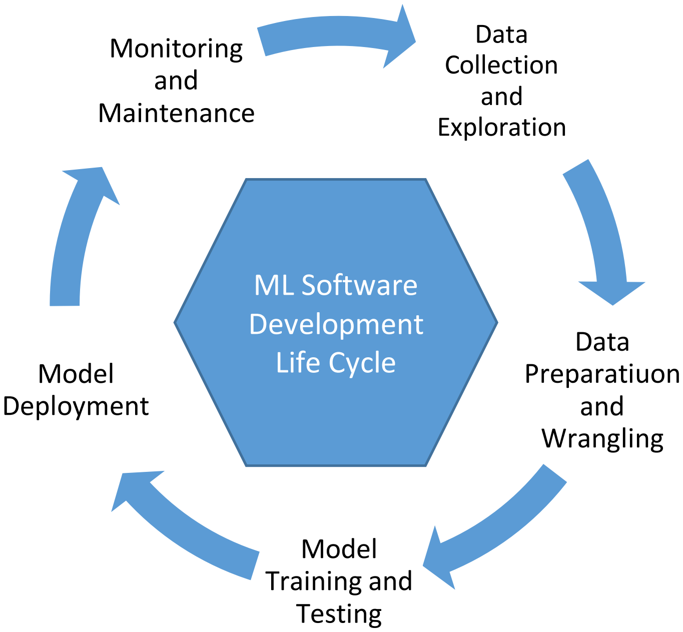
Birla Institute of Technology and Science Pilani, Hyderabad Campus
2^{nd} Semester 2023-24, BITS F464: Machine Learning
Assignment No: 1,   Date Given: 20.01.2024,   Date of Sub: 25.01.2024

(Note: Data Exploration, Pre-processing and Wrangling using Scikit Learn)   Max. Marks: 4

This assignment is about data exploration, pre-processing and wrangling which are some of the important initial steps in the Machine Learning software development life cycle (as shown in Fig.1) before the ML model is built, deployed and maintained.


ML Software Development Life Cycle
Data Collection and Exploration
Data Preparatiuon and Wrangling
Model Training and Testing
Model Deployment
Monitoring and Maintenance

(Fig.1: Machine Learning Software Development Life Cycle)

Raw data is often messy, inconsistent, and incomplete. Data exploration involves understanding the characteristics of your data like finding out no. of non-empty values, maximum value, minimum value, mean, standard deviation, how many values are less than 50% percentile etc. in your dataset.

Data Pre-processing techniques like handling missing values, encoding categorical values (like Yes to 1 and No to 0) etc. and Data wrangling techniques (process of cleaning the data) like removing duplicates and invalid data etc. play an important role in making the data ready for analysis or training and testing using a suitable ML model.

You are given with a CSV file containing India's rainfall data (in millimeters) in the month of June from the year 1901 to 2015, downloaded from data.gov.in/catalog/rainfall-india. Some of the years' data intentionally is removed in the CSV file, so that you can treat those as Missing values (4 rows). Use this data to carry out the following:

(Please Turn Over)

1. Install Anaconda distribution for Python which is free and open source on your local computer (PC or Laptop). Go to the Navigator and launch a Jupyter Notebook as we discussed in the class.
2. Import pandas library for dealing with the data. Read data from the data.csv (given here in this assignment) containing rainfall data into a Panda’s DataFrame. Display how many rows are there in the file and its contents. Reading from the file that is available locally can be done by using:

```
header_names = ['YEAR','Rainfall in June']
data = pd.read_csv(filepath/data.csv, usecols=header_names)
```

3. Descriptive statistics normally include those stats that summarize the central tendency, dispersion and shape of a dataset’s distribution. Generate descriptive statistics using describe method on the pandas DataFrame. Display stats like mean, median and standard deviation.
4. Instead of removing the 4 rows that have missing values in the given CSV file, impute Median of the values (rainfall) by calling the median method on the rainfall attribute of the DataFrame. You could also impute with mean or mode if you wish.
5. Two rows in the CSV file have duplicates. Remove those duplicates by calling drop_duplicates method on the DataFrame.
6. Use Matplotlib to create a histogram plot of the data distribution to see if it is symmetry or skewed to the right or skewed to the left. Verify the behavior by comparing the Mean with Median.
7. Import Scikit Learn (sklearn library) into your code and perform Standard scaling of the data using StandardScalar (as discussed in the class) and MinMaxScalar using the Scikit functions and report their differences in range.
8. Build a Linear Regression model for predicting the future rainfall. Use 80-20 rule to split the data into training and testing.

