***
Clase 1	7	27	14	38	6
Clase 2	30	997	426	226	31
Clase 3	75	747	1923	715	118
Clase 4	77	442	788	1503	139
Clase 5	9	35	63	143	14

La Tabla 35 muestra que el clasificador Random Forest (RF) muestra el valor más alto de accuracy con un 0.57, lo que indica que tiene la capacidad de clasificar correctamente el 57% de las instancias de prueba. Sin embargo, es importante tener en cuenta que el accuracy por sí solo no proporciona una imagen completa del rendimiento del modelo.

Tabla 35: Indicadores de rendimiento, generales en pista 18

Clasificador	***accuracy***	***precision***	***recall***	***specificity***	***f1***
NB	0.57	0.50	0.46	0.88	0.48
TAN	0.58	0.51	0.46	0.88	0.48
DT	0.55	0.46	0.50	0.85	0.48
RF	0.49	0.44	0.59	0.81	0.50

5.4 Análisis de los resultados

En base a los resultados obtenidos en la sección previa y considerando las métricas de rendimiento para cada uno de los cuatro modelos evaluados, se llevará a cabo una comparativa exhaustiva para poder evaluar y obtener conclusiones sólidas acerca de los resultados.

En primer lugar, se realizará una comparativa detallada para la pista 32, con el objetivo de analizar y destacar las diferencias entre el modelo ***CAVOK*** y el modelo de variables generales. Se presentarán los mejores resultados obtenidos de manera conjunta, considerando métricas como la precisión, el recall, la F1-score y cualquier otra medida relevante para la clasificación de la separación de umbral entre aeronaves en esta pista. Además, se realizará un análisis en profundidad de los tiempos de ejecución y cualquier otro aspecto relevante que permita evaluar la eficiencia y eficacia de los modelos aplicados. Los siguientes datos fueron extraídos de las Tablas 29 y 31. Véase la Tabla 36.

Tabla 36. Mejores resultados para la pista 32

Modelo	***accuracy***	***precision***	***recall***	***specificity***	***f1***
CAVOK	0.57	0.51	0.52	0.85	0.52
Gener.	0.55	0.52	0.52	0.85	0.52

Por otro lado, se llevará a cabo un proceso comparativo similar para la pista 18, con el fin de obtener conclusiones sólidas sobre el rendimiento tanto de los algoritmos utilizados como de las métricas de rendimiento. Dado que la pista 18 cuenta con un menor número de observaciones en comparación con la pista 32, se prestará especial atención a cómo este factor afecta el desempeño de los modelos y si se requieren enfoques específicos para abordar este desafío. Los siguientes datos fueron extraídos de las Tablas 33 y 35. Véase la Tabla 37.

Tabla 37: Mejores resultados para la pista 18

Modelo	***Accuracy***	***precision***	***recall***	***specificity***	***f1***
CAVOK	0.49	0.43	0.58	0.81	0.49
Gener.	0.49	0.44	0.59	0.81	0.5099

Basándome en los resultados obtenidos en la pista 32, se concluye que el mejor algoritmo en términos de métricas de rendimiento fue ***Random Forest***, utilizado en conjunto con la técnica de ponderación de pesos. Luego, analizaré ambos resultados de los modelos y llegaré a la conclusión de que ambos se encuentran en un rango similar. Sin embargo, si deseamos evaluar la eficiencia en cuanto al tiempo, se recomendaría optar por el modelo ***CAVOK***, ya que utiliza menos variables y muestra mayor objetividad. Este mismo enfoque se aplicará para la pista 18.

Capítulo 6

Conclusiones y líneas de trabajo

En la industria de la aviación, la seguridad y la eficiencia son aspectos vitales que requieren una atención constante. La capacidad de predecir con exactitud la distancia de separación en el umbral entre las aeronaves desempeña un papel fundamental en la mejora de la capacidad de los aeropuertos y en la garantía de la seguridad de los vuelos. En este Trabajo de Fin de Máster en Ciencia de Datos, nos hemos propuesto el desafío de desarrollar un modelo de aprendizaje automático que aborde este problema de manera precisa y efectiva.

