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

Proyecto Final

Año Lectivo:

2025

INDICE

INTRODUCCIÓN

En la industria aeronáutica, los retrasos en vuelos representan un desafío constante, afectando la planificación operativa, la satisfacción de los pasajeros y la eficiencia económica de las aerolíneas. Aprovechando el gran volumen de datos históricos disponibles sobre vuelos, es posible aplicar técnicas de ciencia de datos para anticipar eventos de retraso antes de que ocurran. Este proyecto busca implementar un modelo predictivo que permita detectar de forma anticipada si un vuelo se retrasará, brindando alertas tempranas que mejoren la toma de decisiones por parte de los actores involucrados.

Para abordar este problema, muchas partes interesadas de la industria de la aviación están recurriendo al análisis predictivo, un enfoque basado en datos que utiliza datos históricos y en tiempo real, modelos estadísticos y algoritmos de aprendizaje automático para pronosticar resultados y eventos futuros. El análisis predictivo puede ayudar con varios aspectos de la gestión de retrasos en los vuelos, como:

- Identificar las causas y patrones de los retrasos en los vuelos.
- Estimación de la probabilidad y duración de los retrasos en los vuelos.
- optimizar las decisiones y acciones para mitigar los retrasos de los vuelos. (FasterCapital, 2025)

Uno de los principales objetivos del análisis predictivo es crear y evaluar modelos de aprendizaje automático que puedan predecir con precisión el resultado de un evento futuro basándose en datos históricos. En el contexto de los retrasos en los vuelos, esto significa utilizar datos de vuelos anteriores para estimar la probabilidad de que un vuelo se retrase más de un determinado umbral, como 15 minutos. Esto puede ayudar a las aerolíneas, los aeropuertos y los pasajeros a planificar con anticipación y mitigar el impacto de los retrasos.

Sin embargo, construir y evaluar modelos predictivos de retrasos en los vuelos no es una tarea trivial. Hay muchos desafíos y consideraciones que deben abordarse, tales como:

- Calidad y disponibilidad de los datos
- Preprocesamiento de datos e ingeniería de características
- Selección y entrenamiento de modelos
- Evaluación y validación del modelo

OBJETIVO

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

El problema de los retrasos aéreos tiene implicaciones reales en la industria de la aviación, por lo que contar con herramientas que permitan anticiparlos es de gran utilidad. La elección de este tema responde a:

- Su aplicabilidad en el mundo real.
- La disponibilidad de datos abiertos y abundantes.
- La posibilidad de integrar técnicas de limpieza, análisis descriptivo, visualización y modelos supervisados.

Predecir retrasos antes de que ocurran no solo mejora la experiencia del usuario, sino que permite a aeropuertos y aerolíneas optimizar la asignación de recursos, ajustar operaciones y reducir pérdidas. En un sector donde cada minuto cuenta, una predicción acertada puede prevenir congestiones, minimizar demoras en cadena y mejorar la eficiencia global del sistema. Este proyecto contribuye a ese objetivo mediante la creación de un modelo de clasificación basado en datos reales y verificables, con enfoque práctico y utilidad 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.

ANÁLISIS PREDICTIVO

- ## DETERMINACIÓN DE LA BASE DE DATOS

Se seleccionó el dataset “Flight Delays and Cancellations” de Kaggle, (https://www.kaggle.com/datasets/usdot/flight-delays/data) que contiene más de 5 millones de registros y 31 columnas: (Kaggle, s.f.)

1. YEAR: 2015. Data Type is int64.
2. MONTH: 1 (Enero) - 12 (Diciembre). Data Type is int64.
3. DAY: Dia de los meses (1 - 31). Data Type is int64.
4. DAY_OF_WEEK: Día de la semana (1 (lunes) - 7 (Domingo)). Data Type is int64.
5. AIRLINE: Código único de aerolínea (código IATA de la compañía aérea) Data Type is object.
6. FLIGHT_NUMBER: Número del Vuelo. Data Type is int64.
7. TAIL_NUMBER: Número de cola del avión: matrícula de la aeronave, identificador único de la aeronave. Data Type is object.
8. ORIGIN_AIRPORT: Origen Código de identificación IATA del aeropuerto. Data Type is object.
9. DESTINATION_AIRPORT: Código de identificación IATA del aeropuerto de destino. Data Type is object.
10. SCHEDULED_DEPARTURE: Hora de salida programada (despegue) (hora local, hhmm). Data Type is int64.
11. DEPARTURE_TIME: Hora de salida real (hora local, hhmm). Hora real de despegue. Data Type is float64.
12. DEPARTURE_DELAY: Retraso en la salida, en minutos. Diferencia entre la hora de despegue prevista y la hora de despegue real. Data Type is float64.
13. TAXI_OUT: Tiempo de salida del taxi, en minutos. Tiempo de rodaje en la maniobra de despegue (hasta que las ruedas del avión se separan del suelo).Data Type is float64.
14. WHEELS_OFF: Inicio del tiempo de vuelo, en minutos El momento en que las ruedas del avión se separan del suelo durante el despegue. Data Type is float64.
15. SCHEDULED_TIME: Tiempo de vuelo programado, en minutos. Data Type is float64.
16. ELAPSED_TIME: Tiempo real de vuelo, en minutos (ELAPSED_TIME = TAXI_OUT + AIR_TIME + TAXI_IN). Data Type is float64.
17. AIR_TIME: Tiempo de vuelo de la aeronave, en minutos. Data Type is float64.
18. DISTANCE: Distancia de vuelo, en millas. Data Type is int64
19. WHEELS_ON: Fin del tiempo de vuelo, en minutos. El momento en que las ruedas del avión tocan el suelo al aterrizar. Data Type is float64.

20. TAXI_IN: Tiempo de maniobra de aterrizaje en taxi, en minutos. Tiempo transcurrido entre el aterrizaje y la llegada a la puerta de embarque del aeropuerto de destino. Data Type is float64.
21. SCHEDULED_ARRIVAL: Hora prevista de llegada (aterrizaje) (hora local, hhmm). Data Type is int64.
22. ARRIVAL_TIME: Hora de llegada real (hora local, hhmm). Hora real de aterrizaje. Data Type is float64.
23. ARRIVAL_DELAY: Retraso en la llegada, en minutos. Diferencia entre la hora de aterrizaje prevista y la real. Se considera que un vuelo es «puntual» si opera con menos de 15 minutos de retraso respecto a la hora de llegada prevista que figura en los sistemas informatizados de reservas (CRS) de la compañía aérea. Data Type is float64.
24. DIVERTED: ¿Se desvió el vuelo? 0 = NO, 1 = YES. Data Type is int64.
25. CANCELLED: ¿Se canceló el vuelo? 0 = NO, 1 = YES. Data Type is int64.
26. CANCELLATION_REASON: Motivo de la cancelación (A = Aerolínea, B = Condiciones meteorológicas extremas, C = Sistema aéreo, D = Seguridad). Data Type is object.
27. AIR_SYSTEM_DELAY: Retraso del Sistema Nacional del Espacio Aéreo (NAS) en minutos. Los retrasos atribuibles al Sistema Nacional del Espacio Aéreo (NAS) pueden incluir las siguientes condiciones: condiciones meteorológicas no extremas, operaciones aeroportuarias, volumen de tráfico intenso y control del tráfico aéreo. Data Type is float64.
28. SECURITY_DELAY: Retraso por motivos de seguridad en minutos. El retraso por motivos de seguridad se debe a la evacuación de una terminal, el reembarque de un avión debido a una infracción de seguridad, el mal funcionamiento del equipo de control y/o las largas colas de más de 29 minutos en las zonas de control. Data Type is float64.
29. AIRLINE_DELAY: Retraso de la aerolínea en minutos. La aerolínea es responsable de este retraso. Algunos ejemplos que pueden causar estos retrasos son: limpieza de la aeronave, daños en la aeronave, espera de la llegada de pasajeros o tripulación en conexión, equipaje, colisión con aves, carga de mercancías, catering, avería del equipo de la aerolínea, legalidad de la tripulación (descanso del piloto o del personal de cabina), daños causados por mercancías peligrosas, inspección técnica, repostaje, asistencia a pasajeros con discapacidad, retraso de la tripulación, limpieza de los aseos, mantenimiento, sobreventa, servicio de agua potable, expulsión de pasajeros conflictivos, embarque o asignación de asientos lentos, almacenamiento del equipaje de mano, retrasos por peso y equilibrio. Data Type is float64.
30. LATE_AIRCRAFT_DELAY: Retraso tardío de la aeronave en minutos. Estos retrasos se producen en el aeropuerto de destino debido a la llegada tardía de la misma aeronave al aeropuerto anterior. El efecto dominó de un retraso

anterior en los aeropuertos posteriores se denomina propagación del retraso. Data Type is float64.

31. WEATHER_DELAY: Retraso por condiciones meteorológicas en minutos. El retraso por condiciones meteorológicas se debe a condiciones meteorológicas extremas o peligrosas que se prevén o se manifiestan en el punto de partida, en la ruta o en el punto de llegada, y que impiden volar. Data Type is float64.

Adicional cuenta con dos bases de datos complementarias las cuales son Airlines.csv (contiene el código de la aerolínea y el nombre de la aerolínea) esto se muestra en la tabla 1 y airport.csv (contiene códigos de la aerolínea, nombre del aeropuerto, e información geográfica como longitud, latitud ciudad, estado) para este estudio no las vamos a utilizar, pero es relevante si queremos llevar nuestro proyecto a herramientas de visualización.

Tabla 1. Aerolíneas en el Dataset asociado con la IATA.

IATA CODE	AIRLINE
UA	United Airlines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Airlines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
EV	Atlantic Southeast Airlines
HA	Hawaiian Airlines Inc.
MQ	American Eagle Airlines Inc.
VX	Virgin America

- **PREPROCESAMIENTO Y LIMPIEZA**

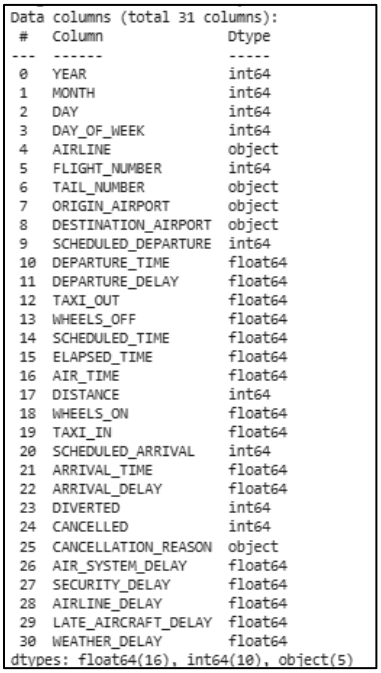
Antes de comenzar con la limpieza del dataset se realizó un análisis exploratorio del mismo, verificando si mantiene valores nulos, como se muestra en la figura 1, también verificamos el tipo de datos como se observa en la figura 2.

```
#Valores Nulos Dataset flight
df.isnull().sum()
```

	0
YEAR	0
MONTH	0
DAY	0
DAY_OF_WEEK	0
AIRLINE	0
FLIGHT_NUMBER	0
TAIL_NUMBER	14721
ORIGIN_AIRPORT	0
DESTINATION_AIRPORT	0
SCHEDULED_DEPARTURE	0
DEPARTURE_TIME	86153
DEPARTURE_DELAY	86153
TAXI_OUT	89047
WHEELS_OFF	89047
SCHEDULED_TIME	6
ELAPSED_TIME	105071
AIR_TIME	105071
DISTANCE	0
WHEELS_ON	92513
TAXI_IN	92513
SCHEDULED_ARRIVAL	0
ARRIVAL_TIME	92513
ARRIVAL_DELAY	105071
DIVERTED	0
CANCELLED	0
CANCELLATION_REASON	5729195
AIR_SYSTEM_DELAY	4755640
SECURITY_DELAY	4755640
AIRLINE_DELAY	4755640
LATE_AIRCRAFT_DELAY	4755640
WEATHER_DELAY	4755640

Figura 1. Valores nulos de todos los datos

Podemos observar varias variables con valores nulos, las cuales tendremos que trabajar en la limpieza



```
Data columns (total 31 columns):
 #   Column               Dtype
---  ------               -----
 0   YEAR                 int64
 1   MONTH                int64
 2   DAY                  int64
 3   DAY_OF_WEEK          int64
 4   AIRLINE              object
 5   FLIGHT_NUMBER        int64
 6   TAIL_NUMBER          object
 7   ORIGIN_AIRPORT       object
 8   DESTINATION_AIRPORT  object
 9   SCHEDULED_DEPARTURE  int64
 10  DEPARTURE_TIME       float64
 11  DEPARTURE_DELAY      float64
 12  TAXI_OUT             float64
 13  WHEELS_OFF           float64
 14  SCHEDULED_TIME       float64
 15  ELAPSED_TIME         float64
 16  AIR_TIME             float64
 17  DISTANCE             int64
 18  WHEELS_ON            float64
 19  TAXI_IN              float64
 20  SCHEDULED_ARRIVAL    int64
 21  ARRIVAL_TIME         float64
 22  ARRIVAL_DELAY        float64
 23  DIVERTED             int64
 24  CANCELLED            int64
 25  CANCELLATION_REASON  object
 26  AIR_SYSTEM_DELAY     float64
 27  SECURITY_DELAY       float64
 28  AIRLINE_DELAY        float64
 29  LATE_AIRCRAFT_DELAY  float64
 30  WEATHER_DELAY        float64
dtypes: float64(16), int64(10), object(5)
```

Figura 2. Tipo de datos de cada variable

En esta figura 2, observamos variables que mantienen distinto tipo por lo que debemos considerar quienes son categóricas y quienes son numéricas.

Generamos una matriz de correlación con todas las variables numéricas de nuestro dataset en primera instancia. La cual observamos en la figura 3.

