Universidad Tecnológica de Panamá
Propuesta de Proyecto Final De la Materia de Modelos Predictivos

Nombre: Angiel Fernández
Cédula: 8-897-592
Grupo: 1AN216
Profesor: Juan Marcos Castillo, PhD

- **NOMBRE DE LA INVESTIGACIÓN**

Este proyecto lleva por nombre: Predicción de Retrasos en Vuelos Comerciales utilizando Modelos de Clasificación

- **DESCRIPCIÓN DE LA BASE DE DATOS**

Se realizó una exploración inicial de la base de datos principal, la cual lleva por nombre flight.csv, cuenta con 5819079 filas y 31 columnas las mismas están descritas en la tabla 1.

Tabla 1: Información General del dataset

Columna		Descripción
YEAR	Año	Año del vuelo
MONTH	Mes	Mes del vuelo (1 a 12).
DAY	Día	Día del mes en que se realizó el vuelo.
DAY_OF_WEEK	Día de la semana	Día de la semana (1 = lunes, 7 = domingo).
AIRLINE	Aerolínea	Código de dos letras que identifica a la aerolínea.
FLIGHT_NUMBER	Número de vuelo	Número asignado al vuelo por la aerolínea.
TAIL_NUMBER	Número de matrícula del avión	Identificación única del avión (como su placa).
ORIGIN_AIRPORT	Aeropuerto de origen	Código IATA del aeropuerto desde donde parte el vuelo.
DESTINATION_AIRPORT	Aeropuerto de destino	Código IATA del aeropuerto al que llega el vuelo.
SCHEDULED_DEPARTURE	Salida programada	Hora programada para la salida (en formato HHMM).
DEPARTURE_TIME	Hora real de salida	Hora en que realmente despegó el avión (formato HHMM).
DEPARTURE_DELAY	Retraso en la salida (minutos)	Minutos de retraso en la salida respecto al horario programado.
TAXI_OUT	Rodaje de salida (minutos)	Tiempo desde que el avión deja la puerta hasta que despega.
WHEELS_OFF	Hora de despegue	Hora exacta en que el avión despegó (ruedas fuera del suelo).
SCHEDULED_TIME	Tiempo de vuelo programado (minutos)	Tiempo estimado de vuelo incluyendo taxi (salida + llegada).
ELAPSED_TIME	Tiempo transcurrido (minutos)	Duración real total del vuelo desde puerta a puerta.
AIR_TIME	Tiempo en el aire (minutos)	Tiempo efectivo que el avión estuvo volando.
DISTANCE	Distancia del vuelo (millas)	Distancia entre aeropuertos en millas.
WHEELS_ON	Hora de aterrizaje	Hora en que el avión tocó tierra.
TAXI_IN	Rodaje de llegada (minutos)	Tiempo desde que aterriza hasta llegar a la puerta del destino.
SCHEDULED_ARRIVAL	Llegada programada	Hora programada de llegada (formato HHMM).
ARRIVAL_TIME	Hora real de llegada	Hora real en que el avión llegó a la puerta.
ARRIVAL_DELAY	Retraso en la llegada (minutos)	Minutos de retraso al llegar respecto al horario programado.
DIVERTED	Desviado	1 si el vuelo fue desviado, 0 si no.
CANCELLED	Cancelado	1 si el vuelo fue cancelado, 0 si no.
CANCELLATION_REASON	Motivo de cancelación	Motivo por el cual se canceló el vuelo (A, B, C o D).
AIR_SYSTEM_DELAY	Retraso del sistema aéreo (minutos)	Retraso por congestión aérea, control de tráfico o gestión del espacio.
SECURITY_DELAY	Retraso por seguridad (minutos)	Retraso por inspección de seguridad u otras causas similares.
AIRLINE_DELAY	Retraso por aerolínea (minutos)	Retraso atribuible a la aerolínea (como logística o personal).
LATE_AIRCRAFT_DELAY	Retraso por llegada tardía del avión	Retraso porque el avión llegó tarde de su vuelo anterior.
WEATHER_DELAY	Retraso por clima (minutos)	Retraso debido a condiciones climáticas adversas.

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) y airport.csv (contiene códigos de la aerolínea, nombre del aeropuerto, e información geográfica como longitud, latitud ciudad, estado)

- **SELECCIÓN DE LOS DATOS**

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 calara 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.
- Podemos aplicar modelos de clasificación como Regresión logística, *Random Forest*, entre otros y compararlos para visualizar su efectividad.
- Es fácil de interpretar y se puede aplicar en lo cotidiano.

Además, es un tema interesante, cercano a mi campo laboral y común a la experiencia de la mayoría de las personas, lo que facilita su análisis y presentación.

- **INTRODUCCIÓN DEL CONTEXTO**

Los retrasos en vuelos representan una gran preocupación tanto para los pasajeros como para las aerolíneas. Identificar los factores que influyen en los retrasos permite mejorar la planificación, asignación de recursos y la experiencia del cliente. Gracias a los avances en ciencia de datos, hoy es posible utilizar modelos predictivos para anticipar si un vuelo pudiese retrasarse en función de múltiples variables históricas.
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.

- **LÍNEA DE TIEMPO DE INVESTIGACIÓN**

La línea de tiempo se presenta en la tabla 2.

Tabla 2: Línea de tiempo de investigación.

Fase	Actividad principal	Fechas propuestas
Fase 1: Propuesta e investigación inicial	*Redacción y entrega de la propuesta del proyecto *Revisión del dataset principal y complemantarios *Definición del objetivo (clasificación: retraso o no)	7 al 9 de julio
Fase 2: Limpieza y preparación de datos	*Carga y limpieza de datos (Eliminación de valores faltantes y columnas innecesarias) en KNIME o Alteryx *Carga en Google Colab *Codificación de variables y creación de variable objetivo *Elaboración del primer Avance	10 al 13 de julio
Fase 3: Entrenamiento de modelos predictivos	*División de datos en entrenamiento y prueba *Entrenamiento con Regresión Logística y Random Forest (principalmente) evaluar si usar otros modelos de clasificacion. *Evaluación de desempeño del modelo	14 al 17 de julio
Fase 4: Visualización y análisis de resultados	*Exportación de resultados *Creación de visualizaciones en Power BI (por aerolínea, aeropuerto, hora, etc.)	18 al 20 de julio
Fase 5: Redacción final y presentación	*Elaboración del reporte *Preparación de la presentación (storystelling) para exposición del proyecto	21 al 23 de julio