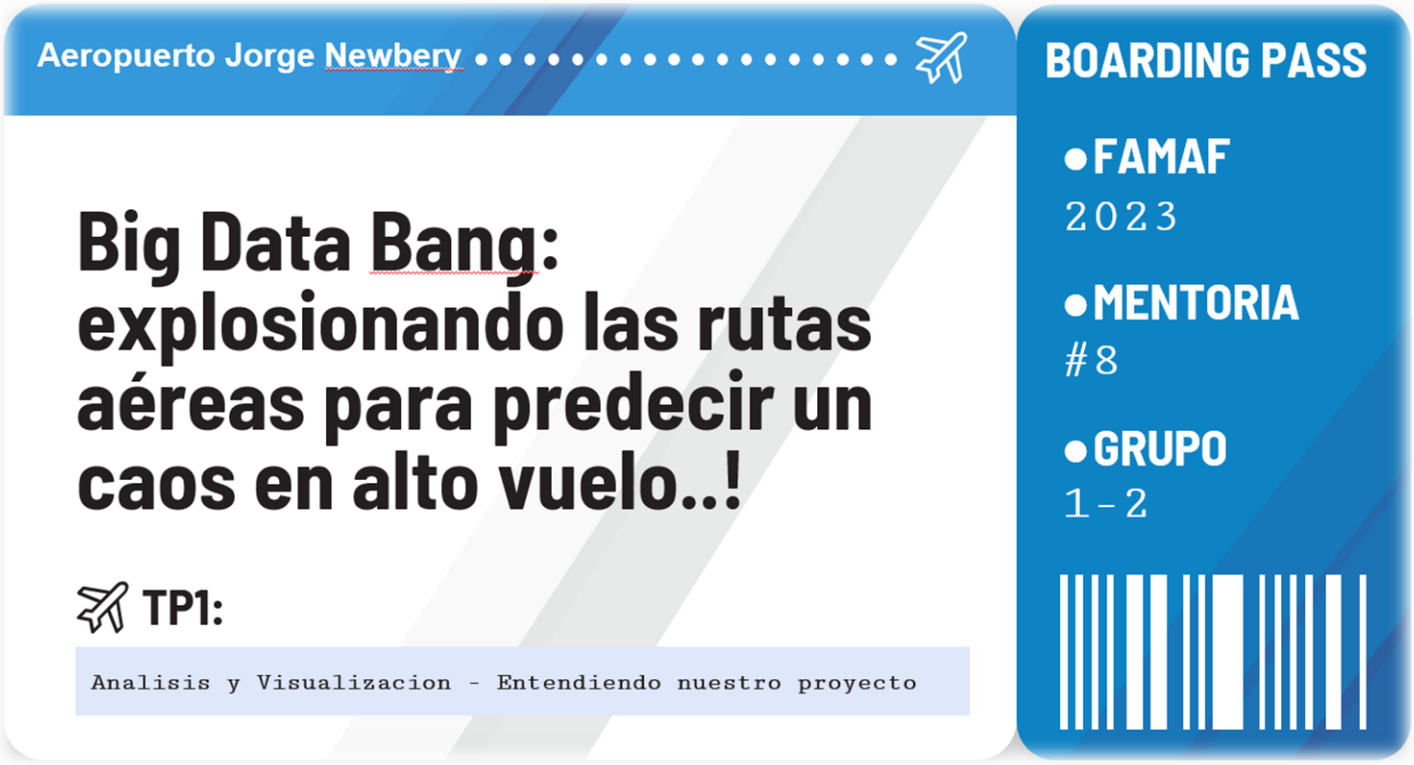

Comenzar un proyecto de Ciencia de datos, requiere mucho trabajo de análisis e investigación y vamos a ir encontrando respuestas, en gran medida, si sabemos lo que queremos lograr?

El objetivo de este Trabajo Práctico es comenzar a conocer nuestro proyecto, realizando una exploración de datos completa sobre el dataset bajo análisis. Esta exploración involucra:

1. Entender el dominio:

   Para lo cual, deberá crearse dentro del notebook, los siguientes apartados/secciones
   a. Abstract
   b. Contexto Comercial
   c. Problema Comercial
   d. Contexto analítico

   Tanto la contextualización, como el resto de los ítems citados, son importantes para el éxito de un proyecto de Machine Learning y deben ser considerados cuidadosamente durante todo el ciclo de vida del proyecto.

