**MACHINE LEARNING PROJECT**

***Project Name***

## Biological Oxygen Demand (BOD) prediction using XG Boost, Bayesian Ridge Regression and Linear Regression

***Submitting to*Vishwakarma Institute of Information Technology, Pune**

***Submitted by***

**Prof. Shreenivas Londhe**

# Abstract of Project:

* Biochemical oxygen demand (BOD) are critically important parameters for water quality assessment as well as for development of various management policies. The present study aims to develop a water quality model for the critical parameter biochemical oxygen demand (BOD) using the influential parameters.
* The data is read and the characteristics are understood followed by developing model using Machine learning technique.
* After preparing the data apply training and testing to the model.
* In this project the machine learning techniques used are: XG Boost, Bayesian Ridge Regression and Linear Regression
* Using sklearn packages importing the machine learning algorithms

and finding the Accuracy of the models.

# Project Summary:

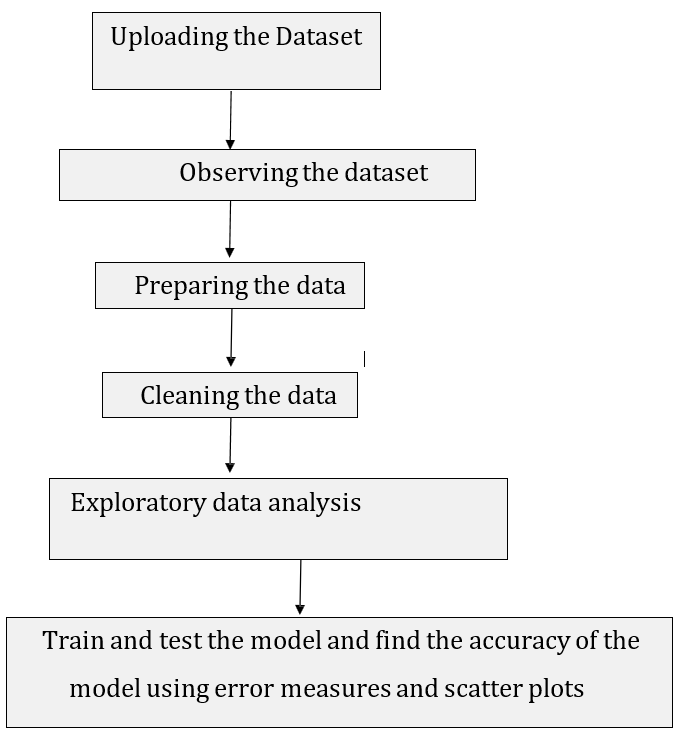
* The present project contains parameters important for BOD prediction: Input parameter- Dissolved oxygen (DO), pH, Electrical conductivity (EC), Alkalinity (ALK), Total Solids (TS), Total Dissolved Solids (TDS), and Nitrite/Nitrate (No3-N, No3-No2). and the Output parameter as BOD of river.
* Using pandas we can visualize the dataset features.
* Prepare the dataset.
* Perform the exploratory data analysis using seaborn and matplotlib Modules.
* Analyze the insights of the dataset using different visualization techniques. Cleaning the dataset if there are any categorical values in the dataset.
* Find the independent and dependent variables after encoding the dataset.
* Splitting the dataset into training and testing from sklearn package import train\_test\_split
* Evaluate the model using linear regression, Random Forest and Support vector regression

# Objectives of Project:

* Predicting BOD of river using XG Boost, Bayesian Ridge Regression and Linear Regression
* Judge the performance of the techniques to predict BOD and compare the same.

# Details of Project developed: (Refer fig.1)

* Firstly, imported all required packages and modules into your required format.
* Read data set file using pandas.
* Check duplicate values and remove those rows from the data set.
* Check null values in the dataset and using describe keyword get all required data.
* Perform the exploratory data analysis using seaborn and matplotlib.
* Clean the dataset using sklearn package import label encoder
* Determine the independent and dependent variables in the dataset.
* Using train\_test\_split, split the dataset into training and testing part. train the model 70% of the data and test the model 30% of the data.
* Evaluating the model using linear regression.
* Find the accuracy of the model using error measures as coefficient of determination and root mean square between Observed concrete strength and predicted concrete strength. Accuracy of model is also found through scatter plot.



