Caso Final Integrador

Análisis de Tráfico Aéreo en el Aeropuerto de San Francisco

María Ángel Lobón

Universidad CEU San Pablo

Big Data

Rubén Juárez Cádiz

10/11/2024

Introudcción

El objetivo de este análisis es introducir los conceptos básicos de la ciencia de datos aplicada al tráfico aéreo, explorando y comprendiendo un conjunto de datos que contiene información valiosa sobre aerolíneas, destinos, y patrones de tráfico. Este análisis permitirá obtener un entendimiento más profundo de los comportamientos del transporte aéreo y apoyar la toma de decisiones informadas en esta industria. El conjunto de datos incluye múltiples columnas, cada una con un propósito específico: el año (Year) y mes (Month) en que ocurre la actividad; la aerolínea operativa (Operating Airline); un resumen geográfico (GEO Summary) que indica si el vuelo es nacional o internacional; la región geográfica (GEO Region) del destino o salida, como Estados Unidos, Europa, y Asia; un código de actividad (Activity Type Code) que clasifica la operación como embarque o desembarque; y el número de pasajeros (Passenger Count), que refleja el tráfico en cada registro. Este análisis inicial permitirá identificar tendencias y patrones en el flujo de pasajeros y en la popularidad de aerolíneas y destinos, generando conocimientos útiles para optimizar decisiones en el sector.

Para lograrlo, se llevarán a cabo diversas etapas: preparación y procesamiento de los datos en entornos distribuidos con la ayuda de PySpark y almacenamiento en Cassandra, además de un análisis descriptivo y predictivo de variables como flujo de pasajeros, destinos y frecuencia de vuelos mediante técnicas de modelado y correlación. En caso de que el tamaño de los datos no sea demasiado grande, se puede emplear Python con sus abundantes librerías, como Scikit-Learn, Matplotlib y Numpy, que permiten realizar análisis y visualización de datos de manera eficiente. También se agregan capas de seguridad de datos mediante la autenticación y autorización con Kerberos, Knox, Ranger y Sentry en Hadoop. Finalmente, los resultados se

visualizarán mediante herramientas interactivas como Tableau, CartoDB y librerías avanzadas como D3.js y Leaflet. En este informe se discutirán los hallazgos del análisis y se propondrán decisiones fundamentadas en los datos.

Metodología

Limpieza de Datos

En este paso del proceso de limpieza de datos, se realizaron dos tareas clave: la eliminación de valores nulos y la eliminación de registros duplicados. Primero, se analizó el conjunto de datos para identificar la cantidad de valores nulos presentes en cada columna, lo que permitió evaluar si existían datos faltantes que pudieran afectar el análisis. Posteriormente, se eliminaron las filas que contenían valores nulos, asegurándose de que el DataFrame resultante estuviera libre de dichos elementos.

```
Registros antes de eliminar duplicados y nulos: 15007
Activity Period                0
Operating Airline              0
Operating Airline IATA Code    54
Published Airline              0
Published Airline IATA Code    54
GEO Summary                    0
GEO Region                     0
Activity Type Code             0
Price Category Code            0
Terminal                       0
Boarding Area                  0
Passenger Count                0
Adjusted Activity Type Code    0
Adjusted Passenger Count       0
Year                           0
Month                          0
dtype: int64
¿Hay valores nulos en el DataFrame?: True

Verificando valores nulos después de eliminarlos:
Activity Period                0
Operating Airline              0
Operating Airline IATA Code    0
Published Airline              0
Published Airline IATA Code    0
GEO Summary                    0
GEO Region                     0
Activity Type Code             0
Price Category Code            0
Terminal                       0
Boarding Area                  0
Passenger Count                0
Adjusted Activity Type Code    0
Adjusted Passenger Count       0
Year                           0
Month                          0
dtype: int64

¿Hay valores nulos en el DataFrame?: False
```

A continuación, se llevó a cabo la eliminación de registros duplicados para garantizar que cada fila del DataFrame representara información única, evitando redundancias que podrían distorsionar los resultados. Finalmente, se verificó la cantidad de registros antes y después de aplicar estos procesos, identificando si se habían eliminado elementos y cuantificando el impacto de la limpieza en el tamaño del conjunto de datos. Este procedimiento asegura que los datos restantes sean consistentes, completos y confiables para su análisis posterior.

