

CODER HOUSE

TRABAJO FINAL

Arango, Johansen & Piana
Data Science - Comisión 29740





PROBLEMA ESPECÍFICO

Necesidad de predecir retrasos en la aviación comercial de Estados Unidos.

OBJETIVO DE LA INVESTIGACIÓN

Mejorar los servicios prestados a los pasajeros mediante la predicción de retrasos; conociendo la aerolinea que opera el vuelo, el día de la semana en la que está programado, su aeropuerto de origen y aeropuerto destino.

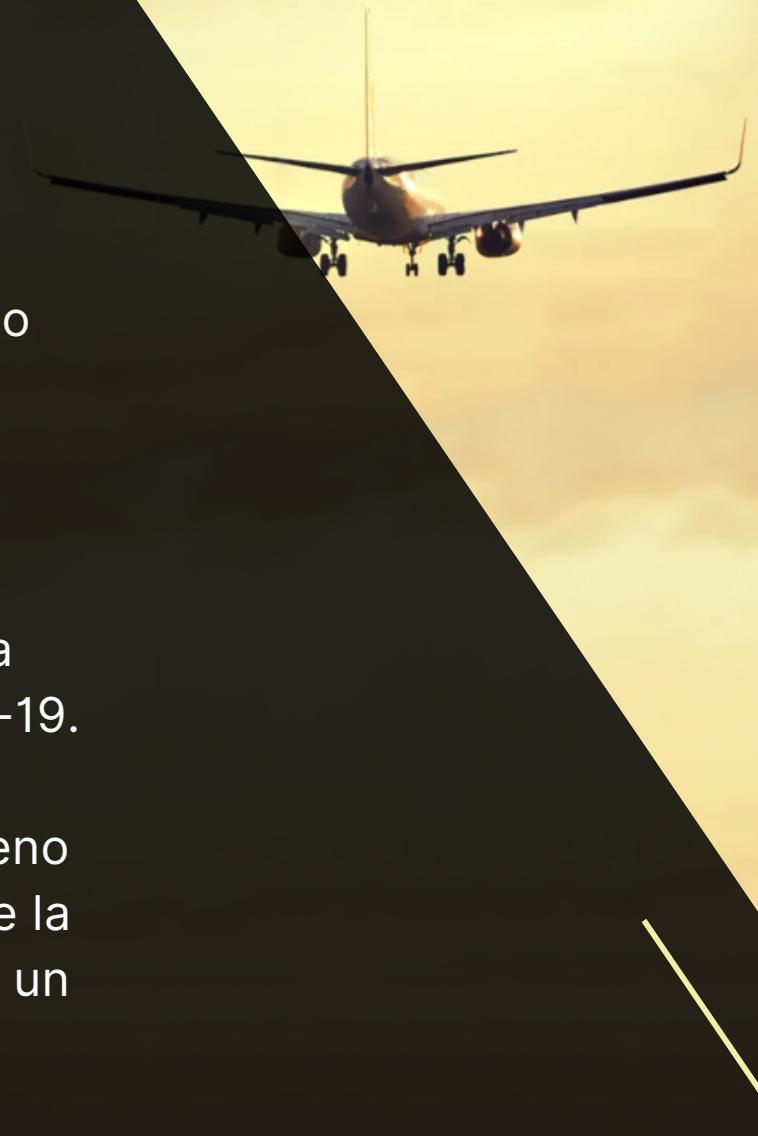
LA FUENTE DE DATOS

El dataset a utilizar se obtuvo de Kaggle.

Se seleccionó este dataset en base a la cantidad de registros y variables disponibles, así como la posibilidad de trabajar con un problema de negocio que nos resultara pertinente en la actualidad y acorde a nuestros intereses particulares.

Los principales aeropuertos de todo el mundo se encuentran experimentando una situación cercana al colapso tras la pandemia causada por la COVID-19.

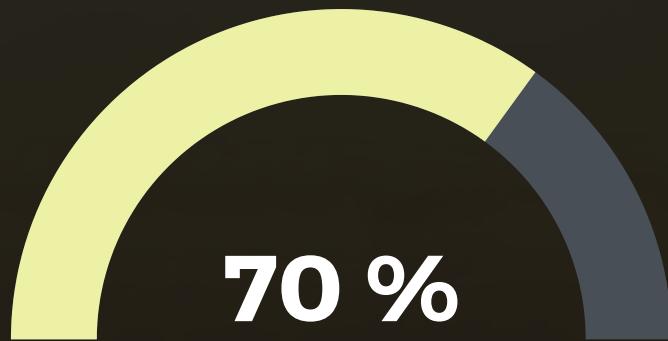
Las consecuencias de esta crisis impactarán de lleno en perjuicio de los usuarios como consecuencia de la reducción de rutas y frecuencias, cancelaciones y un incremento en los precios de los pasajes.



