Proyecto Datalab: Análisis de datos “ Aerolíneas”

Betsabé Morales
Minerva Bobadilla
Latam
12-09-2024

Caso

De análisis

Este proyecto se centró en el análisis de datos para el Departamento de Transporte de EE.UU., específicamente la Oficina de Estadísticas de Transporte. Las aerolíneas que generan al menos el 0,5% de los ingresos por pasajeros de servicios regulares nacionales reportan su puntualidad y las causas de los retrasos. Las aerolíneas informantes incluyen: AS, G4, AA, DL, 9E, MQ, F9, HA, B6, OH, NK, OO, UA, WN y YX, notificando las causas de los retrasos en cinco categorías: carrier, late_aircraft, NAS, weather y security.

El análisis se desarrolló en el contexto de la industria aérea, explorando los datos con el objetivo de segmentar y evaluar el estado de los vuelos. El problema central es la necesidad de automatizar y optimizar el análisis de cancelaciones y retrasos para gestionar eficazmente el riesgo asociado.

Caso

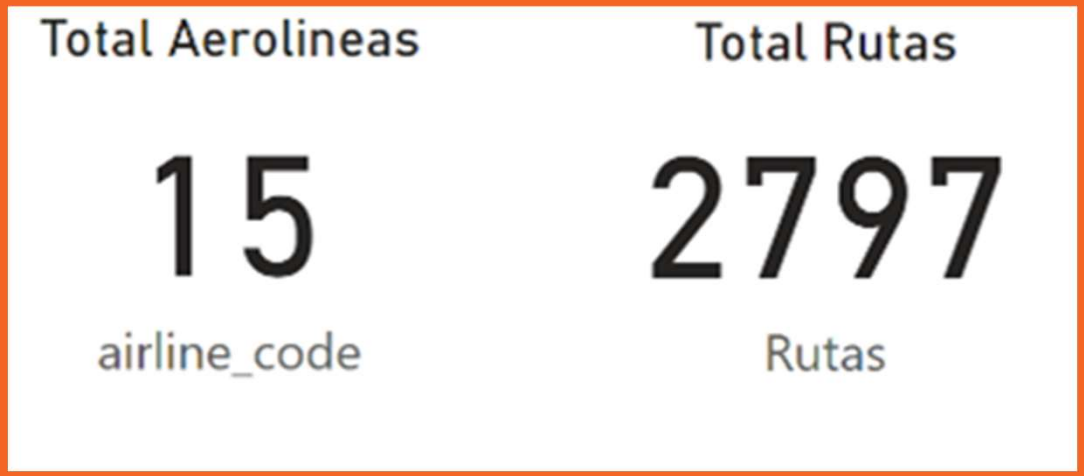
De análisis

El objetivo principal es mejorar la eficiencia y precisión en la evaluación de riesgos de retraso, permitiendo a las aerolíneas tomar decisiones informadas sobre cancelaciones y reducir los retrasos. Se busca validar (confirmar o refutar) hipótesis a través del análisis de datos y proporcionar recomendaciones estratégicas.

El periodo de análisis es enero de 2023.

Riesgo relativo: “Análisis de Aerolíneas”

Para el estudio se cuenta con una base de datos de 15 aerolíneas con 2797 rutas, la cual se proceso con técnica de Análisis de Datos, cálculo de riesgo relativo y desarrollo la metodología de segmentación (Cuartiles) para validación de hipótesis y clasificar por grupos con mayor riesgo relativo.



Riesgo relativo: “Análisis de Aerolíneas”

Las Hipótesis que se deben validar, para saber qué hace que un vuelo presente retraso son las siguientes:

I) Existen rutas que presentan una mayor frecuencia de retrasos en comparación con otras.

II) Es posible calcular el tiempo promedio de retraso por ruta utilizando las variables disponibles.

III) Con los datos disponibles, se puede identificar el principal motivo que genera los retrasos en los vuelos.

IV) Determinados orígenes y destinos contribuyen de manera significativa a los retrasos de los vuelos.

V) La Regresión Lineal permite predecir de manera efectiva el tiempo de retraso de un vuelo.

VI) La Regresión Logística es una herramienta adecuada para determinar el estado (a tiempo o retrasado) de un vuelo.

Riesgo relativo: “Análisis de Aerolíneas”

I) Existen rutas que presentan una mayor frecuencia de retrasos en comparación con otras.