Filtrado y Creación de Subconjunto de Datos

En este bloque de código se realizó un filtrado detallado y la creación de subconjuntos de datos basados en diferentes criterios, lo que permite analizar de manera específica ciertas características del conjunto de datos.

Filtrado 1: Se filtraron los registros correspondientes a la aerolínea "ATA Airlines" durante el año 2005. Este subconjunto ayuda a centrar el análisis en un periodo y aerolínea específicos, facilitando el estudio de su desempeño durante ese año.

```
Filtrar para ATA Airlines en 2005
    Operating Airline  Year   Month  Passenger Count
0         ATA Airlines  2005    July            27271
1         ATA Airlines  2005    July            29131
2         ATA Airlines  2005    July             5415
115       ATA Airlines  2005  August            27472
116       ATA Airlines  2005  August            26535

Filtrar vuelos internacionales hacia Europa
   Operating Airline GEO Region  Passenger Count
7          Air France     Europe            12050
8          Air France     Europe            11638
31   BelAir Airlines     Europe              325
32   BelAir Airlines     Europe              545
33   British Airways     Europe            20632

Extraer datos de 2005 a 2020, solo para vuelos nacionales con más
    Year  Operating Airline  Passenger Count
0   2005       ATA Airlines            27271
1   2005       ATA Airlines            29131
13  2005    Alaska Airlines            36641
14  2005    Alaska Airlines            39379
23  2005  American Airlines           166577
```

Filtrado 2: Se seleccionaron datos de vuelos internacionales hacia Europa. Este filtro se enfocó en registros con "International" como resumen geográfico y "Europe" como región geográfica, permitiendo estudiar los patrones de tráfico aéreo internacional hacia ese continente.

Filtrado 3: Se aplicó un filtro combinado para extraer registros de vuelos domésticos entre 2005 y 2020, con un volumen de pasajeros mayor a 10,000. Este filtro multivariable facilita identificar tendencias significativas en vuelos nacionales de alta demanda durante este periodo.

Además, se realizaron consultas adicionales que combinan múltiples condiciones. Por ejemplo, se seleccionaron vuelos internacionales operados por "ATA Airlines" en 2005, así como vuelos internacionales de "Air France" entre 2005 y 2009. Estas consultas permiten obtener insights específicos sobre el tráfico aéreo de dichas aerolíneas en distintos periodos. Este enfoque

estructurado y detallado garantiza la extracción de subconjuntos de datos significativos, lo que es esencial para realizar análisis enfocados y relevantes.

```
Vuelos Internacionales de ATA Airlines en 2005
    Operating Airline GEO Region  Passenger Count
473      ATA Airlines     Canada              284
474      ATA Airlines     Canada              284

Vuelos Internacionales de Air France entre 2005 y 2009
    Operating Airline    GEO Summary  Passenger Count  Year
7          Air France  International            12050  2005
8          Air France  International            11638  2005
122        Air France  International            11230  2005
123        Air France  International            11731  2005
240        Air France  International            10731  2005
```

Analysis Descriptivo

En este bloque del código se lleva a cabo un análisis descriptivo y se exploran correlaciones entre variables, lo cual es fundamental para obtener una comprensión inicial de los datos y sus patrones.

```
count     14953.000000
mean      29345.619006
std       58398.448380
min           1.000000
25%        5409.000000
50%        9260.000000
75%       21222.000000
max      659837.000000
```

Estadísticas descriptivas: Se calculan métricas como el promedio, mediana, desviación estándar, valores mínimo y máximo, y los percentiles del número de pasajeros. Estas estadísticas ofrecen un panorama general del rango y distribución de los datos, identificando tendencias como la concentración de valores o la presencia de extremos.

Analisis Descriptivo mas extenso: Para intentar entender patrones podemos hacer unas agrupaciones iniciales, las cuales nos pueden guiar de cara a la creación de los modelos. Este análisis descriptivo y de agrupación permite caracterizar los datos, y también sienta las bases para exploraciones más avanzadas, como identificar relaciones significativas o realizar predicciones basadas en estas observaciones iniciales. Un ejemplo puede ser encontrar el

