Tarea #3 - Big Data

Objetivo

Ejecutar el proceso de extracción de datos y entrenamiento de un modelo en Apache Spark de principio a fin.

Resultados esperados

Para esta asignación los estudiantes deberán entregar un Jupyter Notebook donde se entrene **un modelo de clasificación binaria**, basada en algún conjunto de datos de su escogencia (e.g. conjuntos de Kaggle)

El notebook deberá ser autocontenido en su ejecución y análisis de resultados, utilizando una instancia de Postgresql como apoyo, ejecutada a través de un contenedor, como hemos hecho hasta el momento. El notebook deberá ejecutarse desde un contenedor Docker. Se espera que los estudiantes se basen en infraestructura Spark y no en código secuencial hecho en Python. El uso del framework será parte de lo evaluado en la asignación.

Entrega: Archivo comprimido con notebook, guía de ejecución en formato PDF con las instrucciones y los datos por medio del TEC-Digital a más tardar el **lunes 18 de Diciembre de 2023 a las 5:00 PM**

Datos de entrada (5 puntos)

Los estudiantes podrán seleccionar un conjunto de datos de su preferencia. Se espera que se provea una descripción de los datos, donde se detalle cuál es el dominio del problema y una descripción de los diferentes atributos en el conjunto. **Debe incluir explícitamente cuál es la variable de predicción a utilizar.**

Preprocesamiento de datos

Similar al protocolo visto en clase, la primera fase deberá leer y ajustar los datos previo a la fase de entrenamiento. Se espera que los estudiantes cumplan con:

- Cargado y limpieza de datos de archivo de entrada (en formato CSV). Esto implica la definición del "schema" y muestras en el notebook que los datos se han cargado exitosamente **(5 puntos)**

- Gráficos y estadísticas descriptivas previo al entrenamiento. Se espera que los estudiantes muestren estadísticas descriptivas, correlaciones, etc. Esto con el fin de entender el conjunto de datos **(10 puntos)**
- Normalización / Estandarización. Los estudiantes deberán seleccionar alguna estrategia para mitigar los problemas de escala que pueden tener las diferentes columnas del modelo **(10 puntos)**
- Escritura a base de datos. Una vez que los datos hayan sido depurados se espera que los estudiantes escriban a una tabla llamada **tarea3** (con overwrite) el conjunto de datos que se utilizará como base para el entrenamiento de los modelos. Los estudiantes deberán documentar en detalle cualquier instrucción necesaria para poder calificar esta sección. **Los datos escritos en la base de datos no podrán estar almacenados en forma de vector**. Deben ser extraídos a columnas individuales **(10 puntos)**

Entrenamiento de modelos

Se deberá cargar de la base de datos el conjunto de datos limpio y se deberá entrenar dos modelos de clasificación (a escoger por los estudiantes). Se espera que se utilice el protocolo estándar de k-fold cross validation además de dejar un conjunto adicional para validación final.

- Uso de protocolo K-fold cross validation, apoyándose en funciones Spark **(10 puntos)**
- Entrenamiento de dos modelos **(10 cada uno)**
    - En este rubro se incluye analizar métricas sobre el conjunto de datos de entrenamiento (en la siguiente sección se evalúa el conjunto de validación)

Evaluación de conjunto de validación

Para cada uno de los modelos se espera que los estudiantes los evalúen y generen una predicción persistente en base de datos. Como evaluaremos dos modelos deberá crearse tablas llamadas **modelo1** y **modelo2** (con overwrite) en la base de datos, que tendrán las mismas columnas que **tarea3** con una adicional llamada **prediccion**, que mostrará el resultado predicho de cada modelo.

Además, deberá mostrarse un análisis de resultados dentro del notebook para cada modelo, comparando los resultados de cada uno. Debe aportar suficiente detalle en el análisis de los resultados.

- Evaluación y almacenado de modelo1 **(10 puntos)**
- Evaluación y almacenado de modelo2 **(10 puntos)**
- Análisis de resultados **(10 puntos)**

Consideraciones generales

Para obtener el puntaje de cada uno de los rubros los estudiantes deberán mostrar suficiente información en la salida del Jupyter Notebook para demostrar que se cumple con lo pedido. Como el Jupyter Notebook deberá correrse sobre un contenedor, deberá entregarse un archivo comprimido que contenga el análogo al repositorio con un Dockerfile, además de un PDF con todas las instrucciones necesarias para poder ejecutar exitosamente el código del Notebook.