Al analizar la matriz de correlación entre las variables numéricas del dataset, se pueden extraer varias observaciones relevantes para comprender las relaciones entre los datos y orientar la construcción del modelo predictivo:

- Correlaciones fuertes entre variables horarias:
    - ❖ Las variables relacionadas con la programación y ejecución del vuelo muestran alta correlación positiva, como:
        - SCHEDULED_DEPARTURE y DEPARTURE_TIME (0.96)
        - WHEELS_OFF y DEPARTURE_TIME (0.97)
        - SCHEDULED_ARRIVAL y ARRIVAL_TIME (0.86)
        - WHEELS_ON y ARRIVAL_TIME (0.97)

Estas altas correlaciones son esperadas, ya que estas variables representan momentos del mismo proceso de vuelo (salida, en vuelo y llegada).

- Correlación entre retrasos:

- ❖ La variable ARRIVAL_DELAY está fuertemente correlacionada con DEPARTURE_DELAY (0.94), lo cual es lógico: si un vuelo sale tarde, probablemente también llegue tarde.
- ❖ También se observa correlación moderada entre ARRIVAL_DELAY y algunas causas específicas de retraso:
  - AIRLINE_DELAY (0.61)
  - LATE_AIRCRAFT_DELAY (0.52)
  - WEATHER_DELAY, AIR_SYSTEM_DELAY (0.24 y 0.26)

- Variables de tiempo de vuelo (air time, distance, scheduled time):
  - ❖ DISTANCE, AIR_TIME y SCHEDULED_TIME presentan alta correlación entre sí (mayor a 0.9 en algunos casos), ya que todos reflejan de alguna manera la duración o extensión del vuelo.
- Variables de calendario (MONTH, DAY, DAY_OF_WEEK):
  - ❖ Tienen correlaciones muy bajas con otras variables, lo que indica que por sí solas no explican grandes variaciones en los retrasos, aunque podrían aportar valor como parte de un modelo más complejo (por ejemplo, en combinación con otros factores como clima o tráfico aéreo).
- Correlaciones bajas o nulas:
  - ❖ Muchas variables tienen correlaciones bajas entre sí (cerca de 0), lo cual es bueno desde el punto de vista del modelo, ya que evita redundancia excesiva y multicolinealidad.
- Otros:
  - ❖ Algunas variables como CANCELLED, DIVERTED, y SECURITY_DELAY no presentan correlaciones fuertes con otras variables, y muchas aparecen con valores NaN debido a que no siempre están presentes o son muy esporádicas. Estas deben analizarse con cuidado y, en muchos casos, pueden eliminarse del modelo si no aportan valor predictivo.

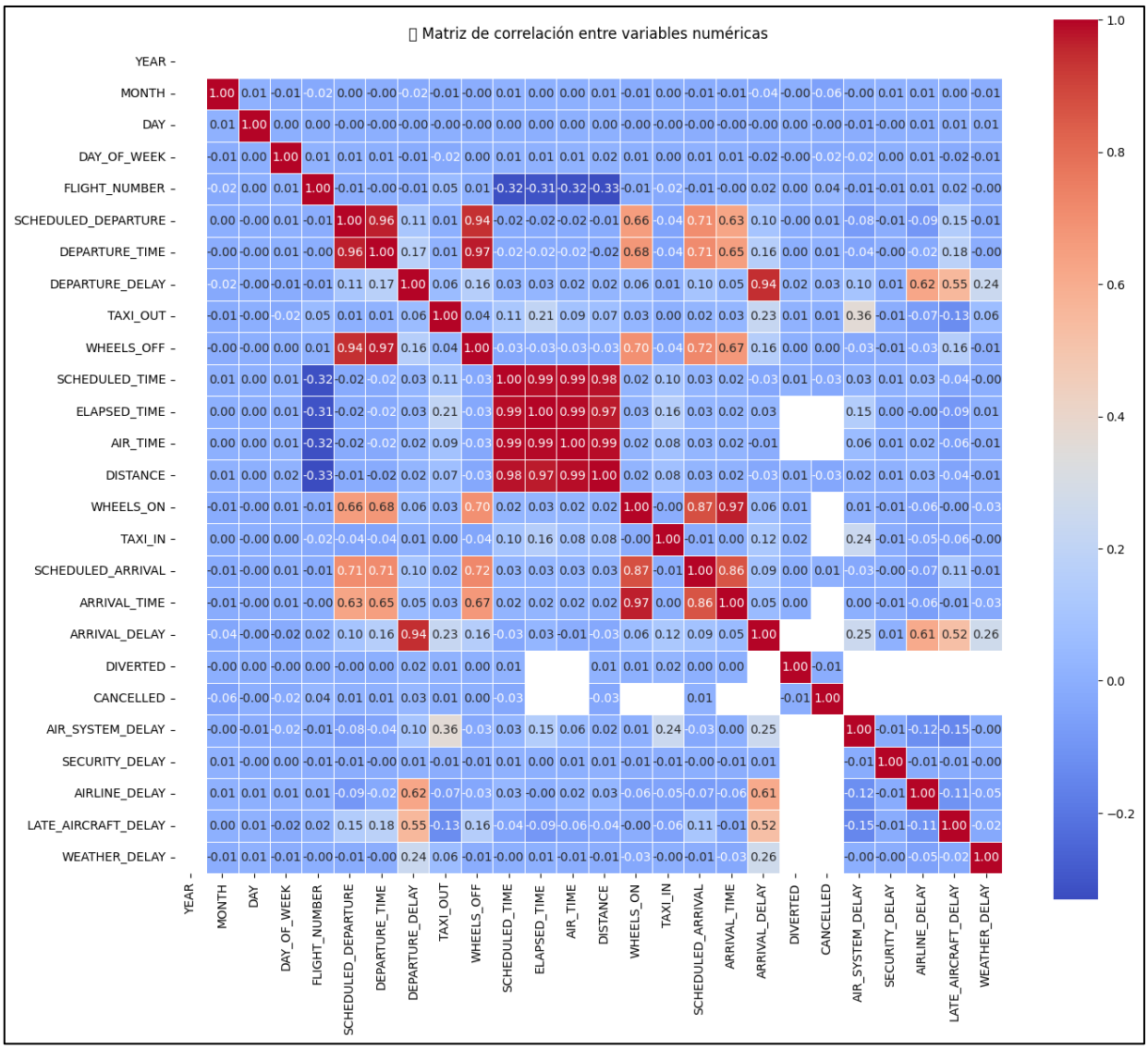


Figura 3. Matriz de correlación de todas las variables

Ya que nuestro objetivo principal es la detección de tempranas de retrasos debemos hacer una limpieza del Dataset, previo al entrenamiento del mismo.

Procedemos a eliminar variables como 'CANCELLED', 'DIVERTED', 'CANCELLATION_REASON', 'DEPARTURE_TIME', 'DEPARTURE_DELAY', 'TAXI_OUT', 'WHEELS_OFF', 'WHEELS_ON', 'ARRIVAL_TIME', 'TAXI_IN', 'ELAPSED_TIME', 'AIR_TIME', 'AIR_SYSTEM_DELAY', 'SECURITY_DELAY', 'AIRLINE_DELAY', 'LATE_AIRCRAFT_DELAY', 'WEATHER_DELAY' 'TAIL_NUMBER', 'YEAR' no tan solo por la cantidad de variables nulas que poseen algunas más del 80%, sino que también algunas son variables que se obtendrían después de que el vuelo haya despegado y con excepción de ARRIVAL_DELAY, no las vamos a necesitar para este estudio.

Volvemos a verificar los valores nulos, y se obtiene lo siguiente que se muestra en la figura 4,

```
#Revisamos si quedan valores Nulos
df.isnull().sum()
```

	0
MONTH	0
DAY	0
DAY_OF_WEEK	0
AIRLINE	0
FLIGHT_NUMBER	0
ORIGIN_AIRPORT	0
DESTINATION_AIRPORT	0
SCHEDULED_DEPARTURE	0
SCHEDULED_TIME	0
DISTANCE	0
SCHEDULED_ARRIVAL	0
ARRIVAL_DELAY	0

Figura 4. Valores nulos después de la eliminación de columnas no relevantes en el dataset.

Seguido creamos la variable Target, misma esta dada por el ARRIVAL_DELAY con la condición de que si con vuelos igual o mayores a 15 min le asignen el valor binario de 1 y si son vuelos menores de 15 min se le otorgue el valor binario de 0 ya que son vuelos a tiempo. Se identificó un ligero desbalance de clases alrededor del 18.6 % de los vuelos presentan algún tipo de retraso significativo. Luego de este paso eliminamos la variable ARRIVAL_DELAY para evitar el sobreajuste.

Luego de esto analizamos algunas variables y consideramos que debemos quitar la variable FLIGHT_NUMBER ya que no aporta mucho para nuestro estudio, también procedemos a codificar las variables SCHEDULED_DEPARTURE Y SHEDULED_ARRIVAL con el fin de de preservar su naturaleza cíclica. Esto permite que el modelo capture patrones horarios de manera más precisa, como la alta congestión en ciertos momentos del día. Luego de esto se procedieron a eliminar las variables SCHEDULED_DEPARTURE Y SHEDULED_ARRIVAL y nos quedamos con las trigonométricas. Además, las estadísticas descriptivas muestran una distribución razonable de los vuelos a lo largo del mes, la semana y el día, con una proporción moderada de vuelos retrasados, lo que indica un dataset levemente desbalanceado.

Luego se procedió a evaluar los modelos seleccionados (ver sección de selección de modelos)

Cabe mencionar que para la evaluación del desempeño optamos por agarrar una muestra de 500mil, esto debido a que al ser más de 5 millones de muestras y evaluar modelos bastantes robustos algunos son muy lentos, por ende, se tomó la decisión de agarra la muestra para optimizar el tiempo, se balancearon de manera correcta

para que siempre fuera 81.4% de vuelos a tiempo y 18.6% de vuelos retrasados con el fin de emular de manera más fiable el dataset total.

Se realizo una división de datos para entrenarlos (train-test-split), se separaron los datos en un 80% para entrenamiento y 20% para prueba, usando estratificación (stratify=y) para mantener la proporción original de clases (evitando que una clase esté sobre representada en un subconjunto).

Se hizo preprocesamiento de Variables, las variables numéricas fueron escaladas con StandardScaler, y las categóricas se codificaron usando OneHotEncoder, ajustando si se usaba un formato esparso según el modelo.

Dado a que la variable objetivo (DELAYED) está desbalanceada, se aplicaron diversas técnicas de balanceado,

Para cada ara cada modelo, se definió un conjunto de hiperparámetros y se evaluaron con GridSearchCV, usando f1 como métrica de evaluación. Esto asegura que se selecciona el modelo con el mejor balance entre precisión y recall.

Una vez entrenados, los modelos se evaluaron con métricas clave como el accurancy, recall, f1 Score, entre otras.

- **ANÁLISIS DESCRIPTIVO**

Luego de limpiar el Dataset procedemos con el análisis de las variables seleccionadas, así como de algunas graficas producidas de dichas variables.

En primera instancia realizamos un primer análisis estadístico con las variables antes de la depuración y selección final, esto se observa en la figura 5.

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 5. Estadísticas descriptivas de las variables depuradas

Y se observa lo siguiente:

- 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.

Seguido realizamos la matriz de correlación de estas variables ya depuradas por primera vez, la cual se observa en la figura 6.

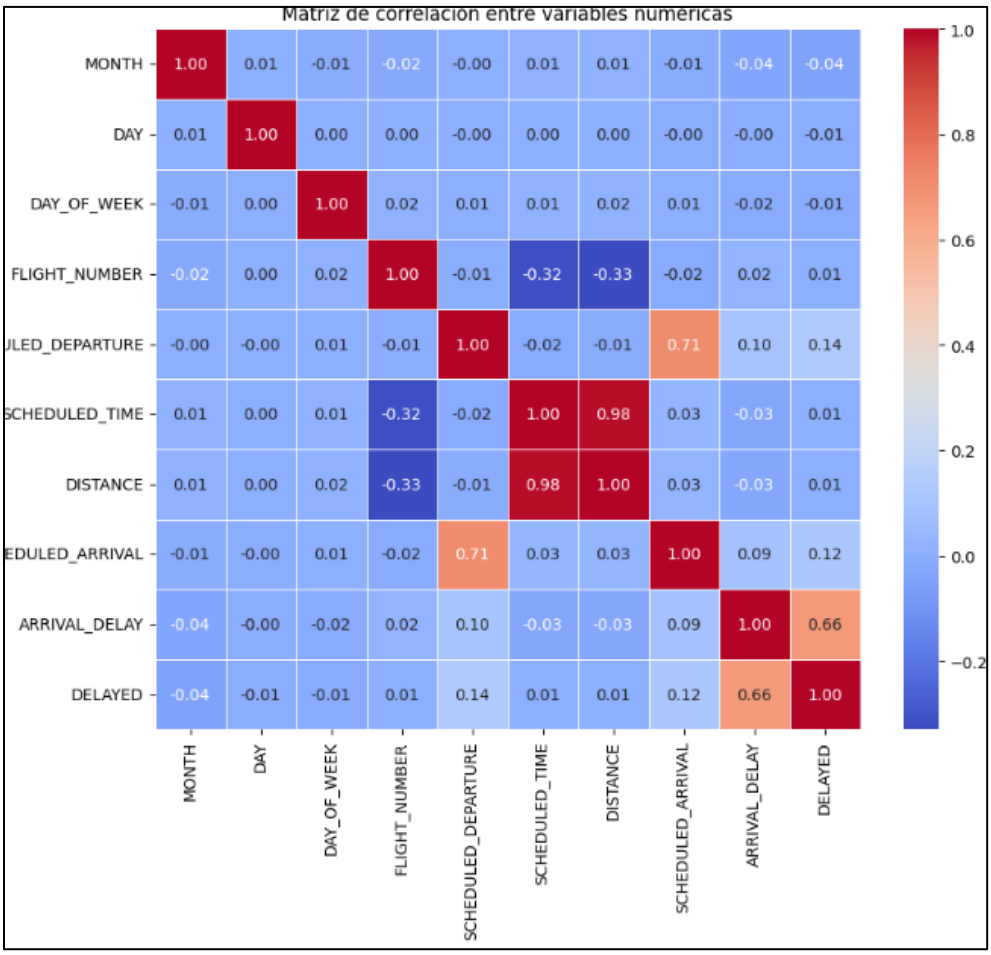


Figura 6. Matriz de Correlación de las Variables numéricas después de la depuración.