promedio de pasajeros por aerolinea, donde se agrupan los datos por aerolínea y se calcula el promedio de pasajeros por cada una, lo que permite identificar cuáles aerolíneas tienen un mayor tráfico promedio de pasajeros. Los resultados se ordenan en orden descendente para destacar las aerolíneas con mayor impacto en términos de volumen de pasajeros. Otro ejemplo es el promedio de pasajeros por aerolinea y destino, donde se compara el número promedio de pasajeros por aerolínea y región geográfica. Este análisis más detallado combina dos variables para identificar patrones regionales en el tráfico aéreo, permitiendo descubrir, por ejemplo, si ciertas aerolíneas tienen un mayor volumen de pasajeros en regiones específicas.

```
Promedio de pasajeros por aerolínea
Operating Airline
American Airlines     127164.389706
Southwest Airlines     81188.158576
Virgin America         74405.353591
United Airlines        72732.058296
Delta Air Lines        68498.497409
```

```
Comparar el número promedio de pasajeros por aerolínea y destino.
Operating Airline                GEO Region
American Airlines                US          127164.389706
United Airlines                  US          125233.556306
Virgin America                   US          124884.199052
United Airlines - Pre 07/01/2013 US          102855.210393
Delta Air Lines                  US           90706.919580
```

Matrices de Correlacion

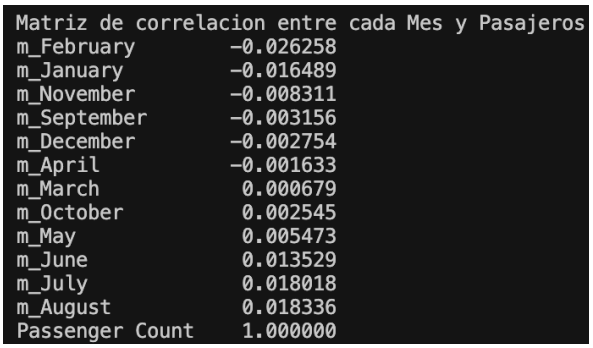
Este bloque de código se enfoca en el análisis de correlaciones para identificar relaciones significativas entre variables, lo que es crucial para detectar patrones y dependencias en los datos. Este análisis de correlaciones proporciona una visión cuantitativa de las relaciones entre las variables del conjunto de datos, sirviendo como base para modelos predictivos o decisiones estratégicas como la optimización de rutas o recursos.

Primera matriz de correlación (año y pasajeros): Se calcula una matriz de correlación básica entre el año y el número de pasajeros, empleando únicamente variables numéricas del DataFrame. Esto permite determinar si existe una relación temporal significativa en el aumento o disminución del tráfico de pasajeros. El coeficiente de correlación entre el Año y el

```
Matriz de correlacion entre Año y Pasajeros
                    Year  Passenger Count
Year            1.000000         0.060917
Passenger Count 0.060917         1.000000
```

número de Pasajeros es **0.060917**, lo cual indica una correlación muy baja y positiva. Esto sugiere que el aumento o cambio en el número de pasajeros no está fuertemente relacionado con el año de forma lineal. En otras palabras, aunque pueda haber ligeras variaciones en el número de pasajeros a lo largo del tiempo, estas no son consistentes ni significativas.

Segunda matriz de correlación (meses y pasajeros): Para analizar la relación entre los meses y el número de pasajeros, las variables categóricas de los meses se convierten en formato numérico mediante one-hot encoding. Esto genera una matriz de correlación más detallada, mostrando cómo cada mes del año influye en la cantidad de pasajeros. Por ejemplo, se puede observar si hay meses con un tráfico consistentemente alto debido a factores como vacaciones o temporadas específicas. La mayoría de los valores están muy cercanos a 0, lo que indica que no hay una relación lineal significativa entre los meses específicos y el número de pasajeros. Esto sugiere que el volumen de pasajeros no está fuertemente afectado por la temporada o el mes, al menos de manera directa. Algunos meses, como **February (-0.026258)**, **January (-0.016489)** y **November (-0.008311)**, tienen correlaciones negativas muy bajas con el número de pasajeros. Esto podría indicar que en estos meses hay una leve tendencia a tener menos pasajeros en comparación con otros meses.



```
Matriz de correlacion entre cada Mes y Pasajeros
m_February       -0.026258
m_January        -0.016489
m_November       -0.008311
m_September      -0.003156
m_December       -0.002754
m_April          -0.001633
m_March           0.000679
m_October         0.002545
m_May             0.005473
m_June            0.013529
m_July            0.018018
m_August          0.018336
Passenger Count   1.000000
```

