Universidad Tecnológica de Panamá
Facultad de Ingeniería Industrial
Facultad de Ingeniería de Sistemas Computacionales

Maestría en Analítica de Datos.

“Predicción de Retrasos en Vuelos Comerciales utilizando Modelos de Clasificación”

Materia:

Modelos Predictivos

Profesor:

Prof. Juan Castillo, PhD

Presentado por:

Fernández, Angiel

8-897-592

Primer Avance

Año Lectivo:

2025

INTRODUCCIÓN

En el sector aeronáutico, los retrasos en los vuelos constituyen un problema persistente que impacta negativamente la organización operativa, la experiencia de los pasajeros y la rentabilidad de las aerolíneas. Gracias a la abundancia de datos históricos sobre vuelos, es posible aplicar herramientas de ciencia de datos para prever retrasos antes de que ocurran. Este proyecto tiene como objetivo desarrollar un modelo de predicción capaz de anticipar si un vuelo sufrirá demoras, proporcionando alertas tempranas que faciliten una mejor toma de decisiones por parte de los distintos actores involucrados.

Para enfrentar esta situación, muchas entidades del sector aéreo están adoptando el análisis predictivo, una metodología basada en datos que emplea tanto información histórica como en tiempo real, junto con modelos estadísticos y algoritmos de aprendizaje automático, para anticipar resultados y eventos futuros. El análisis predictivo puede ser útil en varios aspectos relacionados con la gestión de retrasos, tales como:

- Identificación de las causas y tendencias de los retrasos.
- Estimación de la probabilidad y duración de las demoras.
- Optimización de las decisiones y medidas para reducir el impacto de los retrasos. (FasterCapital, 2025)

Un propósito central del análisis predictivo es construir y probar modelos de machine learning que permitan predecir con precisión eventos futuros basados en datos del pasado. En este contexto, se trata de usar registros de vuelos anteriores para calcular la probabilidad de que un vuelo se retrase más allá de cierto límite, por ejemplo, 15 minutos. Esta información puede resultar clave para aerolíneas, aeropuertos y pasajeros al momento de planificar y reducir los efectos negativos de las demoras.

No obstante, desarrollar y validar modelos de predicción de retrasos en vuelos conlleva una serie de retos importantes, entre ellos:

- La calidad y disponibilidad de los datos.
- El procesamiento y preparación de los datos, así como la creación de variables relevantes.
- La selección adecuada de algoritmos y su entrenamiento.
- La evaluación y validación precisa del modelo.

Este proyecto tiene como objetivo construir un modelo de clasificación que prediga si un vuelo se retrasará o no, utilizando datos históricos del transporte aéreo, con el fin de generar alertas tempranas y apoyar la toma de decisiones en la industria aeronáutica.

JUSTIFICACIÓN

Los retrasos en los vuelos representan un problema significativo dentro del sector aeronáutico, con consecuencias tangibles tanto operativas como económicas. Por esta razón, contar con herramientas que permitan anticipar estas demoras resulta sumamente valioso. La elección de este tema se justifica por varios motivos:

- Su relevancia práctica en contextos reales.
- La existencia de una gran cantidad de datos públicos disponibles.
- La oportunidad de aplicar técnicas de depuración de datos, análisis exploratorio, visualización y modelos de aprendizaje supervisado.

Anticipar retrasos antes de que ocurran no solo mejora la experiencia de los pasajeros, sino que también ayuda a aeropuertos y aerolíneas a gestionar mejor sus recursos, ajustar sus procesos operativos y reducir pérdidas económicas. En una industria donde el tiempo es un factor crítico, una predicción precisa puede evitar cuellos de botella, mitigar efectos en cadena y aumentar la eficiencia general del sistema. Este proyecto se alinea con ese propósito al desarrollar un modelo de clasificación, fundamentado en datos reales y comprobables, con un enfoque orientado a la aplicación práctica y a la mejora operativa.

ANTECEDENTES

En los últimos años, el empleo de modelos predictivos en el sector aeronáutico ha experimentado un notable crecimiento, impulsado principalmente por la disponibilidad de grandes volúmenes de datos abiertos. Diversas investigaciones han evidenciado que variables como el horario de salida, el día de la semana, la aerolínea operadora y la distancia del vuelo influyen de manera significativa en la probabilidad de que se presenten retrasos.

Para el desarrollo de este proyecto, se ha optado por utilizar el conjunto de datos titulado *"2015 Flight Delays and Cancellations"*, disponible en la plataforma Kaggle. Este dataset incluye más de cinco millones de registros de vuelos comerciales realizados en Estados Unidos durante el año 2015, proporcionando una base sólida para la identificación de patrones relevantes y la construcción de un modelo predictivo basado en clasificación binaria.

Cabe destacar que el Departamento de Transporte de los Estados Unidos (U.S. Department of Transportation) recopila y publica de manera sistemática información detallada sobre vuelos, demoras, cancelaciones y desvíos, lo cual ha permitido el desarrollo de numerosos estudios que abordan la predicción de retrasos mediante técnicas de aprendizaje automático. Entre los algoritmos comúnmente empleados se encuentran la regresión logística, los árboles de decisión y las redes neuronales.

Si bien el objetivo final de este proyecto es explorar la posibilidad de aplicar estas soluciones en el contexto panameño, actualmente no se dispone de datos abiertos equivalentes a los utilizados en los estudios internacionales. Por esta razón, se ha optado por emplear el dataset anteriormente mencionado como punto de partida para el diseño, entrenamiento y validación del modelo predictivo.

DEFINICIÓN DEL PROBLEMA

La puntualidad en los vuelos es un aspecto clave para el buen funcionamiento del sector aeronáutico. No solo afecta la eficiencia operativa de aerolíneas y aeropuertos, sino también la experiencia de los pasajeros, quienes dependen cada vez más de un servicio confiable y predecible.