*Fig.1 Details of Project Developed*

## System Requirement Used:

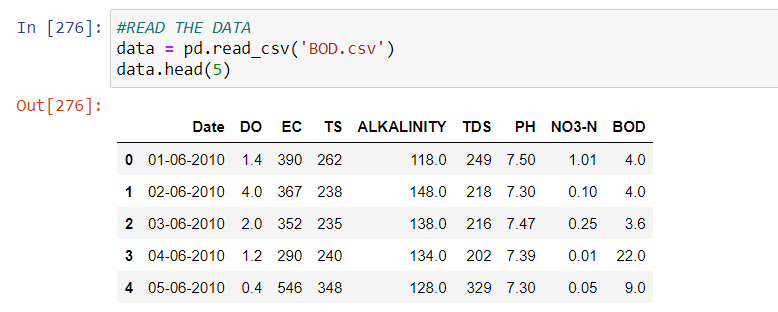
* Windows 10
* Python 3
* Jupyter Notebook

# Input Output Datasets / screenshots

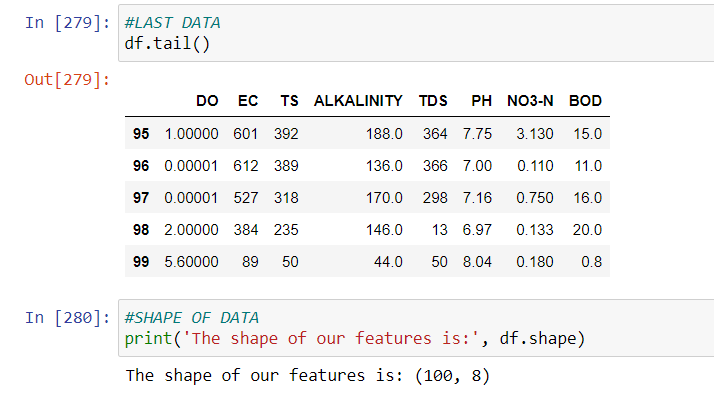
* *Import all required packages*

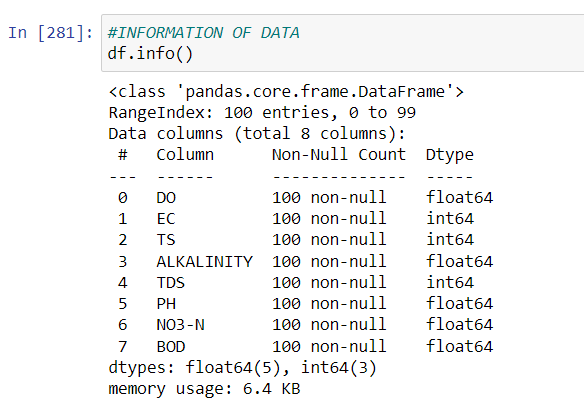
**

* *Uploading the dataset*