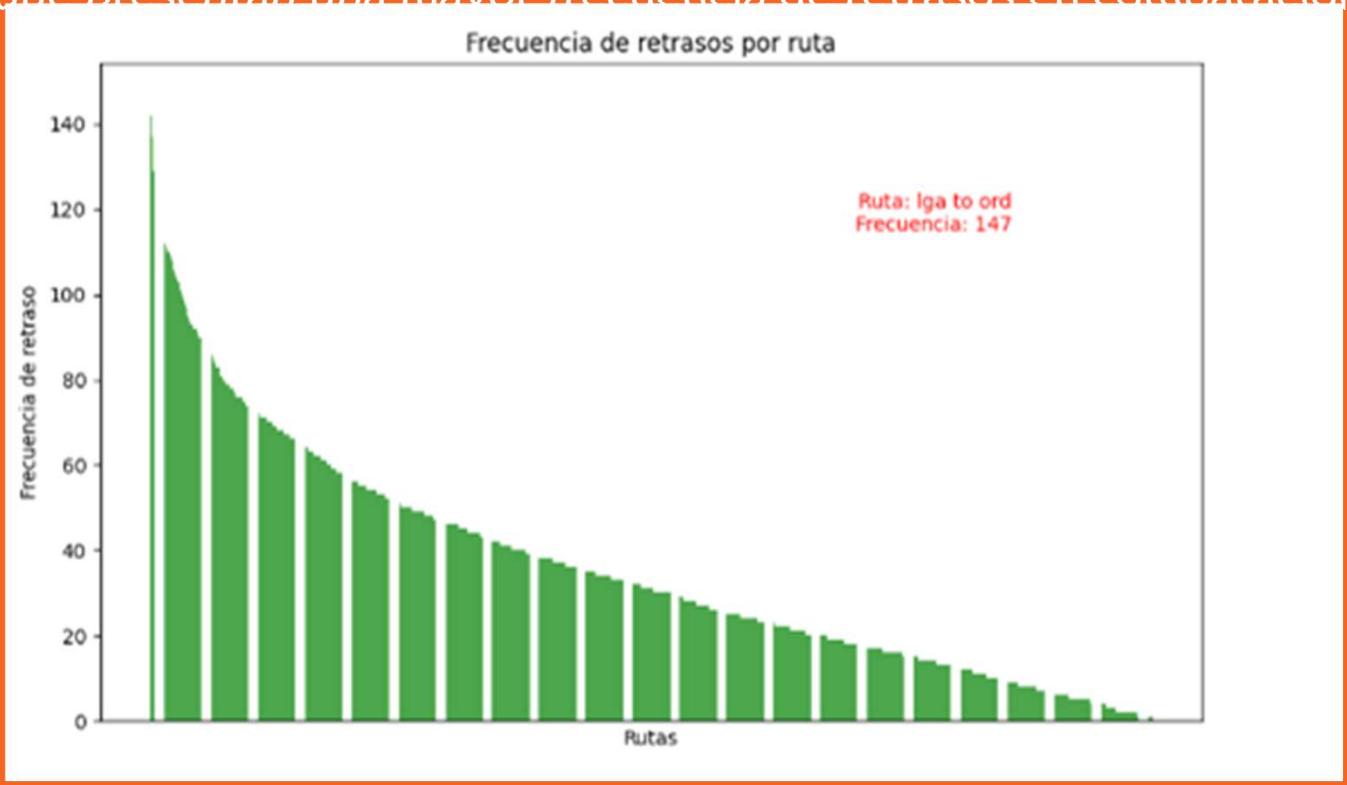


Figura N° 1: Frecuencia de retrasos por ruta (arr_delay_winsorized).

Riesgo relativo: “Análisis de Aerolíneas”

II) Es posible calcular el tiempo promedio de retraso por ruta utilizando las variables disponibles.

Tabla N° 1: Tiempo promedio de retraso por ruta

ruta	Promedio de arr_delay_winsorized
aus to srq	208,00
bos to vps	208,00
fll to pie	208,00
lga to mtj	208,00
cos to hou	158,50
iah to lit	138,00
hou to rsw	124,50
jax to lax	124,50
btr to iah	114,75
ida to pdx	114,00
btr to dca	112,00
aus to tys	104,00
fll to orf	95,00
Total	**14,33**

Con los resultados de la tabla N° 1 podemos indicar que existen 4 rutas que presentan el mayor valor promedio de retraso (AUS to SRQ, BOS to VPS, FLL to PIE y LGA to MTJ.

Riesgo relativo: “Análisis de Aerolíneas”

II) Es posible calcular el tiempo promedio de retraso por ruta utilizando las variables disponibles.