A partir de esto, surge la siguiente pregunta que da origen a este proyecto: **¿Podemos anticipar si un vuelo se retrasará, utilizando únicamente los datos disponibles antes del despegue?**

Responder esta pregunta permitiría desarrollar un sistema que clasifique los vuelos en dos categorías: a tiempo (0) o retrasado (1). Para este estudio, se considerará como “retrasado” cualquier vuelo cuya salida se postergue más de 15 minutos, siguiendo criterios estándar en la industria.

El objetivo es poder generar alertas tempranas, antes del abordaje, que ayuden a los actores involucrados —aerolíneas, aeropuertos y pasajeros— a anticiparse y tomar decisiones oportunas. Esto podría traducirse en una mejor asignación de recursos, reducción de retrasos en cadena y mayor satisfacción del usuario.

Para lograrlo, se evaluarán distintos modelos de clasificación supervisada, como regresión logística, CatBoost, Random Forest, entre otros. Cada modelo será entrenado con datos reales y comparado con base en métricas como precisión, recall, F1-score y AUC, con el fin de seleccionar la alternativa más efectiva y aplicable en un entorno operativo real.

AVANCE DE ANÁLISIS PREDICTIVO

Se seleccionó el dataset “Flight Delays and Cancellations” de Kaggle, que contiene más de 5 millones de registros y 31 variables. Hasta el momento se han realizado los siguientes pasos:

- Limpieza del dataset:
  - ❖ Se eliminaron columnas con más del 80% de valores nulos.
  - ❖ Se filtraron vuelos cancelados y desviados.
  - ❖ Se eliminaron columnas irrelevantes para el modelo, como número de vuelo o número de cola.
- Se creó la variable objetivo DELAYED, con valor 1 si el vuelo se retrasó igual o más de 15 minutos, y 0 si fue puntual (menos de 15 min)
- Se identificó un ligero desbalance de clases: alrededor del 18.6 % de los vuelos presentan algún tipo de retraso significativo.
- Se eliminaron las variables que no tienen relevancia con el vuelo antes del despegue, dejando así las variables:
  - ❖ MONTH: evidencian variaciones estacionales y semanales en los patrones de retraso, reflejando diferencias en la demanda y condiciones climáticas según la época del año.
  - ❖ DAY: reflejan variaciones estacionales y semanales, ya que ciertos días pueden tener mayor tráfico o condiciones operativas diferentes.
  - ❖ DAY_OF_WEEK: muestran fluctuaciones a lo largo de la semana, donde algunos días presentan tasas más altas de retrasos debido a la concentración de vuelos o eventos específicos.
  - ❖ AIRLINE: algunas aerolíneas presentan tasas de retraso sistemáticamente mayores, lo que puede estar relacionado con su gestión operativa, rutas o recursos disponibles.
  - ❖ FLIGHT_NUMBER: está en revisión, ya que puede reflejar patrones específicos de vuelo, pero su utilidad como predictor puede variar según la consistencia de la operación.
  - ❖ ORIGIN_AIRPORT: el aeropuerto de origen influye en la probabilidad de retrasos debido a factores como congestión, infraestructura y condiciones climáticas locales.
  - ❖ DESTINATION_AIRPORT: el aeropuerto de destino también impacta los retrasos, especialmente si existen limitaciones operativas o meteorológicas que afectan la llegada.
  - ❖ SCHEDULED_DEPARTURE: el horario programado de salida incide en la probabilidad de retraso, dado que ciertos períodos del día son más propensos a congestiones o condiciones adversas.

- ❖ SCHEDULED_TIME: la duración planificada del vuelo puede afectar el margen para absorber demoras y, por ende, la probabilidad de llegar retrasado.
- ❖ DISTANCE: la distancia del trayecto es un factor clave, ya que vuelos más largos suelen tener mayor capacidad para recuperar retrasos, mientras que los cortos son más vulnerables a demoras acumulativas.
- ❖ SCHEDULED_ARRIVAL: la hora prevista de llegada puede relacionarse con la congestión en aeropuertos y ventanas horarias críticas que afectan la puntualidad.
- ❖ DELAYED: variable objetivo binaria que indica si el vuelo se retrasó 15 minutos o más (1) o si llegó a tiempo (0).

- Se trabajó en el análisis exploratorio con todos los datos.

Estadísticas descriptivas de las variables.

Se realizó un análisis estadístico básico sobre las variables seleccionadas del dataset, con un total de 5,714,008 registros válidos como se observa en la figura 1.

Estadísticas descriptivas:

	MONTH	DAY	DAY_OF_WEEK	FLIGHT_NUMBER	\
count	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06	
mean	6.547799e+00	1.570759e+01	3.932643e+00	2.164384e+03	
std	3.397421e+00	8.774394e+00	1.985967e+00	1.754706e+03	
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	
25%	4.000000e+00	8.000000e+00	2.000000e+00	7.280000e+02	
50%	7.000000e+00	1.600000e+01	4.000000e+00	1.681000e+03	
75%	9.000000e+00	2.300000e+01	6.000000e+00	3.211000e+03	
max	1.200000e+01	3.100000e+01	7.000000e+00	9.320000e+03	

	SCHEDULED_DEPARTURE	SCHEDULED_TIME	DISTANCE	SCHEDULED_ARRIVAL
count	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06
mean	1.328907e+03	1.418940e+02	8.244569e+02	1.493187e+03
std	4.835251e+02	7.531400e+01	6.086620e+02	5.069011e+02
min	1.000000e+00	1.800000e+01	3.100000e+01	1.000000e+00
25%	9.160000e+02	8.500000e+01	3.730000e+02	1.110000e+03
50%	1.325000e+03	1.230000e+02	6.500000e+02	1.520000e+03
75%	1.730000e+03	1.740000e+02	1.065000e+03	1.917000e+03
max	2.359000e+03	7.180000e+02	4.983000e+03	2.400000e+03

	ARRIVAL_DELAY	DELAYED
count	5.714008e+06	5.714008e+06
mean	4.407057e+00	1.861109e-01
std	3.927130e+01	3.891961e-01
min	-8.700000e+01	0.000000e+00
25%	-1.300000e+01	0.000000e+00
50%	-5.000000e+00	0.000000e+00
75%	8.000000e+00	0.000000e+00
max	1.971000e+03	1.000000e+00