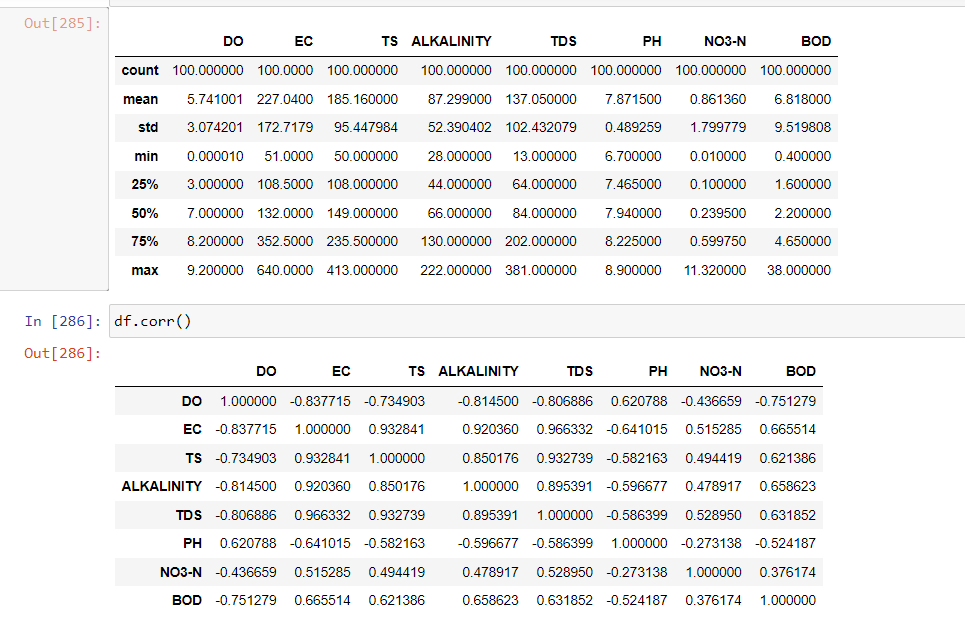
**

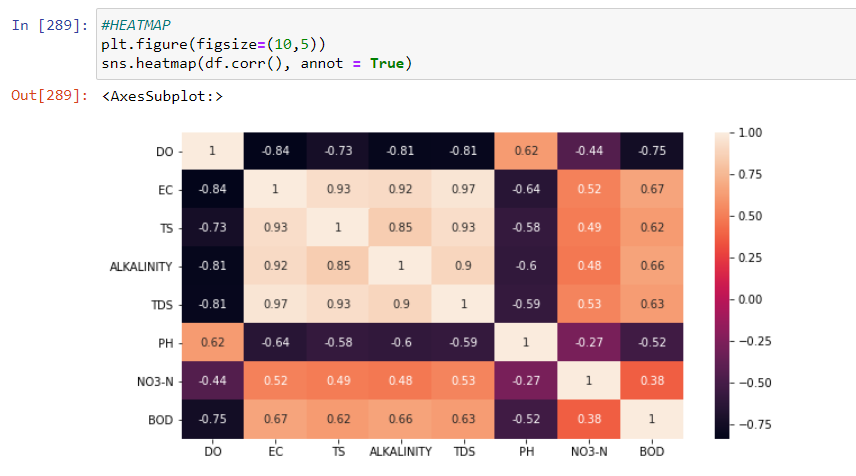
* *Details of Dataset*

**

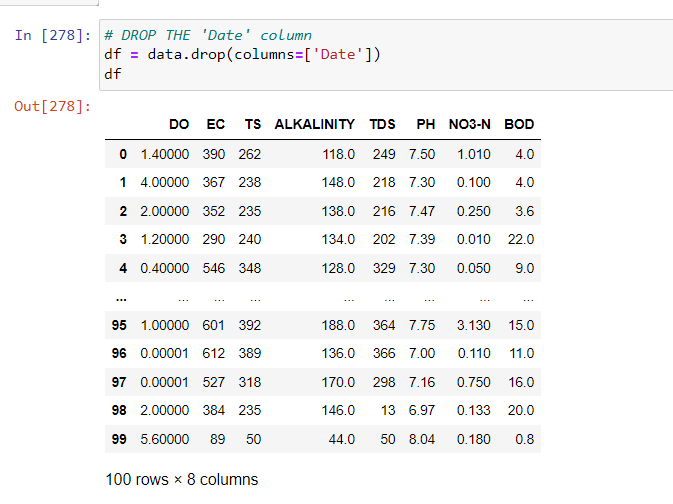


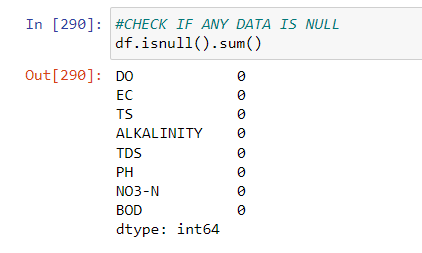
* *Observing the dataset features*

**

**

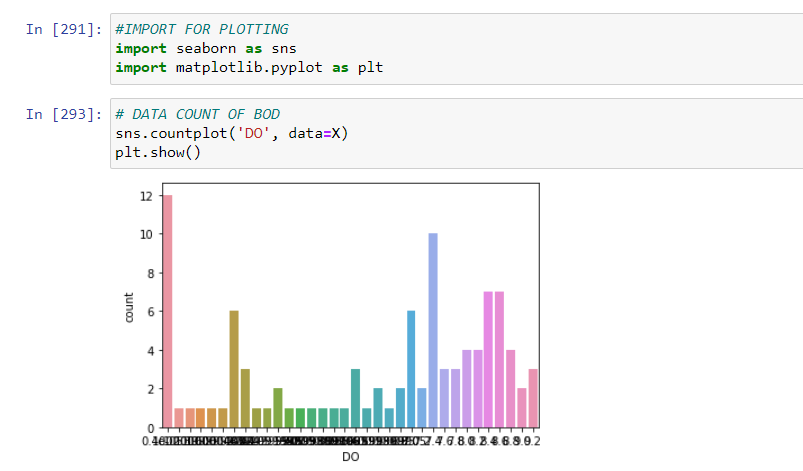
* *Drop any data if not required and check for null*