OBJETIVO DE LOS DATOS

Aspiramos a que este conjunto de datos nos permita encontrar correlaciones entre las diferentes variables disponibles para predecir aquellas aerolíneas, rutas aéreas y/o aeropuertos dónde los pasajeros serán más propensos a sufrir retrasos en sus vuelos.

El objetivo que establecimos es utilizar el 70% de los datos para entrenar el modelo, y el 30% restante para testearlo.



CARACTERÍSTICAS DE LOS DATOS

Contamos con **539.383 observaciones**,
6 variables y sin valores faltantes.

Columnas del archivo:

1. Airline: Nombre de la aerolínea.
2. Flight: Número de vuelo.
3. Airport From: Aeropuerto de salida.
4. Airport To: Aeropuerto de llegada.
5. DayOfWeek: Día de la semana del vuelo.
6. Time: Duración del vuelo (minutos).
7. Length: Distancia del vuelo.
8. Delay: Indica si el vuelo tuvo demoras o no.



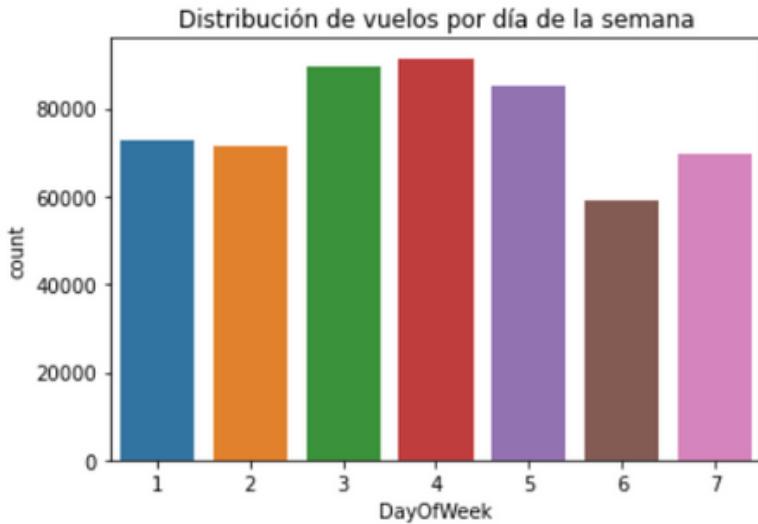
DATA WRANGLING

- Eliminar registros con distancia 0 (Length = 0).
- Convertir los registros de las columnas 'Time' y 'Length' a horas.
- Crear agrupación para la columna 'Time':
 - Mañana
 - Mediodía
 - Tarde
 - Noche
- Crear agrupación para la columna 'Length':
 - '<= 1.35'
 - '1.36 - 1.92'
 - '1.93 - 2.70'
 - '>= 2.71'
- Incorporar al dataset información sobre los aeropuertos.
- Reemplazar los valores "NaN" de aeropuertos por "Other".

