



SMART FLIGHTS

Mejorando la Experiencia de Vuelo con Datos

Comisión
75690

Alumna
Patricia Alonso Castillo



Introducción

La gestión de demoras y cancelaciones en vuelos es un desafío crucial para las aerolíneas, aeropuertos y, en particular, para las agencias de viajes, que buscan ofrecer experiencias positivas a sus clientes.

En este contexto,

el análisis de datos históricos sobre vuelos retrasados puede proporcionar insights valiosos para optimizar la planificación de viajes, mejorar la satisfacción del cliente y desarrollar estrategias para mitigar los impactos de estas interrupciones.

Este trabajo

se centra en explorar un conjunto de datos detallados sobre demoras y cancelaciones de vuelos en los Estados Unidos durante el año 2019, con el objetivo de identificar patrones y construir un modelo predictivo que permita anticipar demoras y cancelaciones.

Descripción del problema

Las demoras y cancelaciones en vuelos generan costos significativos tanto para las aerolíneas como para los pasajeros, además de afectar la reputación de las agencias de viajes. La capacidad de anticipar estos eventos permite a las agencias de viajes tomar decisiones más informadas, como sugerir horarios alternativos o aerolíneas con menores riesgos de retrasos. Sin embargo, la predicción de estos incidentes es un desafío debido a la influencia de factores diversos, como las condiciones climáticas, el tráfico aéreo y las operaciones aeroportuarias. Analizar los datos disponibles y desarrollar modelos predictivos puede ayudar a identificar los factores clave que contribuyen a las interrupciones en los vuelos y mejorar la experiencia de viaje.

Objetivo



El objetivo principal de este proyecto es construir un modelo de clasificación que permita predecir si un vuelo tendrá una demora significativa o será cancelado, basándose en datos históricos de vuelos.

Para ello, se llevará a cabo un análisis exploratorio de los datos, se identificarán patrones clave relacionados con las demoras y cancelaciones, y se evaluará el rendimiento de diversos modelos de clasificación. Los resultados obtenidos pueden ser utilizados por agencias de viajes para diseñar estrategias de ventas más efectivas y personalizadas, minimizando el impacto de interrupciones en los itinerarios de los clientes.

Fuente

Los datos utilizados en este análisis fueron obtenidos del portal Kaggle y están disponibles en el siguiente enlace: 2019 Airline Delays and Cancellations

<https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations> Este conjunto de datos incluye información detallada sobre vuelos, aerolíneas, aeropuertos, demoras en salidas y llegadas, así como las razones detrás de las cancelaciones. Su análisis permitirá generar insights significativos y construir un modelo predictivo robusto para anticipar demoras y cancelaciones, beneficiando tanto a las agencias de viajes como a los viajeros.



Variables

MONTH

Número del mes (1-12, donde 1 = enero y 12 = diciembre)

DAY_OF_WEEK

Día de la semana (1-7, donde 1 = lunes y 7 = domingo).

DEP_DEL15

Indicador de retraso en la salida superior a 15 minutos (1 = retraso, 0 = a tiempo).

DEP_TIME_BLK

Bloque de tiempo de salida en intervalos de 59 minutos (por ejemplo, '0001-0059' para salidas entre las 00:01 y las 00:59).

DISTANCE_GROUP

Grupo de distancia del vuelo.

SEGMENT_NUMBER

Número de segmento que indica la posición del vuelo para la aeronave en el día.

CONCURRENT_FLIGHTS

Número de vuelos concurrentes que salen del aeropuerto en el mismo bloque de tiempo de salida.

NUMBER_OF_SEATS

Número de asientos en la aeronave.

CARRIER_NAME

Nombre de la aerolínea.

AIRPORT_FLIGHTS_MONTH

Número de vuelos desde el aeropuerto por mes.

AIRLINE_FLIGHTS_MONTH

Número de vuelos de la aerolínea por mes.