2. Comenzar a respondernos algunas preguntas que nos podemos hacer como Data Scientist:

a. Importar y analizar el dataset para ver con qué datos contamos, en cantidad y calidad: indagar no sólo el tipo de cada dato (categóricos, numéricos) si no su naturaleza (si son demográficos, económicos, temporales, etc).
b. ¿Entendimos correctamente el contexto del negocio?
c. ¿Consideramos que el objetivo podría ser cumplido con los datos disponibles?
d. ¿Tenemos forma de identificar de manera única cada registro? ¿Cuántas observaciones tenemos en el dataset?
e. ¿Qué periodos de fechas tenemos en el dataset?
f. ¿Tenemos datos nulos? Pensando en el histórico de datos, ¿tenemos datos "suficientes" para pensar en realizar un modelo predictivo?
g. ¿Tenemos la columna target (necesaria en problemas de aprendizaje supervisado)?

Hasta ahora usamos una buena cantidad de tiempo trabajando en el proyecto, sin escribir ni una línea de código, ni tampoco mirando los datos! Pensar, planificar y documentar es una parte importante de los proyectos, que frecuentemente es pasada por alto, pero como resultado de ese research inicial, obtenemos los conocimientos necesarios del contexto y luego que tenemos bien definida la pregunta o problema a resolver, estamos en condiciones de pasar a una serie de acciones que van a ayudarnos a seguir poniendo el foco en interpretar el comportamiento de los datos y entender lo más que se pueda el negocio.

3. Revisar los datos & comenzar proceso de Feature Engineering:

El siguiente paso es mirar a los datos con los que trabajaremos y hacer el correspondiente mapeo. Aún los set de datos "limpios" pueden tener errores y es vital trabajar con esos errores antes de comenzar el análisis. Generalmente buscamos responder a las siguientes preguntas:

¿Hay algún problema con los datos?
¿Hay defectos en los datos?
¿Necesitamos arreglar o eliminar algún dato?

Esto, involucra realizar acciones sobre todas las columnas de:

a. Revisión, eliminación o imputación de datos Nulos
b. Análisis de Nulos "no claros" (undefined? ceros? etc)
c. Evaluación de datos de outliers y análisis de los mismos, para decidir qué acción tomar.
d. Transformación de fechas, strings a numéricos y/o categórico y otras, en caso de ser necesario.
e. Limpieza de datos duplicados
f. Eliminación de datos innecesarios

Hasta aquí iría el TP obligatorio en relación a código.

Opcionales valorados:

4. Podemos comenzar a plantear algunas preguntas relevantes para el negocio, como por ejemplo:
    a. ¿Cómo varía el volumen total de arribos según el mes? ¿y por día de la semana?
    b. ¿Cómo varía el volumen de pasajeros por mes?
    c. ¿Existe alguna tendencia ó fluctuación cíclica en función de la estacionalidad? ¿Hay días especiales en el año?
    d. ¿Cual es la distribución de vuelos provenientes de los distintos orígenes?
    e. y otras similares.

5. Construcción de nuevas columnas, vinculando otras columnas.
    a. Construcción de columnas de información valiosa. Por ejemplo, crear una columna donde solo este el mes, para hacer un conteo de vuelos por meses, , etc.
    b. Idealmente, construir columna target suponiendo predicción de retraso. Para esto, se debería tomar como …
    c. En caso de construir dicha columna, evaluar el balance (proporción de 1 vs proporción de 0) de la misma, y la distribución de las otras variables contra el target.
    d. Selección de variables relevantes para predecir los retrasos

6. Elaboración de un proceso de tratamiento de datasets (supongamos que llega un dataset nuevo, el proceso debería realizar el mismo procesamiento realizado en esta oportunidad, para obtener un dataset limpio y prolijo)

TP1 – Análisis y Visualización

Características que debe cumplir el entregable:

- Un proyecto de Ciencia de datos, es un proyecto que tiene que estar fuertemente estructurado. Se recomienda usar la PEP8, (guía que indica las **convenciones estilísticas** a seguir para escribir código https://ellibrodepython.com/python-pep8)
- 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.

Input:

https://github.com/NoeliaFerrero/Proyecto_MentoriaFAMAF_2023/blob/bca35cab78763d906f8cbaf28c650ff1d1036135/DataSet%20Aeropuerto%20Jorge%20Newery.csv

Fecha de Entrega: 23/06

¡Feliz comienzo!