


Aeropuerto Jorge Newbery


Big Data Bang: explosionando las rutas aéreas para predecir un caos en alto vuelo..!

 **TP2:**

Analisis Exploratorio y Curacion de Datos

BOARDING PASS

- **FAMAF**
2023
- **MENTORIA**
#8
- **GRUPO**
1 - 2



Una vez que se han identificado las variables, es importante realizar una exploración de los datos para comprender mejor la distribución y las características de los mismos. Esto puede incluir gráficos, tablas, estadística descriptiva y otros métodos de visualización y exploración. Todo esto pensando en un aprendizaje más magro (más puro) de los datos bajo estudio.

1. Entendiendo el Customer Journey Map:

- ¿Cómo varía el volumen total de arribos según el mes? ¿y por día de la semana?
- ¿Cómo varía el volumen de pasajeros por mes?
- ¿Existe alguna tendencia ó fluctuación cíclica en función de la estacionalidad?
¿Hay días especiales en el año?
- ¿Cual es la distribución de vuelos provenientes de los distintos orígenes?
- Durante un año, que cabecera es la que más se utilizó? ¿y que posición?

2. Traduciendo los retrasos

- ¿Qué porcentaje de vuelos experimentan un retraso? Entre esos vuelos, ¿cuál fue el tiempo promedio de retraso (en minutos)?

- b. ¿Cómo varía el % de vuelos retrasados a lo largo del año? ¿Se puede calcular el porcentaje de retraso según la estación del año?
- c. ¿Hay zonas geográficas que tienden a presentar más demoras que otras? ¿Se pueden predecir demoras por ruta? (retrasos por lugar de origen)
- d. ¿Se puede identificar algún patrón relevante que origine demoras?
- e. ¿Qué aerolíneas parecen ser más y menos confiables, en términos de salidas a tiempo?
- f. Ponderación del porcentaje de retraso por Aerolíneas según el número de vuelos que realizan (Aerolíneas con mayor porcentaje de retraso según el número de vuelos realizado vs. Aerolíneas con menor porcentaje de retraso según el número de vuelos realizado)
- g. Discretizar el cumplimiento de las aerolíneas en: adelantado/cumplido/demorado

3. Correlaciones:

Verificar mediante una matriz de correlación la correlación entre cada variable y la columna target.

- a. ¿Hay datos fuertemente correlacionados con los retrasos? ¿y si sumamos un dataset con información asociada al pronóstico (viento y precipitación) para obtener la imagen completa?
- b. Eliminar las features fuertemente correlacionadas (una de cada par), ya que mantener columnas altamente correlacionadas, puede ocasionar un comportamiento no deseado en los modelos de clasificación.

4. Encoding:

Pasar las variables categóricas (strings) a numéricas. Analizar diferentes métodos para elegir el más adecuado (One hot encoding, Label encoding, Getdummies).

5. Escalamiento de los datos:

Transformar las features para que tengan distribuciones más cercanas a la normal (elegir qué método es más conveniente: logaritmica, normalizar, estandarizar). En el caso de usar PCA, este paso debe realizarse posteriormente de aplicarlo, para que en las componentes PCA las variables sean “pesadas” de manera similar, y no tenga alto impacto la varianza de las columnas originales

Características que debe cumplir el entregable:

- ✔ Generar un dataset “limpio”, con todos los pasos aplicados, ya que será el que utilizaremos en los siguientes TP para los modelos de clasificación. Concluir luego de ésta “limpieza” cuántos registros hemos mantenido/eliminado, con el fin de no quedarnos con muy pocos registros para avanzar más adelante con algún modelo de clasificación.
- ✔ Se debe ir desarrollando cada punto en la misma notebook donde se escriba el código. Dicho notebook debe contar con un índice, con sus diferentes apartados y el código debe ser fácil de leer, estar probado y comentado (esto último, en función de la necesidad).
- ✔ Se debe enviar el link directo del archivo .ipynb ó alternativamente subir el entregable a un repositorio GitHub mediante la integración con Google Colab. Recordar que al compartir el notebook, queden habilitados los permisos de edición, para poder dejar comentarios/correcciones.
- ✔ Tener en cuenta que si bien, pueden realizar diversos análisis y visualizaciones, se debe dejar en el entregable sólo aquello que sea relevante.
- ✔ Luego de cada análisis es importante poder obtener una conclusión de lo observado y/o breve interpretación de los resultados.

Fecha de Entrega: 17/07