Figura 1. Estadísticas descriptivas de las variables depuradas

A continuación, se resumen los principales hallazgos:

- **Variables temporales:**
    - *MONTH* tiene un promedio cercano a 6.5, con valores entre 1 y 12, representando la distribución anual de vuelos.
    - *DAY* presenta una media de 15.7, abarcando todos los días del mes (1 a 31).
    - *DAY_OF_WEEK* varía de 1 a 7, con un promedio de 3.93, reflejando la distribución semanal de los vuelos.
- **Características del vuelo:**
    - *FLIGHT_NUMBER* muestra una amplia variabilidad, con un promedio de 2164 y un rango de 1 a 9320, lo que indica una gran diversidad de rutas y operaciones.
    - *SCHEDULED_DEPARTURE* y *SCHEDULED_ARRIVAL* presentan valores que corresponden al formato horario (en formato 24 horas, por ejemplo, 1325 representa las 13:25). Sus medias son

aproximadamente 1329 y 1493 respectivamente, indicando la concentración de vuelos en distintos momentos del día.

    - *SCHEDULED_TIME* (duración programada del vuelo) tiene una media de 142 minutos, con valores que van desde 18 hasta 718 minutos, lo que abarca vuelos cortos y largos.
    - *DISTANCE* varía ampliamente desde 31 hasta casi 5000 millas, con una media alrededor de 824 millas.

- **Variable objetivo y retrasos:**
    - *ARRIVAL_DELAY* tiene una media positiva de 4.4 minutos, pero con una gran desviación estándar (39.3), indicando que aunque muchos vuelos llegan a tiempo o temprano, existen retrasos muy largos en ciertos casos. El rango va desde -87 minutos (llegadas adelantadas) hasta casi 2000 minutos de retraso.
    - La variable binaria *DELAYED* indica que aproximadamente un 18.6% de los vuelos se retrasaron 15 minutos o más.

Este análisis preliminar confirma la diversidad y riqueza del dataset, y proporciona una base sólida para continuar con la selección de características, tratamiento de datos y construcción de modelos predictivos.

GRÁFICAS DE VISUALIZACIÓN

En la figura 2, se observa la gráfica Arrival_delay por Aerolinea (antes de binarizar), la misma muestra los retrasos en la llegada por aerolínea, en valores continuos (antes de transformarlo en puntual/retrasado). Y podemos observar que:

- Aerolíneas como WN (Southwest), AA (American Airlines) y DL (Delta) presentan altas concentraciones de retrasos acumulados.
- En cambio, HA (Hawaiian Airlines) y VX (Virgin America) muestran una mejor puntualidad.
- Esto puede reflejar eficiencia operativa, ubicación de hubs (aeropuertos base), o estrategias de planificación de vuelos.

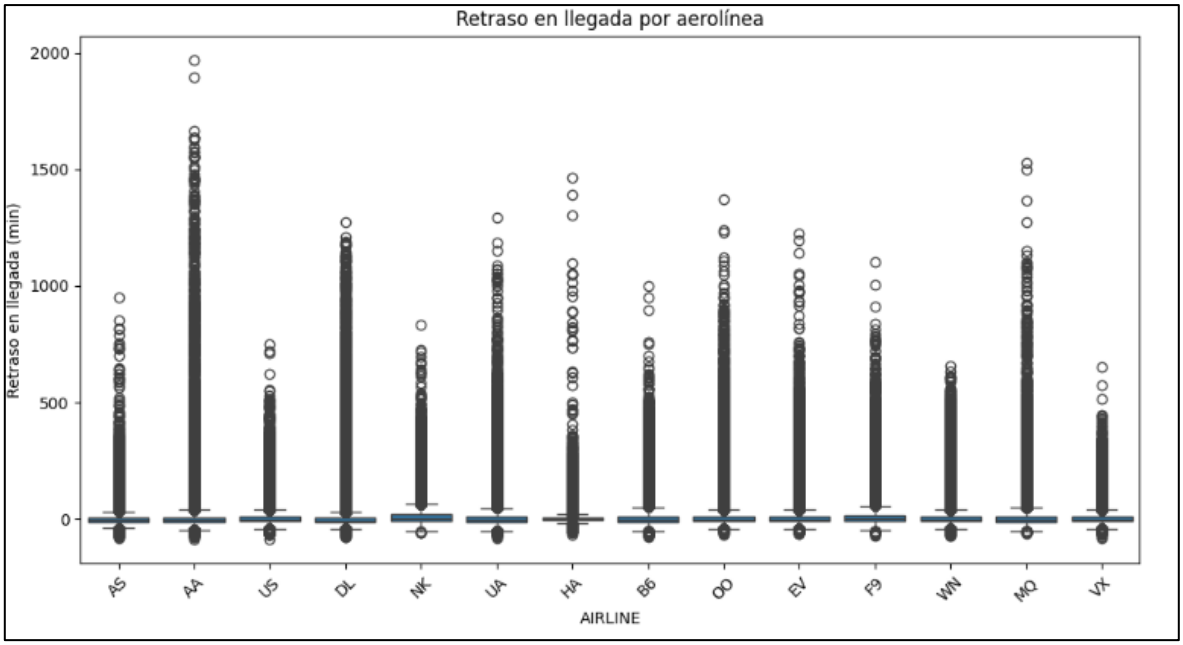

Figura 2. Grafica de arrival delay por aerolínea (antes de binarizar).

En la figura 3, observamos la relación entre la hora de salida y retraso en la llegada esto nos muestra como la hora del día en que sale un vuelo afecta su probabilidad de llegar retrasado. Podemos observar que:

- Vuelos temprano en la mañana tienen menor probabilidad de retrasarse, ya que inician el ciclo operativo del día.
- A medida que avanza el día, aumentan los retrasos, probablemente por efecto acumulado de demoras anteriores.