Podemos destacar de la misma que

- Variables más relacionadas con los retrasos

  - ❖ (DELAYED y ARRIVAL_DELAY) con una correlación fuerte de 0.66. esto tiene sentido ya que, si un vuelo tiene una gran demora, es muy probable que esté clasificado como retrasado (DELAYED = 1).
  - ❖ SCHEDULED_DEPARTURE y DELAYED con una correlación moderada 0. 137 lo que indica que la hora de salida programada influye un poco en los retrasos. Posiblemente hay más demoras en ciertos horarios (por ejemplo, vuelos en horas pico).
  - ❖ SCHEDULED_ARRIVAL también tiene una correlación leve (0.12) con DELAYED.

Las variables horarias sí tienen cierto efecto sobre los retrasos, aunque no son determinantes por sí solas.

- Variables de relación entre distancia, duración y tiempo programado
  - ❖ DISTANCE y SCHEDULED_TIME: correlación muy alta de 0.98. esto tiene sentido ya que vuelos más largos en distancia tienen una mayor duración programada.
  - ❖ FLIGHT_NUMBER tiene correlación negativa moderada con DISTANCE y SCHEDULED_TIME (alrededor de -0.32), posiblemente los vuelos con números más bajos son regionales (cortos), y los altos son de mayor distancia.
- Variables de Estacionalidad: Mes, día y día de la semana: baja correlación Todas las variables temporales (MONTH, DAY, DAY_OF_WEEK) tienen correlaciones cercanas a 0 con DELAYED. Esto sugiere que el mes o el día en sí no influye demasiado en los retrasos, al menos de forma lineal. Puede haber patrones no lineales, pero no se reflejan aquí.

Al analizar la matriz de correlación, observamos que las variables con mayor influencia lineal en los retrasos (DELAYED) son principalmente ARRIVAL_DELAY (con una fuerte correlación de 0.66), seguida por SCHEDULED_DEPARTURE (0.13) y SCHEDULED_ARRIVAL (0.12). Esto indica que los horarios programados tienen cierto efecto en la probabilidad de un vuelo retrasarse. Por otro lado, variables como el mes o el día tienen una correlación muy baja, por lo que no parecen influir directamente en los retrasos. También se confirma una fuerte relación entre la distancia del vuelo y el tiempo programado (0.98), lo cual valida la coherencia del dataset.

Después de la depuración final obtuvimos el siguiente análisis estadístico que se muestra en la figura 7,

	MONTH	DAY	DAY_OF_WEEK	SCHEDULED_TIME	DISTANCE	DELAYED	SCHEDULED_DEPARTURE_sin	SCHEDULED_DEPARTURE_cos	SCHEDULED_ARRIVAL_sin	SCHEDULED_ARRIVAL_cos
count	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06
mean	6.547799e+00	1.570759e+01	3.932643e+00	1.418940e+02	8.244569e+02	1.861109e-01	-1.288751e-01	-3.584280e-01	-3.000328e-01	-2.230113e-01
std	3.397421e+00	8.774394e+00	1.985967e+00	7.531400e+01	6.086620e+02	3.891961e-01	7.361489e-01	5.594689e-01	6.272467e-01	6.832335e-01
min	1.000000e+00	1.000000e+00	1.000000e+00	1.800000e+01	3.100000e+01	0.000000e+00	-1.000000e+00	-1.000000e+00	-1.000000e+00	-1.000000e+00
25%	4.000000e+00	8.000000e+00	2.000000e+00	8.500000e+01	3.730000e+02	0.000000e+00	-8.549119e-01	-8.660254e-01	-8.660254e-01	-8.549119e-01
50%	7.000000e+00	1.600000e+01	4.000000e+00	1.230000e+02	6.500000e+02	0.000000e+00	-3.090170e-01	-5.000000e-01	-4.809888e-01	-4.265687e-01
75%	9.000000e+00	2.300000e+01	6.000000e+00	1.740000e+02	1.065000e+03	0.000000e+00	6.427876e-01	6.123234e-17	1.822355e-01	3.826834e-01
max	1.200000e+01	3.100000e+01	7.000000e+00	7.180000e+02	4.983000e+03	1.000000e+00	1.000000e+00	9.999905e-01	1.000000e+00	1.000000e+00

Figura 7. Análisis estadístico después de la depuración final con las variables trigonométricas.

En ella se observa que:

- Variables temporales: MONTH, DAY, DAY_OF_WEEK

- ❖ MONTH: El promedio es 6.54, es decir, la mayoría de los vuelos suceden en junio/julio.
- ❖ DAY: El promedio es 15.7, lo que sugiere que los vuelos están bien distribuidos a lo largo del mes.
- ❖ DAY_OF_WEEK: La media es 3.93, cerca del miércoles/jueves, así que no hay un sesgo claro hacia fines de semana o lunes.

Los vuelos están bastante bien repartidos a lo largo del calendario.

- Sobre los vuelos: SCHEDULED_TIME, DISTANCE
  - ❖ SCHEDULED_TIME (tiempo de vuelo programado) tiene una media de 142 minutos.
  - ❖ DISTANCE tiene una media de 824 km, pero hay vuelos muy largos (hasta casi 5.000 km), lo que indica que el dataset incluye vuelos nacionales largos o internacionales.

La mayoría de los vuelos son medianos (unas 2 horas y entre 300 y 1.000 km), pero hay algunos muy largos.

- DELAYED
  - ❖ DELAYED es una variable binaria (0: no retrasado, 1: sí retrasado).
  - La media es 0.186, lo que significa que solo el 18.6% de los vuelos sufrieron retrasos.

Eso es bueno para las aerolíneas ya que, indica que la mayoría de los vuelos llegaron a tiempo.

- Variables trigonométricas: SCHEDULED_DEPARTURE_sin, cos, SCHEDULED_ARRIVAL_sin, cos

Estas variables fueron transformadas con funciones seno y coseno para representar horarios del día como ciclos (como un reloj de 24 horas). Porque las horas son cíclicas (después de las 23:59 viene 00:00) y modelos de machine learning no lo entienden bien si solo les damos el número "hora". Entonces, se convierten en coordenadas circulares (seno y coseno) para que los modelos capten mejor patrones según la hora del día (por ejemplo: más retrasos a la tarde o noche).

Adicional se generan otras graficas como, por ejemplo:

En la figura 8, 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%)

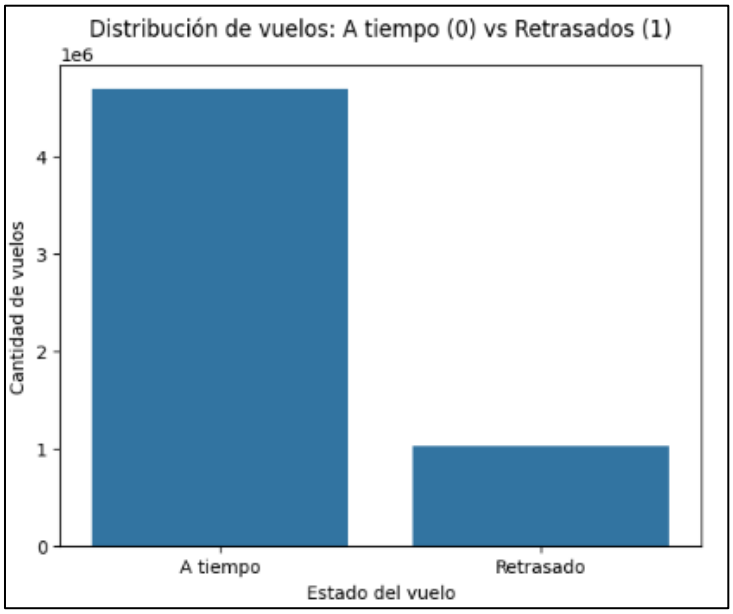


Figura 8. Gráfica de distribución de vuelos

En la figura 9, 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.

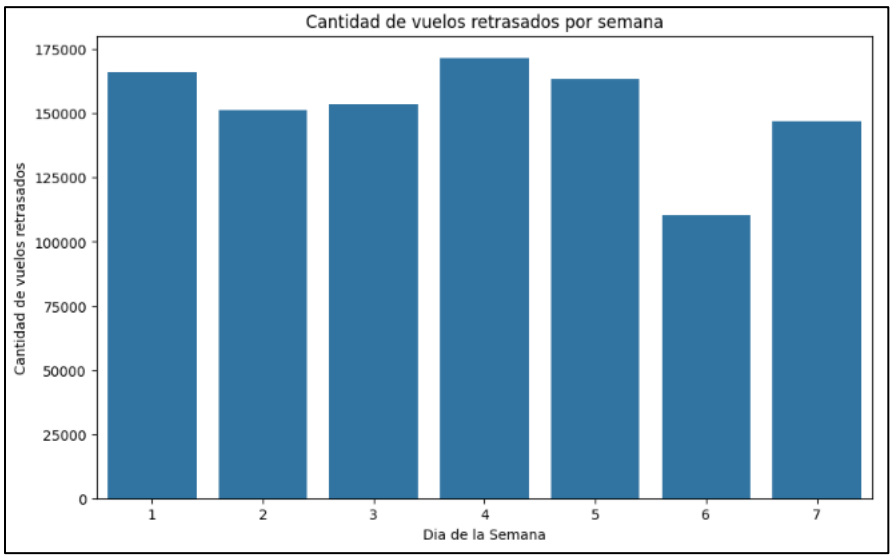


Figura 9. Cantidad de vuelos retrasados por semana.

En la figura 10, 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.

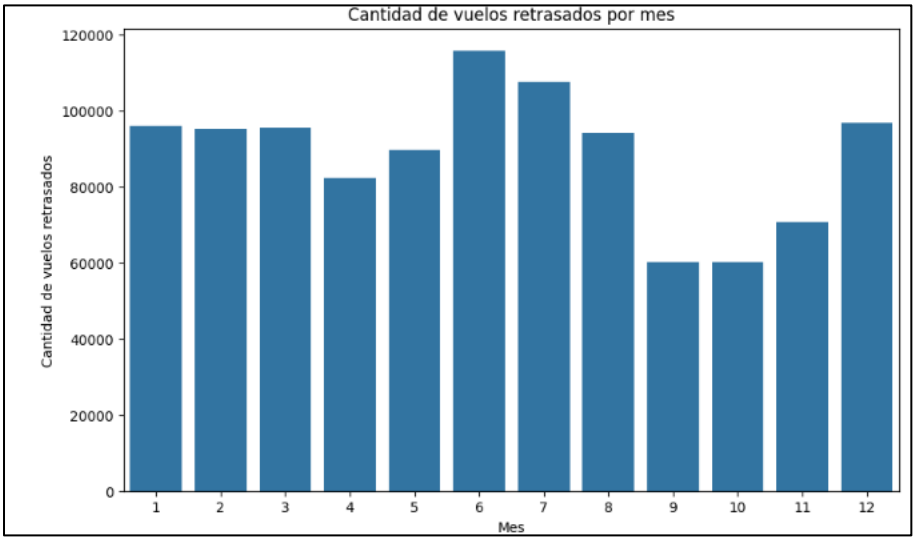


Figura 10. Cantidad de vuelos retrasados por mes.

En la Figura 11, 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.

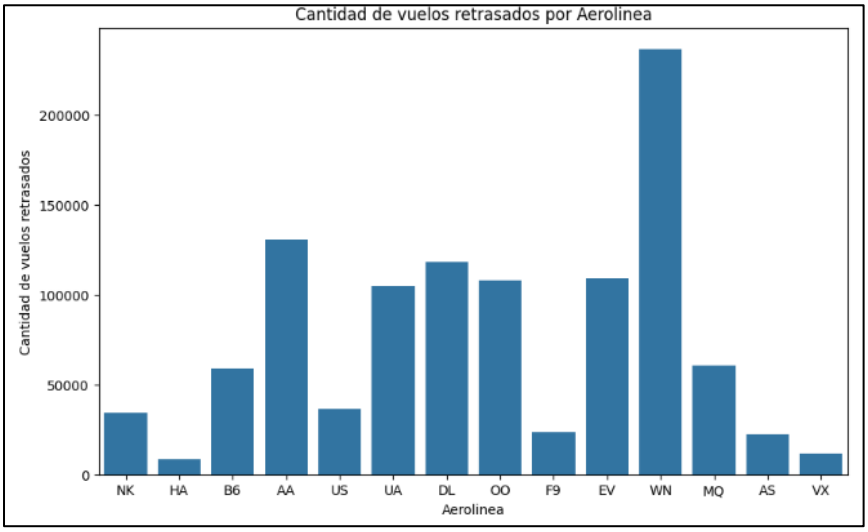


Figura 11. Cantidad de vuelos retrasados por Aerolinea.

En la figura 12, se muestra la distribución de la hora programada de salida según los retrasos, no es más 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.