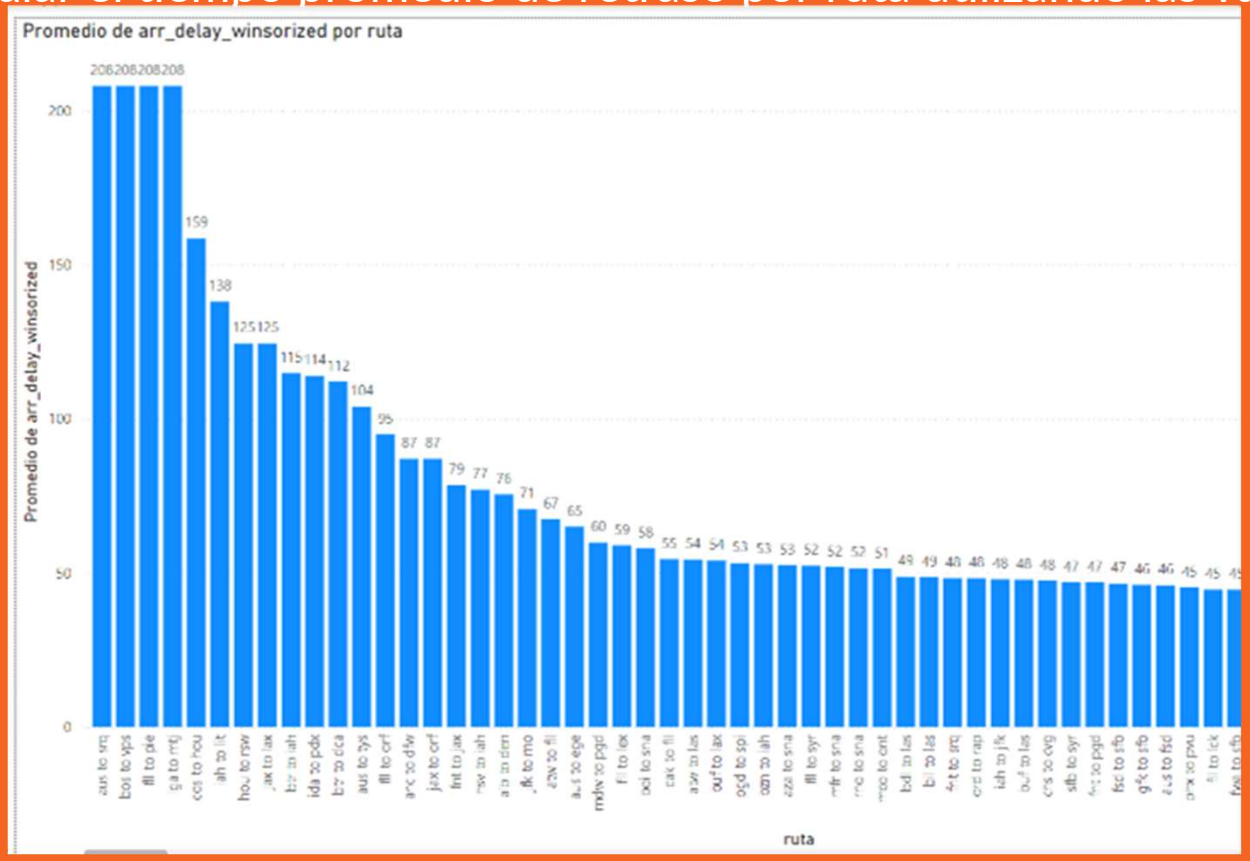


Figura N° 2: Tiempo promedio de retraso por ruta

Riesgo relativo: “Análisis de Aerolíneas”

III) Con los datos disponibles, se puede identificar el principal motivo que genera los retrasos en los vuelos.

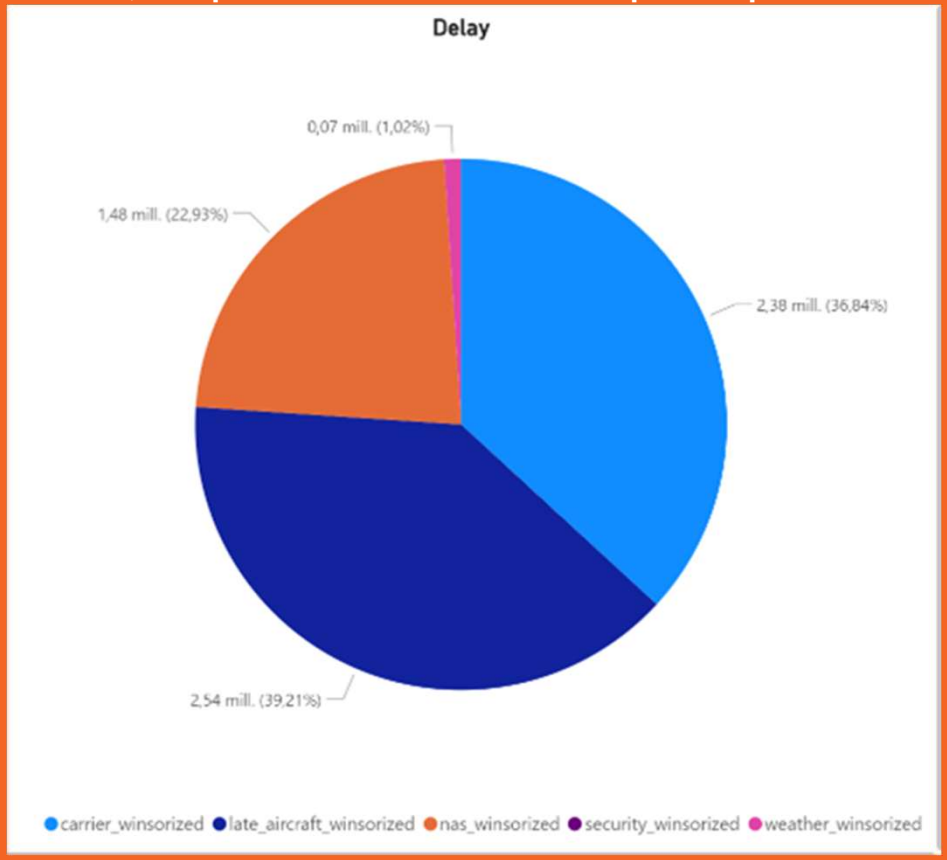


Figura N° 3: Motivos que generan retraso por ruta

Riesgo relativo: “Análisis de Aerolíneas”

IV) Determinados orígenes y destinos contribuyen de manera significativa a los retrasos de los vuelos.