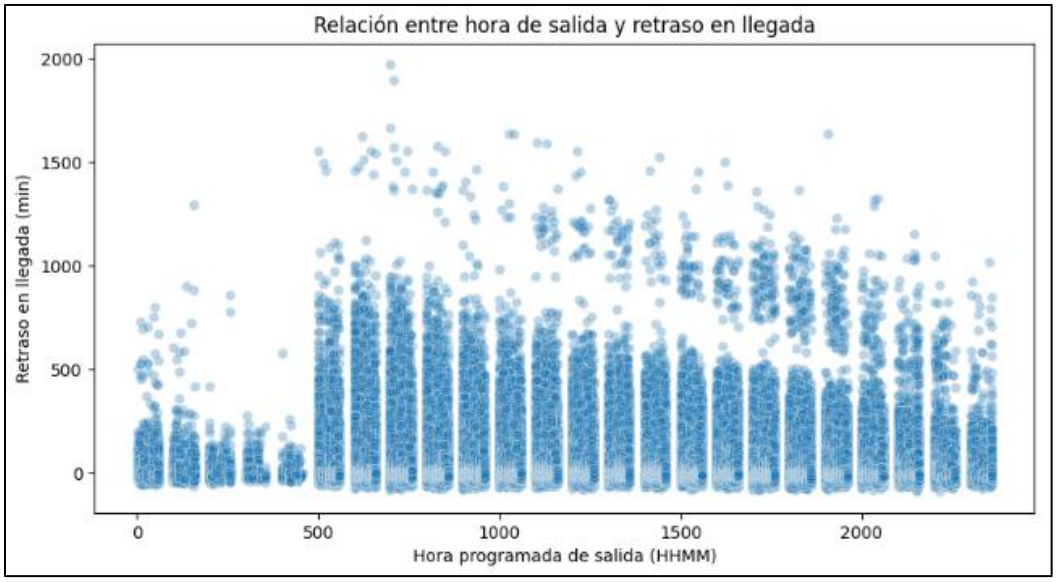

Figura 3. Relación entre la hora de salida y retraso en llegada

En la figura 4, se encuentra la gráfica de relación entre la distancia y el retraso en la llegada, la misma muestra cómo se relaciona la distancia del vuelo con el nivel de retraso en la llegada. Se observa que:

- Vuelos más largos pueden presentar mayores retrasos acumulados, pero también tienen margen para recuperar tiempo.
- Vuelos cortos, aunque parecen más fáciles de manejar, pueden verse más afectados por congestión aeroportuaria o condiciones climáticas locales.

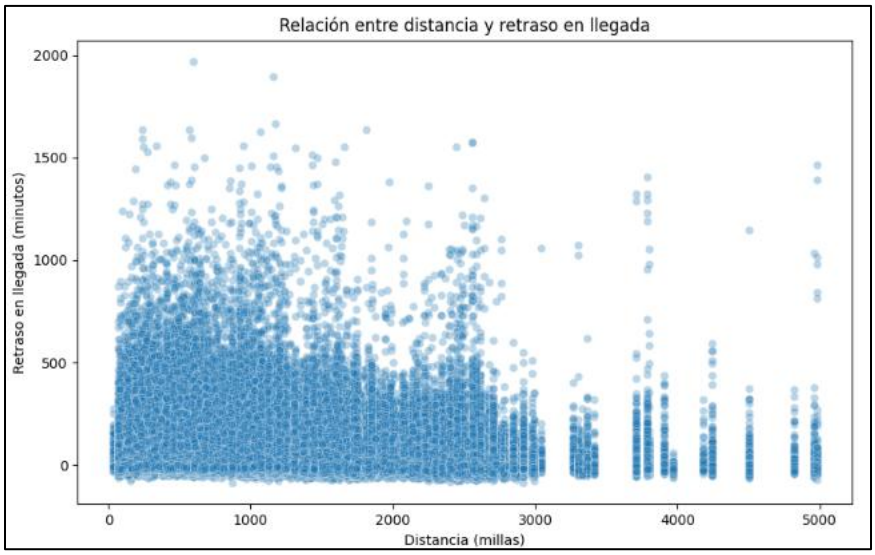

Figura 4. Relación entre la distancia y retraso en llegada.

En la figura 5, se muestra la gráfica de distribución de vuelos, nos indica como están distribuidos los vuelos en el dataset al analizar esta grafica podemos ver un leve desbalance entres los vuelos que se encuentran a tiempo (81.4%) y los retrasados (18.6%)

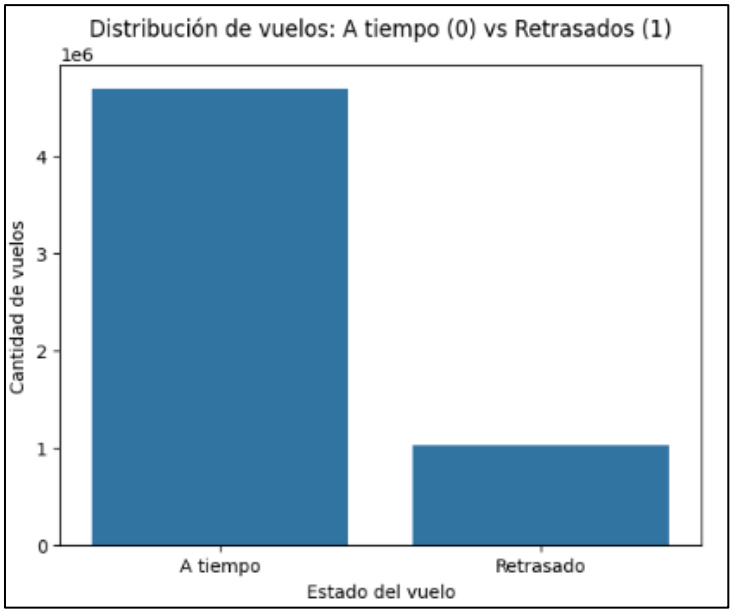

Figura 5. Gráfica de distribución de vuelos

En la figura 6, podemos visualizar los vuelos retrasados por día de la semana siendo el jueves el día con más vuelos retrasados, posiblemente por acumulación de operaciones semanales, y el sábado es el día con menor retraso, lo cual concuerda con una reducción en vuelos comerciales corporativos ese día.