**

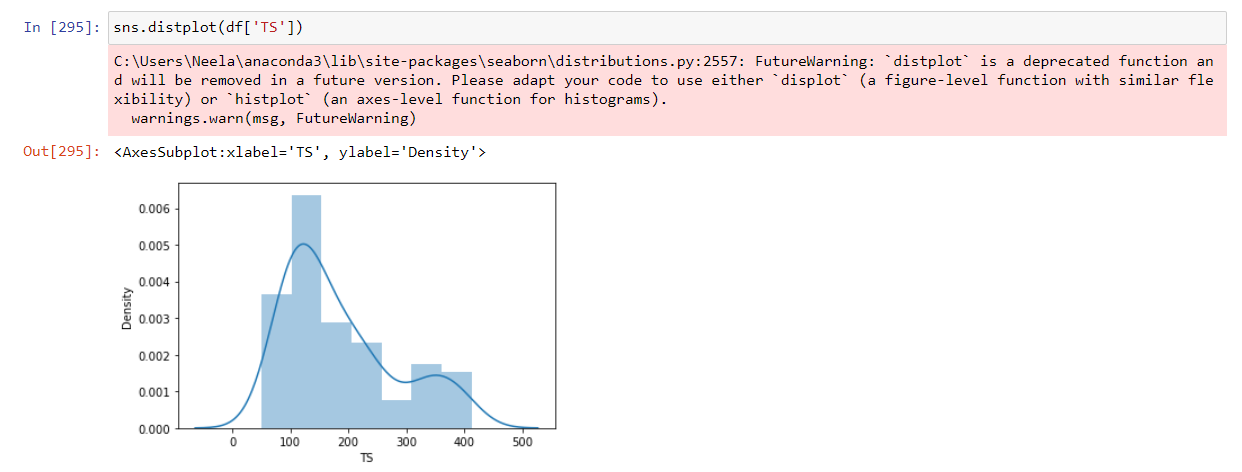
**

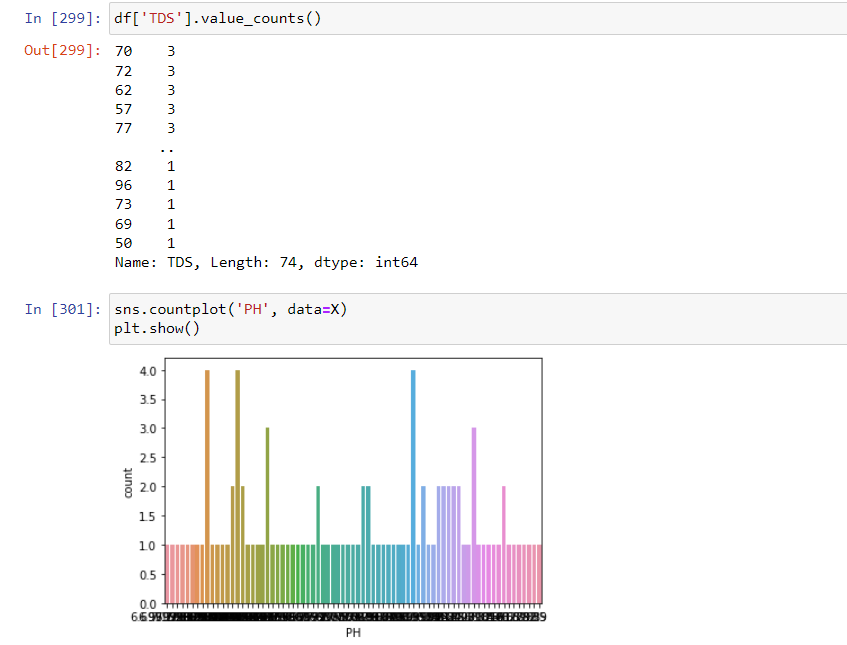
### 

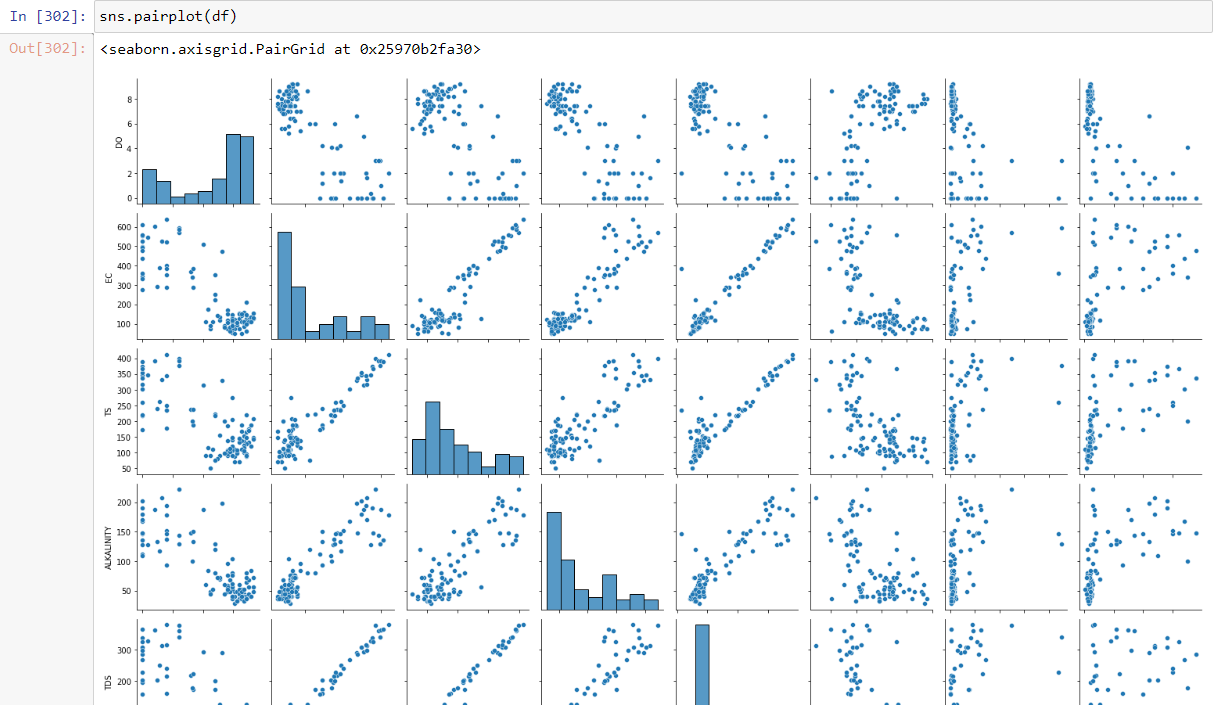
### EXPLORATORY DATA ANALYSIS

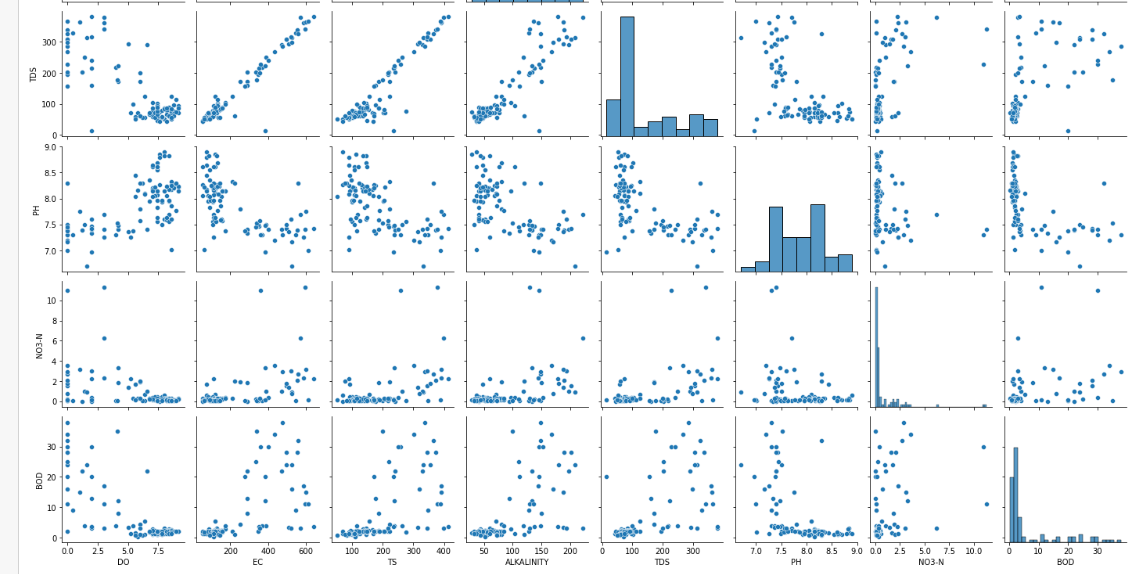