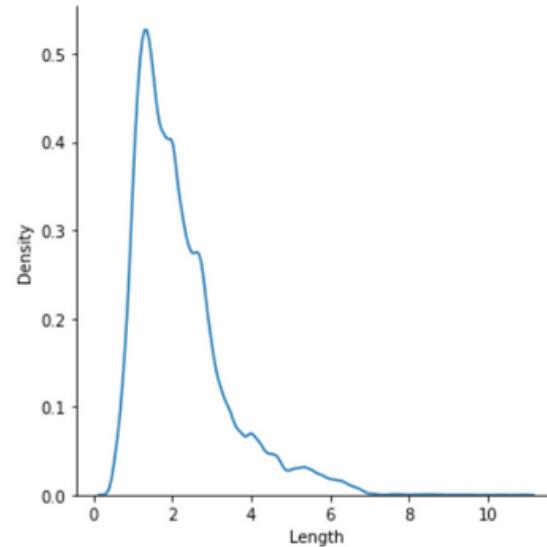


CONCLUSIONES

ANÁLISIS UNIVARIADO



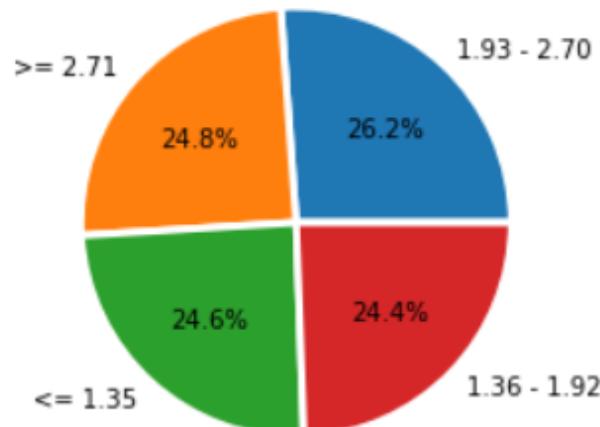
Los días con mayor cantidad de vuelos corresponden a los miércoles, jueves y viernes. En segundo lugar, los días lunes y martes. Finalmente, los sábados y domingos son los días con menos cantidad de vuelos registrados.



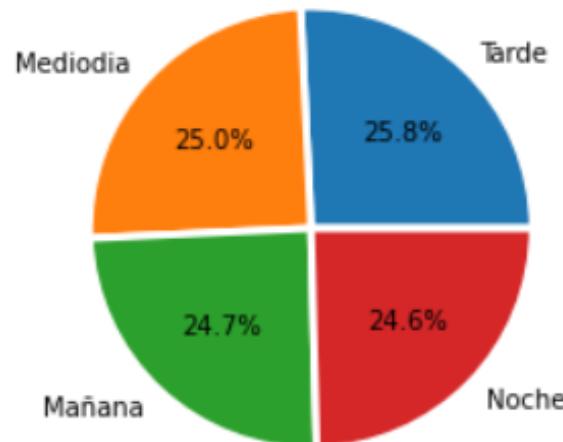
La variable 'Length' cuenta con una distribución asimétrica izquierda. Se evidencia que la duración de los vuelos se concentraron entre 0.5 y 3 horas. Con un valor mínimo de 0 y un máximo de 10.92. Un total de 4 registros con una duración igual a cero fueron excluidos del análisis.

CONCLUSIONES

ANÁLISIS UNIVARIADO



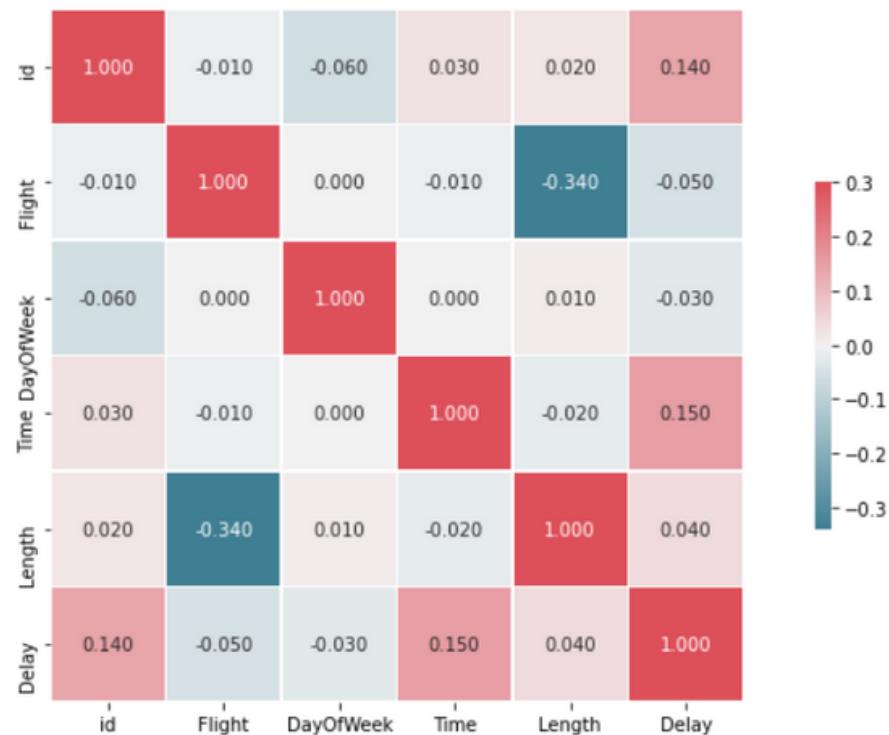
La distribución de la cantidad de vuelos, en base a la duración, es bastante equitativa a lo largo de las 4 agrupaciones realizadas.