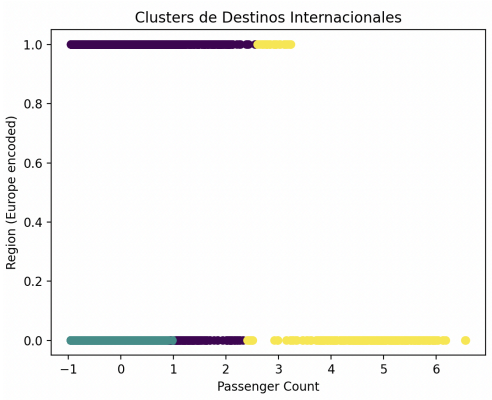
Tercera matriz de correlación (destinos y pasajeros): Se realiza un proceso similar al de los meses, pero esta vez para los destinos. Mediante la codificación one-hot, se genera una matriz que permite analizar qué regiones geográficas tienen una mayor correlación con el volumen de pasajeros. Esto es útil para identificar los destinos más relevantes en términos de tráfico aéreo. De los resultados podemos concluir: **GEO_US (0.398198)** tiene una correlación

positiva moderada con el número de pasajeros. Esto sugiere que los vuelos hacia destinos dentro de los Estados Unidos tienden a tener un mayor número de pasajeros en comparación con otras regiones. Esto es consistente con la posible alta demanda de vuelos domésticos en EE.UU., debido a su gran tamaño, población y economía. Todas las demás regiones tienen correlaciones negativas con el número de pasajeros, aunque los valores son bajos. Por ejemplo, **GEO_Asia (-0.144164)** tiene la correlación negativa más alta. Esto podría indicar que los vuelos hacia Asia representan una proporción menor del volumen total de pasajeros. **GEO_Europe (-0.113674)** y **GEO_Canada (-0.108458)** también tienen correlaciones negativas, aunque algo menores, lo que sugiere que los vuelos hacia estas regiones tienen un impacto más limitado en el volumen total de pasajeros.

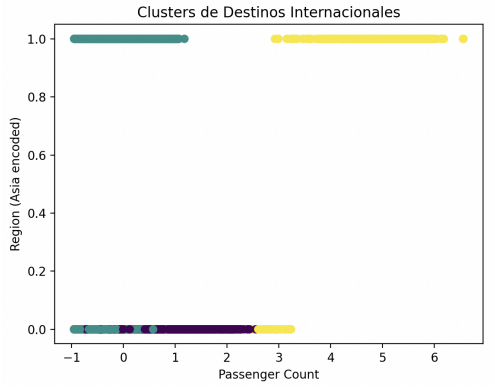
```
Matriz de correlacion entre cada Destino y Pasajeros
GEO_Asia                  -0.144164
GEO_Europe                -0.113674
GEO_Canada                -0.108458
GEO_Mexico                -0.107775
GEO_Australia / Oceania   -0.089400
GEO_Central America       -0.056790
GEO_Middle East           -0.042686
GEO_South America         -0.035392
GEO_US                     0.398198
Passenger Count            1.000000
```

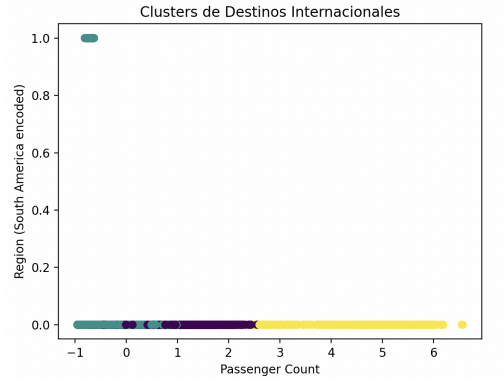
Clusters

Este bloque de análisis utiliza clustering para agrupar destinos internacionales en función del número de pasajeros y la región geográfica. Primero, los datos se preparan filtrando vuelos internacionales y seleccionando las variables relevantes, como el número de pasajeros y la región geográfica. El número de pasajeros se normaliza utilizando StandardScaler para evitar que las diferencias en las escalas numéricas afecten el análisis, y las regiones se convierten en variables dummy mediante pd.get_dummies. Luego, se aplica el



algoritmo K-Means con 3 clusters, asignando a cada registro un número de cluster que identifica a qué grupo pertenece. Finalmente, se generan gráficos de dispersión para visualizar los clusters, utilizando el número de pasajeros en el eje X y las regiones codificadas como eje Y, con colores que representan los clusters asignados.





