

Presentado por Juan Martin Morano

# Flight Delay Predictor

*Anticipando las Perdidas*



A Coderhouse

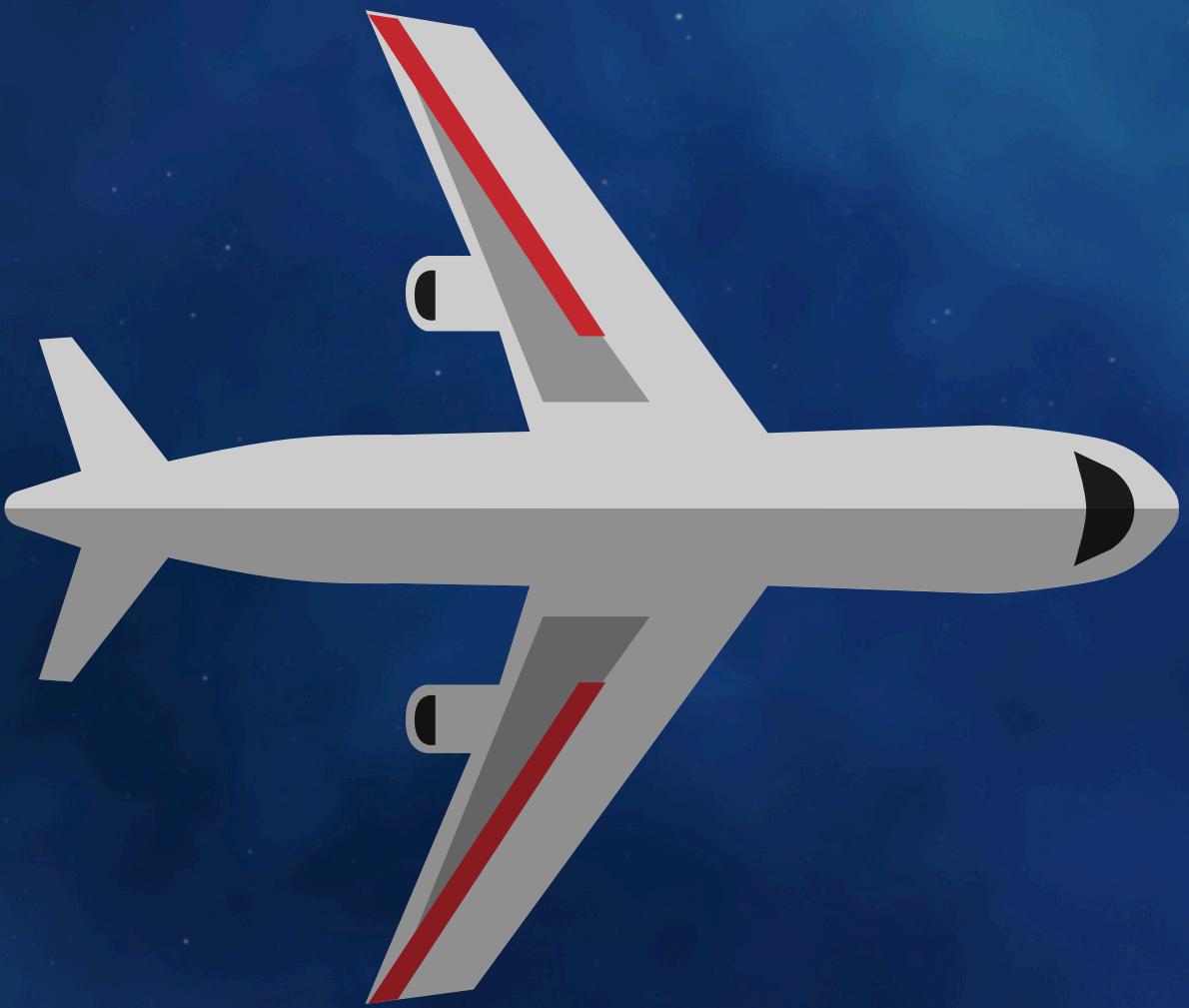
# Introducción

## A partir del Punto de Dolor...

Se concentró información de los vuelos de la empresa, llevando en cada registro el vuelo efectuados a cada una localidades específica.

Se exploraron estos datos de forma intensiva para encontrar ciertos patrones que nos permitirán mediante un modelo de machine learning optimizado predecir nuestro si un vuelo tendrá o no una demora...