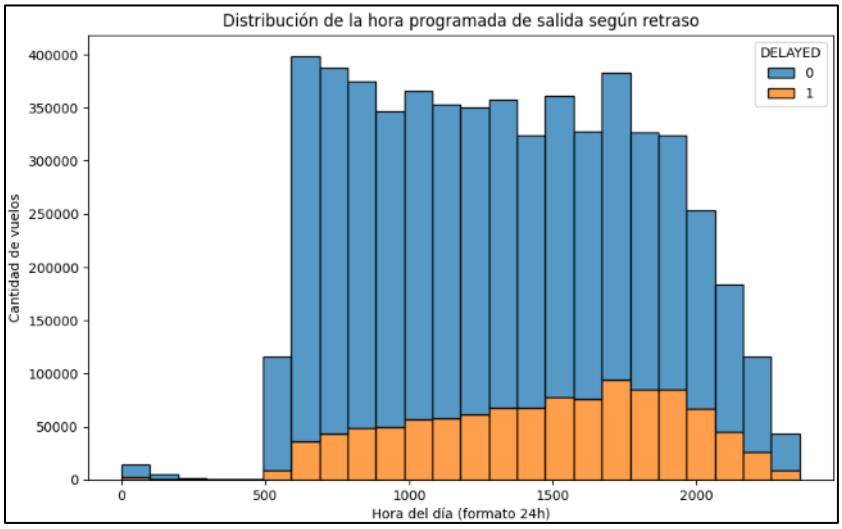


Figura 12. Distribución de la hora programada de salida según retrasos.

En la gráfica 13, 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.

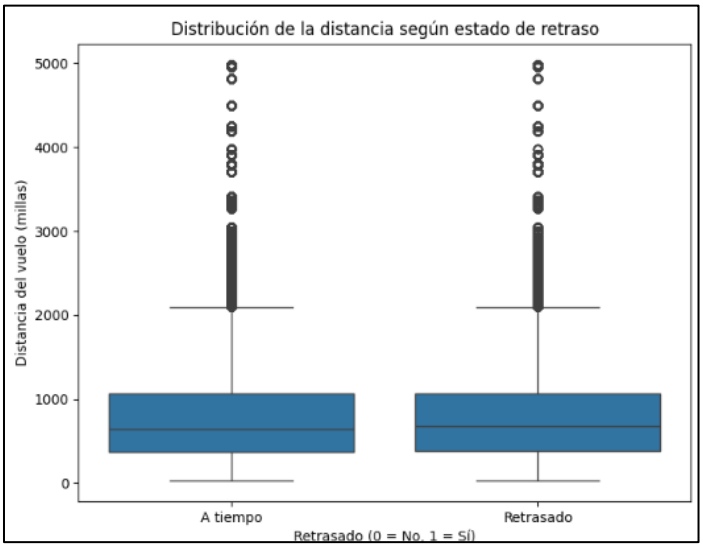


Figura 13. Grafica de distancia vs. El estado del vuelo

En la figura 14, observamos la gráfica 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.

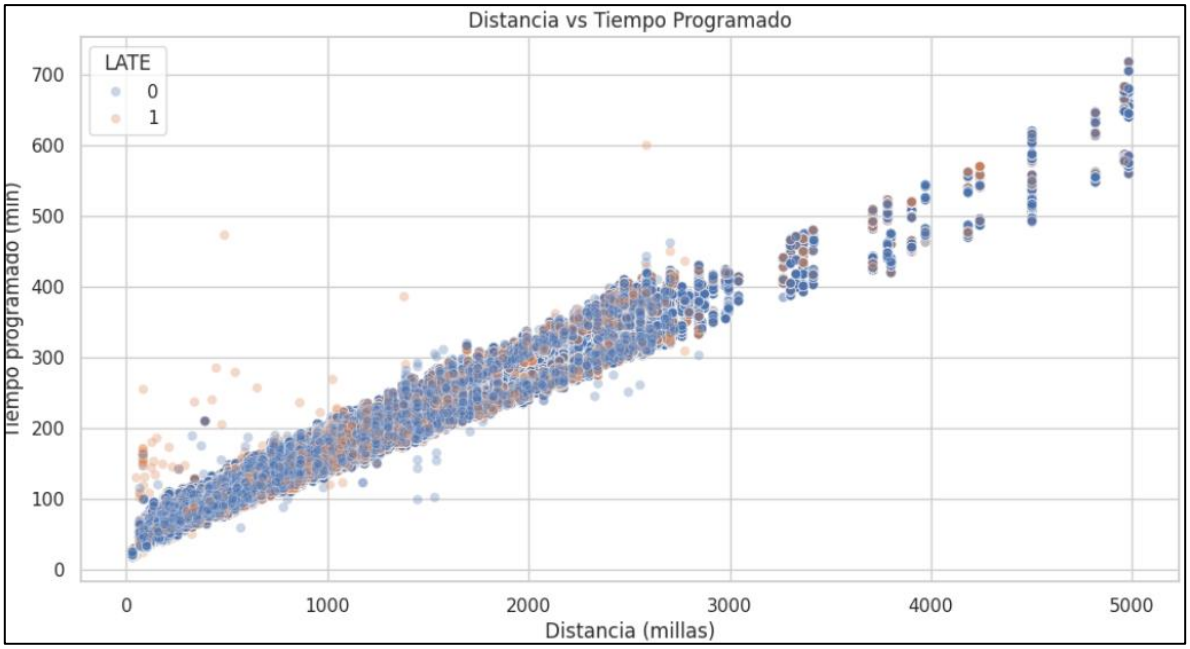


Figura 14. Grafica distancia versus tiempo programado

En la figura 15, se observa la gráfica de distribución de la duración programada del vuelo, nos muestra cómo 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.

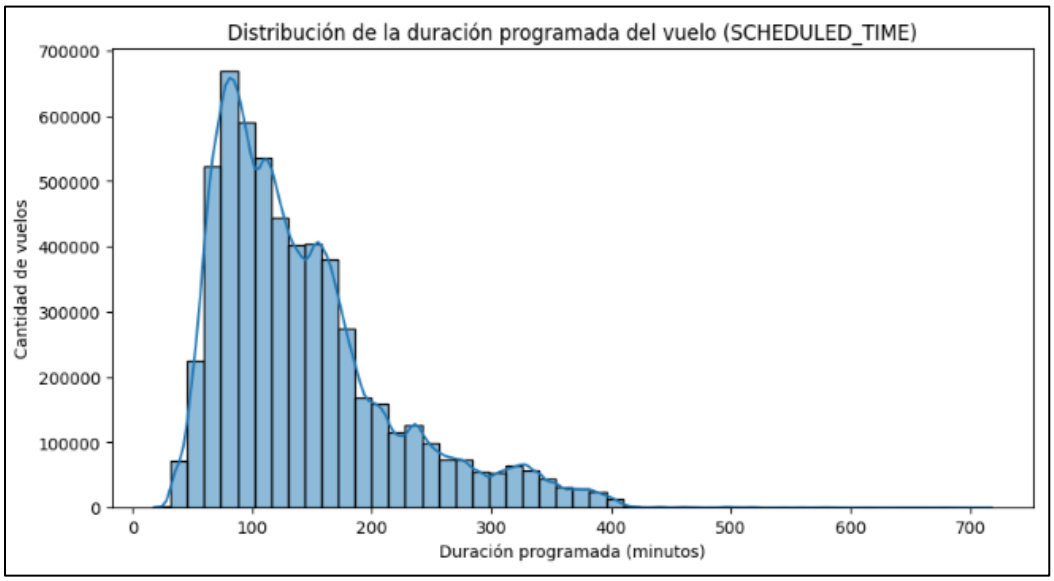


Figura 15. Distribución de la duración programada del vuelo

En la figura 16, 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.

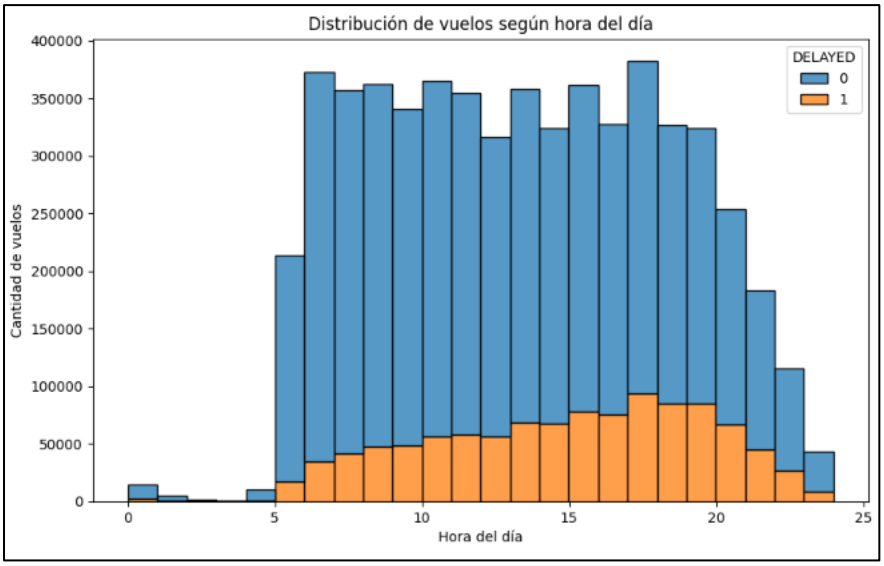


Figura 16. Distribución de vuelos según la hora del día.

En la figura 15, 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.

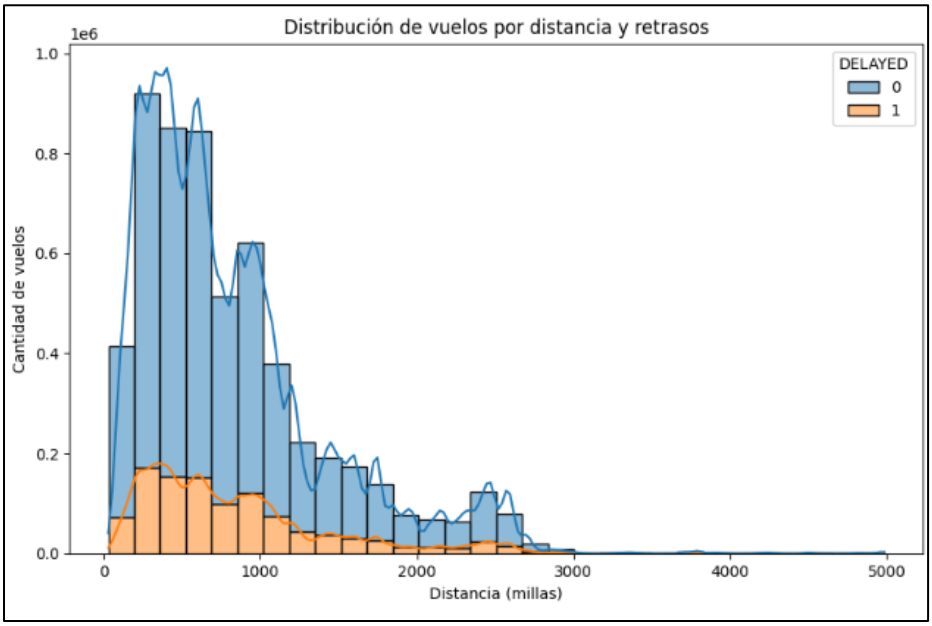


Figura 15. Distribución de vuelos por distancia y retrasos

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.

• SELECCIÓN DE VARIABLES

El retraso de vuelos es un problema común y costoso para las aerolíneas (implica costo, reputación), pasajeros (afecta la experiencia) y aeropuertos. Este tipo de problema es perfecto para aplicar modelos predictivos ya que:

- Como vamos a analizar retrasos, hay una clara variable objetivo (si el vuelo se retrasa o no)
- La base de datos principal contiene muchas variables relevantes y diversas para realizar nuestro análisis como por ejemplo hora, aerolínea, hora de llegada, entre otras.

Para la elaboración de este proyecto se seleccionaron las principales variables relacionadas con con el vuelo antes del despegue, las cuales son:

- 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: nace de ARRIVAL_DELAY, variable objetivo binaria que indica si el vuelo se retrasó 15 minutos o más (1) o si llegó a tiempo (0).

• SELECCIÓN DE MODELOS

Para nuestro análisis hemos optado por la comparación de diversos modelos de Clasificación los mismos fueron seleccionados algunos por simplicidad y otros por su robustes a la hora de enfrentar datos grandes:

- Regresión Logística: Modelo lineal simple pero efectivo para problemas de clasificación binaria. Su interpretabilidad lo hace una buena línea base. [Hosmer et al., 2013]
- Hist Gradient Boosting: Variante de Gradient Boosting optimizada para grandes volúmenes de datos, usa histogramas para acelerar el entrenamiento. [Ke et al., 2017]
- Árbol de decisión: Algoritmo no paramétrico que divide el espacio de decisiones mediante reglas simples, fácil de visualizar e interpretar. [Quinlan, 1986]
- Light GBM (con OHE y sin OHE): Modelo de boosting que utiliza histogramas y crecimiento por hojas (leaf-wise), logrando alta velocidad y precisión, ideal para grandes datasets. Se evaluó con y sin codificación One-Hot Encoding (OHE). [Ke et al., 2017]
- XGBoost: Variante robusta del boosting con regularización L1 y L2, excelente manejo de datos tabulares y gran desempeño en competiciones. [Chen & Guestrin, 2016]
- Catboost: Optimizado para variables categóricas, evita la necesidad de codificación explícita. Su tratamiento nativo de categorías lo hace especialmente útil para datasets mixtos. [Prokhorenkova et al., 2018]
- Random Forest: Ensamble de árboles que reduce el sobreajuste al combinar predicciones de múltiples árboles de decisión entrenados con bagging. [Breiman, 2001]

Adicional analizaremos los siguientes parámetros que contribuirán en la a la evaluación del desempeño de nuestros modelos:

- Accuracy: Proporción de predicciones correctas sobre el total. No es ideal en presencia de clases desbalanceadas.