La distribución de la cantidad de vuelos, en base al horario de partida, es bastante equitativa a lo largo de las 4 agrupaciones realizadas.

CONCLUSIONES

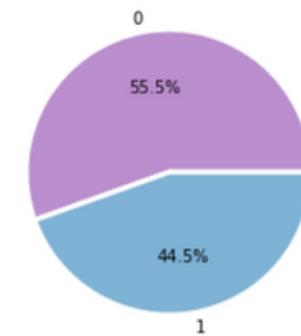
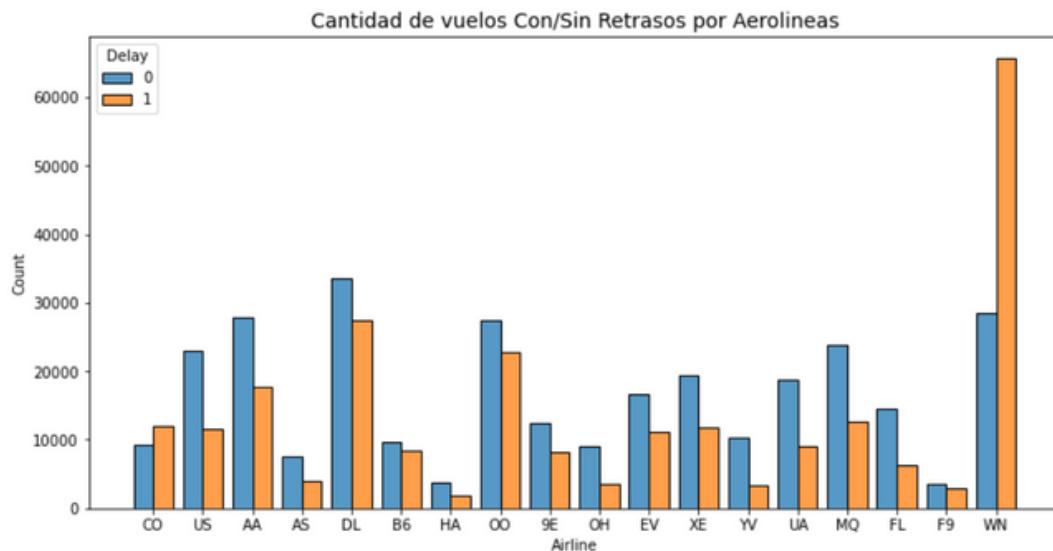
ANÁLISIS UNIVARIADO



Podemos observar que no hay fuertes correlaciones entre las variables.

CONCLUSIONES

ANÁLISIS UNIVARIADO: DELAY



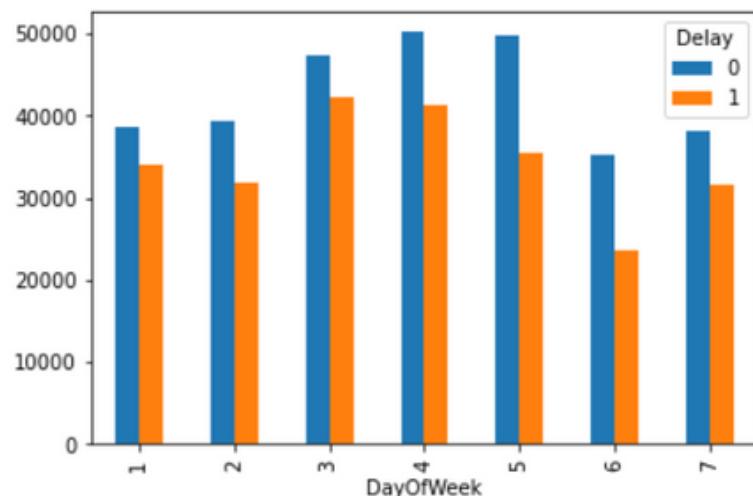
El 44.5% de los vuelos salieron retrasados.

La aerolínea WN si bien es la que mayor cantidad de vuelos realiza, también es la que peor servicio ofrece en cuanto a la puntualidad de los horarios de partida, teniendo más del doble de vuelos retrasados que vuelos puntuales. A su vez, junto con WN, CO es la única otra compañía que tiene más vuelos retrasados que puntuales.