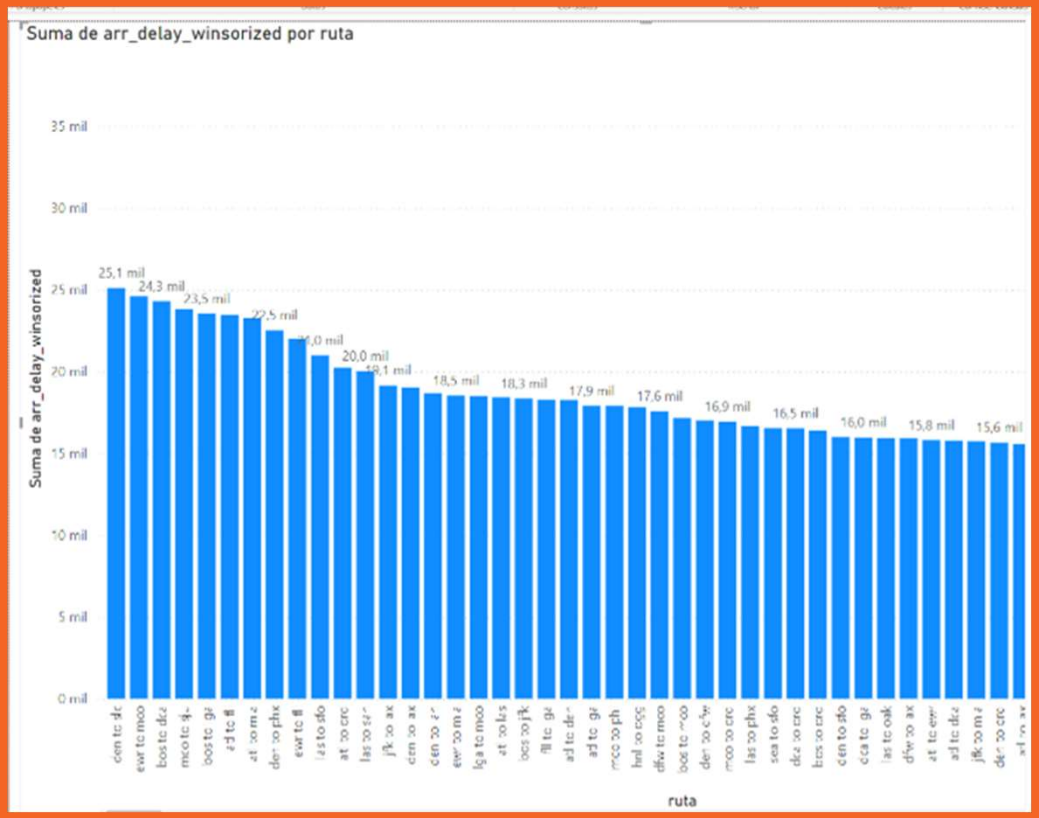


Figura N° 4: Orígenes y destinos (Rutas) que contribuyen a los retrasos ruta

Riesgo relativo: “Análisis de Aerolíneas”

V) La Regresión Lineal permite predecir de manera efectiva el tiempo de retraso de un vuelo.