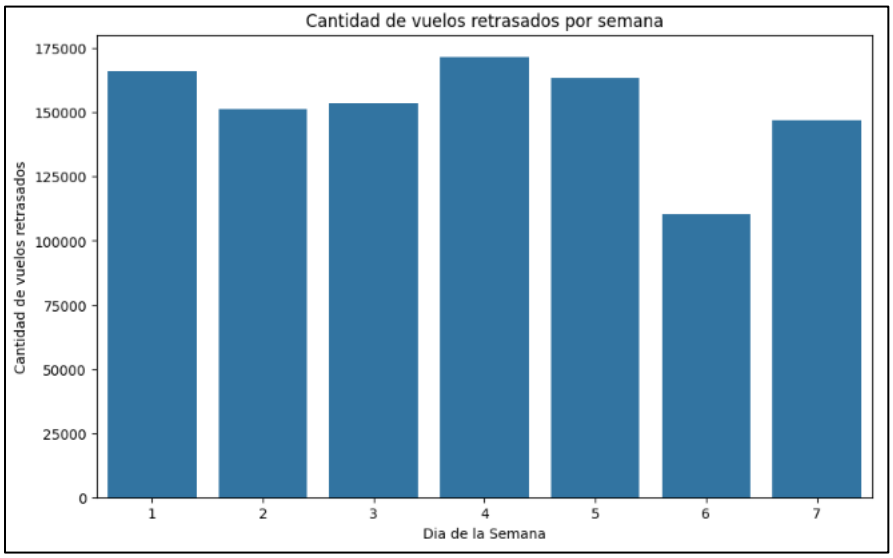

Figura 6. Cantidad de vuelos retrasados por semana.

En la figura 7, se observan los vuelos retrasados por mes, los mismos muestran que en Verano (junio-julio) registra más retrasos, por el incremento de tráfico vacacional y condiciones climáticas adversas; en Otoño (septiembre-octubre) es más estable, con menor volumen de vuelos y menos afectaciones por clima. La estacionalidad es evidente, lo que permite ajustar las expectativas del modelo según el mes.

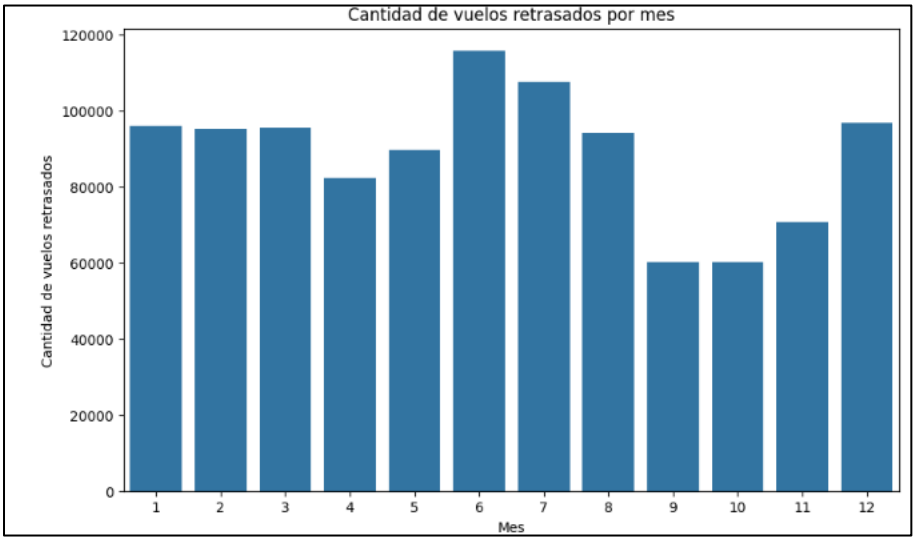

Figura 7. Cantidad de vuelos retrasados por mes.

En la Figura 8, observamos la cantidad de vuelos retrasados por aerolínea, las aerolíneas más grandes y con más vuelos (WN(Southwest Airlines Co.), AA (American Airlines Inc.)) naturalmente tienen más retrasos. También refleja posibles ineficiencias operativas o hubs congestionados. Aerolíneas pequeñas como HA (Hawaiian Airlines Inc.) operan en regiones menos congestionadas, lo cual mejora su puntualidad. Esta variable tiene gran potencial como predictor.

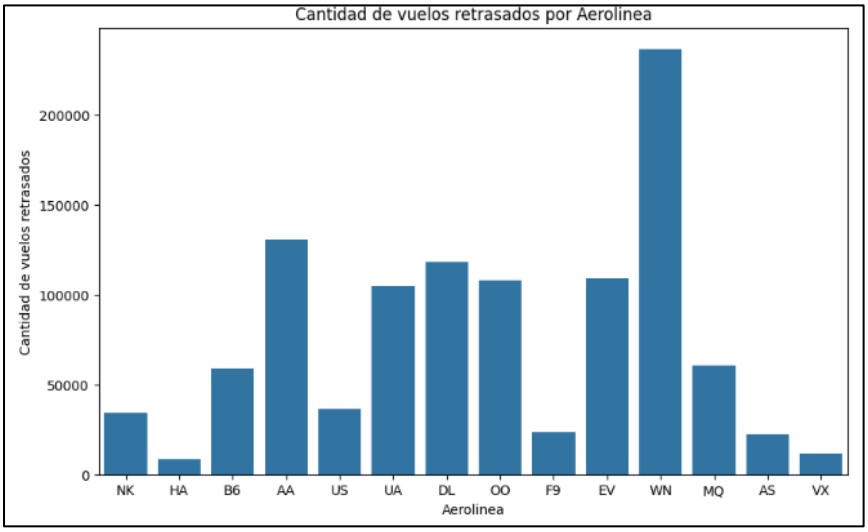

Figura 8. Cantidad de vuelos retrasados por Aerolinea.