AIRLINE_AIRPORT_FLIGHTS_MONTH

Número de vuelos específicos de la aerolínea desde el aeropuerto por mes.

AVG_MONTHLY_PASS_AIRPORT

Promedio mensual de pasajeros en el aeropuerto.

AVG_MONTHLY_PASS_AIRLINE

Promedio mensual de pasajeros de la aerolínea.

FLT_ATTENDANTS_PER_PASS

Número de asistentes de vuelo por pasajero.

GROUND_SERV_PER_PASS

Personal de servicio en tierra por pasajero.

PLANE_AGE

Edad de la aeronave en años.

DEPARTING_AIRPORT

Código del aeropuerto de salida

LATITUDE

Latitud del aeropuerto de salida.

LONGITUDE

Longitud del aeropuerto de salida.

PREVIOUS_AIRPORT

Código del aeropuerto previo (indica el último aeropuerto en el que estuvo la aeronave; "NONE" si no hubo uno previo).

PRCP

Precipitación en pulgadas para el día.

SNOW

Nieve caída en pulgadas para el día.

SNWD

Profundidad de la nieve en el suelo en pulgadas para el día.

TMAX

Temperatura máxima en grados Fahrenheit para el día.

AWND

Velocidad máxima del viento en millas por hora para el día.

Hipótesis planteadas

- Las condiciones meteorológicas adversas aumentan significativamente la probabilidad de retrasos y cancelaciones, sobre todo las precipitaciones.
- Los vuelos en meses de vacaciones (como julio y diciembre) tienen más pasajeros, lo que podría influir en los tiempos de operación y generar retrasos
- Los viernes tienen una mayor incidencia de retrasos debido al mayor volumen de tráfico aéreo.



Clustering de Aeropuertos (K-Means)

Se aplicó la técnica de aprendizaje no supervisado **K-Means** con el objetivo de agrupar aeropuertos de EE.UU. en función de sus características operativas y climáticas:

◆ Variables utilizadas:

- Vuelos Concurrentes (congestión operativa)
- Precipitación (PRCP)
- Temperatura máxima (TMAX)
- Velocidad del viento (AWND)
- Latitud y Longitud

◆ Resultados:

- Se identificaron **3 clusters** bien diferenciados:
 - **Cluster 0**: Aeropuerto aislado (Puerto Rico), comportamiento atípico. Se trata como outlier.
 - **Cluster 1**: Aeropuertos del Este. Alta congestión, clima húmedo.
 - **Cluster 2**: Aeropuertos del Oeste. Mayor temperatura, clima seco, alta actividad.

✖ Las métricas de evaluación del clustering fueron:

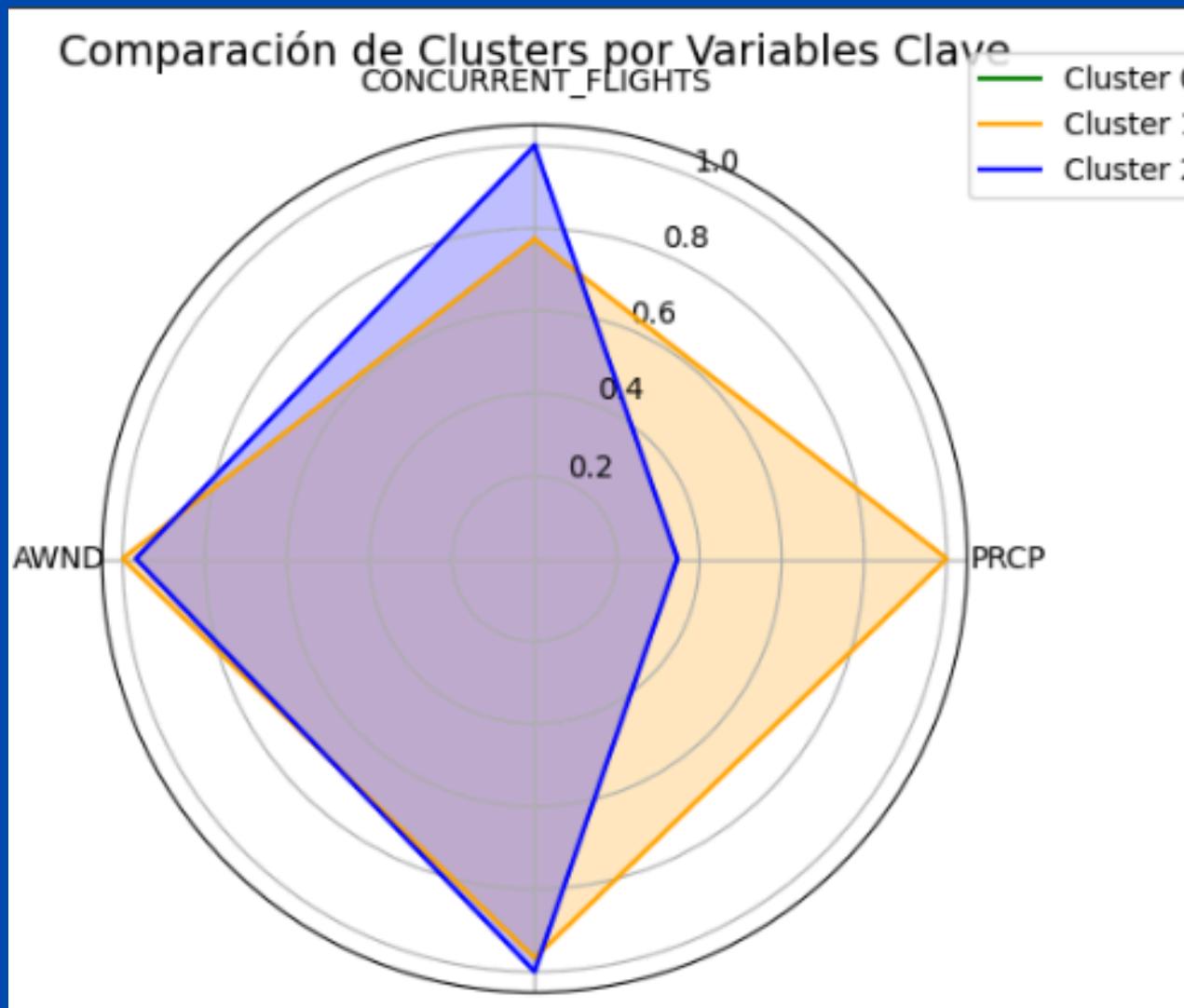
- Silhouette Score: 0.296
- Calinski-Harabasz Index: 26.42
- Davies-Bouldin Index: 1.048

Estas métricas indican una estructura moderada pero válida, y los clusters capturan diferencias reales entre zonas del país.

Comparación de Clusters por Variables Clave

El gráfico radar permite visualizar cómo se diferencian los clusters identificados en el análisis no supervisado (K-Means), en relación con las variables más representativas:

- CONCURRENT_FLIGHTS: mide la congestión operativa del aeropuerto
- PRCP: representa la precipitación promedio
- TMAX: temperatura máxima (no visible si fue excluida por superposición)
- AWND: velocidad promedio del viento



Interpretaciones clave:

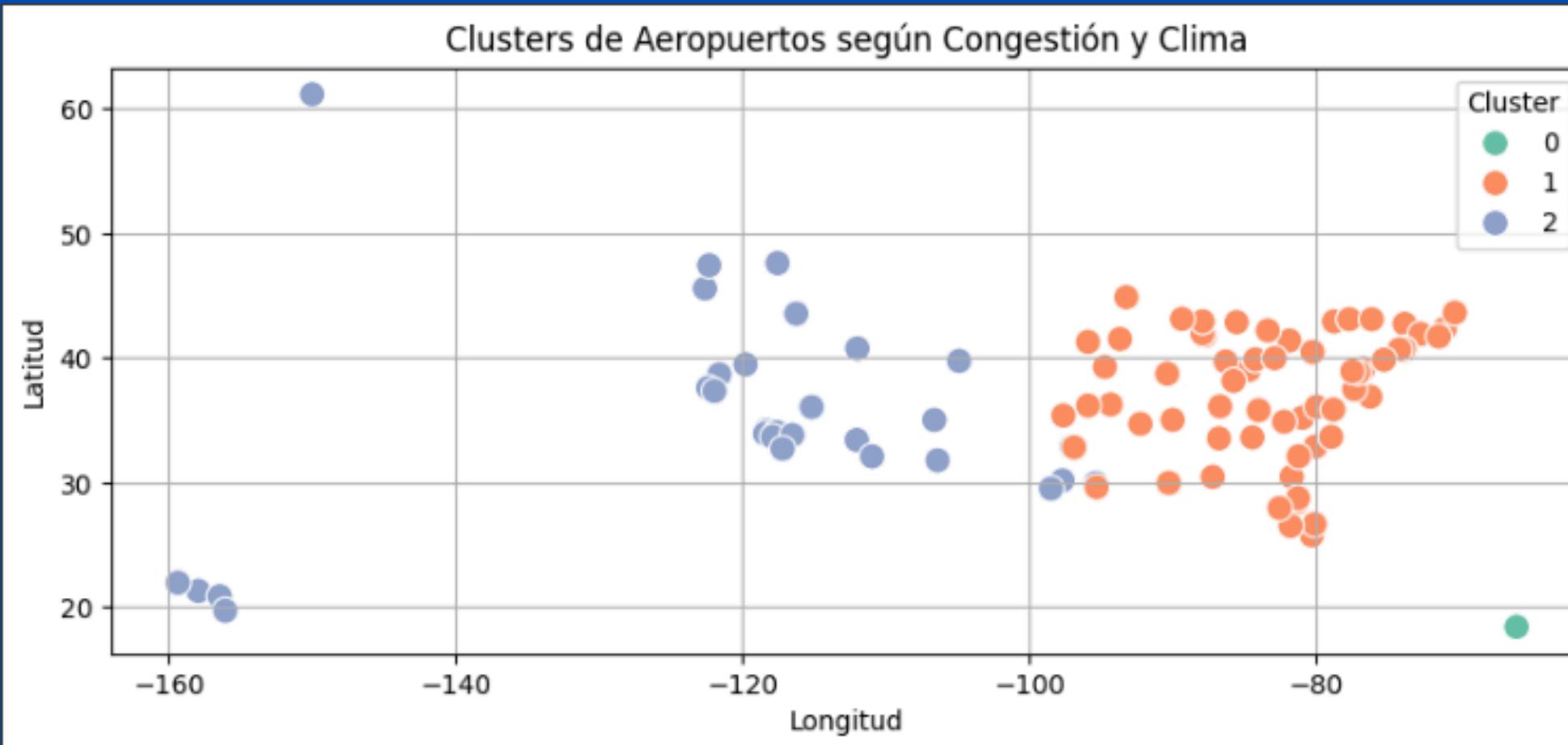
- Cluster 2 (azul) muestra el valor más alto en vuelos concurrentes, indicando aeropuertos muy activos, típicos del oeste del país.
- Cluster 1 (naranja) presenta la mayor precipitación promedio, lo que coincide con zonas más húmedas como el este.
- Cluster 0 (verde) mantiene valores intermedios o muy bajos, posiblemente influenciado por el outlier (Puerto Rico), y debe ser interpretado con cautela.
- Las diferencias entre los picos muestran cómo cada grupo está marcado por una variable dominante: clima o congestión.

Visualización de clusters geográficos

El gráfico muestra la distribución espacial de los aeropuertos agrupados por el algoritmo K-Means, utilizando su longitud y latitud como referencia.