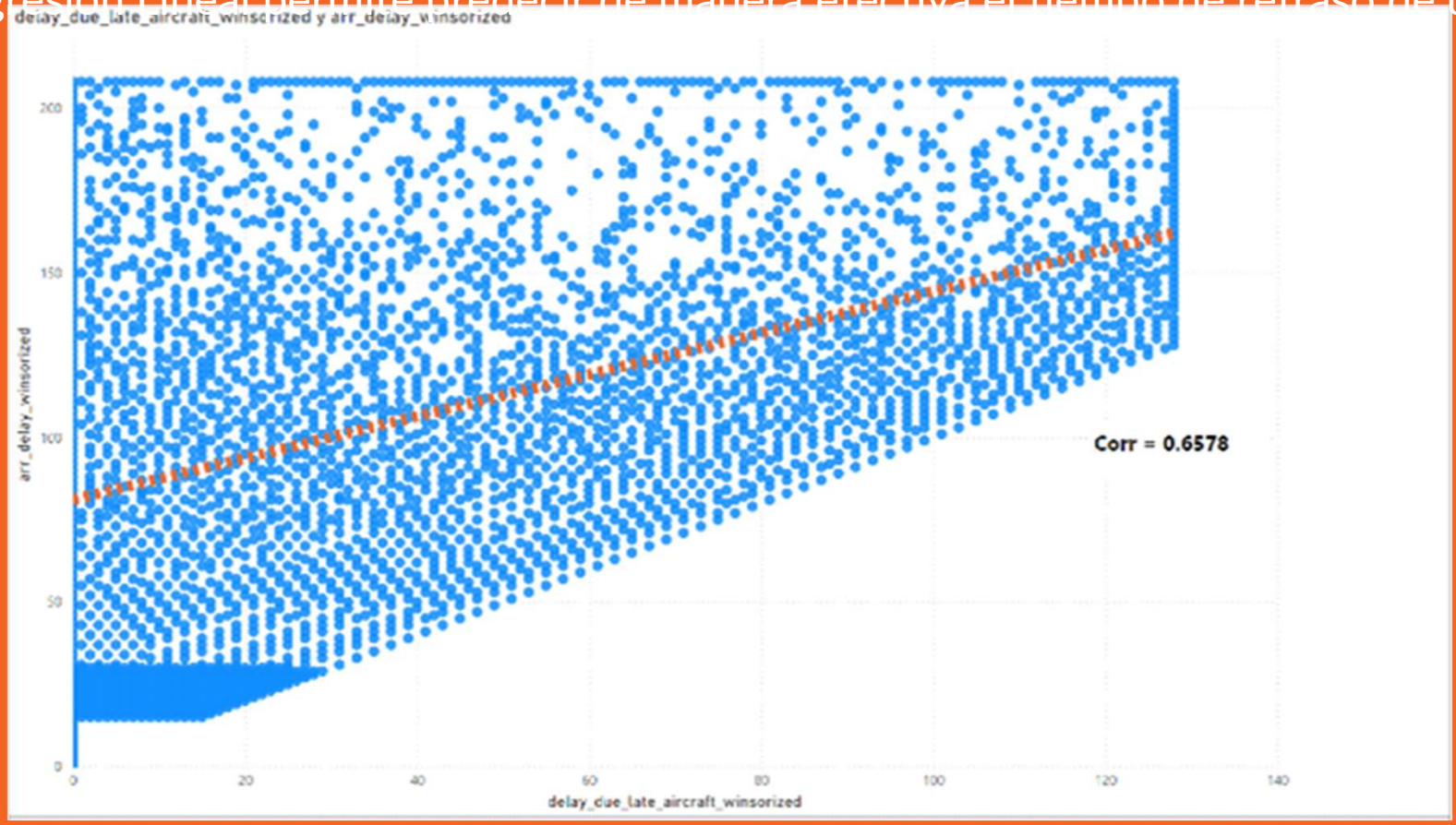


Figura N° 5: Correlación entre variables delay_due_late_aircraft y arr_delay.

Riesgo relativo: “Análisis de Aerolíneas”

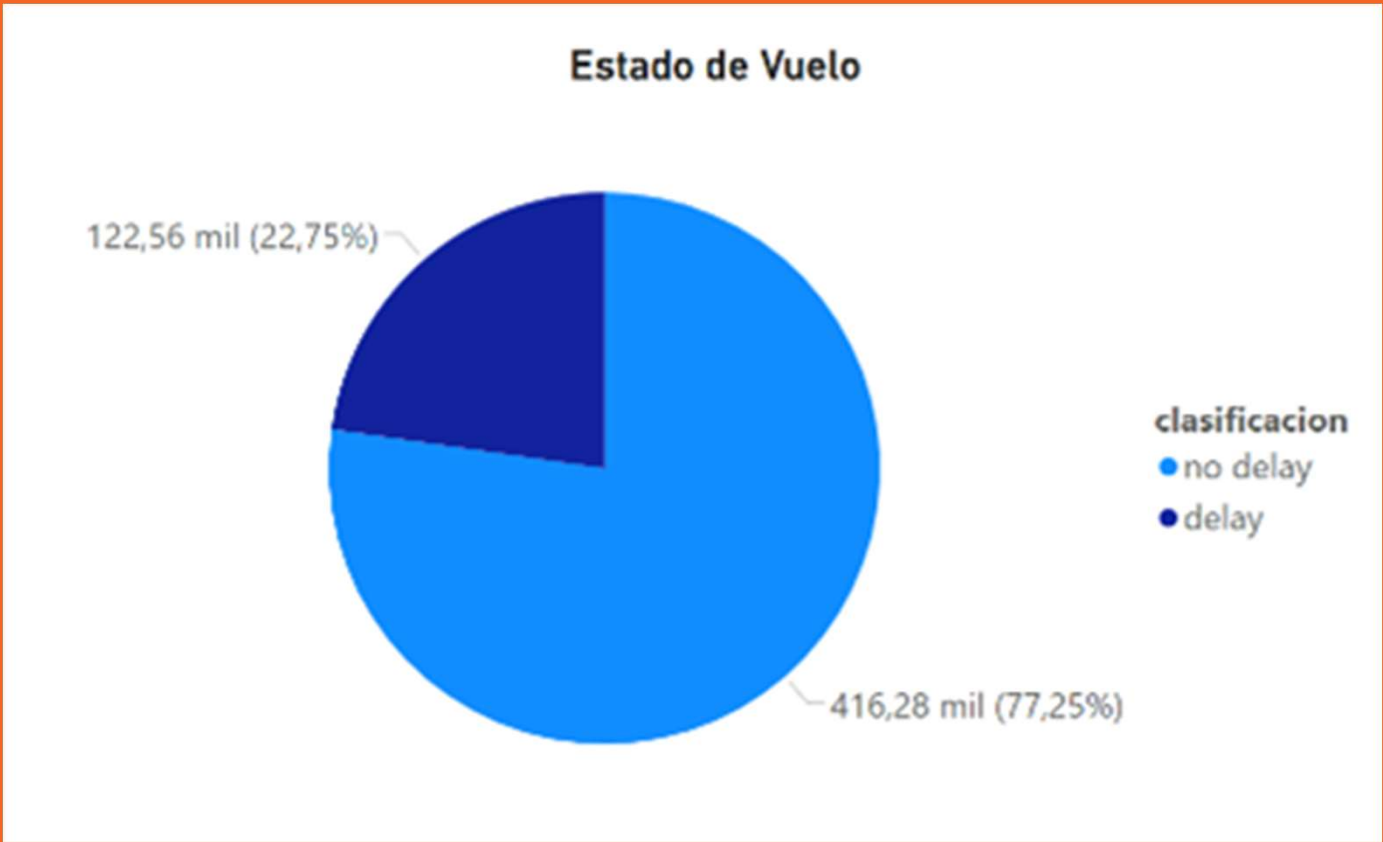


Figura N° 6: Clasificación del estado del vuelo de acuerdo a análisis de Riesgo relativo, cumpliendo con más de tres condiciones de riesgo (carrier, late_aircraft, NAS, weather y security).