En la figura 9, se muestra la distribución de la hora programada de salida según los retrasos, no es mas que la distribución de vuelos retrasados/puntuales según su hora programada. Con esta gráfica se confirma que los vuelos matutinos son más puntuales. Además de la probabilidad de retraso aumenta significativamente en la tarde y noche. Esta variable es altamente predictiva y operacionalmente útil.

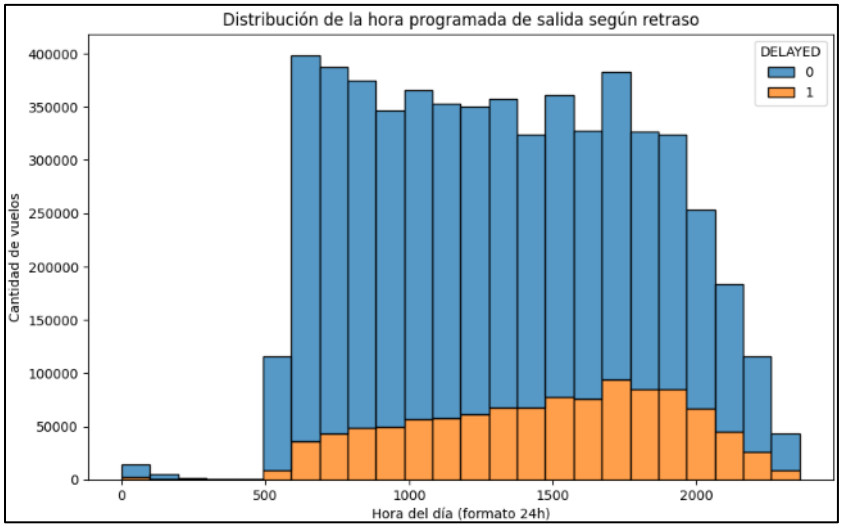

Figura 9. Distribución de la hora programada de salida según retrasos.

En la grafica 10, vamos el Boxplot de distribución de la distancia en relación a el estado del vuelo, mejor dicho si la distancia del vuelo está relacionada con estar retrasado o no, la misma nos podría indicar que ciertos rangos de distancia (intermedios) son más propensos a retrasos. A su vez es útil para identificar tramos de ruta vulnerables o sujetos a congestión específica.

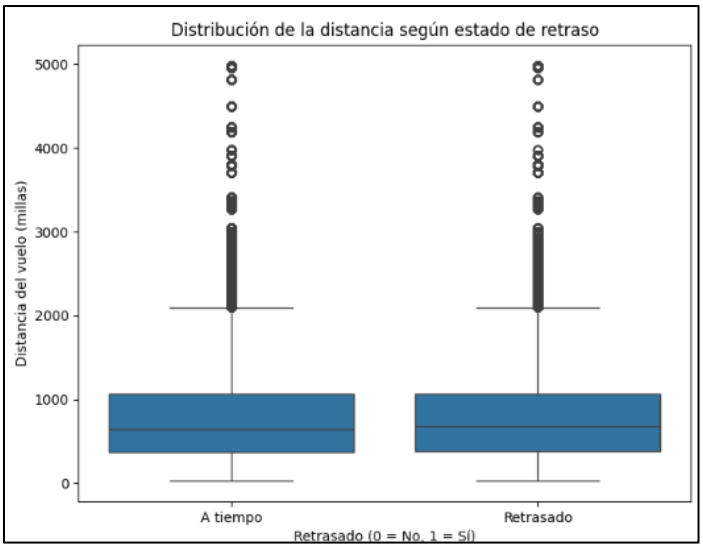

Figura 10. Grafica de distancia vs. El estado del vuelo

En la figura 11, observamos la grafica de distancia versus el tiempo programado la misma muestra la relación entre la distancia del vuelo y la duración programada, es una relación lineal a mayor distancia, mayor duración, con ella se verifica la coherencia interna del dataset.

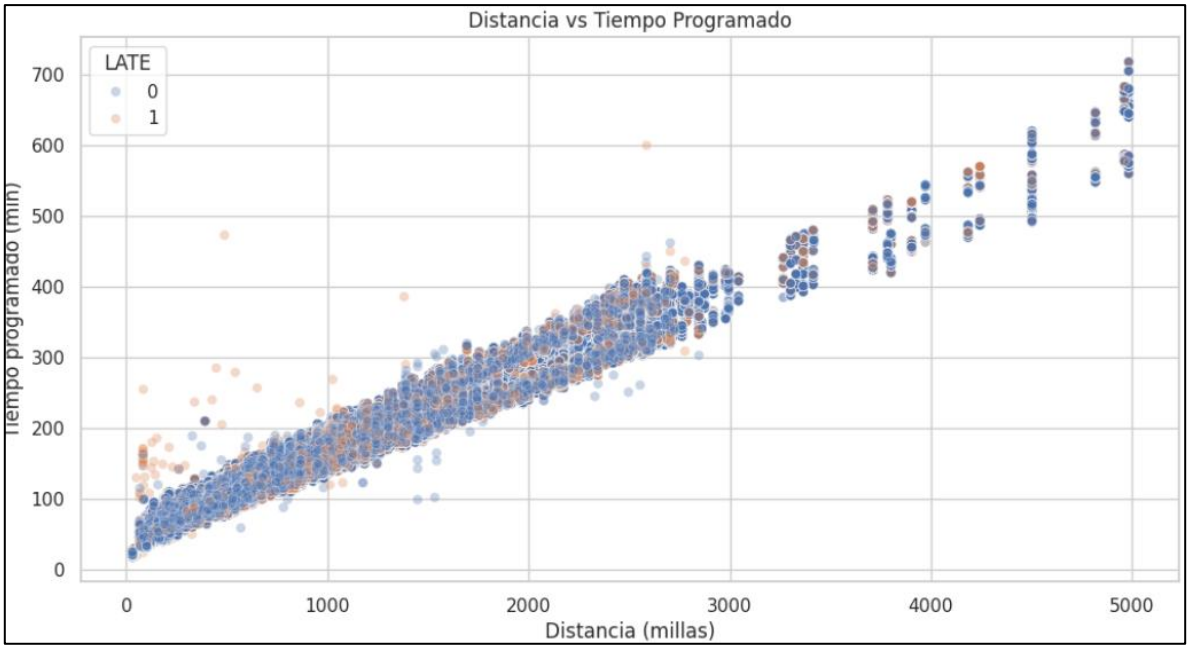

Figura 11. Grafica distancia versus tiempo programado

En la figura 12, se observa la gráfica de distribución de la duración programada del vuelo, nos muestra como se distribuyen los vuelos según si duración estimada, la mayoría de los vuelos tienen duraciones entre 60 y 180 min, esto permite ajustar el análisis para vuelos cortos, medios y largos de ser necesario.