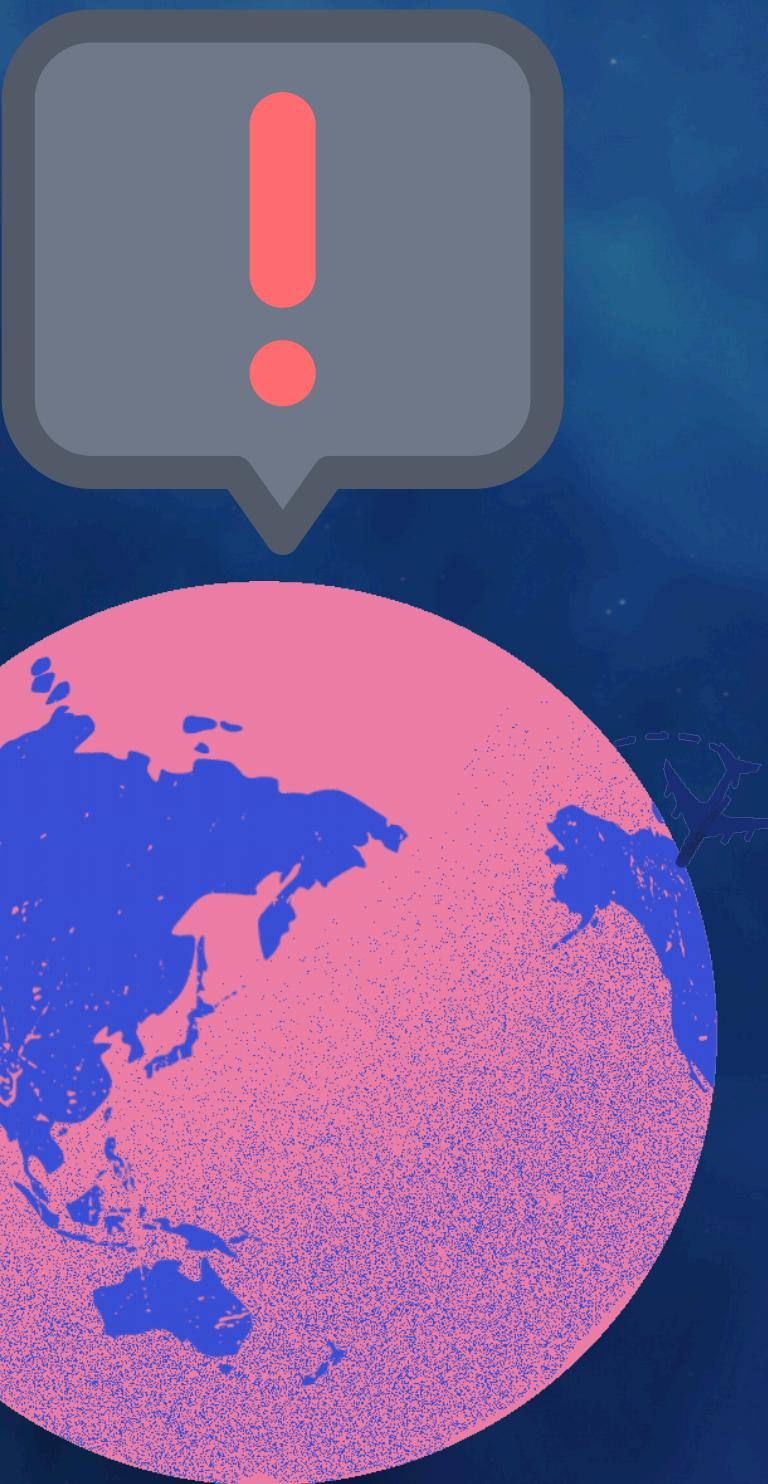
***¡Antes de que suceda!***



# Punto de Dolor Hallado

Encontramos efectos nocivos en la satisfacción de los clientes para con la aerolínea y también las perdidas por devoluciones (para aquellos vuelos que tuvieron esta opción) sobre *DEMORAS INESPERADAS*.

Conocer si un vuelo saldrá demorado o no permite trabajar y/o modificar y mejorar la planificación operativa, atendiendo a que, el poder atender esos puntos específicos de la operación normal del negocio permitirá disminuir estos efectos.