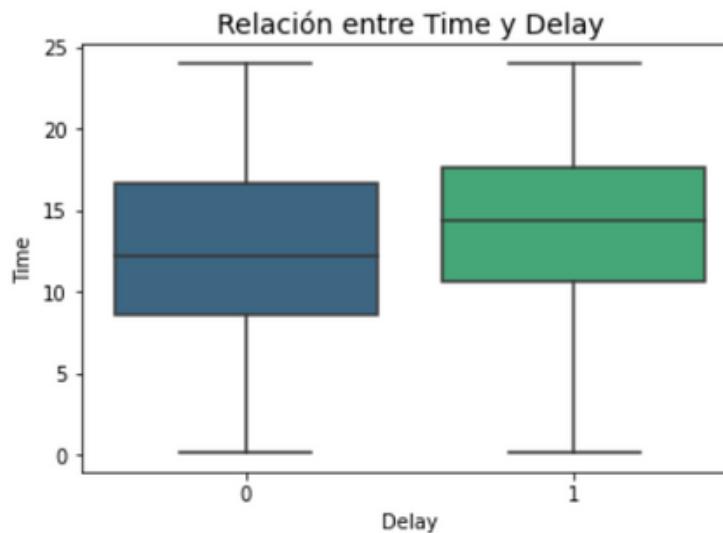
En 16 de las 18 aerolíneas la cantidad de vuelos sin retrasos es superior a la cantidad de vuelos con retrasos. Llamativamente, encontramos la situación inversa en WN, la aerolínea con mayor cantidad de vuelos, y en CO, una de las aerolíneas con menor cantidad de vuelos registrados.

CONCLUSIONES

ANÁLISIS UNIVARIADO: DELAY



El día de la semana no tiene una gran influencia sobre los retrasos en los vuelos. Analizando los retrasos por día de la semana no se ha identificado ninguna diferencia significativa. Todos cuentan con aproximadamente entre un 40 y 47% de vuelos retrasados.

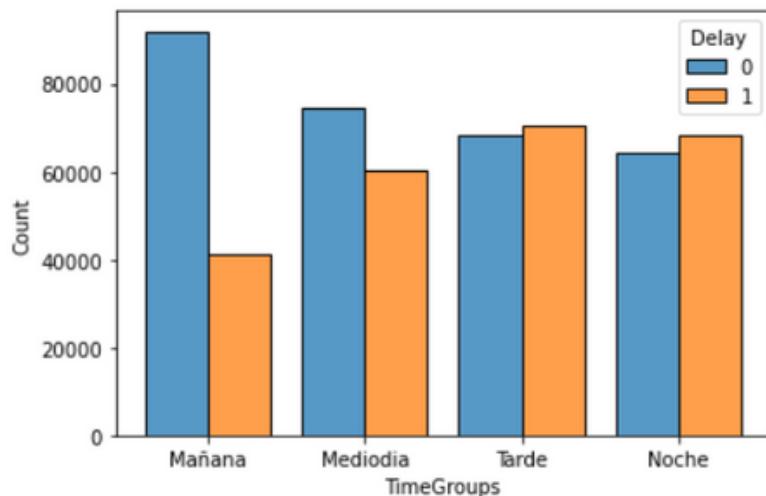


Advertimos una ligera tendencia de que los vuelos con retrasos suelen tener un horario de partida un poco más tarde que los vuelos sin retrasos.

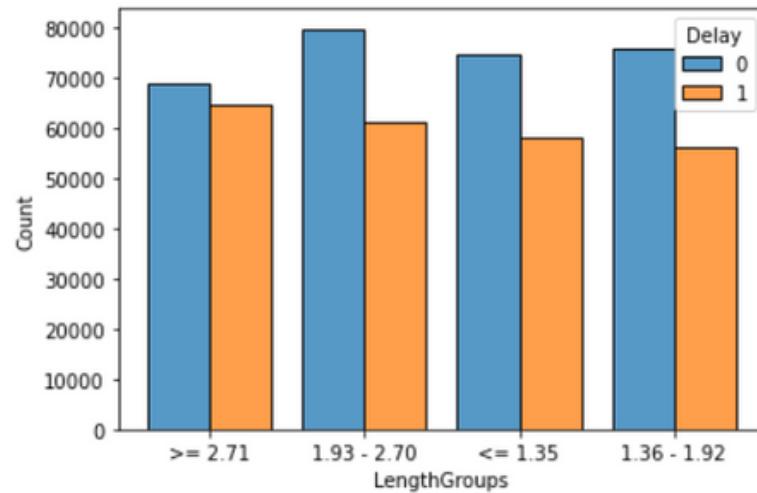
Sin embargo, no es evidencia suficiente para tal afirmación y se debería realizar un test de hipótesis para su confirmación.