**

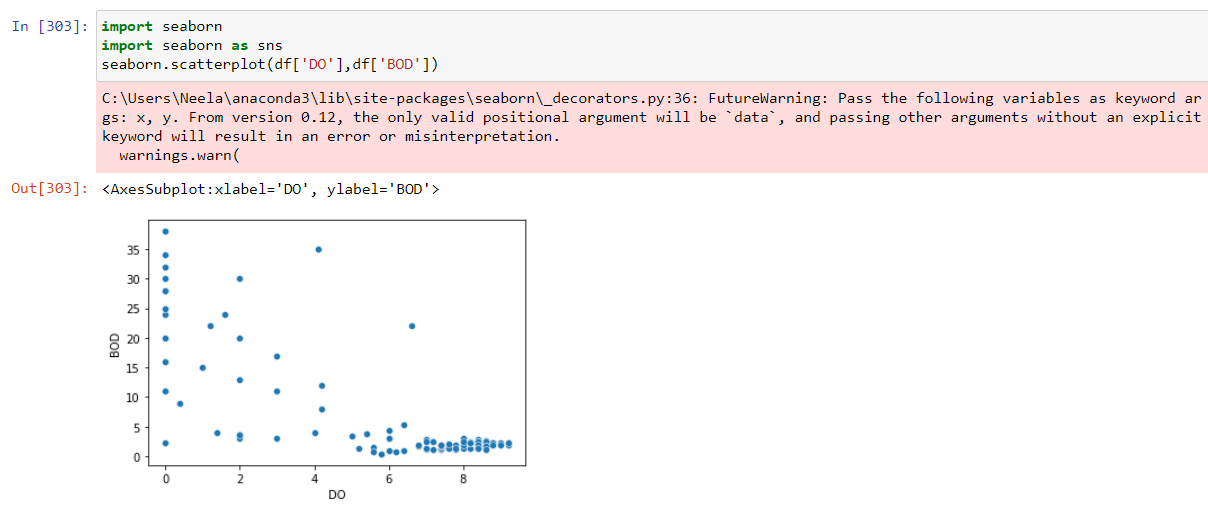
**

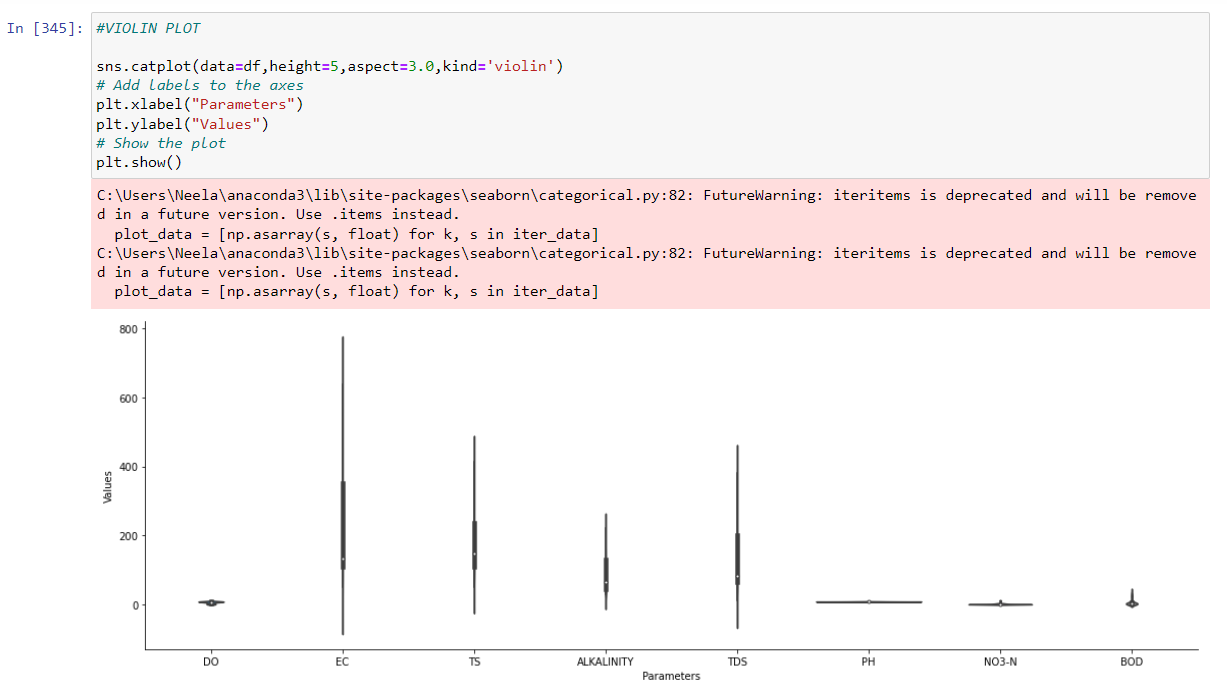
**

**

**

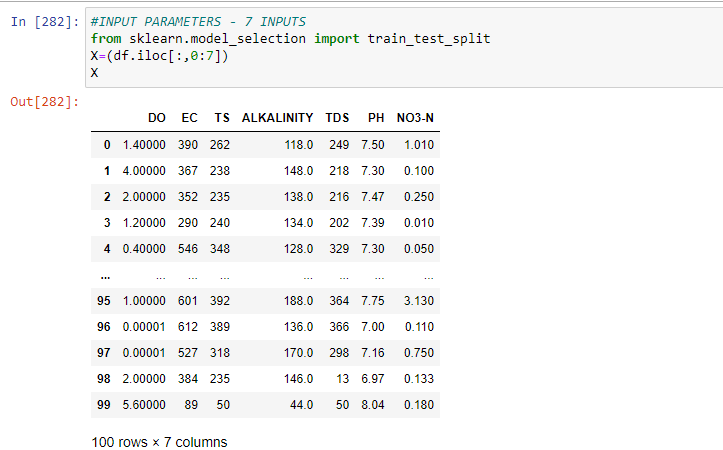
**

**

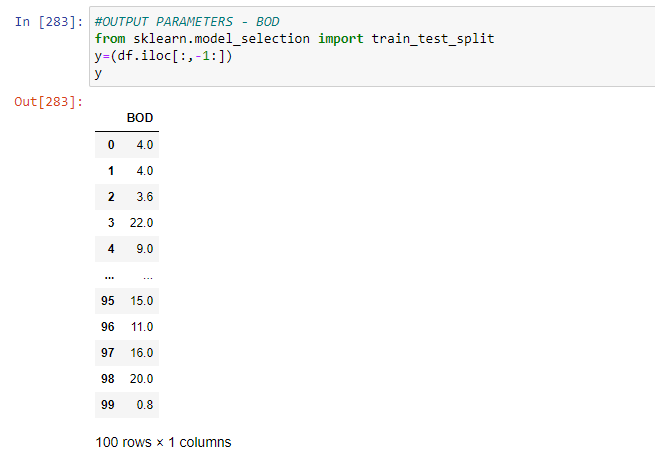