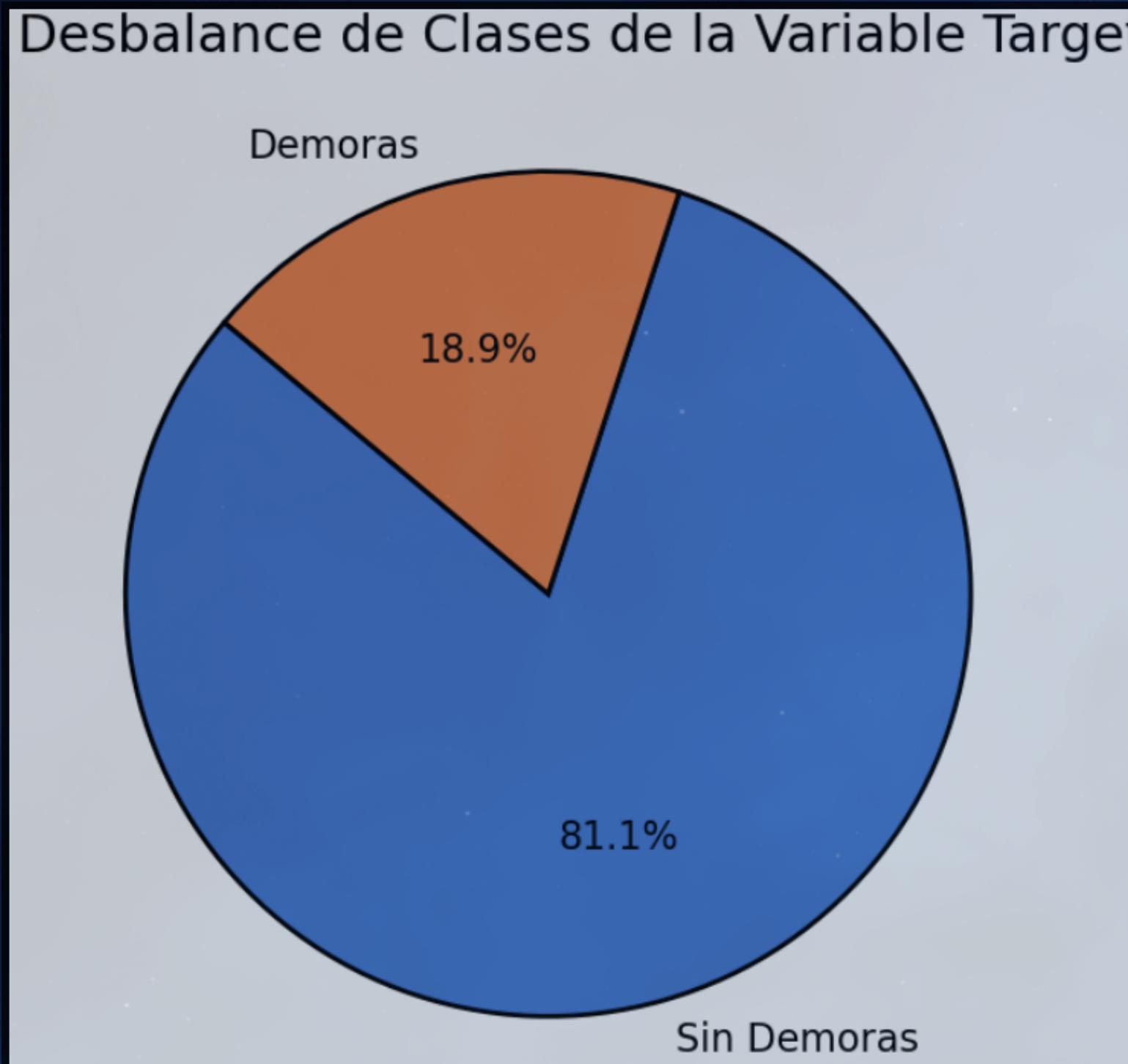




## Preparación y Pre-análisis

Analizando el dataframe, no encontramos nulos y este mismo concentra un porcentaje muy bajo de registros duplicados (0,44% en porcentaje y 28.473 en valores absolutos). Siendo que este porcentaje es demasiado bajo y que su eliminación no afectará la integridad del mismo.