Riesgo relativo: “Análisis de Aerolíneas”

VI) La Regresión Logística es una herramienta adecuada para determinar el estado (a tiempo o retrasado) de un vuelo.

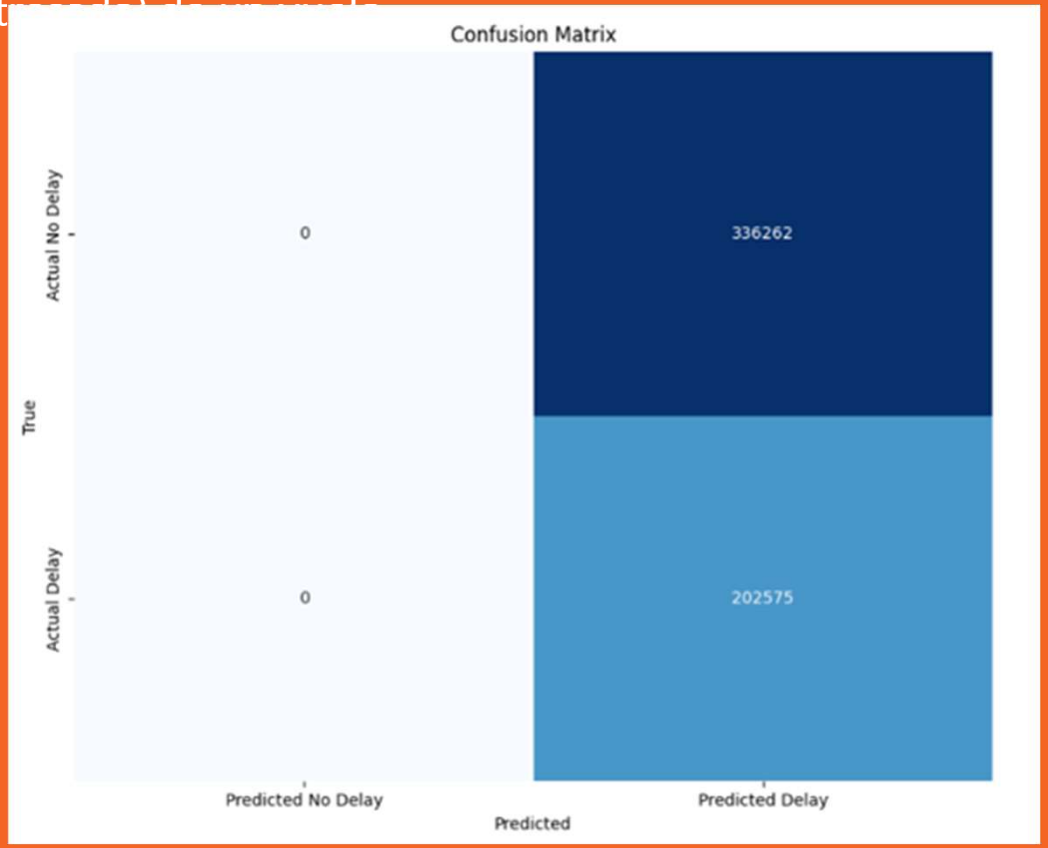


Figura N° 7: Matriz de Confusión.

Tabla N° 2: Métricas de matriz de confusión.

```
Métricas para la Clase 0 (No Delay):
  Precision: 0.0000
  Recall: 0.0000
  F1 Score: 0.0000

Métricas para la Clase 1 (Delay):
  Precision: 0.3759
  Recall: 1.0000
  F1 Score: 0.5465

Exactitud (Accuracy): 0.3759
```

El modelo tiene un rendimiento deficiente en la clasificación de la Clase 0 (No Delay), pero un rendimiento muy bueno en la clasificación de la Clase 1 (Delay).

Riesgo relativo: “Análisis de Aerolíneas”

VI) La Regresión Logística es una herramienta adecuada para determinar el estado (a tiempo o retrasado) de un vuelo.