**

***REDICTION MODEL DEVELOPMENT***

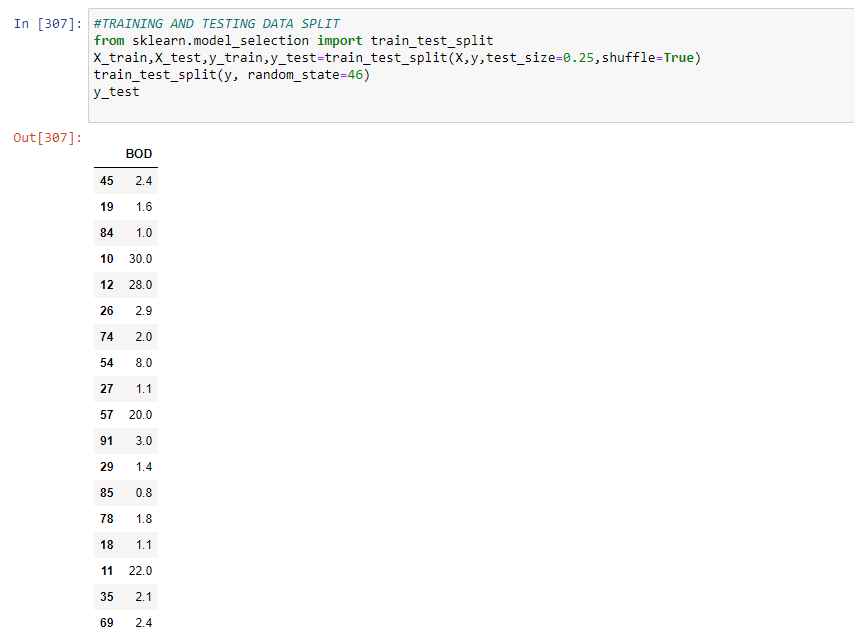
* *Input data*

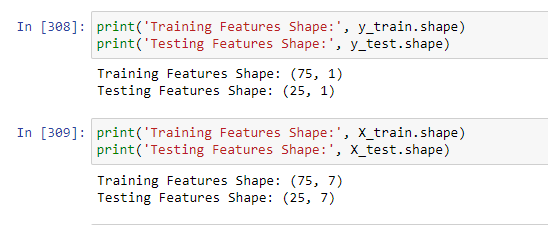
**

* *Output Parameter*

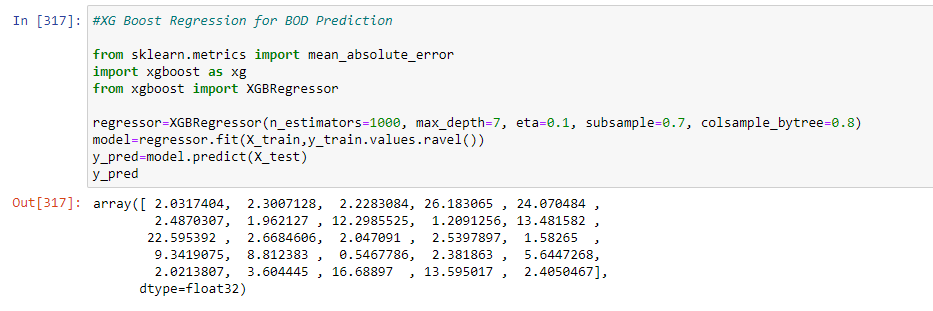
**

* *Split data*

