



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

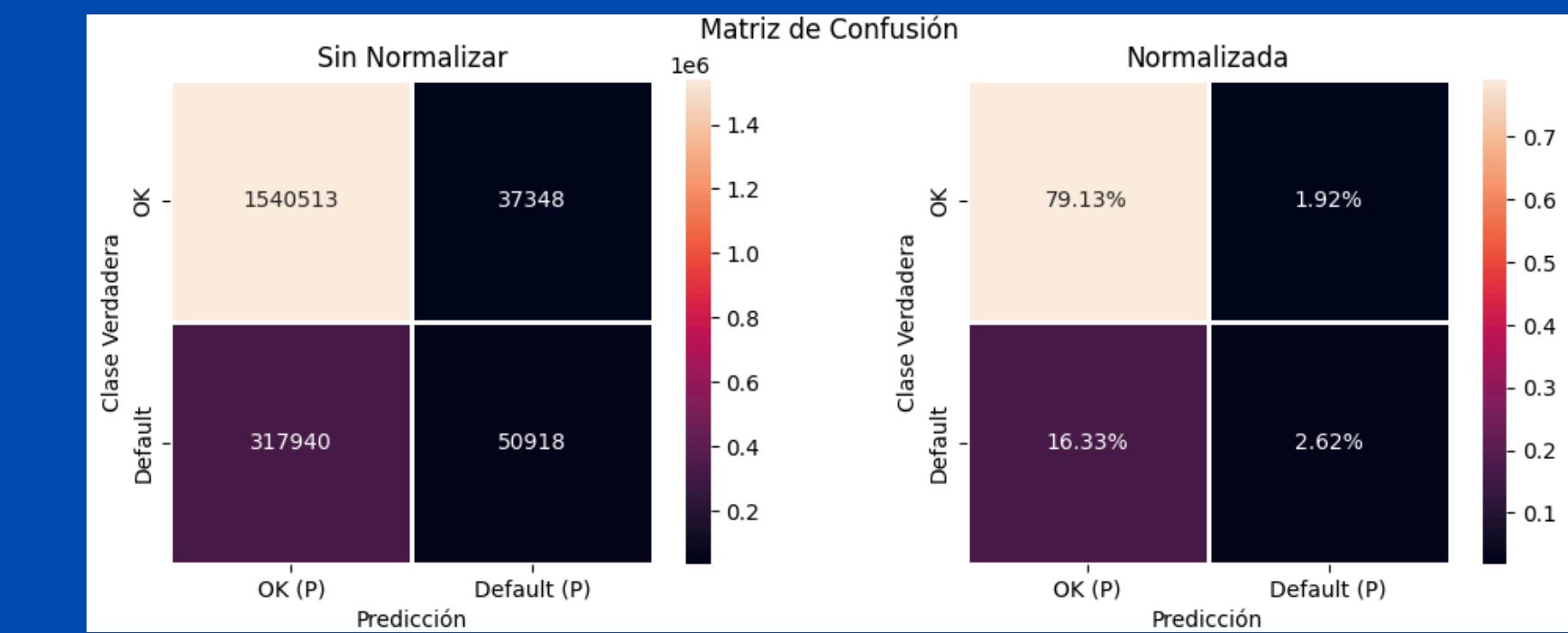
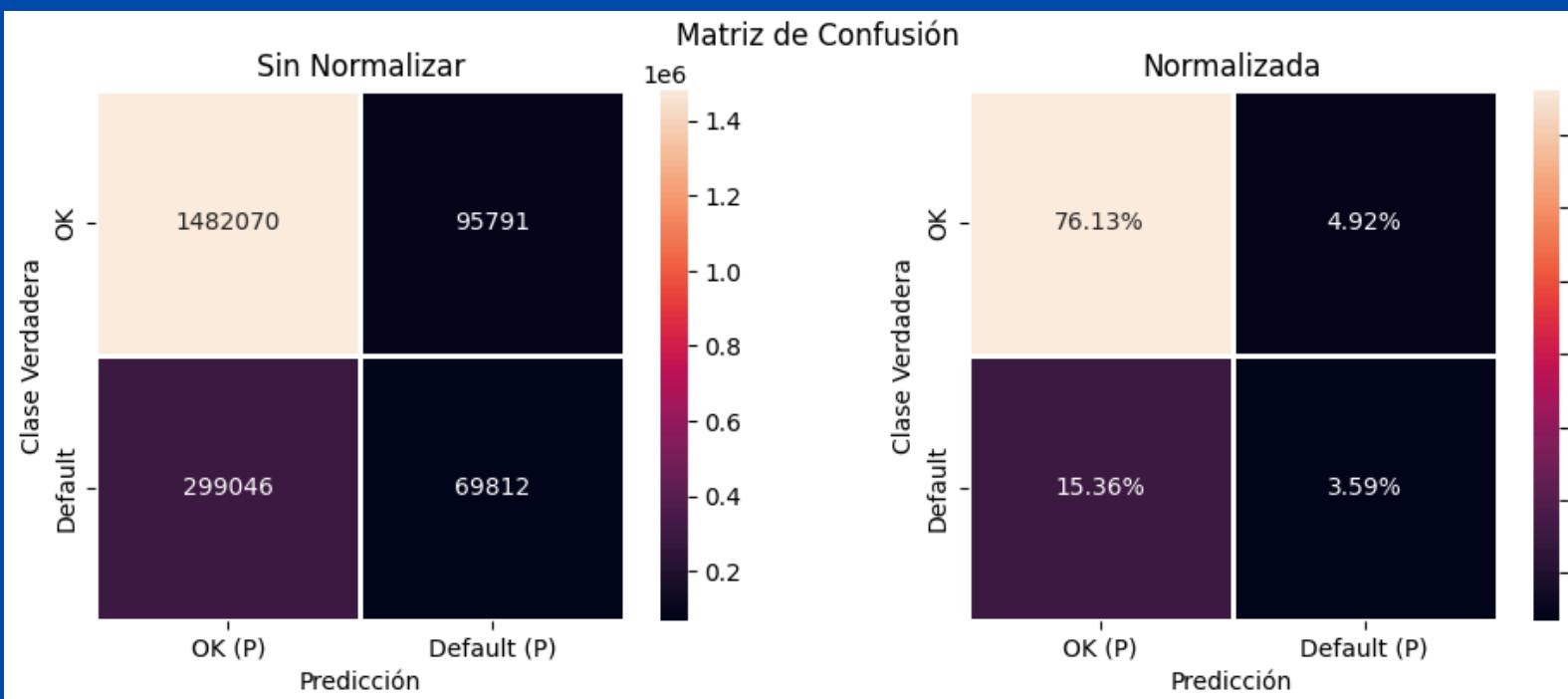
- Las condiciones meteorológicas adversas aumentan significativamente la probabilidad de retrasos y cancelaciones, sobre todo las precipitaciones.
- Los vuelos en meses de vacaciones (como julio y diciembre) tienen más pasajeros, lo que podría influir en los tiempos de operación y generar retrasos
-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