CONCLUSIONES

ANÁLISIS UNIVARIADO: DELAY



Mientras más tarde es el horario de salida, mayor es el porcentaje de vuelos retrasados. A partir del horario tarde, hay más vuelos retrasados que puntuales.



Los vuelos de una duración larga (mayor a 2.71 horas) suelen tener mayor probabilidad de retraso que las otras 3 agrupaciones.

PCA

Pasos realizados:

- Carga de datos
- Exploración inicial
- Preparación de datos
- Definición de variables *target*:
 - En base al objetivo planteado de poder predecir que vuelos puede presentar demoras, nuestra variable objetivo será "Delay"
- Dividimos nuestro dataset en 2 partes:
 - X: contendrá las variables con las cuales vamos a construir el modelo. Dichas variables se consideran independientes.
 - Y: contendrá la variable Delay y se considera nuestra variable dependiente.
- Dividimos el dataset.
 - El 70% del mismo para el entrenamiento de los distintos algoritmos.
 - El 30% restante como test de los algoritmos.



ANÁLISIS DE COMPONENTES

Luego de convertir todas nuestras variables independientes en numéricas nos encontramos con un dataset que contiene 607 columnas o variables de análisis.

Con el fin de verificar si es posible reducir la dimensionalidad aplicamos el método PCA o Análisis de Componentes Principales.

Escalado de las variables

El proceso de PCA identifica las direcciones con mayor varianza.

Como la varianza de una variable se mide en sus mismas unidades elevadas al cuadrado, si antes de calcular las componentes no se estandarizan todas las variables para que tengan media cero y desviación estándar de uno, aquellas variables cuya escala sea mayor dominarán al resto. De ahí que sea recomendable estandarizar siempre los datos.



PCA

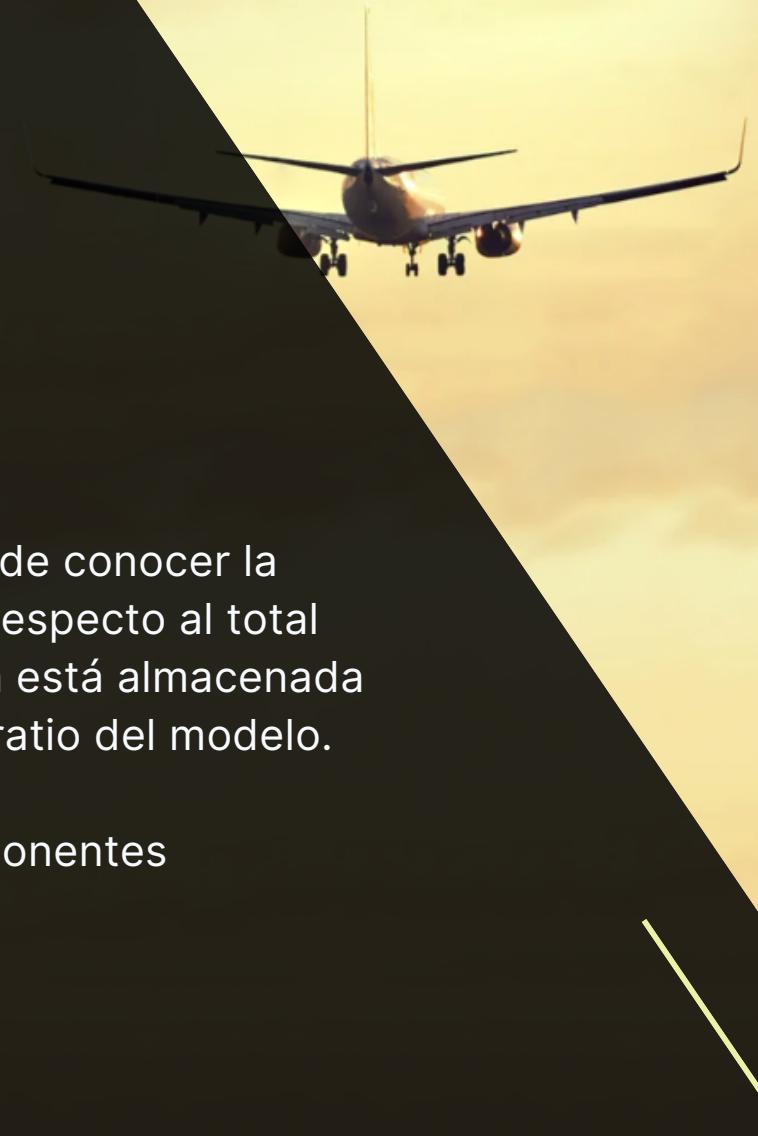
INTERPRETACIÓN

Una vez entrenado el objeto PCA, pude accederse a toda la información de las componentes creadas.