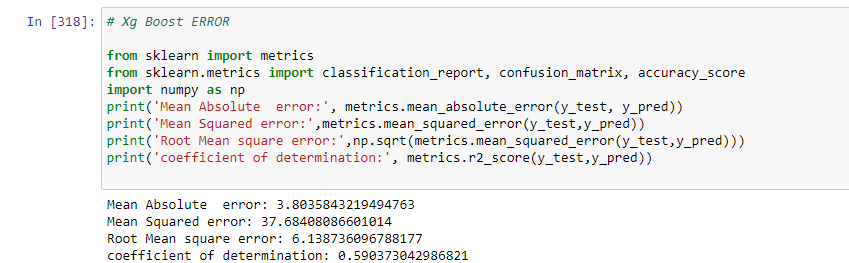
**

**

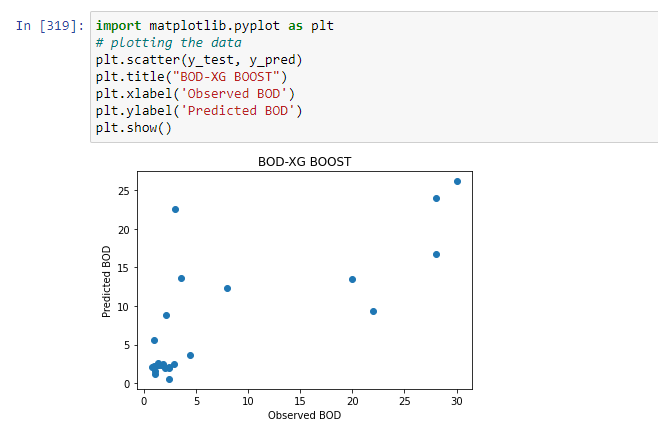
* *BOD Strength prediction using XG Boost Regression*

**

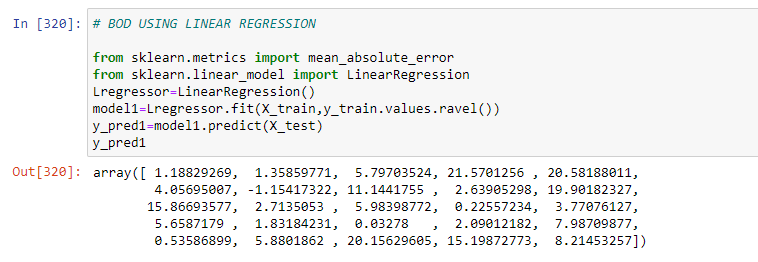
* *Error*

