

Proyecto:

***SIN BAJAR LA GUARD.IA ~ CONSTRUYENDO CONOCIMIENTO AL SERVICIO
DE LA SALUD***



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 completa de los datos 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/Problema Sanitario
- c. 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 de todo el proyecto.

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

- a. Importar y analizar los datasets 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 caso de estudio?
- 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. Pensando en el histórico de datos, ¿tenemos datos "suficientes" para pensar en realizar un modelo de ML?
- f. ¿Tenemos alguna columna target/objetivo (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 que forme parte de un algoritmo! 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 al máximo, el contexto.

3. Revisar los datos:

El siguiente paso es mirar a los datos con los que trabajaremos y hacer el correspondiente mapeo. Aún los sets 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

En este punto, se sugiere fusionar los datasets de establecimientos y especialidades médicas (normalizar formatos).

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.

Repositorio:

https://github.com/NoeliaFerrero/Proyecto_MentoriaFAMAF_2024

Deadline: 01/07

¡Feliz comienzo!