Por otro lado, se tiene un desbalance de clases bastante grande con un 81,1% de registros SIN DEMORAS y un 18,9% de registros CON DEMORAS. Este pasará a ser nuestro BASELINE de aca en adelante.



# Exploración Analítica de los Datos

01

El mismo incluye datos operacionales (como número de asientos y número de vuelos), datos meteorológicos (como precipitación y temperatura), y datos específicos del vuelo (como retrasos y aeropuertos involucrados).

02

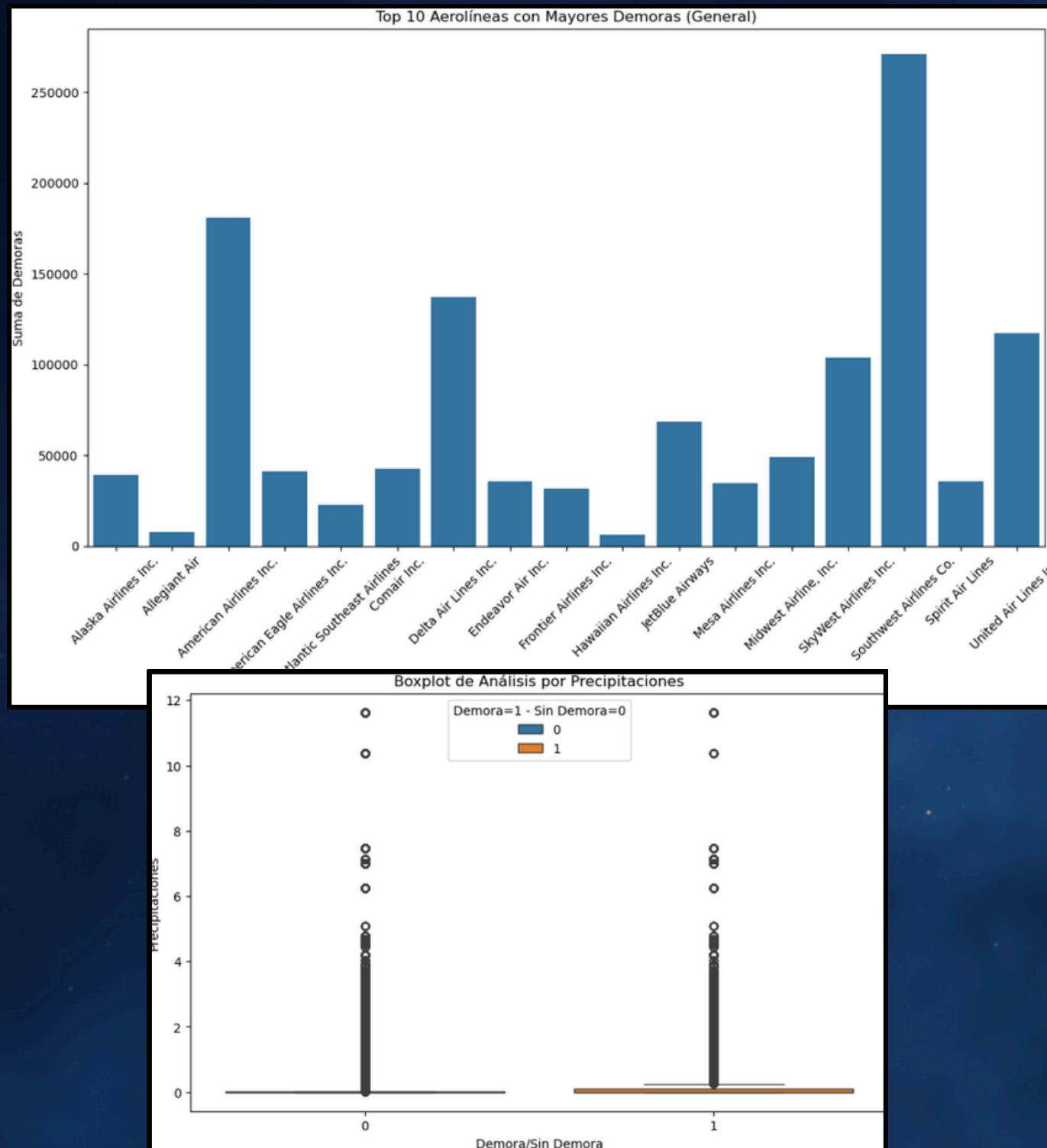
Siguiendo una serie de supuestos preestablecidos, establecemos la posibilidad de estructurar el modelo de clasificación con una performance entre 83% y 90% de accuracy SIN optimización y entre 90% - incluido - y 95% CON optimización. Esto considerando el Benchmark del 81% dado por el desbalance de clases.

