# Manual Proyecto Big Data

## Estudiante: María Yorleni Alfaro Alfaro

En el presente documento se encuentran las instrucciones respectivas para ejecutar tanto el programa principal de Python como el Jupyter Notebook.

### Pasos para la ejecución:

1. Primero se debe descomprimir el archivo proyecto.zip entregado.

Los siguientes son los archivos que se encuentran en dicha carpeta y son necesarios para la correcta ejecución:

* **proyecto.py**: es el programa principal de Python, desde el cual se realiza el llamado a las diferentes funciones para el cargado y preprocesamiento de los datos, y su posterior escritura a base de datos (con las fuentes de datos ya limpias).
* **proyecto\_funciones.py**: en este archivo se encuentra el código de cada una de las funciones desarrolladas (que son llamadas desde el programa principal “proyecto.py”)
* **test\_ proyecto.py**: contiene una serie de pruebas unitarias que permitan corroborar la correctitud de las diferentes funciones internas al programa.
* **BigData-Proyecto\_MaríaYorleniAlfaroAlfaro.ipynb**: es el Jupyter Notebook donde se desarrollan los 2 modelos de predicción. En este archivo también se encuentra el análisis de resultados.
* **Archivos csv**: conjuntos de datos fuente utilizados para el desarrollo del proyecto (estos se encuentran almacenados en la carpeta “Fuentes\_de\_datos”)
  + **megabaseprimaria\_2015.csv**: contiene una serie de características de las escuelas en Costa Rica
  + **IDS\_distrital\_cantonal.csv**: contiene información del índice de desarrollo social distrital de Costa Rica
* **postgresql-42.2.14.jar:** este archivo es requerido para el correcto funcionamiento con postgresql.
* **Archivo de configuración**:
  + \_\_init\_\_.py
  + conftest.py
  + Dockerfile

1. Abrir Power Shell o línea de comandos
2. Navegar al directorio donde se descomprimieron los archivos del código fuente
3. Realizar el build de la imagen con el comando: **docker build --tag proyecto .**
4. Dado que se requiere una instancia de Postgresql, ejecutada a través de un contenedor, se deben ejecutar los siguientes comandos:
   1. **docker rm bigdata-db** (esta instrucción solo es necesaria en caso de que ya exista)
   2. **docker run --name bigdata-db -e POSTGRES\_PASSWORD=testPassword -p 5433:5432 -d postgres**
5. Ejecutar la imagen del Docker con el comando: **docker run -p 8888:8888 -i -t proyecto /bin/bash**
6. Para la ejecución del programa principal, ejecutar el siguiente comando:

“**spark-submit --driver-class-path ./postgresql-42.2.14.jar proyecto.py**”.

Como producto de la ejecución anterior, se crearán (o se sobreescribirán) en la base de datos PostgreSQL llamada “postgres”, las siguientes tablas:

* **escuelas**: contiene la información ya preparada de las escuelas (una vez que se ha realizado el preprocesamiento de los datos).
* **ids**: contiene la información ya preparada del índice de desarrollo social distrital (una vez que se ha realizado el preprocesamiento de los datos).
* **escuelas\_ids**: contiene la información ya cruzada de ambas fuentes de datos (escuelas e ids).

Nota importante: la estructura de cada una de las tablas se encuentra almacenada en la carpeta “Estructuras tablas sql” (específicamente en los archivos “escuelas.sql”, “ids.sql” y “escuelas\_ids.sql” respectivamente). Al final de este documento hay una sección llamada “Descripción detallada de los datos” donde se encuentra la descripción de cada una de las columnas de estas tablas.

1. Para la ejecución de las pruebas, ejecutar el siguiente comando:

“**pytest test\_proyecto.py -v**”.

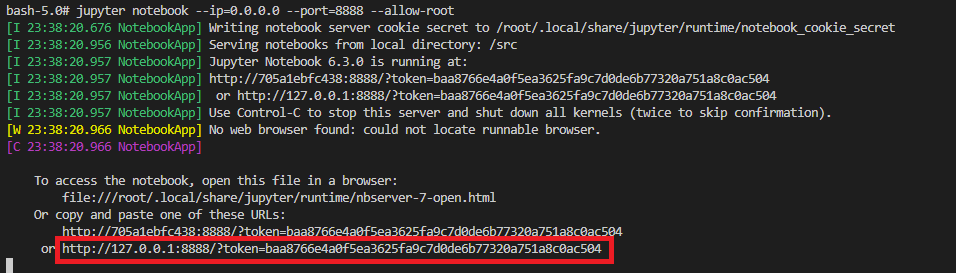
Con la ejecución del comando anterior se muestra el resultado de cada una de las pruebas.

Una vez ejecutados todos los pasos anteriores, se procede a abrir el Jupyter Noteboook donde se desarrollan los 2 modelos de predicción. Para esto siga las siguientes instrucciones.

1. Ejecutar el siguiente comando para cargar el jupyter notebook:

**jupyter notebook --ip=0.0.0.0 --port=8888 --allow-root**

Esto le devolverá un resultado como se muestra en la siguiente imagen:



1. Navegue a la dirección marcada en rojo. Esto le abrirá Jupyter, como se muestra en la siguiente imagen:



1. Abra el archivo “BigData-Proyecto\_MaríaYorleniAlfaroAlfaro.ipynb”
2. Ejecute cada uno de los pasos del jupyter notebook para validar su funcionamiento.
3. La siguiente es la información requerida para conectarse a la base de datos (ya sea a través de pgAdmin, o bien a través de una terminal con el comando psql) y revisar las tablas generadas:
   1. Host: localhost
   2. Puerto: 5433
   3. Usuario: postgres
   4. Password: testPassword
   5. Base de datos: postgres

Las tablas creadas con la ejecución del jupyter notebook son las siguientes:

1. tarea3: almacena el conjunto de datos limpio que se utiliza como base para el entrenamiento
2. modelo1: contiene las mismas columnas que la tabla “tarea3” con una columna adicional llamada “predicción”, que muestra el resultado predicho del modelo 1.
3. modelo2: contiene las mismas columnas que la tabla “tarea3” con una columna adicional llamada “predicción”, que muestra el resultado predicho del modelo 2.