**Objective:**

- To apply **Machine Learning model** for the given dataset.
- To prepare a jupyter notebook or Google Colab to build, train and evaluate a Machine Learning models using **MLlib - PySpark DataFrames on Databricks** for the given dataset.
- To provide appropriate analysis for the same and do the prediction for the test data and display the results for the inference.

**Please read the instructions carefully.**

**Dataset -** https://drive.google.com/file/d/1F7kSOWXJKeZX44NXCXhwysfWT5wXfxSP/view?usp=sharing

Consider the dataset of COVID-19 . The data is a collection of number of cases (positive, recovered, died) per day per geographical area recorded for the response of COVID- 19.

1. **Import Libraries/Dataset**
   a. Download the dataset
   b. Import the required libraries

2. **Data Visualization and Exploration**
   a. Print at least 5 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   b. Print the description and shape of the dataset.
   c. Provide appropriate visualization to get an insight about the dataset.
   d. Try exploring the data and see what insights can be drawn from the dataset.

3. **Data Pre-processing and cleaning**
   a. Do the appropriate preprocessing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Apply appropriate feature engineering techniques for them.
   b. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset.
   c. Do the correlational analysis on the dataset. Provide a visualization for the same.
4. **Data Preparation**
   a. Do the final feature selection and extract them into Column X and the class label into Column into Y.
   b. Split the dataset into training and test sets.

5. **Model Building**

**a.** Perform Model Development using at least three models, separately. You are free to apply any Machine Learning Models on the dataset by using **MLlib- PySpark**. **Deep Learning Models are strictly not allowed.**

b. Train the model and print the training accuracy and loss values.

6. **Performance Evaluation**

   a. Print the confusion matrix. Provide appropriate analysis for the same.

   b. Do the prediction for the test data and display the results for the inference.

Reference:
https://docs.databricks.com/getting-started/dataframes-python.html
https://www.kaggle.com/code/towhidultonmoy/end-to-end-pyspark-project
https://www.kaggle.com/code/tientd95/advanced-pyspark-for-exploratory-data-analysis

------