03

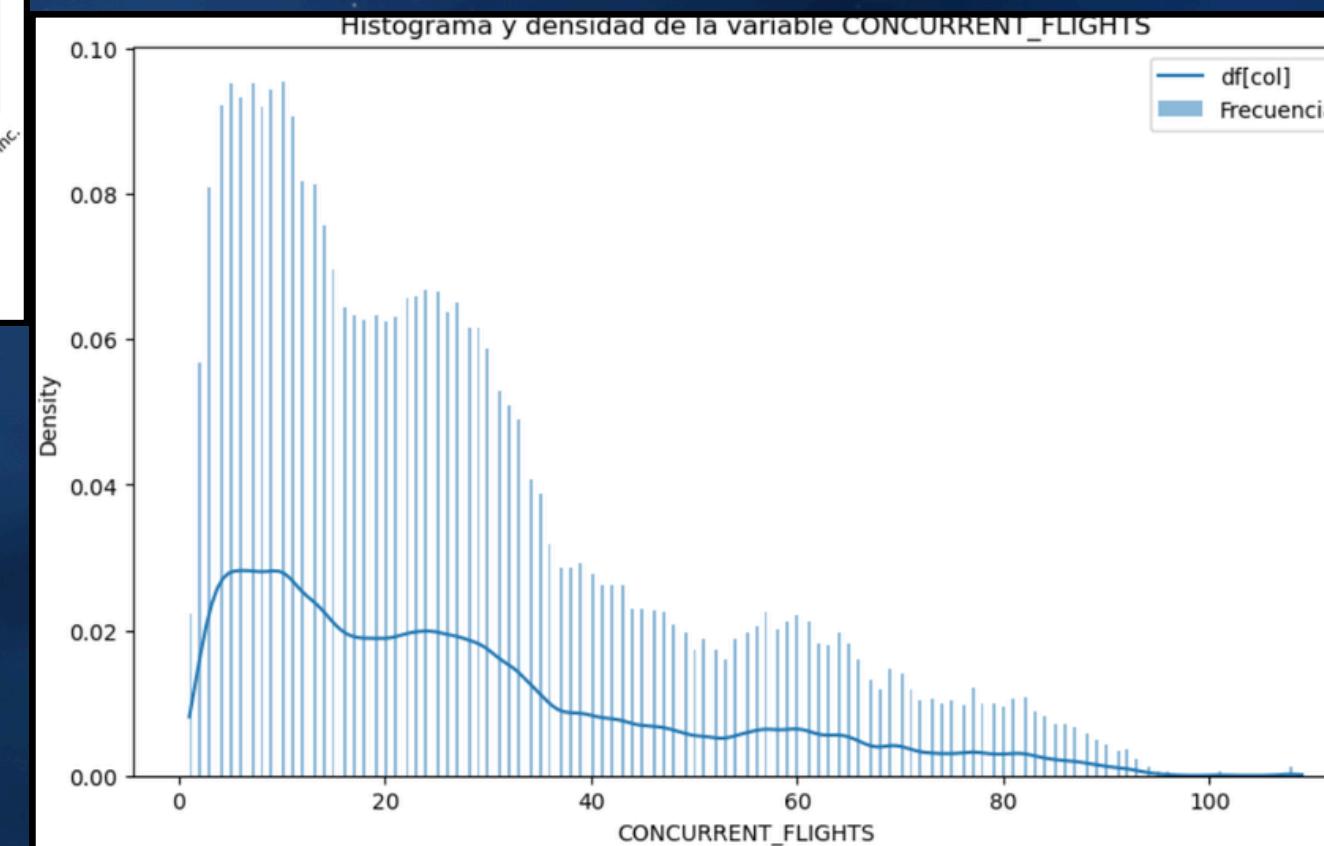
Se verifican grandes desvíos sobre el promedio de las variables, lo que nos induce a pensar que hay una gran variabilidad en los datos y es algo que hay que tener en cuenta para el modelo.