Un buen ejemplo de clustering se observa cuando los datos se agrupan claramente por colores, como en la visualización de Asia, lo que indica que los destinos comparten características similares, como niveles de tráfico. Estos resultados pueden interpretarse para segmentar destinos en grupos con bajo, moderado o alto tráfico de pasajeros, ofreciendo información útil para decisiones operativas, como ajustar frecuencias de vuelo, capacidad o estrategias comerciales. Si los puntos de los clusters estuvieran mezclados, podría indicar que el número de clusters no es adecuado o que las variables seleccionadas no distinguen suficientemente los datos.

Modelos Predictivos

En este bloque, se utilizan varios modelos de regresión lineal para predecir el número de pasajeros en vuelos según diferentes factores. El primer modelo (Modelo 1) utiliza las variables 'Year' y 'Month' como predictoras para estimar el número de pasajeros ('Passenger Count'). Se realiza un preprocesamiento de las variables categóricas mediante la codificación one-hot para el mes, y luego se dividen los datos en conjuntos de entrenamiento y prueba (80%-20%). El modelo

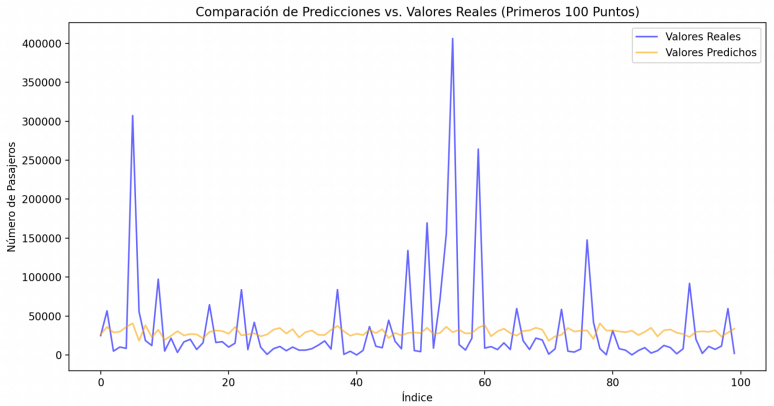
de regresión lineal es entrenado con el conjunto de entrenamiento y luego se predicen los valores en el conjunto de prueba. El segundo modelo (Modelo 2) también utiliza el 'Year' y 'GEO Region' como variables predictoras, para predecir el número de pasajeros en destinos específicos. Similar al primer modelo, se aplican transformaciones de variables categóricas usando one-hot encoding para la región geográfica. Después, los datos se dividen en conjuntos de entrenamiento y prueba, y se entrena un modelo de regresión lineal que genera predicciones. Finalmente, el tercer modelo (Modelo 3) utiliza las variables 'Year', 'Month', 'GEO Summary' (nacional o internacional), y 'Operating Airline' para predecir el número de pasajeros. Las variables categóricas se transforman mediante one-hot encoding, y el modelo se entrena con los datos de entrenamiento y realiza predicciones sobre el conjunto de prueba. En todos los casos, el parámetro objetivo es el 'Passenger Count' y se utilizan los mismos pasos de preprocesamiento y división de datos en entrenamiento y prueba para cada modelo.

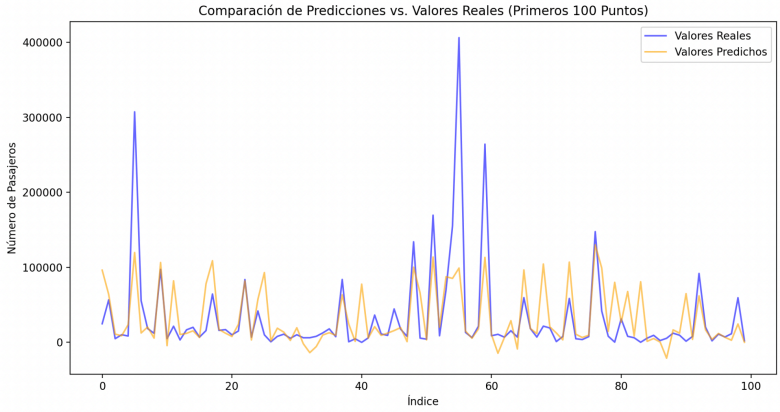
Comparación y Valores

El análisis comparativo de los tres modelos de regresión lineal utilizados para predecir el número de pasajeros en vuelos muestra que, aunque todos los modelos presentan un rendimiento similar en términos de predicción, existen diferencias significativas en cuanto a la precisión y eficiencia de sus predicciones, que se pueden evaluar utilizando métricas como el Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE).