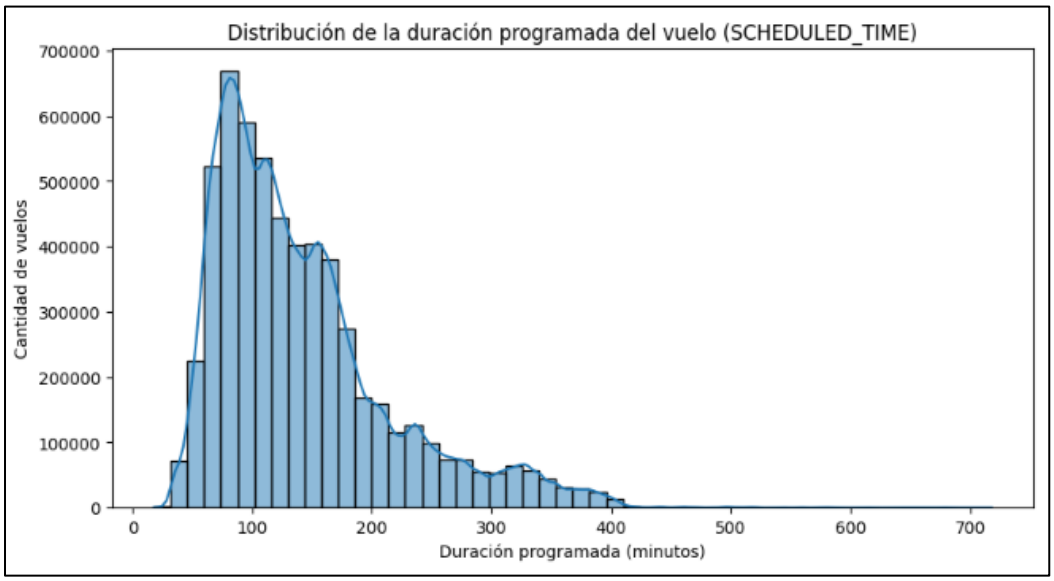

Figura 12. Distribución de la duración programada del vuelo

En la figura 13, se muestra la distribución de vuelos por hora del día, vemos en que horarios se concentra la mayor cantidad de vuelos, hora pico o: mañana (6-9 a.m.) y tarde (4-7 p.m.). Esto coincide con mayores índices de retraso durante estos periodos.

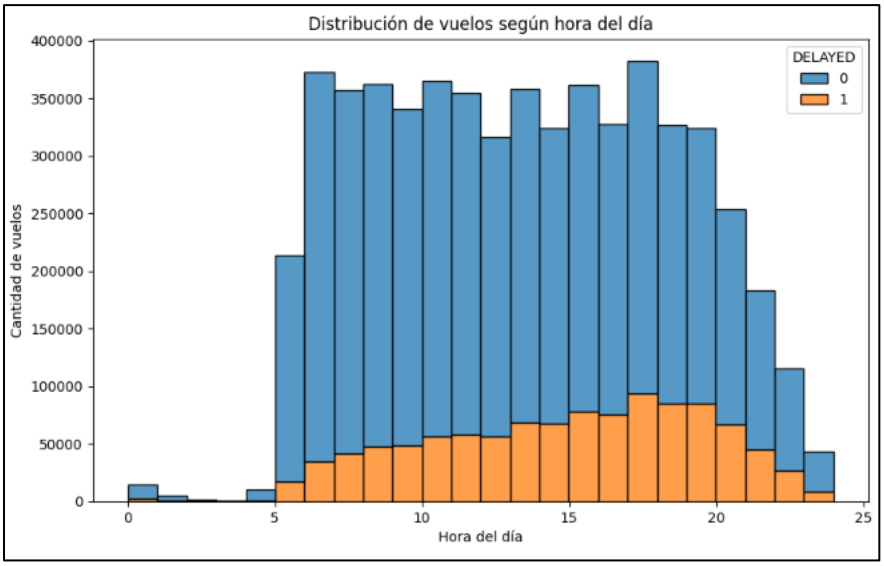

Figura 13. Distribución de vuelos según la hora del día.

En la figura 14, se observa la distribución de distancia y retrasos, básicamente es como se distribuyen los retrasos según la distancia del vuelo, ciertos rangos de distancia (por ejemplo, entre 500 y 1000 millas) podrían tener mayor concentración de retrasos. También permite entender qué tipo de rutas son más propensas a problemas operativos.