# Análisis Estadístico Destacado

## Densidad + Outliers + Categóricas



Se destacan algunas variables categóricas, las cuales tienen relevancia relación a las horas planificadas de vuelo (con más demoras hacia la tarde noche) y por las particularidades de cada aerolínea. Las distribuciones estadísticas de algunas variables de interés como los vuelos concurrentes ("Concurrents Flights" y el número de vuelos de la aerolínea en un aeropuerto específico durante el mes, junto con algunas Climáticas) nos dejan una cierta normalidad en los datos, con algún valor atípico a tratar para el entrenamiento de los modelos.

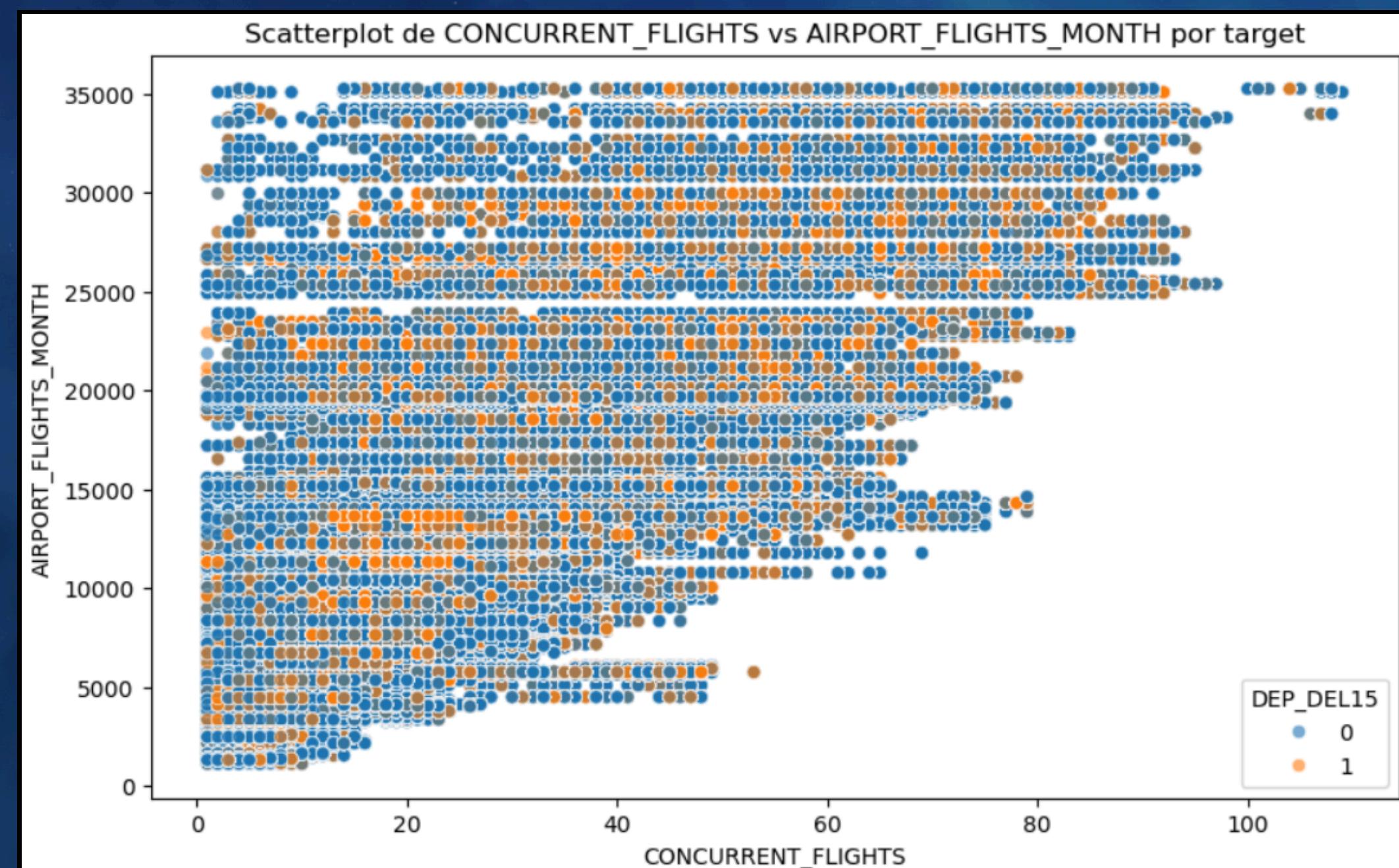


# Otros Análisis Estadísticos...

## *Dispersión + Correlación*

Se destacan algunas variables correlacionadas, las cuales tienen relevancia para el modelo. Entre estas tenemos una cierta justificación lógica tal como es el de vuelos concurrentes junto con la media de cantidad de pasajeros y el número total de vuelos que operan en el aeropuerto durante el mes.