- Precision: Proporción de verdaderos positivos sobre todos los positivos predichos. Mide la calidad del modelo al clasificar la clase positiva.
- Recall: Proporción de verdaderos positivos sobre los positivos reales. Evalúa qué tan bien detecta el modelo los casos positivos
- F1 Score: Media armónica entre precision y recall. Útil cuando hay un desbalance entre clases.
- AUC-ROC: Área bajo la curva ROC. Resume la capacidad del modelo para distinguir entre clases en todos los umbrales posibles.
- Matriz de confusión: Permite observar el detalle de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

También como se observó cierto desbalance en los datos los analizaremos los datos balanceándolos con las siguientes técnicas:

- Sin balanceo: Permite observar el comportamiento natural del modelo frente a los datos originales.
- Baslanceado con Class Weight: Ajusta el peso de cada clase en la función de pérdida, para penalizar más los errores en la clase minoritaria.
- Balanceado con Undersampling: Reduce la cantidad de ejemplos de la clase mayoritaria para igualarla con la minoritaria. Puede llevar a pérdida de información.
- Balanceado con SMOTE **(Synthetic Minority Over-sampling Technique)**: Genera nuevas muestras sintéticas de la clase minoritaria, creando un dataset más equilibrado sin perder datos de la clase mayoritaria. [Chawla et al., 2002]

RESULTADOS

En la tabla 2 se observa la evaluación de los parámetros de los distintos modelos que se han evaluado.

Tabla 2. Evaluación de los modelos de Clasificación.

Analisis predictivo					
Modelos	balanceo	Accuracy	Presicion	Recall	F1 Score
Regresion Logistica	NO	0.814	0.414	0	0
	Class Weight	0.583	0.251	0.627	0.359
	Undersampling	0.582	0.251	0.628	0.359
	Smote	0.581	0.251	0.631	0.359
Hist Gradient Boosting	NO	0.818	0.662	0.047	0.087
	Class Weight	NA	NA	NA	NA
	Undersampling	0.648	0.296	0.647	0.406
	Smote	0.817	0.535	0.111	0.184
Árbol de decision	NO	0.725	0.279	0.301	0.29
	Class Weight	0.62	0.272	0.625	0.379
	Undersampling	0.604	0.262	0.622	0.369
	Smote	0.61	0.243	0.518	0.331
LigtGBM (con OHE)	NO	0.818	0.669	0.049	0.091
	Class Weight	0.66	0.303	0.635	0.41
	Undersampling	0.649	0.296	0.642	0.405
	Smote	0.812	0.476	0.093	0.156
LigtGBM (sin OHE)	NO	0.816	0.659	0.021	0.041
	Class Weight	0.639	0.289	0.642	0.398
	Undersampling	0.637	0.287	0.64	0.396
	Smote	0.814	0.496	0,08	0.137
XGBoost	NO	0.816	0.656	0.022	0.043
	Class Weight	0.639	0.289	0.644	0.399
	Undersampling	0.63	0.285	0.652	0.396
	Smote	0.803	0.403	0.121	0.186
Catboost	NO	0.814	0.651	0.005	0.011
	Class Weight	0.665	0.308	0.638	0.415
	Undersampling	0.612	0.273	0.653	0.385
	Smote	0.814	0.501	0.037	0.068
Random Forest	NO	0.806	0.413	0.095	0.155
	Class Weight	0.595	0.261	0.642	0.371
	Undersampling	0.607	0.272	0.664	0.386
	Smote	0.629	0.267	0.567	0.363

Podemos notar lo siguiente:

1. Regresión Logística:

    - Sin balanceo: Alta precisión general (accuracy: 0.814), pero completamente incapaz de detectar vuelos retrasados (recall: 0), lo que indica que el modelo predice casi exclusivamente la clase mayoritaria (on-time).
    - Class Weight: Aumenta significativamente la capacidad de detectar retrasos (recall: 0.627), aunque sacrifica accuracy. F1 score aceptable (0.359).
    - Undersampling: Resultados similares a class weight, con ligera disminución de accuracy pero mejora en balance de clases.
    - SMOTE: También con desempeño similar, muestra que el modelo mejora mucho con técnicas de balanceo.

Necesita balanceo. Es simple y explicativo, pero con rendimiento limitado.

2. HistGradientBoosting

    - Sin balanceo: Alta precisión (accuracy: 0.818), pero bajísima sensibilidad (recall: 0.047), lo que indica que casi nunca predice vuelos retrasados.
    - Class Weight: No disponible para este modelo.
    - Undersampling: Baja accuracy (0.648), pero buena capacidad para detectar retrasos (recall: 0.647), y mejor F1 (0.406).
    - SMOTE: Recupera algo de precisión pero sigue con baja sensibilidad (recall: 0.111). F1 bajo.

Con undersampling es competitivo. Mejor que regresión logística.

3. Árbol de Decisión

    - Sin balanceo: Bajo desempeño general (F1: 0.29).
    - Class Weight: Mejora importante en recall (0.625) y F1 (0.379).
    - Undersampling: Rendimiento casi igual que Class Weight.
    - SMOTE: Baja considerablemente en recall (0.518) y F1 (0.331), por lo que no es tan útil aquí.

Interpretable pero poco eficaz frente a otros modelos.

4. LightGBM (con OneHotEncoding)

    - Sin balanceo: Alta accuracy (0.818) pero bajísimo recall (0.049).
    - Class Weight: Buen compromiso entre recall (0.635) y F1 (0.410).
    - Undersampling: Buen desempeño balanceado, F1 (0.405).
    - SMOTE: Muy baja sensibilidad (recall: 0.093) y F1 (0.156).

Muy buen modelo con class weight o undersampling.

5. LightGBM (sin OneHotEncoding)

- Sin balanceo: Accuracy alto (0.816), pero recall bajísimo (0.021).
- Class Weight: Mejora sensible del recall (0.642) y F1 (0.398).
- Undersampling: Resultados similares al anterior.
- SMOTE: Algo de mejora en recall (0.08) pero bajo F1 (0.137).

Rinde mejor con codificación OHE.

6. XGBoost

- Sin balanceo: Alta precisión (accuracy: 0.816), pero nulo recall (0.022).
- Class Weight: Balance adecuado entre recall (0.644) y F1 (0.399).
- Undersampling: Muy similar al anterior.
- SMOTE: Baja en recall y F1 (0.186).

Potente con balanceo. Mejor cuando se ajusta con class_weight.

7. CatBoost

- Sin balanceo: Muy bajo recall (0.005), aunque accuracy alto.
- Class Weight: Excelente combinación de recall (0.638) y F1 (0.415).
- Undersampling: Buen resultado también (F1: 0.385).
- SMOTE: Bajo recall (0.037) y F1 pobre (0.068).

Muy competitivo. Mejor resultado general con class weight.

8. Random Forest

- Sin balanceo: F1 bajo (0.155) y bajo recall.
- Class Weight: Mejor balance, recall (0.642) y F1 (0.371).
- Undersampling: Buen desempeño en recall (0.664) y F1 (0.386).
- SMOTE: Menor desempeño que los anteriores.

Modelo equilibrado con buen recall y resultados sólidos al balancear.

Después de comparar todos los modelos y técnicas de balanceo, el mejor desempeño lo ofrece CatBoost con Class Weight, logrando el F1 score más alto (0.415), junto a un recall de 0.638, lo cual es clave dado que nuestro objetivo es detectar retrasos (clase minoritaria). Otros modelos como LightGBM (con OHE) y XGBoost con class weight también obtienen buenos resultados, pero CatBoost destaca por su robustez y rendimiento general.

Ahora es ¿Es viable en un entorno real?
Sí, es viable, siempre y cuando se mantenga una estrategia de balanceo adecuada y se actualice el modelo con datos recientes. Aunque la precisión global es

moderada, la capacidad de identificar vuelos con probabilidad de retraso puede ser muy útil para generar alertas tempranas y tomar decisiones proactivas en la industria aeronáutica. Un recall alto significa que muchos retrasos reales sí son anticipados, que es el objetivo principal del modelo.

Para la evaluación del AUC-ROC y la Matriz de confusión de obtuvieron los siguientes resultados presentados en la tabla 3.

Tabla 3. Cuadro comparativo de los AUC-ROC y la matriz de confusión de los modelos de clasificación

Modelos	balanceo	AUC-ROC	Matriz de Confusion
Regresion Logistica	NO	0.637	[[101729 7] [23259 5]]
	Class Weight	0.637	[[58274 43462] [8676 14588]]
	Undersampling	0.637	[[58205 43531] [8665 14599]]
	Smote	0.636	[[57998 43738] [8589 14675]]
Hist Gradient Boosting	NO	0.709	[[101182 554] [22177 1087]]
	Class Weight	NA	NA
	Undersampling	0.701	[[65932 35804] [8217 15047]]
	Smote	0.697	[[99488 2248] [20678 2586]]
Árbol de decision	NO	0.562	[[83621 18115] [16257 7007]]
	Class Weight	0.663	[[62917 38819] [8725 14539]]
	Undersampling	0.653	[[61013 40723] [8804 14460]]
	Smote	0.595	[[64133 37603] [11207 12057]]
LigtGBM (con OHE)	NO	0.709	[[101177 559] [22133 1131]]
	Class Weight	0.706	[[67699 34037] [8485 14779]]
	Undersampling	0.703	[[66151 35585] [8317 14947]]
	Smote	0.675	[[99349 2387] [21095 2169]]
LigtGBM (sin OHE)	NO	0.696	[[101481 255] [22771 493]]]
	Class Weight	0.694	[[64931 36805] [8327 14937]]
	Undersampling	0.692	[[64750 36986] [8378 14886]]
	Smote	0.677	[[99850 1886] [21409 1855]]
XGBoost	NO	0.695	[[101467 269] [22750 514]]
	Class Weight	0.694	[[64838 36898] [8278 14986]]
	Undersampling	0.693	[[63587 38149] [8089 15175]]
	Smote	0.659	[[97573 4163] [20451 2813]]
Catboost	NO	0.68	[[101668 68] [23137 127]]
	Class Weight	0.711	[[68323 33413] [8422 14842]]
	Undersampling	0.683	[[61267 40469] [8074 15190]]
	Smote	0.665	[[100890 846] [22413 851]]
Random Forest	NO	0.665	[[98584 3152] [21045 2219]]
	Class Weight	0.657	[[59422 42314] [8326 14938]]
	Undersampling	0.679	[[60418 41318] [7816 15448]]
	Smote	0.648	[[65485 36251] [10069 13195]]

En las siguientes figuras 16, figura 17, figura 18, figura 19, figura 20, figura 21, figura 22, y figura 23 se observan las graficas de la matriz de correlación y la curva bajo la curva para los diferentes modelos.

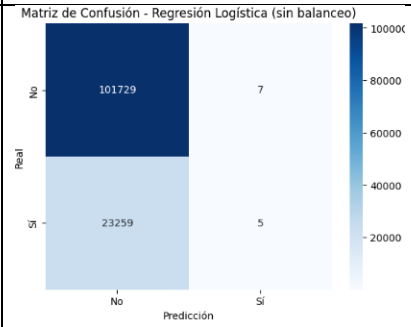
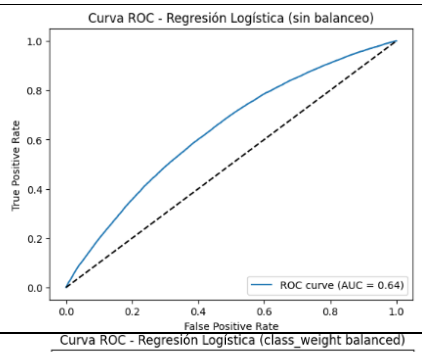
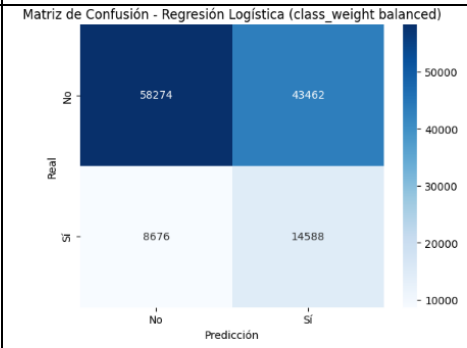
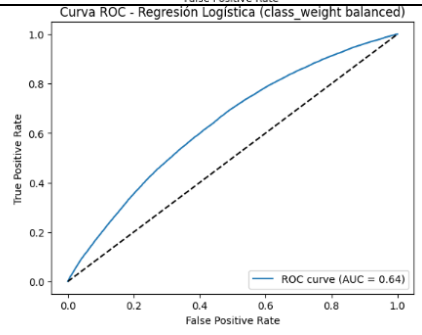
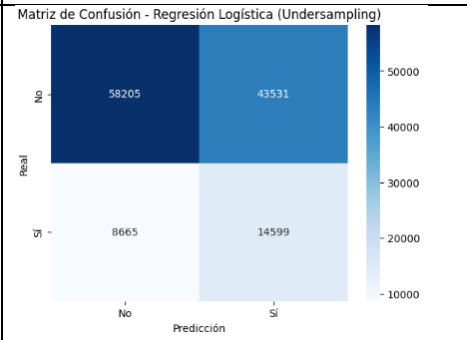
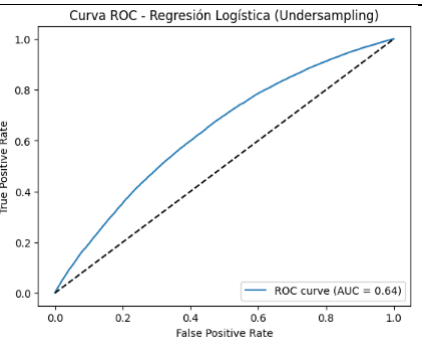
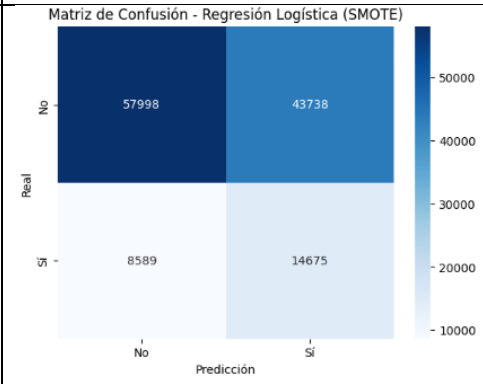
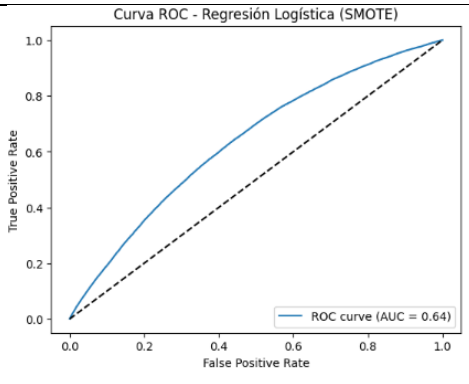
Regresión Logística		
	Matriz de Confusión	Curva ROC
Sin Balanceo		
Con Class_Weight		
Undersampling		
SMOTE		