En nuestro estudio, hemos explorado la aeronáutica y el impacto de la meteorología en el contexto de la predicción de la separación de umbral entre aeronaves. Además, hemos utilizado algoritmos de aprendizaje automático para mejorar el control del tráfico aéreo, proporcionando una herramienta eficiente y precisa para los controladores aéreos en la toma de decisiones.

Durante el desarrollo de este proyecto, hemos establecido una valiosa colaboración con expertos de la industria. En particular, nos hemos centrado en analizar los casos extremos identificados en la matriz de confusión de nuestro modelo. Hemos enviado cuatro casos extremos a estos expertos para su análisis detallado y explicación. Actualmente, estamos a la espera de su respuesta y continuaremos trabajando estrechamente con ellos para comprender mejor los desafíos y mejorar nuestros resultados.

Un resultado destacado de esta colaboración ha sido la incorporación de una nueva variable relevante en nuestro modelo: la intensidad de tráfico esperada en los siguientes 15 minutos. Esta adición tiene como objetivo mejorar aún más la capacidad predictiva del modelo al considerar un factor adicional que puede influir en la distancia de separación en el umbral entre las aeronaves.

En base a los resultados obtenidos y las métricas de rendimiento evaluadas utilizando la matriz de confusión, hemos observado que el modelo de random forest ha obtenido las mejores métricas, destacando un ***accuracy*** de 0.49621, ***recall*** de 0.59415, ***specificity*** de 0.81708 y ***f1-score*** de 0.50991.

En conclusión, este estudio representa un avance significativo en la predicción de la separación de umbral entre aeronaves. Además, la colaboración con expertos de la industria y la incorporación de la variable de intensidad de tráfico en los siguientes 15 minutos demuestran nuestro compromiso continuo en buscar soluciones efectivas y mejorar los resultados. Continuaremos trabajando en conjunto con los expertos para abordar los casos extremos identificados y seguir avanzando en la predicción precisa de la separación de umbral entre aeronaves en el campo aeronáutico.

Bibliografía

[1] «Basic Regulation,» European Union Aviation Safety Agency (EASA), 2018. [En línea]. Available: https://www.easa.europa.eu/en/regulations/basic-regulation.

[2] «Airports,» Agencia Estatal de Seguridad Aérea (AESA), 2008. [En línea]. Available: https://www.seguridadaerea.gob.es/en/ambitos/aeropuertos.

[3] C. Qing y H. J. ANG, «Collision risk assessment of reduced aircraft separation minima in procedural airspace using advanced communication and navigation.,» *Chinese Journal of Aeronautics,* pp. 315-337, 2023.

[4] L. Ren y J.-P. Clarke, «Flight-test evaluation of the tool for analysis of separation and throughput,» *Journal of Aircraft 45,* pp. 323-332, 2008.

[5] AEMET, «Guía MET: Información Meteorológica Aeronáutica.,» 2021. [En línea]. Available: https://www.aemet.es/documentos/es/conocermas/aeronautica/AU-GUI-0102.pdf.

[6] R. Patriarca y S. Francesco, «Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering.,» *Expert Systems with Applications,* vol. 213, pp. 119-210, 2023.

[7] A. Alfaro Díez, «Predicción de la Distancia de Separación en Umbral entre Aeronaves mediante Aprendizaje Automático,» Universidad Politécnica de Madrid, Madrid, 2022.

[8] A. Saraf y et al., «Capacity Finder: A Machine Learning-Based Decision Support Tool for Integrated Metroplex Departure Traffic Management.,» *Digital Avionics Systems Conference,* vol. 40, pp. 1-11, 2021.

[9] M. Sánchez Piñeiro, «Predicción del impacto de regulaciones aéreas sobre los flujos de aeronaves utilizando técnicas de aprendizaje automático,» ETSI Informatica, Marid, 2020.

[10] H. Li, «Clustering discretization methods for generation of material performance databases in machine learning and design optimization,» *Computational Mechanics ,* vol. 64, pp. 281-305, 2019.

[11] S. Grabbe y S. Banavar, «Clustering days and hours with similar airport traffic and weather conditions.,» *Journal of Aerospace Information Systems,* vol. 11, pp. 751-763, 2014.

[12] K. Murphy, Machine learning: a probabilistic perspective, London: MIT press, 2012.

