Para los documentos y el desarrollo del proyecto se encontraran en la carpeta llamada “Proyecto_Agustin_Vargas”, dentro de esta carpeta se encuentra todo el desarrollo del proyecto, primero en la carpeta “data” se encuentran los datasets que se utilizaran para cruzarlos y poder cumplir con el objetivo del proyecto, dentro de la carpeta “data_processing” se encuentran las funciones para tratar cada uno de los datasets y las funciones utilizadas para poder realizar el cruce de los datasets y realizar algunos cálculos para el objetivo predictivo, por ultimo en la carpeta “tests” se encuentran las pruebas unitarias para verificar el correcto funcionamiento de las funciones para tratar y cruzar los datasets, en la carpeta base se encuentran los archivos “DockerFile”, “postgresql-42.2.14.jar”, el archivo .sh el cual es para correr lo que se encuentra en el archivo programaestudiante.py y el jupyternotebook donde se desarrollara el proyecto.

Para poder ejecutar la primera parte del proyecto la cual es realizar una limpieza de los datasets, cruzar los datasets se deberá de correr el comando:

- spark-submit --driver-class-path postgresql-42.2.14.jar --jars postgresql-42.2.14.jar programaestudiante.py

Esto se debe de hacer en la raíz de la carpeta ya que aquí es donde se encuentra el archivo .sh.

Para correr las pruebas unitarias se deberá de estar en la raíz y correr el comando:

- pytest

Para ejecutar lo que se encuentra en el archivo jupyter solo se debe de tener la base de datos corriendo y haber ejecutado el archivo .sh para realizar todo el procesamiento de los datos.