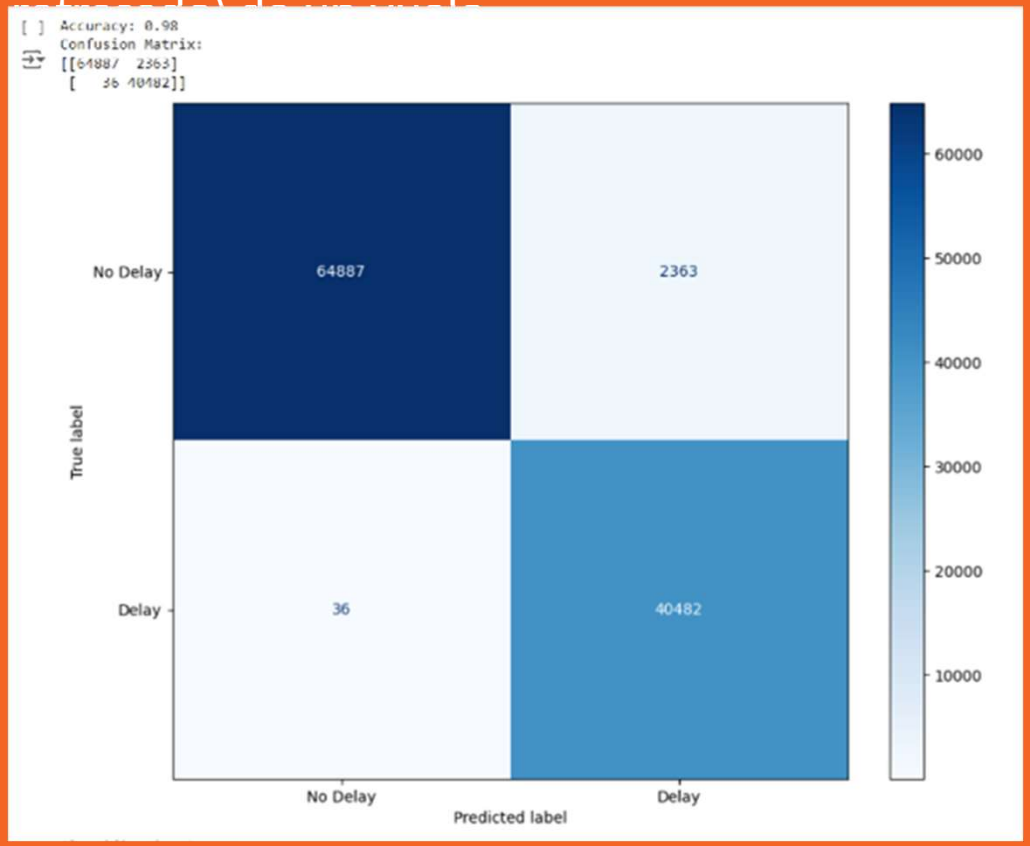


Figura N° 8: Matriz de Confusión de regresiòn logística.

Tabla N° 3: Métricas de matriz de confusión de regresión Logística.

Classification Report:	precision	recall	f1-score	support
0.0	1.00	0.96	0.98	67250
1.0	0.94	1.00	0.97	40518
accuracy			0.98	107768
macro avg	0.97	0.98	0.98	107768
weighted avg	0.98	0.98	0.98	107768

Las métricas proporcionadas muestran un modelo de regresión logística con un rendimiento muy alto en ambas clases. La Clase 0 (No Delay) y la Clase 1 (Delay).

Riesgo relativo: “Análisis de Aerolíneas”

VI) La Regresión Logística es una herramienta adecuada para determinar el estado (a tiempo o retrasado) de un vuelo