Cada color representa un cluster identificado a partir de variables operativas y climáticas. La separación geográfica evidencia que el modelo capturó patrones regionales consistentes.

Interpretaciones clave:

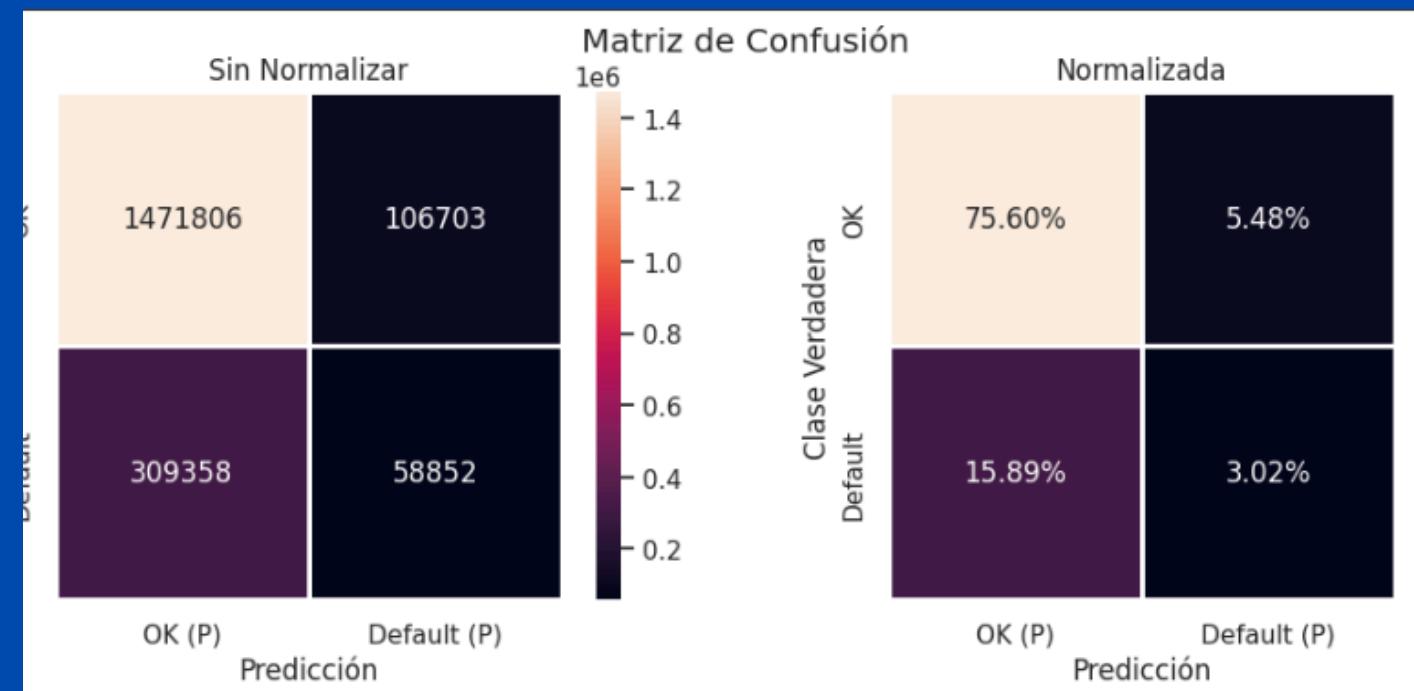


- Cluster 1 (naranja): Agrupa aeropuertos ubicados mayormente en el este de EE.UU., caracterizados por mayor congestión y clima más húmedo.
- Cluster 2 (azul): Corresponde a aeropuertos del oeste y centro, con temperaturas más elevadas y menor precipitación.
- Cluster 0 (verde): Representa un caso atípico y aislado, correspondiente a un aeropuerto ubicado en una región tropical (Puerto Rico), con valores extremos o nulos. Este grupo se considera un outlier operativo.
- mayor precipitación promedio, lo que coincide con zonas más húmedas como el este.
- Cluster 0 (verde) mantiene valores intermedios o muy bajos, posiblemente influenciado por el outlier (Puerto Rico), y debe ser interpretado con cautela.
- Las diferencias entre los picos muestran cómo cada grupo está marcado por una variable dominante: clima o congestión.