Comparación de modelos y desempeño predictivo

	Modelo	Accuracy	Precision	Recall	ROCAUC	F1-Score	Tiempo
0	KNN (k=5)	0.7972	0.4216	0.1893	0.5643	0.2612	21.5620
1	Árbol de Decisión	0.8175	0.5769	0.1380	0.5572	0.2228	36.6370
2	XGBoost (GPU)	0.8175	0.6644	0.0739	0.5326	0.1330	6.8781
3	Regresión Logística	0.8104	0.4769	0.0060	0.5022	0.0118	17.4968

KNN



KNN y Árbol de Decisión fueron los modelos con mejor desempeño en detección de vuelos demorados. La tabla y la matriz permiten visualizar el equilibrio entre aciertos, errores y tiempo de ejecución.

Insights obtenidos

- La mayoría de los vuelos no presentan demoras, pero identificarlos es clave.
- El modelo KNN logró el mayor Recall (18.9%), seguido por Árbol de Decisión. Los no presentan demoras, pero identificarlos es clave.
- La mayoría de los vuelos no presentan demoras, pero identificarlos es clave. Las demoras aumentan en días viernes y meses con mayor tráfico (julio/diciembre).
- La mayoría de los vuelos no presentan demoras, pero identificarlos es clave. Las variables más relevantes fueron: horario de salida (DEP_TIME_BLK), día de la semana y condiciones climáticas (como PRCP y TMAX)..



Conclusiones y recomendación

El objetivo del análisis fue construir un modelo que permita anticipar demoras en vuelos, beneficiando a agencias de viajes y pasajeros.

Tras evaluar múltiples modelos, KNN ($k=5$) resultó el más adecuado para detectar casos de demora, priorizando Recall y F1-Score.

Se recomienda este enfoque para tareas donde sea más importante detectar potenciales problemas a tiempo, aunque con cierto margen de error.

Futuras mejoras podrían incluir nuevas variables (clima en destino, historial por aerolínea) y optimización con técnicas como SMOTE o GridSearch.