Maestría en Ciencia de Datos

Data Mining en Economía y Finanzas

Comisión Jueves

Competencia Kaggle

DMEyF 2024 Tercera

Robinson Galvis Marin

robins_gama@mail.com

UBA

Instructivo para la ejecución y carga de la tercera competencia en Kaggle

Se deben seguir los siguientes pasos, con el fin de llevar a cabo la correcta ejecución, predicción y carga del archivo en Kaggle.

1. Se debe contar con el dataset inicial, llamado **competencia_03_crudo.csv.gz**

2. Ejecutar en Python el script llamado **script_clase_ternaria.ipynb** para la obtención del atributo **clase_ternaria**.
    a. Aquí se debe ajustar la ruta en la que se almacenó el archivo indicado en el paso 1.
    b. Con este script se calcula la clase ternaria y se genera el archivo **competencia_03.csv.gz**, el cual se utilizará para los pasos posteriores y debe cargarse en el bucket creado en **Cloud Storage (GCP)** en la carpeta **datasets/**.

3. Los procesos que vienen a continuación requieren de una máquina optimizada para tareas complejas, es decir un equipo de alto rendimiento y alto poder de cómputo, por lo que se recomienda la creación de una máquina virtual. Para esto se puede seguir el siguiente documento guía para trabajar sobre un entorno de nube (**GCP - Google Cloud Platform**):
    a. https://storage.googleapis.com/open-courses/dmeyf2024-b725/GoogleCloud_DMEyF2024.pdf

4. Asimismo, la ejecución se apoya en la plataforma **WorkFlow UBA**, la cual consiste en una metodología de trabajo con una serie de scripts brindados por la cátedra. Para el uso adecuado de esta plataforma tener en cuenta lo siguiente:
    a. Los scripts requeridos para la correcta ejecución se encuentran en el siguiente GitHub: https://github.com/robinson-galvis/dmeyf2024/tree/main/src
    b. El documento guía con información relevante sobre WUBA es https://storage.googleapis.com/open-courses/dmeyf2024-b725/WorkFlow_UBA.pdf

5. Finalmente, se debe acceder a R y ejecutar el script **909_run_orden227.r**, ubicado en la carpeta **dmeyf2024\src\workflows** y que llama al orquestador **990_wf_semillero_rf_a25_cn_r1_d0_sem881269_comp3_usmp_02.r**, el cual, a su vez, realiza el llamado a los demás scripts asociados a las diferentes etapas consideradas para la obtención del resultado final:
    a. Se selecciona el envío **KA-0023_01_013_r1_11000.csv**

6. **Adicional:** Para la selección del envío final se realizó el siguiente proceso:
    a. En Python, ejecutar el archivo **script_selección_semilla Competencia3.ipynb**, con el cual se genera la gráfica anexada más adelante.
    b. La selección del mismo se da con base en las indicaciones recibidas en la cátedra, tratando de dejar un punto medio entre:
        i. Un envío que estuviera cerca al promedio general de los demás envíos.
        ii. El corte óptimo se encuentra entre 10500 y 11000.
        iii. Identificar un valle en los envíos generados por los diferentes cortes de un mismo modelo.