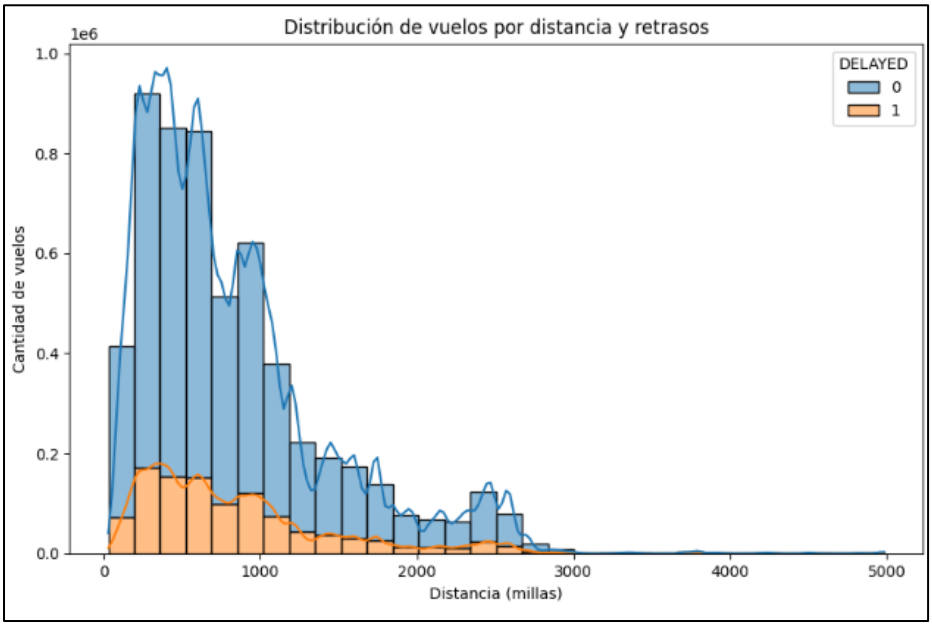

Figura 14. Distribución de vuelos por distancia y retrasos

Por último en la figura 15, se muestra la matriz de correlación después de la depuración de variables podemos observar una fuerte correlación entre departure delay y arrival delay (esperado), variables como scheduled departurey distance también muestran relaciones útiles. Esta matriz ayuda a seleccionar variables predictoras clave para el modelo.

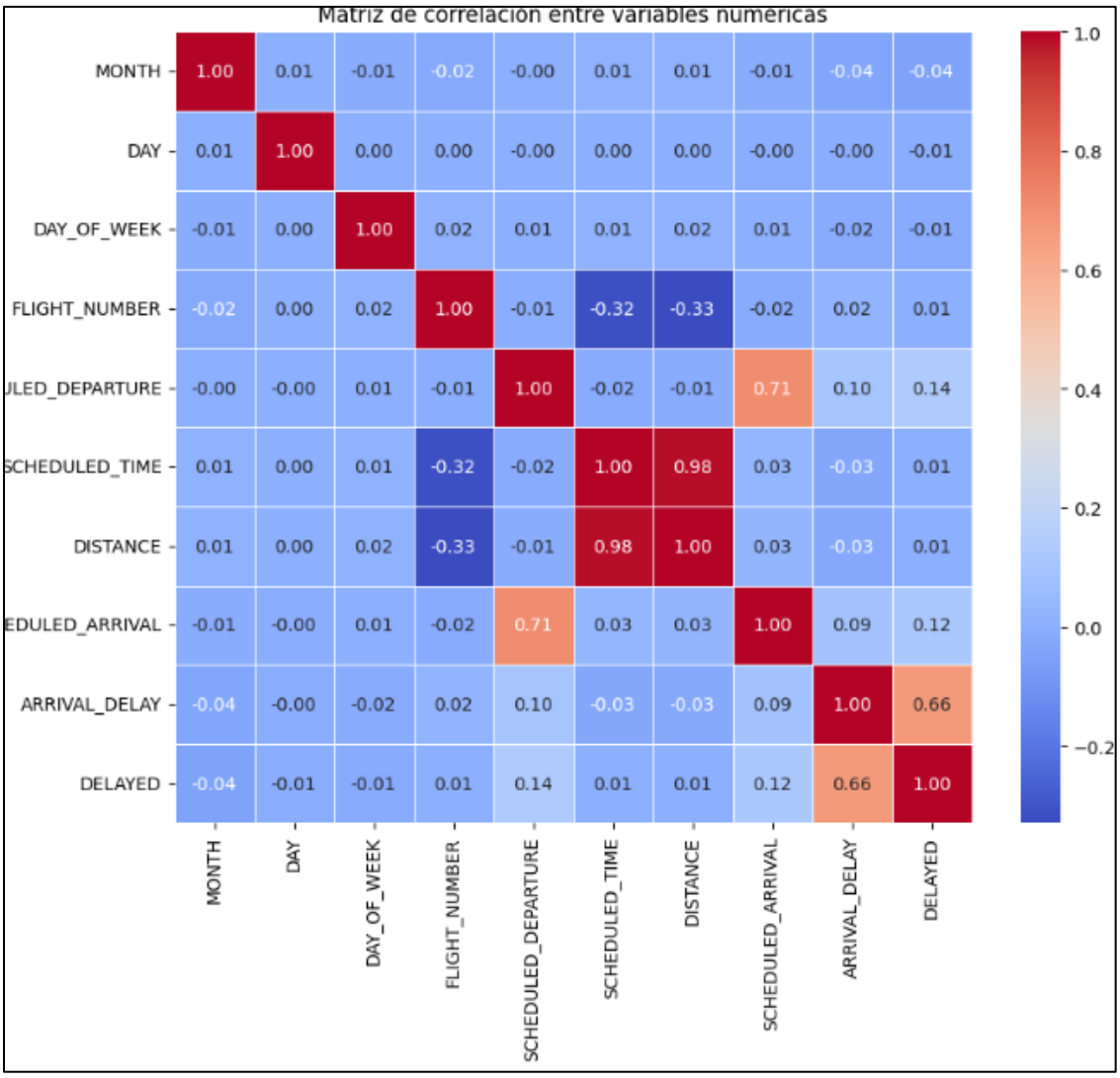

Figura 15. Matriz de Correlación de las Variables numéricas después de la depuración.

Las gráficas muestran patrones claros y consistentes en el comportamiento de los retrasos:

- Hora de salida, día de la semana, mes, aerolínea y distancia son variables fuertemente asociadas con los retrasos.

- El análisis visual respalda la viabilidad de construir un modelo de clasificación binaria robusto y útil operativamente.

Pasos a Seguir

- Entrenar Modelo
- Evaluar los diferentes modelos sin balancear o balanceados
- Elaborar Reporte