Comparación de modelos y desempeño predictivo

	Modelo	Accuracy	Precision	Recall	ROCAUC	F1-Score	Tiempo
0	KNN (k=5)	0.7863	0.3555	0.1598	0.5461	0.2205	11.1061
1	Árbol de Decisión	0.8003	0.3939	0.1039	0.5333	0.1644	27.2473
2	XGBoost (GPU)	0.8119	0.5747	0.0217	0.5090	0.0418	2.8940
3	Regresión Logística	0.8109	0.0000	0.0000	0.5000	0.0000	16.9588

KNN



KNN resultó ser el modelo con mayor desempeño para el negocio

Insights obtenidos

Desbalance de clases

- El 81% de los vuelos del dataset no presentaron demoras, lo que generó un fuerte desbalance que impactó en el rendimiento de los modelos supervisados. Esta condición obligó a priorizar métricas como recall y F1-Score, más que la accuracy general.

KNN fue el modelo más sensible a las demoras

- Si bien su accuracy fue la más baja, el modelo KNN ($k=5$) logró el mayor recall (15.98%) y F1-Score (0.2205), detectando más vuelos demorados que el resto. Esto lo convierte en el mejor modelo si el objetivo es emitir alertas preventivas.

El Árbol de Decisión mostró un buen equilibrio

- Detectó más del 10% de las demoras reales con un nivel aceptable de precisión, sin requerir aceleración por GPU. Es una alternativa sólida cuando se busca balance entre métricas.



XGBoost fue el más eficiente en tiempo

- Aunque no se destacó por su recall, XGBoost en GPU fue el más veloz y mostró alta precisión. Resulta útil cuando se prioriza velocidad de ejecución sobre sensibilidad.moras reales con un nivel aceptable de precisión, sin requerir aceleración por GPU. Es una alternativa sólida cuando se busca balance entre métricas.

La regresión logística no detectó ningún vuelo demorado

- Su performance fue muy baja frente al desbalance de clases, lo que la descarta como opción válida para este problema.