Figura 16. Matriz de Confusión y Curva ROC para Regresión Logística

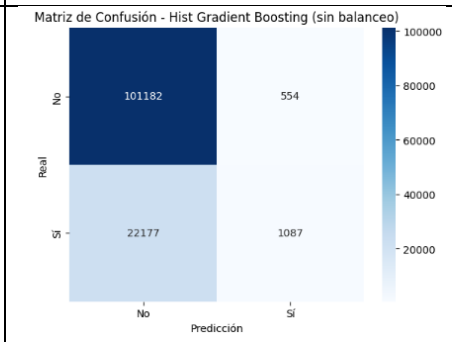
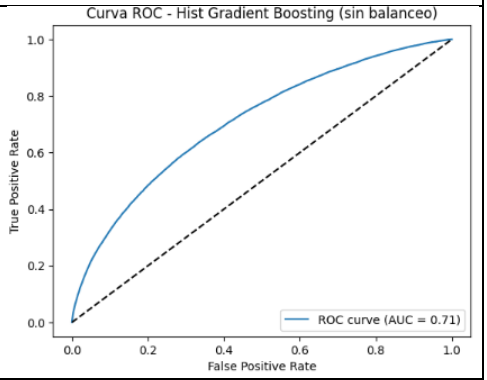
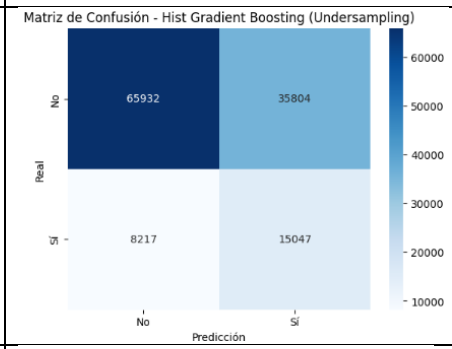
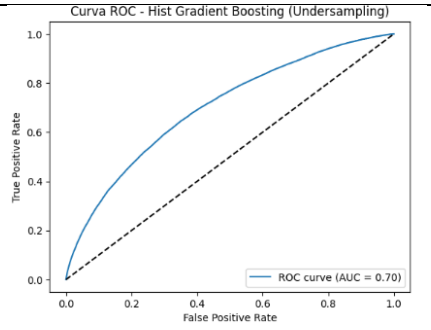
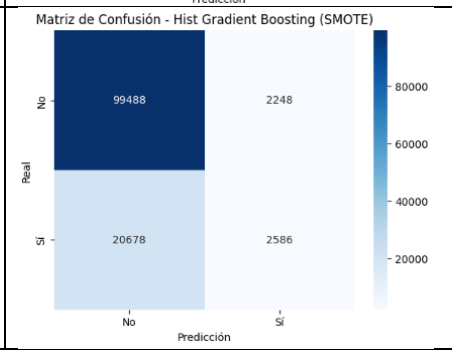
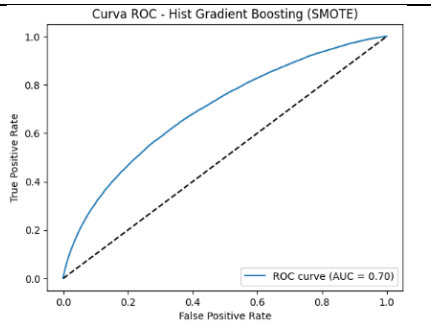
Hist Gradient Boosting		
	Matriz de Confusión	Curva ROC
Sin Balanceo		
Con Class_Weight	NA	NA
Undersampling		
SMOTE		

Figura 17. Matriz de Confusión y Curva ROC para Hist Gradient Boosting

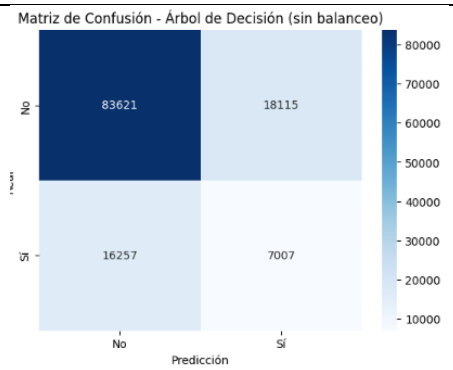
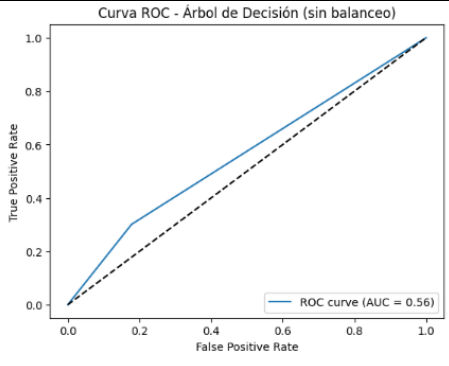
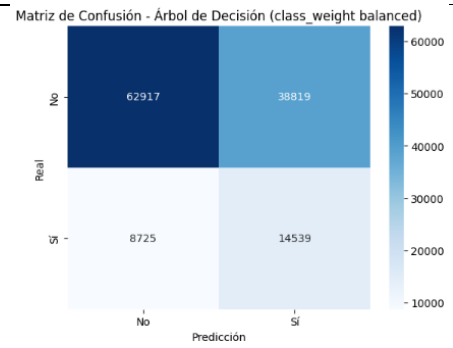
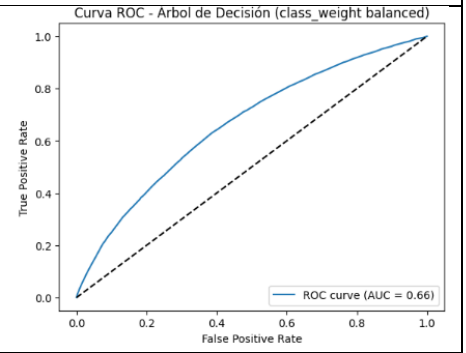
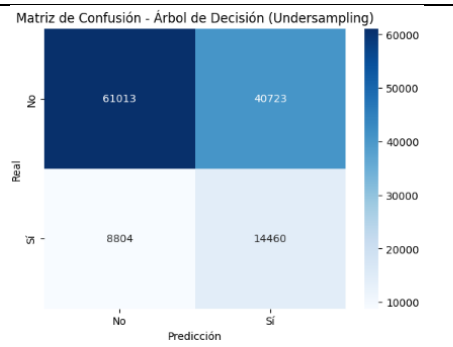
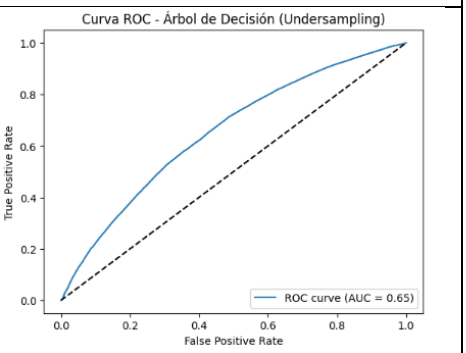
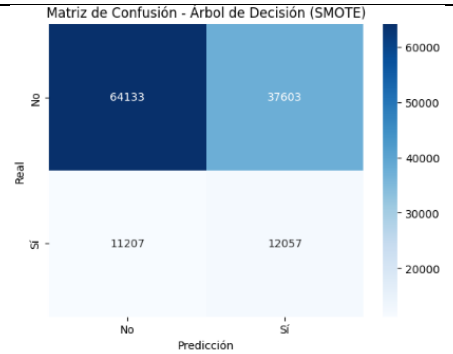
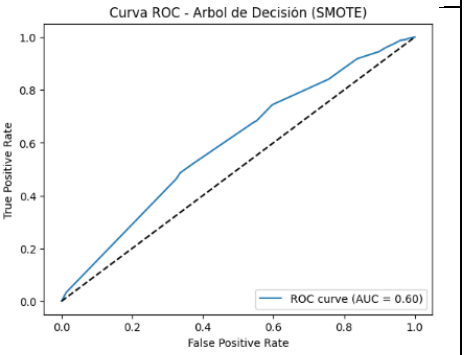
Árbol de Decisión		
	Matriz de Confusión	Curva ROC
Sin Balanceo		
Con Class_Weight		
Undersamplin g		
SMOTE		

Figura 18. Matriz de Confusión y Curva ROC para Árbol de Decisión

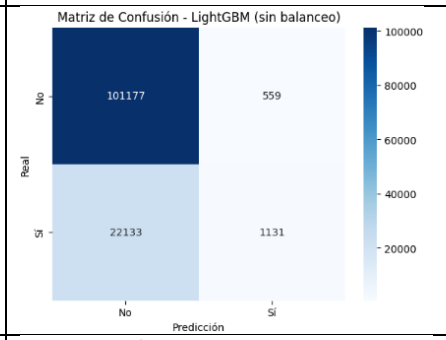
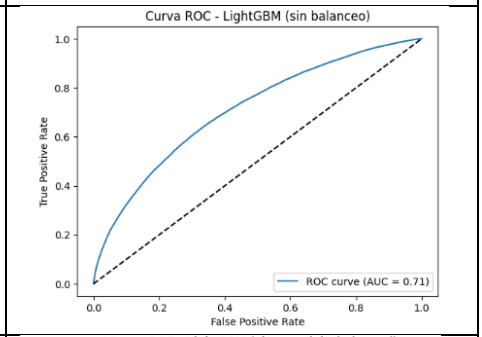
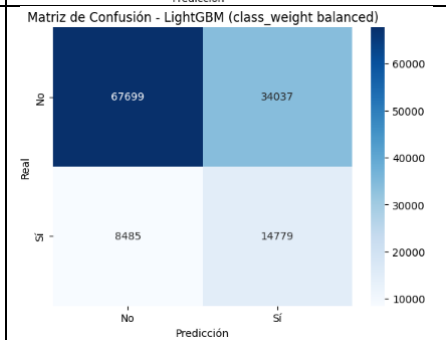
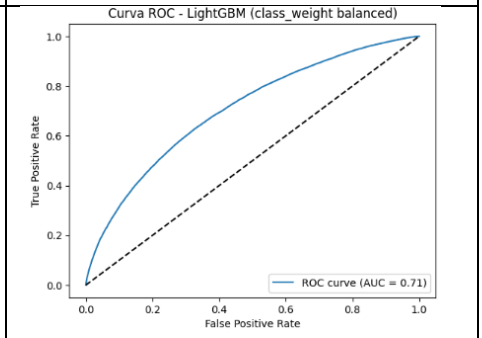
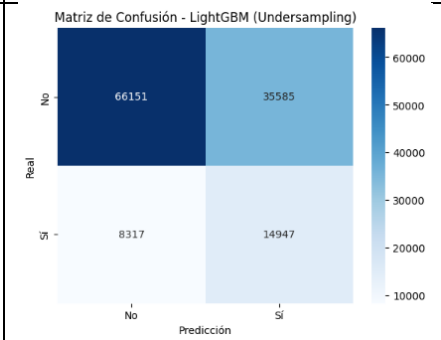
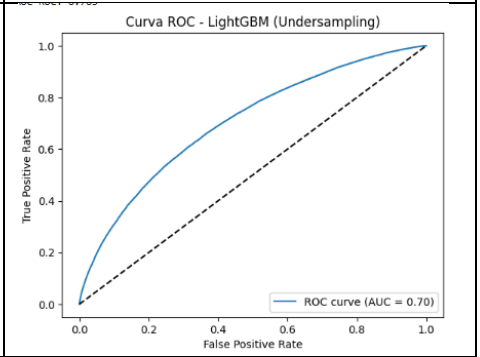
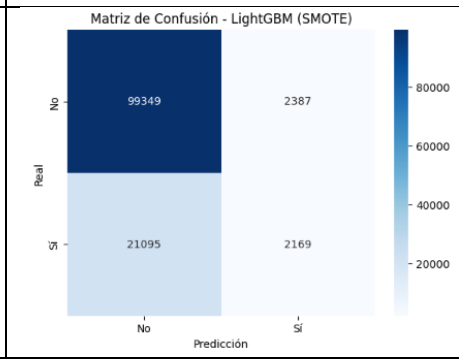
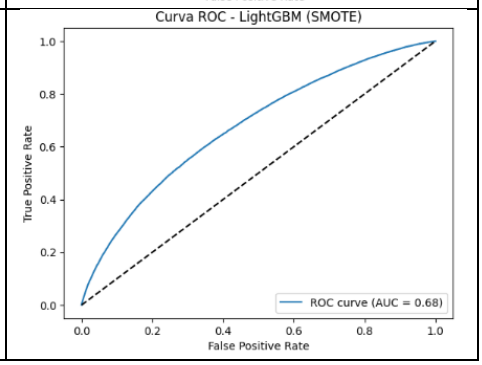
LGBM con OHE		
	Matriz de Confusión	Curva ROC
Sin Balanceo		
Con Class_Weight		
Undersamplin g		
SMOTE		

Figura 19. Matriz de Confusión y Curva ROC para LGBM con OHE.