[13] H. Hachiya y S. Masashi, «Feature selection for reinforcement learning: Evaluating implicit state-reward dependency via conditional mutual information.,» *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* pp. 474-489, 2010.

[14] Y. Saheed y H. Moshood, «Customer churn prediction in telecom sector with machine learning and information gain filter feature selection

algorithms.,» *International Conference on Data Analytics for Business and Industry,* vol. 2021, pp. 208-213, 2021.

[15] O. Snisarevska y L. Sherry, «Balancing throughput and safety: An autonomous approach and landing system (AALS),» *Integrated Communications, Navigation, Surveillance Conference (ICNS),* pp. 3B1-1, 2018.

[16] C. Bielza y P. Larrañaga, Data-driven computational neuroscience: machine learning and statistical models., Madrid: Cambridge University Press, 2020.

[17] M. Stathis y T. Kontogiannis, «Malakis, Stathis, et al. "Classification of air traffic control scenarios using decision trees: insights from a field study in terminal approach radar environment,» *Cognition, Technology & Work 22,* pp. 159-179, 2020.

[18] J. J. Rebollo y H. Balakrishnan, «Characterization and prediction of air traffic delays,» *Transportation research part C: Emerging technologies,* vol. 44, pp. 231-241, 2014.

[19] R. Alligier y D. Gianazza, «Learning aircraft operational factors to improve aircraft climb prediction: A large scale multi-airport study.,» *Learning aircraft operational factors to improve aircraft climb prediction: A large scale multi-airport study.,* vol. 96, pp. 72-95, 2018.

[20] Z. Wang y L. Man, «A hybrid machine learning model for short-term estimated time of arrival prediction in terminal manoeuvring area.,» *Transportation Research Part C: Emerging Technologies,* vol. 95, pp. 280-294, 2018.

[21] L. Moreira, «On evaluating data preprocessing methods for machine learning models for flight delays.,» *International Joint Conference on Neural Networks,* pp. 1-8, 2018.

[22] N. Chakrabarty, «A data mining approach to flight arrival delay prediction for american airlines.,» de *9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, 2019.

[23] Jiawei y Zhunga, «Conditional self-attention generative adversarial network with differential evolution algorithm for imbalanced data classification,» *Chinese Journal of Aeronautics,* pp. 303-315, 2023.

[24] E. Mangortey, «Classification, analysis, and prediction of the daily operations of airports using machine learning.,» *AIAA Scitech,* p. 1196, 2020.

[25] Jeddi, Babak G., J. F. Shortle y L. Sherry, «Statistical separation standards for the aircraft-approach process.,» *Digital Avionics Systems Conference,* pp. 1-13, 2006.

[26] A. Maxwell, T. Warner y F. Fang, «Implementation of machine-learning classification in remote sensing: An applied review,» *nternational Journal of Remote Sensing,* 2018.

[27] S. Choi y S. Briceno, «Prediction of weather-induced airline delays based on machine learning algorithms.,» *Digital Avionics Systems Conference (DASC),* pp. 1-6, 2016.

[28] M. Schultz y S. Reitmann, «Predictive classification and understanding of weather impact on airport performance through machine learning.,» *Transportation Research Part C: Emerging Technologies,* pp. 103-119, 2021.

[29] A. Hajar y L. Moumoun, «Aircraft Performance and Time of the Day on Flight Arrival Delays Prediction in the United States: a Machine Learning Classification.,» *ITM Web of Conferences,* vol. 48, 2022.

[30] Z. Zhipeng, «Development and application of a Bayesian network-based model for systematically reducing safety risks in the commercial air transportation system.,» *Safety science 157,* 2023.

[31] Y. Rajat y R. Shukla, «A Hybrid Model Integrating Adaboost Approach for Sentimental Analysis of Airline Tweets,» *Revue d'Intelligence Artificielle,* p. 519, 2022.

[32] J. K. Young y S. Choi, «Artificial neural network models for airport capacity prediction.,» *Journal of Air Transport Management,* vol. 97, pp. 102-146, 2021.

Anexos

Se pone a disposición el repositorio de GitHub donde se pueden encontrar todos los resultados y código generados durante el desarrollo de este Trabajo de Fin de Máster.

https://github.com/saavedrAndrei/Separacion-Umbral