`components_` contiene el valor de los loadings ϕ que definen cada componente. Las filas se corresponden con las componentes principales (ordenadas de mayor a menor varianza explicada). Las filas se corresponden con las variables de entrada.

Una vez calculadas las componentes principales, se puede conocer la varianza explicada por cada una de ellas, la proporción respecto al total y la proporción de varianza acumulada. Esta información está almacenada en los atributos `explainedvariance` y `explained_varianceratio` del modelo.

Aplicamos el modelo **RandomForest** con todas las componentes obtenidas al aplicar PCA y vemos como funciona.



PCA

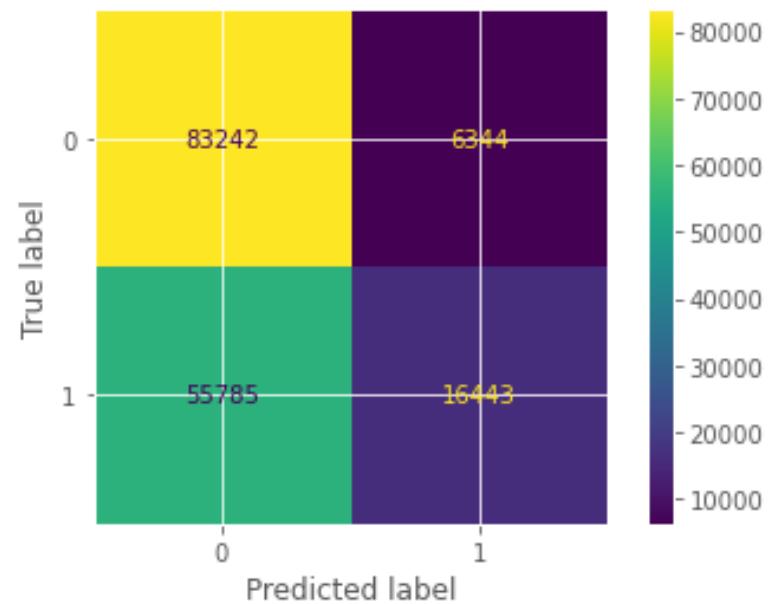
INTERPRETACIÓN

Probamos aplicar nuevamente el método PCA pero esta vez considerando 400 componentes principales (de un total de 600 aproximadamente).

Según el análisis anteriormente planteado sobre la varianza acumulada al aplicar PCA, los primeros 400 componentes representan aproximadamente un 75% del peso del modelo.



MATRIZ DE CONFUSIÓN



En la matriz de confusión podemos observar: Verdaderos Positivos: 83338 Verdaderos Negativos: 16273 Falsos Positivos: 55955 Falsos Negativos: 6248.

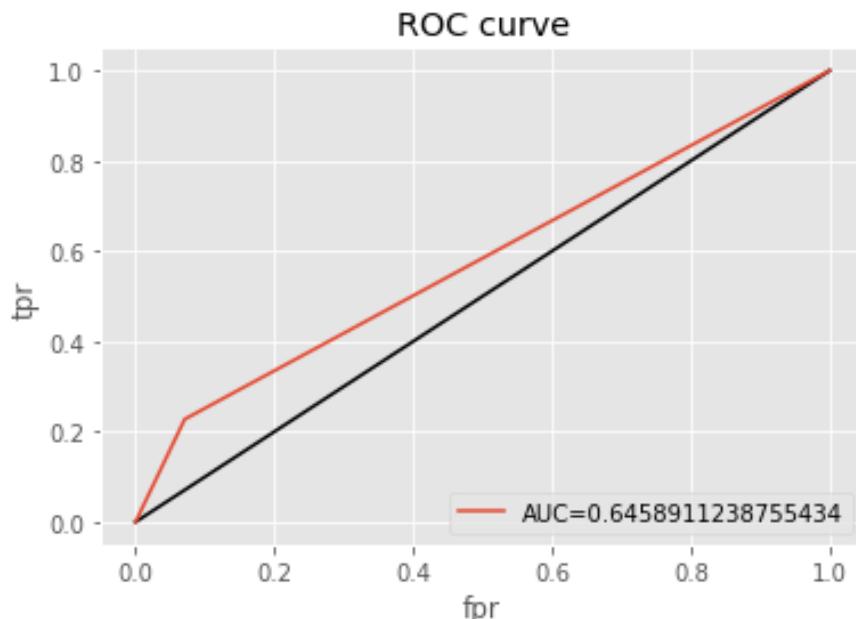
El % de aciertos sobre el set de evaluación es: 0.6160468191874622