En el caso del Modelo 1, que utiliza las variables 'Year' y 'Month' como predictores, se observa que el error cuadrático medio (RMSE) es de 53,778.71, mientras que el error absoluto medio (MAE) es de 31,165.68. El RMSE relativamente alto sugiere que el modelo no captura con precisión las variaciones en los datos, y el MAE también indica un margen considerable de

error promedio por cada predicción. Además, el intercepto es un valor extremadamente grande, lo que indica que el modelo podría estar teniendo problemas con el ajuste de los datos debido a la escala de los valores. Por otro lado, el Modelo 2, que se centra en la variable 'GEO Region' junto con 'Year', tiene un RMSE de 48,612.47 y un MAE de 24,701.56. A pesar de que estos errores son menores que los del Modelo 1, todavía muestran un margen de mejora considerable. El intercepto es más razonable en comparación con el Modelo 1, lo que sugiere una mejor capacidad de ajuste del modelo, pero el RMSE aún sigue siendo relativamente alto, lo que implica que la predicción de pasajeros aún tiene una cantidad importante de variabilidad no explicada. Finalmente, el Modelo 3, que utiliza las variables 'Year', 'Month', 'GEO Summary', y 'Operating Airline', muestra el mejor rendimiento en términos de precisión. El RMSE es 42,402.47 y el MAE es 19,893.92, los cuales son significativamente más bajos que los de los otros dos modelos. Este modelo parece capturar mejor las relaciones entre las variables y los pasajeros, lo que se refleja en su menor error cuadrático medio y error absoluto medio. Además, el intercepto, aunque negativo, es más razonable que en el Modelo 1, lo que sugiere un mejor ajuste.





En resumen, el Modelo 3 es el mejor modelo predictivo, ya que tiene los errores más bajos en ambas métricas (RMSE y MAE), lo que indica que realiza las predicciones con mayor precisión y menor margen de error. Este modelo, al incluir más variables relevantes, parece ser capaz de capturar una mayor cantidad de variabilidad en los datos, lo que lo hace más robusto y confiable para predecir el número de pasajeros en vuelos. A pesar de que el Modelo 2 también muestra una mejora respecto al Modelo 1, el Modelo 3 es claramente superior, lo que sugiere que la inclusión de más características (como el 'Operating Airline' y el 'GEO Summary') en el modelo tiene un impacto positivo en su capacidad predictiva.

Conclusiones

En conclusión, este proyecto ha proporcionado una comprensión integral de los datos relacionados con el tráfico aéreo, empleando diversas técnicas de análisis de datos, modelado predictivo y clustering para explorar, analizar y predecir patrones en el flujo de pasajeros y la dinámica del transporte aéreo. A través de la limpieza y preparación de los datos, se lograron eliminar duplicados y valores nulos, asegurando la calidad de la información utilizada para los modelos de análisis y predicción. El análisis descriptivo inicial permitió identificar tendencias clave en el tráfico aéreo, como las fluctuaciones mensuales y regionales en el número de pasajeros, y las correlaciones entre variables como las aerolíneas y los destinos. El uso de técnicas de clustering, específicamente el algoritmo K-Means, permitió segmentar los destinos internacionales en clusters con características similares, lo que proporciona valiosa información para la toma de decisiones estratégicas en la industria aérea. Esta segmentación de destinos por características del tráfico de pasajeros ayuda a identificar patrones de demanda y preferencias geográficas, lo cual es útil para la optimización de rutas y recursos.

En cuanto a la predicción del tráfico de pasajeros, se aplicaron modelos de regresión lineales para estimar los flujos futuros de pasajeros. A través de la comparación de tres modelos diferentes, se determinó que el Modelo 3, que consideraba múltiples variables como el año, el mes, la región geográfica y la aerolínea operativa, fue el más eficaz, con los menores errores de predicción (RMSE y MAE), lo que sugiere que la inclusión de variables adicionales mejora significativamente la precisión del modelo.

Este proyecto ha demostrado cómo los análisis estadísticos, el clustering y los modelos predictivos pueden ser herramientas poderosas para extraer información valiosa de conjuntos de datos complejos en la industria del transporte aéreo. Los resultados obtenidos ofrecen perspectivas que pueden ser aplicadas en la toma de decisiones empresariales, optimización de rutas, y la predicción del comportamiento del mercado, proporcionando una base sólida para futuras investigaciones y mejoras en la gestión de tráfico aéreo.