INSTRUCTIVO PARA EJECUCIÓN DE LOS NOTEBOOKS DEL PROYECTO

El proyecto fue desarrollado en cuatro notebooks.

En todos los notebooks se encuentran integrados los códigos necesarios para extraer los datos automáticamente. Por consiguiente, ninguna carga de datos deberá efectuarse para ejecutar los notebooks.

El orden en el cual deberán ser ejecutados es el siguiente:

Nota: Los inputs y outputs presentados en este documento hacen referencia a archivos.

00. creacion_análisis_y_tratamiento_datos.ipynb

Desde el repositorio LuisPortela/ELO se descarga un archivo tipo Json, el cual es un archivo de configuración con credenciales que permite acceder directamente a los conjuntos de datos del Challenge de Kaggle desde Colab. Adicionalmente desde el repositorio se obtienen un conjunto de archivos con formato csv que son necesarios para ejecutar el notebook.

Síntesis: ejecutando este notebook se realizará inicialmente la exploración de los conjuntos de datos provenientes del challenge de kaggle, seguido a esto se realizará la unificación de los conjuntos de datos en un único dataset. Una vez obtenido este dataset, se realizará las siguientes tareas: tratamiento de datos, análisis descriptivo, experimentación de los cluster provenientes del notebook Anexo_Creación_Cluster_Merchants_.ipynb y finalmente la estandarización de variables cuantitativas y codificación de variables cualitativas.

Input:

- Train.cvs
- new_merchant transactions.csv
- historical_transactions.csv
- DF_Merchants_GMM_k18_DB.csv
- DF_Merchants_GMM_k22_AIC_BIC.csv
- DF_Merchants_Kmeans_k2_SLH.csv
- DF_Merchants_Kmeans_k3_DB.csv
- DF_Merchants_Kmeans_k8_Entropia.csv

Output:

- 01_DB.csv

Anexo_Creación_Clusters_Merchants_.ipynb

Este notebook es producto de la experimentación con el conjunto de datos **merchants,** sin embargo, según los resultados obtenidos, se concluye que no hará parte de la línea central del

proyecto. Por lo mencionado anteriormente, este puede ser ejecutado en cualquier orden sin afectar los resultados o su interpretación.

Para la ejecución de este notebook se importan dos conjuntos de datos en formato csv, el primer archivo se obtiene directamente del API de Kaggle y el segundo se obtiene del repositorio LuisPortela/ELO

Síntesis: Al ejecutar este notebook se cargará el conjunto de datos merchants, al cual se le realizará un preprocesamiento de los datos, con el objetivo de implementar diferentes técnicas de clustering asi agrupar los comercios por características similares.

En este notebook se experimentaron dos técnicas de cluster (GMM- Kmeans)y múltiples métricas de error (Inercia, Davies Bouldin, entre otras), las cuales generan tipos de agrupamiento distinto.

Input:

- merchant_id_DF.csv
- merchants.csv

Output:

- DF_Merchants_GMM_k18_DB.csv
- DF_Merchants_GMM_k22_AIC_BIC.csv
- DF_Merchants_Kmeans_k2_SLH.csv
- DF_Merchants_Kmeans_k3_DB.csv
- DF_Merchants_Kmeans_k8_Entropia.csv

Notebooks de modelado

Los notebooks 01. Modelado_Regresión y 02 Modelado_Clasificación son independientes y se diseñaron según la tarea que se esté resolviendo. Ambos notebooks tienen la lectura de datos en común. Con respecto a la carga de datos; a partir del repositorio LuisPortela/ELO se descarga el archivo 01_DB.csv, este archivo contiene el conjunto de datos proveniente de la integración y procesamiento de los conjuntos de datos en el notebook **00. creacion_análisis_y_tratamiento_datos.ipynb**

01. Modelado_Regresión

Síntesis de colab: Al ejecutar este notebook se leerá el conjunto de datos 01_BD.csv, con el cual se evaluarán dos técnicas de reducción de dimensionalidad(row reduction and features reduction), de este proceso se obtiene un dataset reducido con el fin de optimizar el tiempo de ejecución durante el modelado. Posteriormente, se crearán los modelos para la tarea de regresión, realizando una búsqueda aleatoria de hiper parámetros en un rango amplio de valores y luego una búsqueda exhaustiva de hiper parámetros acotados para cada modelo. En este notebook se evaluarán siete modelos de regresión diferentes (Linear Regressor, Decision Tree, Regressor, Radom Forest, entre otros) y se presentarán las métricas de error (R^2 y RMSE) tanto para el conjunto de datos de train como para el conjunto de datos test.

Input: 01_DB.csv

Output: N/A

02. Modelado_Clasificación

Síntesis de colab: Ejecutando este notebook se leerá el conjunto de datos 01_BD.csv, con el cual se evaluarán dos técnicas de reducción de dimensionalidad(row reduction and features reduction), de este proceso se obtiene un dataset reducido con el fin de optimizar el tiempo de ejecución durante el modelado de la tarea de clasificación. Posteriormente, se crearán los modelos para la tarea de clasificación, realizando una búsqueda aleatoria de hiper parámetros en un rango amplio de valores y luego una búsqueda exhaustiva de hiper parámetros acotados para cada modelo. En este notebook se evaluarán seis modelos de clasificación diferentes (Logisct regression, DecisionTreeClassifier, Radom Forest Classifier, entre otros), y se presentarán la métrica de error (ROC Score) tanto para el conjunto de datos de train como para el conjunto de datos test.

Input: 01_DB.csv

Output: N/A