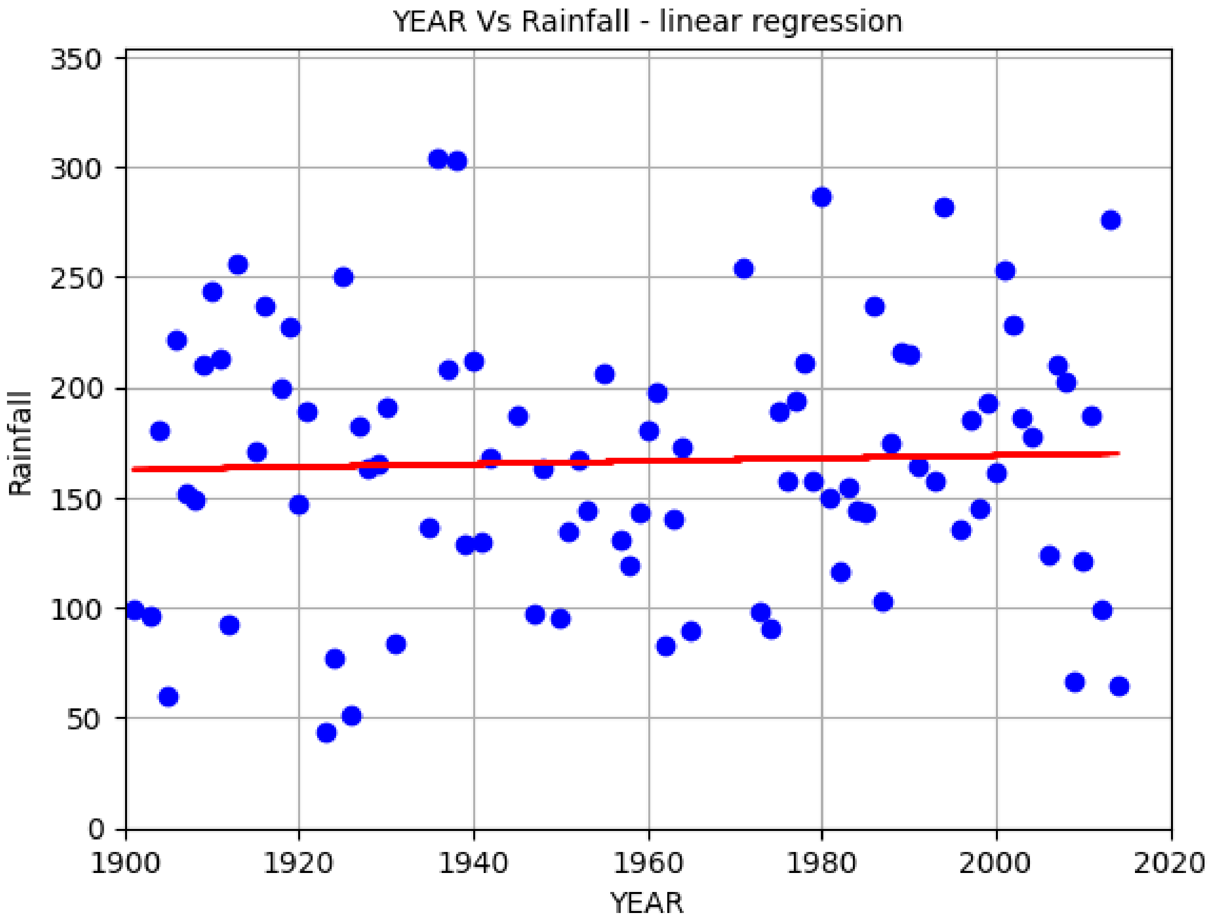
   ```
   model = LinearRegression().fit(x_train,y_train)
   ```

   Regression line on the training data might look similar to the one as shown in Fig.2. Report the MAE, MSE and RMSE as evaluation metrics of your model showing how has it performed. Below shown a sample code for Mean Squared Error (MSE) loss function. You may write similar codes for RMSE and MAE metrics.

   ```
   print(f"Mean Squared Error (MSE) of the test data:
   {metrics.mean_squared_error(y_test,model.predict(x_test))}")
   ```

9. Rerun Step no. 8 by changing the train: test spilt to 70:30 and observe the performance metrics as defined in the previous step.

(Please Turn Over)



(Fig.2 Regression plot on Training rainfall data)

Submission Instructions:

You are free to form your own group of maximum three students as informed in the class. Clean your Notebook code (ipynb) and name your submission file as IDNo.ipynb. Write a readme.text with your group members name and ID nos. Compress these two files into a single Zip, and name your Zip file using your idno (in lowercase). Make only one submission on behalf of your group in Google Class Assignment Submission page. Deadline for submitting your work is 25th Jan 2024 midnight.

References:

1. https://www.w3schools.com/python/pandas/default.asp
2. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
4. https://www.statology.org/matplotlib-distribution-plot/
5. https://docs.anaconda.com/free/anaconda/index.html

--~ --