También, como era de esperarse, sucedio lo mismo con las condiciones climáticas. Hay relación entre varias mediciones meteorológicas realizadas sobre cada registro. Aunque igualmente éstas no son tan significativas como las anteriores.



# Conclusiones de la Exploración

01.

De las variables categóricas, las más importantes son: la aerolínea en particular que tiene la demora, con sus eficiencias y deficiencias correspondientes, y el horario determinado del vuelo, recordemos con una tendencia ascendente hacia la tarde, noche. La variable de estacionalidad (por mes) tiene alguna variación entre los distintos trimestres del año, pero ésta es lo bastante leve, por el momento, como para asumirla variable a ingresar al modelo.

02.

De las variables operacionales y climáticas hay varias correlacionadas (más en las operacionales que en las climáticas). La mayoría tiene una distribución log-normal a normal, salvo la de Temperatura Máxima que tiene más una distribución con asimetría por izquierda. Las demoras se suelen concentrar más en el centro de los datos para las operativas que para las climáticas, que están más hacia afuera. Este diferencial de comportamiento, junto con la variable de "Edad del Avión", nos permitirá concluir el feature selection al día de hoy.

# Entrenamiento y Prueba de Modelos

01.

**PRIMER MODELO:**  
ÁRBOL DE DECISIÓN  
El modelo mantuvo levemente por encima de la performance del Baseline (81%). Aunque no hubo Overfitting u Underfitting no es útil.

02.

**SEGUNDO MODELO:**  
DISTANCE (KNN).  
El modelo se mantuvo por arriba de la performance de test, generando Overfitting, por lo cual, tampoco es útil.

03.

**TERCER MODELO:**  
RANDOM FOREST  
El modelo mantuvo la misma performance que el Baseline (81%). Aunque no hubo Overfitting u Underfitting tampoco es útil.

04.

**CUARTO MODELO:**  
ADABOOST  
La prueba mantuvo la misma performance que el Baseline (81%). Aunque no hubo Overfitting u Underfitting tampoco es útil.

05.

**QUINTO MODELO:**  
XGBOOST  
Se logró mejorar la performance en un 5% sobre el accuracy de los datos de Training, y hay una precisión aceptable en vuelos sin demoras.



**5.000.000 x 5% =  
¡250.000 VUELOS GANADOS!**

# Optimización del Modelo de Boosting **XG-BOOST**

Al intentar optimizar los algoritmos de boosting, XG-Boost siendo más específicos, los resultados empeoraron sobre el modelo original nomenclado como “XG01”, del cuál se hizo un análisis riguroso sobre la elección de los hiperparámetros para poder obtener un resultado aceptable (86% de accuracy). Por esta razón, decidimos adoptar este primero como principal para comenzar las pruebas antes de la puesta en producción.

# Conclusión

## Hacia un futuro con mejor satisfacción

En primer lugar podemos concluir que se obtuvieron grandes avances que nos permiten quedarnos conformes con un 5% ganado en el modelo del XG-BOOST. Este trabajo no fue fácil, requirió realizar una feature selection sobre variables operacionales y climáticas importantes sumada la edad del avión (como característica de estos), y un trabajo de feature engineering, normalizando ciertas variables asimétricas tanto por izquierda como por derecha.

Dados estos resultados del modelo, podemos aclarar que la performance buscada en el trabajo sigue siendo bajo. Se espera a un futuro continuar mejorando el número con fines a poder llegar a estar entre un 90% y un 95% (claramente sin Over ni UnderFitting). Esto se traducirá en beneficios aún mayores para la compañía, tanto en la ganancia de ventas como en la reducción de costos operativos (mejorando la planificación operativa), incluso también ampliando su market share para el caso del aumento de ventas dado su alto nivel de satisfacción retroalimentada por sus clientes actuales.

Presentado por el  
Licenciado  
Juan Martin Morano

# ¡Muchas gracias!

FlyDelayPredictor