El % de precisión sobre la evaluación es: 0.7215956466406285

Recall o la Sensibilidad del algoritmo es: 0.2276540953646785

La especificidad del modelo es: 0.34611377150976164

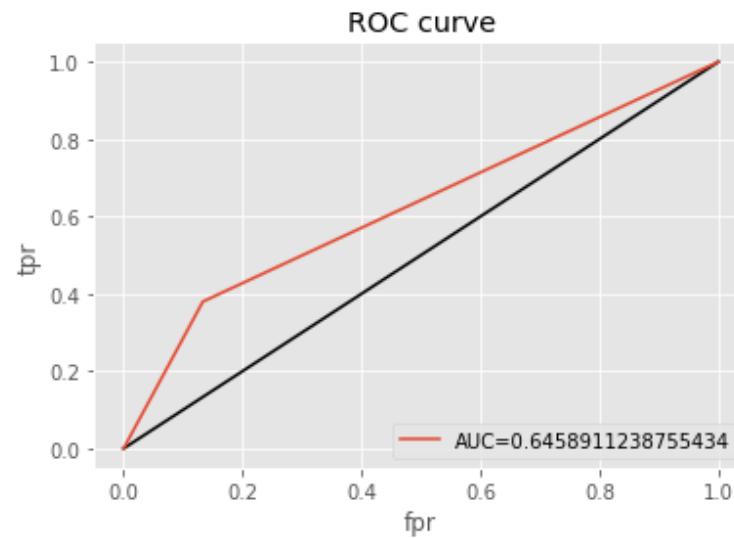
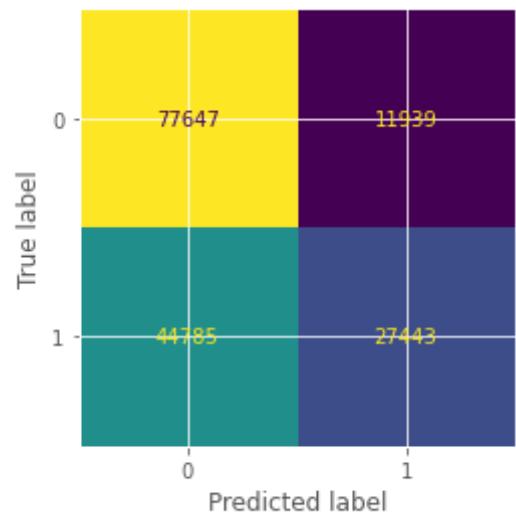
CURVA ROC



La curva AUC-ROC es una métrica de rendimiento que se utiliza para medir el rendimiento del modelo de clasificación en diferentes valores de umbral. Cuanto mayor sea el valor de AUC (Área bajo la curva), mejor será nuestro clasificador para predecir las clases. AUC-ROC se utiliza principalmente en problemas de clasificación binaria.

La curva ROC se traza entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR), es decir, TPR en el eje y y FPR en el eje x. AUC es el área bajo la curva ROC. Un clasificador excelente tiene un valor AUC cercano a 1, mientras que un clasificador de bajo rendimiento tiene un valor AUC cercano a 0. Un clasificador con una puntuación AUC de 0.5 no tiene ninguna capacidad de separación de clases.

PARAMETRIZANDO DE OTRA MANERA



El % de aciertos sobre el set de evaluación es: 0.6494493677926508

El % de precisión sobre la evaluación es: 0.6968411964857041

Recall o la Sensibilidad del algoritmo es: 0.3799496040316775

La especificidad del modelo es: 0.49176597079114776

Podemos observar que los modelos no son muy efectivos ya que sólo tenemos un poco más de probabilidad de acierto que tirando una moneda.