Data Processing Domain

Anotaciones de ayuda para el desarrollo de los notebooks

06.- Data Adequacy

Lo realizaré todo sin excepción, e incluso aquí deberé incluir cosas que no se reflejan y que me aparezcan en mi caso / fichero

07.- Data Sampling

No será necesario seguramente en ningún caso de los que tenéis vosotros para los datos que vosotros manejáis, pero debo justificar el porqué … eso lo podré en Main Conclusión

Pero después, seguramente sí que puede que haga falta, cuando los datos crezcan. De esta forma, y dado que el código lo tenéis ya dado, debéis hacerle una pasada con el fichero en cuestión que ahora tenéis y lo dejáis listo para cuando ese mismo fichero crezca, independientemente de que se concluya que con los datos actuales vuestros no hace falta.

Y de paso así aprendéis de que va esta parte que es muy muy importante en casos con datos reales en los que será necesario seguramente.

08.- Data Split

- ¿Debo hacer Split de mis datos?. … solo si hago machine Learning obviamente. El Split se usa para entrenar modelos. Si uso algoritmos genéticos, pues no, obviamente, no es una técnica de Machine Learning. En este sentido abajo debéis justificar el echo de hacer o no Split.
- Si lo debo hacer ¿como afronto esto?. Veréis mucho código y no fácil de entender para algunos.
    - Una primera aprox es justificar abajo que de momento usaremos 80/20, lo estándar y que a posteriori se realizará este proceso cuando se tengan claros las métricas, los modelos, etc. Solo estará justificado cuando esté claro que no tenemos ni idea todavía de las métricas, modelos, etc. Esto no se sostiene si ya sé que voy a predecir y sé la variable (precio por ejemplo). Raro es que no sepamos ya esto.
    - Segunda aprox si ya sabemos que la variable a predecir por ejemplo. Cojo simplemente un par de modelos y un par de métricas y pruebo para ir teniendo una idea.
    - Aprox final: si o si este notebook al final tengo que volver a él y hacerle bien cuando ya sepa todas las técnicas a usar para elegir el Split correcto.

09.- Data Cleansing

Sí o sí todo y para todos los ficheros. ESTE ES VITAL PARA NO ARRASTRAR DATOS INCORRECTOS.

10.- Imbalanced Analysis

¿Debo hacerlo en mis datos?. Aunque sea intuitivo ver como son los datos, y lo vea a simple vista, soy formal y hago la primera parte que me muestra si están o no balanceados, y

posteriormente ya en función de lo que tenga debo tratar o no esos datos no balanceados de la forma correcta.

Es muy importante que si tenéis datos no balanceados vuestro responsable de caso os diga y oriente en como tratarlo ya que no es un tema menor.

Siempre justificaré abajo todo en un sentido u otro.

11.- ECA - Exploratory Causal Analysis

No lo abordamos en nuestros casos.

12.- EDA

Importantísimo. Me permite sacar conclusiones (que obviamente expondré en cada Apartado y abajo) sobre los datos.

Hay varios notebooks particularizados, de forma que seguiré esta línea de trabajo:

TODOS

12.1.- Univariate Analysis.ipynb

12.2.- Normality Analysis.ipynb

12.3.- Bi-variate Analysis.ipynb

SI HAGO REGRESIÓN EN MI CASO

12.6.- Regression Analysis.ipynb

SI HAGO CLUSTERING EN MI CASO

12.10.- EDA by Clustering.ipynb

SI USO MODELOS DE SERIES TEMPORALES EN MI CASO

12.7.- Stationary Analysis.ipynb

PARA LOS DE PRIMERA FILA DE LA CLASE

12.8.- Parametric Tests.ipynb

12.9.- Non-Parametric Test.ipynb

DEJO DE MOMENTO A NO SER UE ME ATREVA CON ALGO DE ESTO

12.4.- Hypotesis Test.ipynb

12.5.- Homogeneity Analysis.ipynb

13.- Data Visualization

Lo usaré como complemento y apoyo al 12.- EDA

Debéis meter cuanto más análisis visual mejor, peor que tenga lógica obviamente ….