Maestría en Ciencia de Datos

Data Mining en Economía y Finanzas

Comisión Jueves

Competencia Kaggle

DMEyF 2024 Segunda

Robinson Galvis Marin

robins_gama@mail.com

UBA

Instructivo para la ejecución y carga de la segunda 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_02_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 dejó el archivo indicado anteriormente.
    b. Con este script se calcula la clase ternaria y se genera el archivo **competencia_02_Sin_Baja1.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_sinB1_sem881269.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. **KA-0005_01_021_r1_11500.csv**