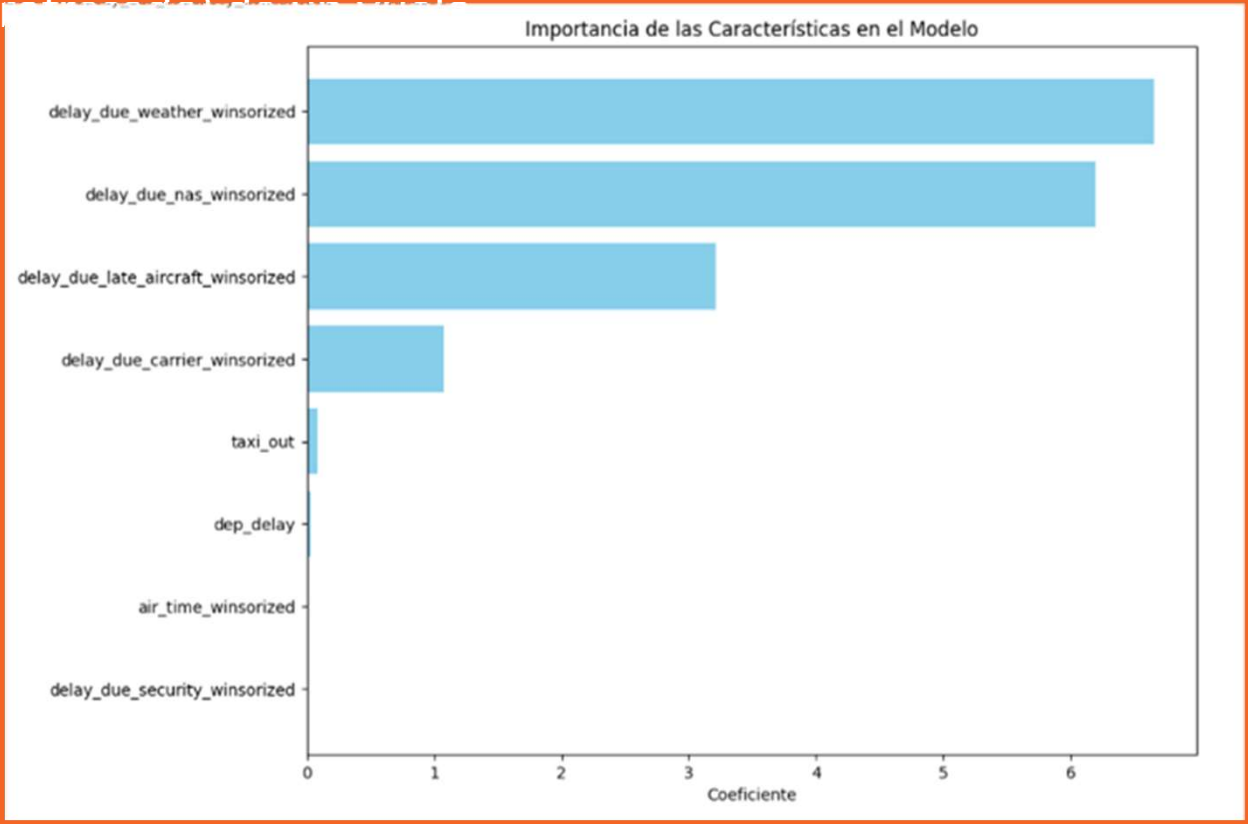


Figura N° 9: Coeficientes del modelo de RL indican la influencia de cada variable sobre la probabilidad de que ocurra un evento, en este caso, un retraso en el vuelo ("Delay").

Tabla N° 4: Coeficientes del modelo de RL

	Feature	Coefficient
4	delay_due_weather_winsorized	6.655996
5	delay_due_nas_winsorized	6.194951
7	delay_due_late_aircraft_winsorized	3.213898
3	delay_due_carrier_winsorized	1.074022
1	taxi_out	0.086081
0	dep_delay	0.028123
2	air_time_winsorized	0.005646
6	delay_due_security_winsorized	0.000000

Factores más importantes: Los factores que más influyen en la probabilidad de que un vuelo se retrase son los relacionados con el clima, el sistema nacional del espacio aéreo (NAS), y la llegada tardía de la aeronave. Estos son factores externos y operativos que no están bajo el control directo del vuelo actual.

Conclusiones

Durante el desarrollo del proyecto se utilizó el proceso de Análisis de Datos, cálculo de riesgo relativo y la metodología de segmentación (Cuartiles) para validación de hipótesis. Se plantearon seis hipótesis, para saber qué hace que un vuelo presente retraso. Estas hipótesis se detallan a continuación y de esto se puede concluir lo siguiente:

I) Existen rutas que presentan una mayor frecuencia de retrasos en comparación con otras.

Se valida la hipótesis con los resultados obtenidos en Figura N° 1: Frecuencia de retrasos por ruta (arr_delay_winsorized) que muestra que la ruta LGA to ORD tiene una frecuencia de 147 retrasos en el periodo analizado.

II) Es posible calcular el tiempo promedio de retraso por ruta utilizando las variables disponibles.

Se valida la hipótesis con los resultados de la tabla N° 1 podemos indicar que existen 4 rutas que presentan el mayor valor promedio de retraso (AUS to SRQ, BOS to VPS, FLL to PIE y LGA to MTJ. y el valor promedio de retraso de la muestra es de 14.33 minutos.

Conclusiones

III) Con los datos disponibles, se puede identificar el principal motivo que genera los retrasos en los vuelos.

Se valida la hipótesis con los resultados de la Figura N° 3: Motivos que generan retraso por ruta, siendo late_aircraft el principal motivo que genera los retrasos con un 39,21 %.

IV) Determinados orígenes y destinos contribuyen de manera significativa a los retrasos de los vuelos.

Se valida la hipótesis con los resultados de Figura N° 4: Orígenes y destinos (Rutas) que contribuyen a los retrasos ruta, los orígenes y destinos que contribuyen en los retrasos de los vuelos son LAS to LAX, LGA to ORD, LAX to SFO, DEN to LAS. Quienes se repiten son LAS Y LAX con más frecuencia y con mayor valores en el total de arr_delay.

Conclusiones

V) La Regresión Lineal permite predecir de manera efectiva el tiempo de retraso de un vuelo.

No es posible validar esta hipótesis dado que segùn la Figura N° 5: Correlación entre variables delay_due_late_aircraft y delay_due_late_aircraft, su correlaciòn es de 0, 6578.Sin embargo El modelo tiene un rendimiento deficiente en la clasificación de la Clase 0 (No Delay), pero un rendimiento muy bueno en la clasificación de la Clase 1 (Delay).

VI) La Regresión Logística es una herramienta adecuada para determinar el estado (a tiempo o retrasado) de un vuelo.

Se valida esta Hipótesis con los resultados de la Tabla N° 3: Métricas de matriz de confusión de regresión Logística, que nos indica que Las métricas proporcionadas muestran un modelo de regresión logística con un rendimiento muy alto en ambas clases. La Clase 0 (No Delay) y la Clase 1 (Delay).

Recomendaciones

Este modelo de regresión logística (RL) está logrando un rendimiento sobresaliente tanto en la clase "No Delay" como en la clase "Delay". Tiene una excelente capacidad para identificar correctamente ambas clases con muy pocos errores, como lo reflejan los altos valores de precisión, recall y F1 score. La exactitud del 98% es también muy indicativa de un modelo que generaliza bien y tiene un muy bajo número de errores. Sin embargo, es importante monitorear cualquier sesgo de clase, aunque aquí parece estar bien balanceado.

En general, el modelo de RL parece estar bien estructurado para reflejar los factores clave que influyen en los retrasos, y resalta cómo los problemas externos (como el clima y las operaciones del NAS) juegan un papel fundamental, se recomienda su uso.

Muchas gracias