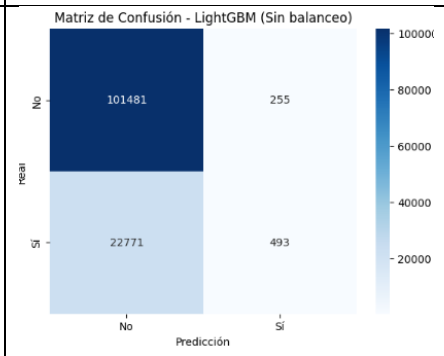
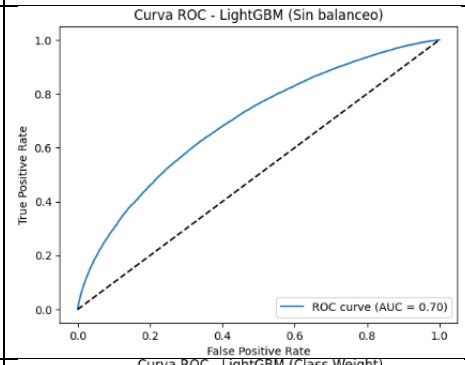
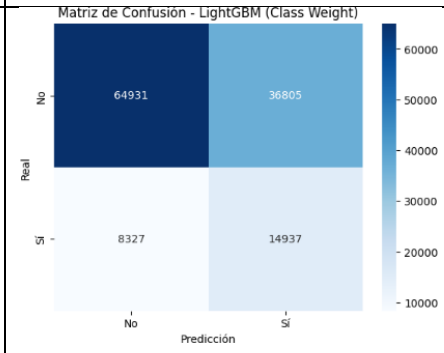
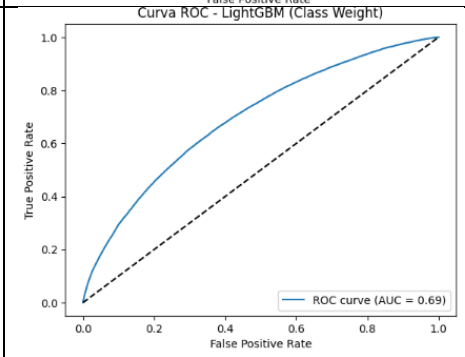
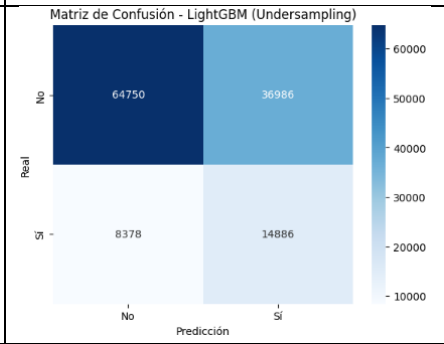
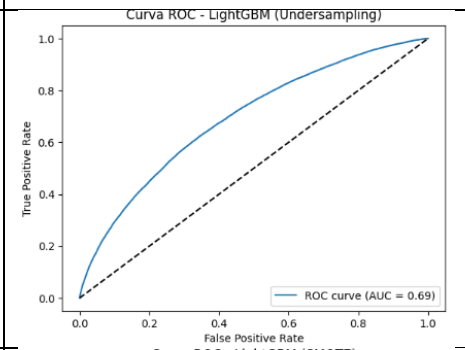
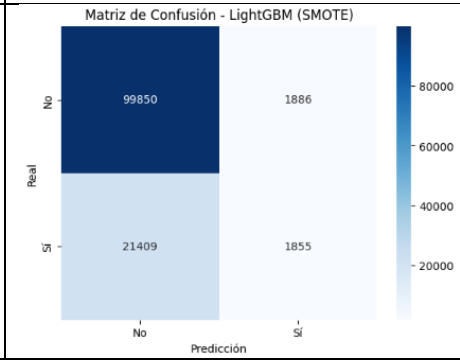
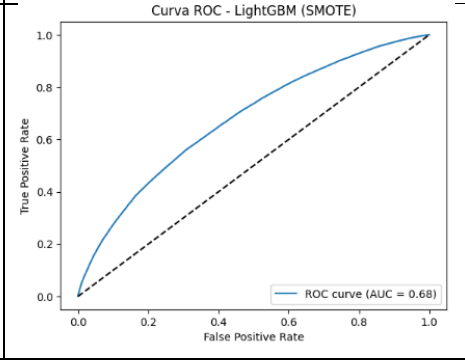
LGBM sin OHE		
	Matriz de Confusión	Curva ROC
Sin Balanceo		
Con Class_Weight		
Undersamplin g		
SMOTE		

Figura 20. Matriz de Confusión y Curva ROC para LGBM sin OHE.

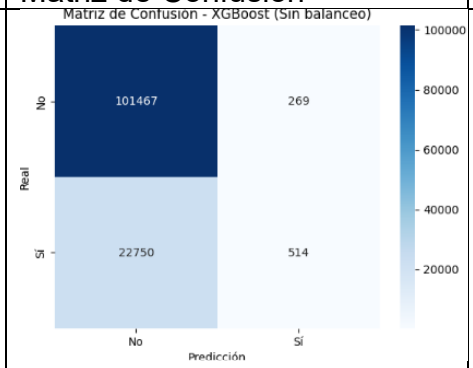
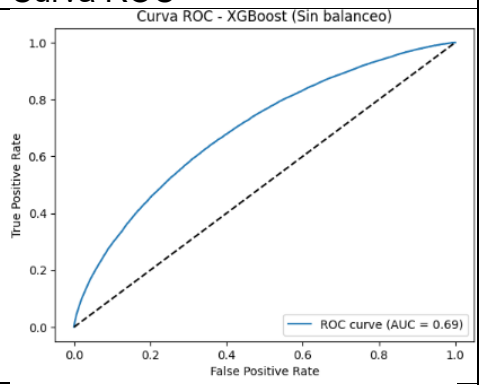
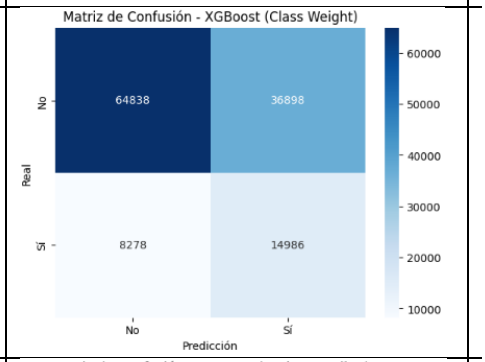
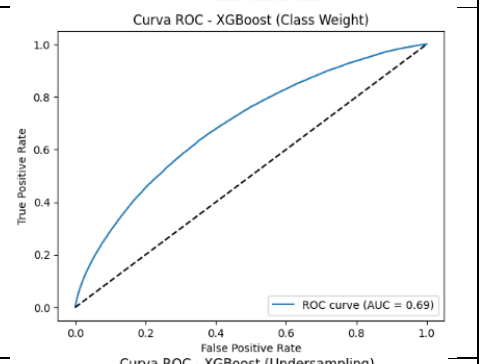
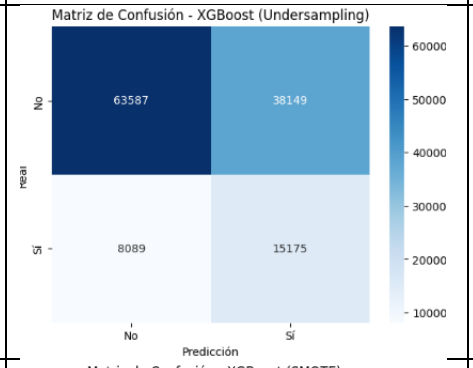
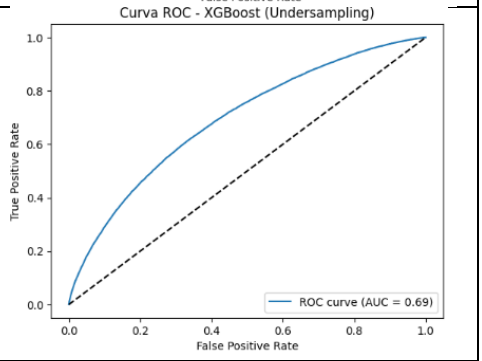
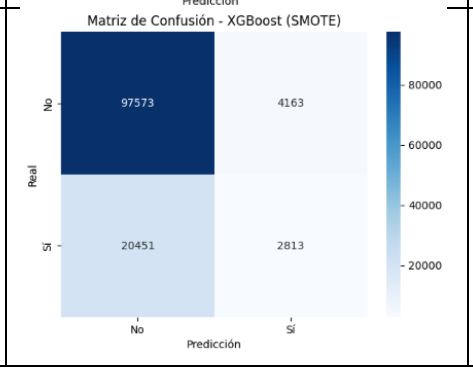
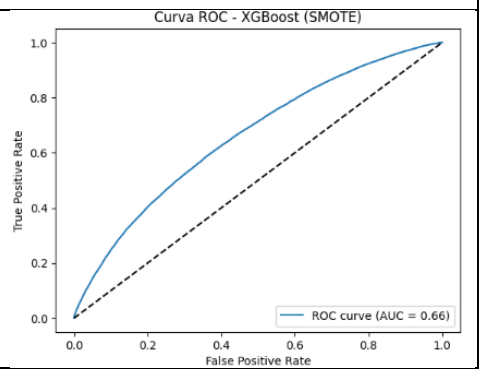
XBoost.		
	Matriz de Confusión	Curva ROC
Sin Balanceo		
Con Class_Weight		
Undersampling		
SMOTE		

Figura 21. Matriz de Confusión y Curva ROC para XBoost.

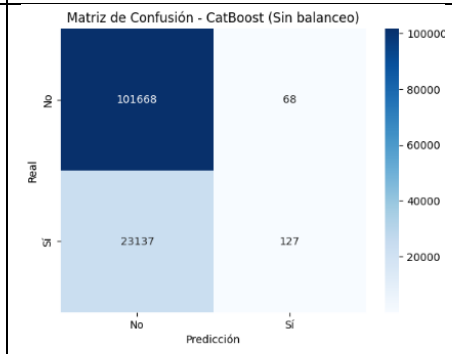
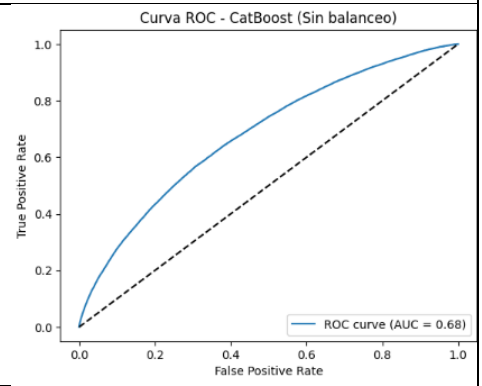
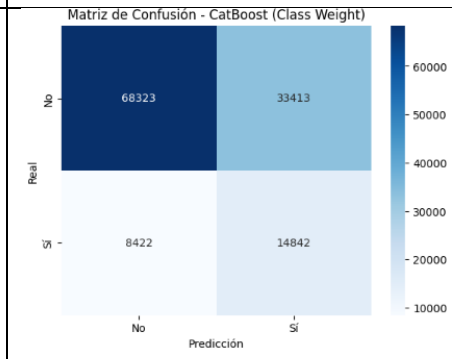
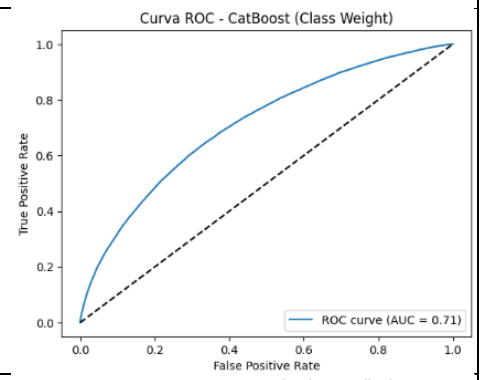
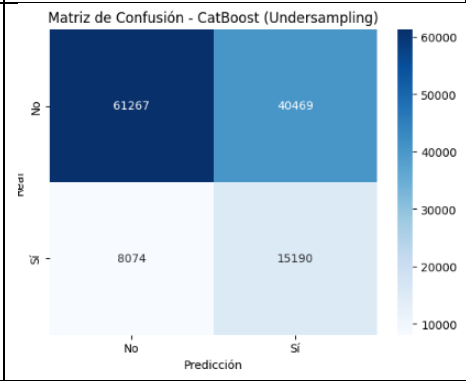
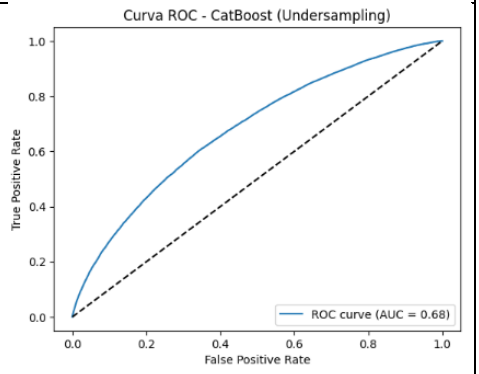
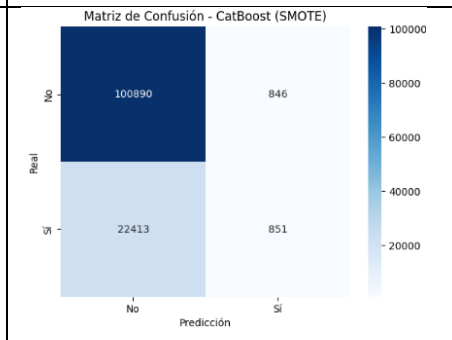
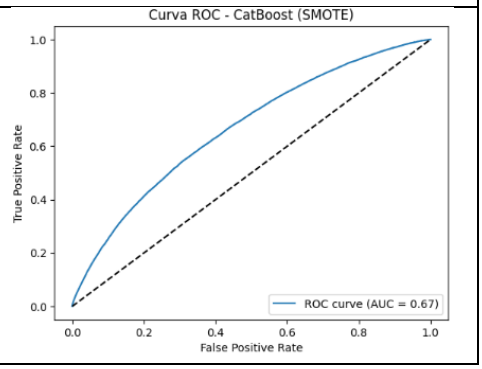
CastBoost		
	Matriz de Confusión	Curva ROC
Sin Balanceo		
Con Class_Weight		
Undersampling		
SMOTE		

Figura 22. Matriz de Confusión y Curva ROC para CastBoost.