Clustering reveló tipologías de aeropuertos

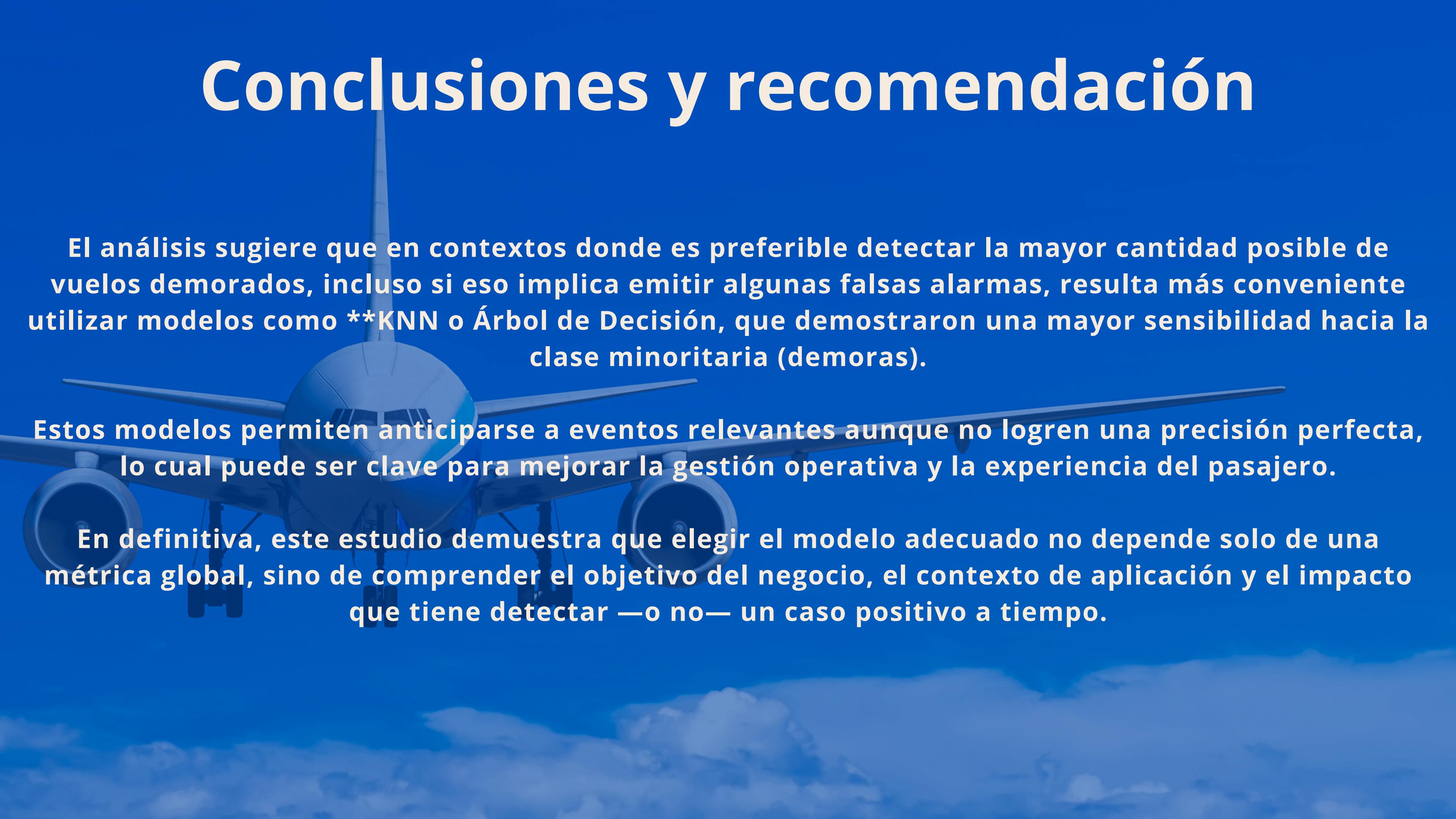
- El análisis no supervisado agrupó aeropuertos según clima, ubicación y congestión. Se identificaron tres grupos diferenciados, lo cual permite entender cómo las condiciones operativas varían regionalmente y podrían asociarse a diferentes riesgos de demora.

Importancia del contexto operativo y estacional

- El análisis exploratorio mostró que las demoras aumentan en meses de alta demanda (julio/diciembre) y que variables como congestión y condiciones climáticas (precipitación y viento) podrían influir indirectamente en la aparición de demoras.



Conclusiones y recomendación

A large airplane is shown from a low angle, flying towards the viewer. The aircraft's body is a light grey color, and its wings and engines are visible. The background is a clear blue sky with scattered white clouds.

El análisis sugiere que en contextos donde es preferible detectar la mayor cantidad posible de vuelos demorados, incluso si eso implica emitir algunas falsas alarmas, resulta más conveniente utilizar modelos como **KNN o Árbol de Decisión, que demostraron una mayor sensibilidad hacia la clase minoritaria (demoras).

Estos modelos permiten anticiparse a eventos relevantes aunque no logren una precisión perfecta, lo cual puede ser clave para mejorar la gestión operativa y la experiencia del pasajero.

En definitiva, este estudio demuestra que elegir el modelo adecuado no depende solo de una métrica global, sino de comprender el objetivo del negocio, el contexto de aplicación y el impacto que tiene detectar —o no— un caso positivo a tiempo.