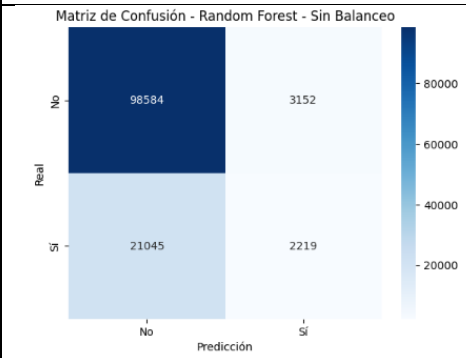
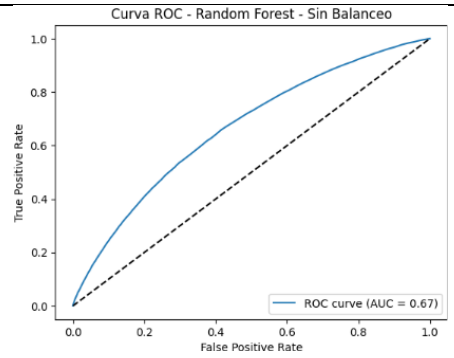
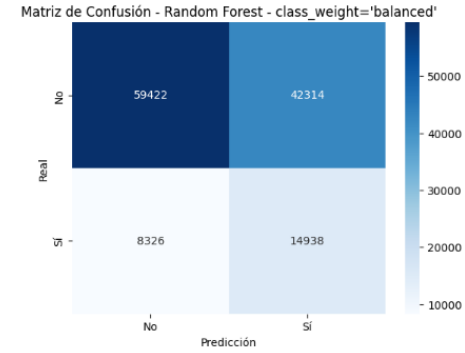
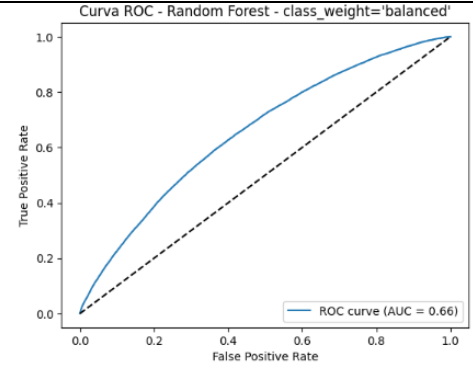
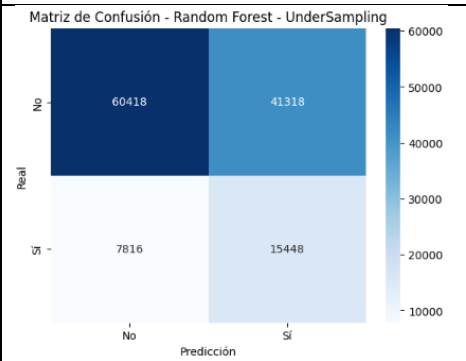
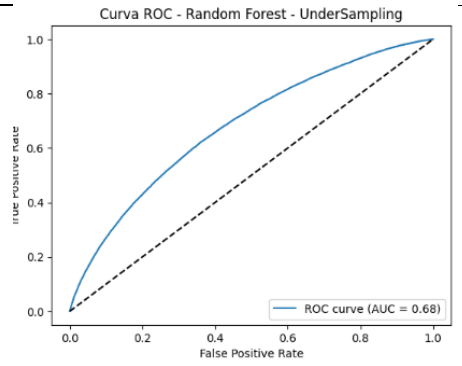
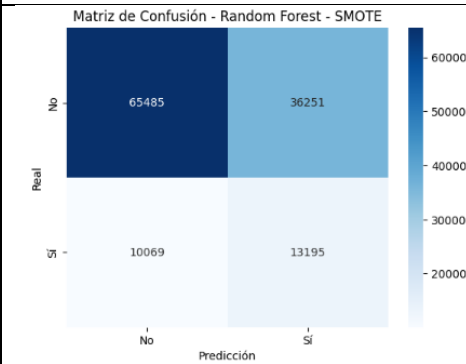
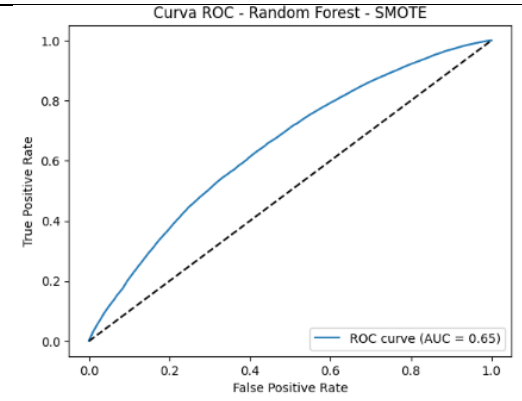
Random Forest		
	Matriz de Confusión	Curva ROC
Sin Balanceo		
Con Class_Weight		
Undersampling		
SMOTE		

Figura 23. Matriz de Confusión y Curva ROC para Random Forest

Podemos notar lo siguiente:

1. Regresión Logistica

- Sin balanceo: El modelo es muy conservador y predice casi siempre la clase mayoritaria (no retraso). Esto se refleja en que hay muchísimos verdaderos negativos (101729) y casi ningún verdadero positivo (5). Por eso, el AUC es bajo (0.637).
- Con balanceo: Al aplicar *class weight*, *undersampling* o *SMOTE*, el modelo aumenta significativamente la cantidad de verdaderos positivos detectados (alrededor de 14,500), lo que indica que ahora sí identifica mejor los vuelos retrasados. Sin embargo, también aumenta mucho la cantidad de falsos positivos (más de 43,000), lo que significa que predice que muchos vuelos que no se retrasan, sí se retrasarán (alertas falsas).
- AUC estable: El AUC no mejora realmente, porque aunque el modelo detecta más positivos, también confunde mucho más las clases.

2. Hist Gradient Boosting

- Sin balanceo: Tiene un buen AUC (0.709), y aunque no detecta muchos positivos (1087), es mucho mejor que regresión logística sin balanceo. La matriz muestra un buen equilibrio entre falsos positivos y negativos.
- Undersampling: Reduce la cantidad total de datos para balancear clases. El AUC baja ligeramente (0.701), pero mejora la detección de verdaderos positivos (15047), aunque aumenta muchos falsos positivos (35804).
- SMOTE: Similar a undersampling, pero genera datos sintéticos. Detecta menos VP que undersampling, pero mantiene buen balance entre FP y VP.

3. Árbol de Decisión

- El árbol de decisión simple tiene el peor desempeño entre los modelos, con AUC muy bajo (0.56 sin balanceo).
- Con balanceo (class weight y undersampling) mejora el AUC (0.65), y aumenta la detección de retrasos (más verdaderos positivos).
- SMOTE no mejora mucho, con un AUC de 0.595.
- La matriz muestra que este modelo comete muchos errores y no es tan fiable para este problema.

4. LightGBM (con y sin One-Hot Encoding)

- El rendimiento es muy parecido con o sin One-Hot Encoding.
- Sin balanceo, LightGBM tiene el mejor AUC (0.709).
- Balancear con class weight o undersampling mantiene un AUC similar (0.69-0.70), con aumento de verdaderos positivos.
- SMOTE disminuye un poco el AUC y genera más falsos positivos.
- LightGBM es uno de los mejores modelos para este problema, con o sin OHE. Balancear ayuda a mejorar recall.

5. XGBoost

    - Rendimiento similar a LightGBM.
    - Sin balanceo detecta pocos retrasos, pero con buena precisión.
    - Balanceo mejora sensibilidad, pero aumenta falsos positivos.
    - SMOTE reduce AUC considerablemente.

6. Catboost

    - Mejor AUC con class weight (0.711), que es el valor más alto en tu tabla.
    - Detecta más positivos con balanceo, aunque incrementa falsos positivos.
    - SMOTE no mejora y reduce el AUC.

7. Random Forest

    - Rendimiento inferior comparado con boosting.
    - AUC relativamente bajo, balancear mejora un poco con undersampling pero con muchos falsos positivos.
    - Muchos falsos positivos y menos precisos que boosting.

CONCLUSIONES

Con el desarrollo de este proyecto se obtuvieron las siguientes conclusiones:

- Los retrasos en vuelos comerciales pueden ser anticipados con una precisión aceptable mediante modelos de clasificación binaria, especialmente utilizando variables disponibles antes del despegue, como hora programada, aerolínea y distancia del vuelo.
- El modelo CatBoost con class_weight fue el que presentó el mejor desempeño general, con un F1-score de 0.415 y un recall de 0.638, siendo capaz de identificar con eficacia los vuelos que probablemente se retrasen.
- Se confirmó que factores como el horario de salida, la aerolínea, el día de la semana y la distancia influyen significativamente en la probabilidad de retraso, lo cual fue respaldado por los análisis descriptivos y gráficos.
- El análisis visual y la matriz de correlación revelaron que algunas variables, como ARRIVAL_DELAY y SCHEDULED_DEPARTURE, tienen correlaciones importantes con los retrasos, mientras que otras como el mes o el día presentan una relación más débil.
- A pesar del desbalance en la variable objetivo (solo el 18.6% de los vuelos fueron retrasados), las técnicas de balanceo como class_weight y undersampling permitieron mejorar significativamente el desempeño de los modelos.

RECOMENDACIONES

1. Actualizar el modelo con datos recientes: los patrones de retraso pueden cambiar con el tiempo debido a nuevas políticas, rutas o eventos externos. Un modelo actualizado periódicamente será más efectivo.
2. Incluir variables adicionales como clima y tráfico aéreo si están disponibles, ya que estas pueden aumentar considerablemente la capacidad predictiva del modelo.
3. Implementar el modelo en un entorno de prueba real, como un dashboard con alertas tempranas, para evaluar su valor práctico en operaciones diarias de aerolíneas o aeropuertos.
4. Monitorear continuamente el rendimiento del modelo, especialmente el recall, para asegurar que se mantenga útil en la detección de vuelos que efectivamente se retrasan.
5. Capacitar al personal operativo sobre el uso del modelo y la interpretación de sus predicciones, a fin de convertir las alertas en acciones concretas que reduzcan los retrasos. Esto es en caso de llevarlo a producción

FUTUROS ESTUDIOS

1. Aplicar el modelo a otros países, como Panamá, cuando se disponga de datasets locales, para evaluar si el comportamiento de los retrasos es similar o distinto.
2. Incorporar modelos más avanzados como redes neuronales recurrentes (RNN) si se tiene acceso a datos secuenciales o series de tiempo, para mejorar la predicción.
3. Integrar variables meteorológicas y de congestión aeroportuaria, usando APIs externas, para enriquecer la predicción.
4. Explorar explicabilidad del modelo con herramientas como SHAP o LIME, para entender mejor por qué el modelo predice que un vuelo se retrasará.

REFERENCIAS

- FasterCapital. (23 de julio de 2025). *FasterCapital*. Obtenido de Datos de aviacion retrasos en los vuelos y analisis predictivo un enfoque basado en datos: https://fastercapital.com/es/contenido/Datos-de-aviacion--retrasos-en-los-vuelos-y-analisis-predictivo--un-enfoque-basado-en-datos.html#Por-qu--son-importantes-los-retrasos-en-los-vuelos-y-c-mo-puede-ayudar-el-an-lisis-predictivo-
- Kaggle. (s.f.). *2015 Flight Delays and Cancellations*. Obtenido de Kaggle: https://www.kaggle.com/datasets/usdot/flight-delays/data
- Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. In *Advances in Neural Information Processing Systems* (Vol. 30). https://papers.nips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: Unbiased boosting with categorical features*. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 6639–6649). https://papers.nips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf
- Quinlan, J. R. (1986). *Induction of decision trees*. Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251

ANEXOS

Todo el material se encuentra en la plataforma de GitHub

Bajo el link: https://github.com/Angiel120995/Proyecto_Modelos_Predictivos

Que contiene:

- La propuesta
- Avances
- Trabajo Final
- Excel de Las Variables total del dataset con su descripción
- Excel de las Evaluaciones de los Modelos
- Carpeta con los Dataset Originales
- Presentación del Story Telling
- Códigos del Proyecto
    - ❖ ProyectoMP_EDA_Limpieza (en esta sección se encuentra la primera parte de la limpieza de variables)
    - ❖ Limpieza Adicional: Segunda parte de la limpieza y variables predictivas finales
    - ❖ Unbalance_GS: (Evaluación de 4 modelos sin balanceo)
    - ❖ Class_Balance_GS: (Evaluación de 3 modelos con balanceo class_weight)
    - ❖ Undersampling_GS: (Evaluación de 4 modelos con balanceo Undersampling)
    - ❖ SMOTE_GS: (Evaluación de 3 modelos con balanceo SMOTE)
    - ❖ Random_Forest: Evaluacion del modelo RF sin el balanceo y con los balanceos correspondiente
    - ❖ LightunOHE, Catboost, XGoost: Evaluación de Modelos sin OHE