**

* *Scatter plot between Observed and Predicted*

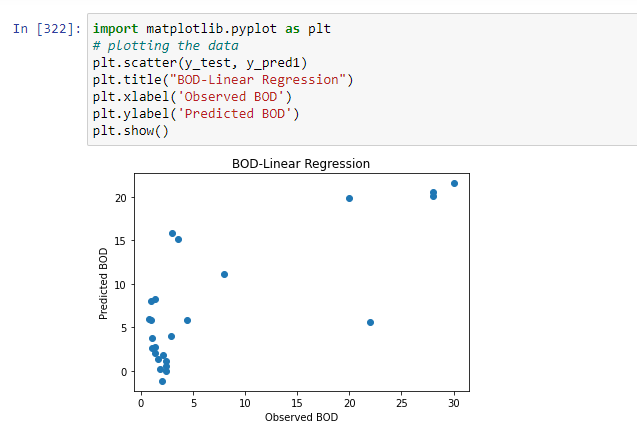
**

* ***BOD using Linear Regression***

**

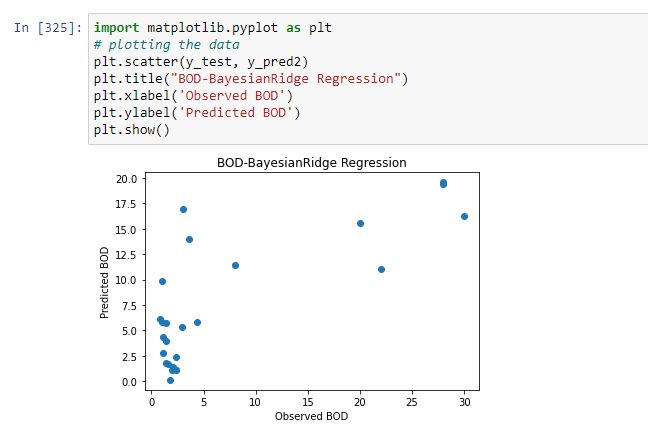
* *Error and Coefficients*

**

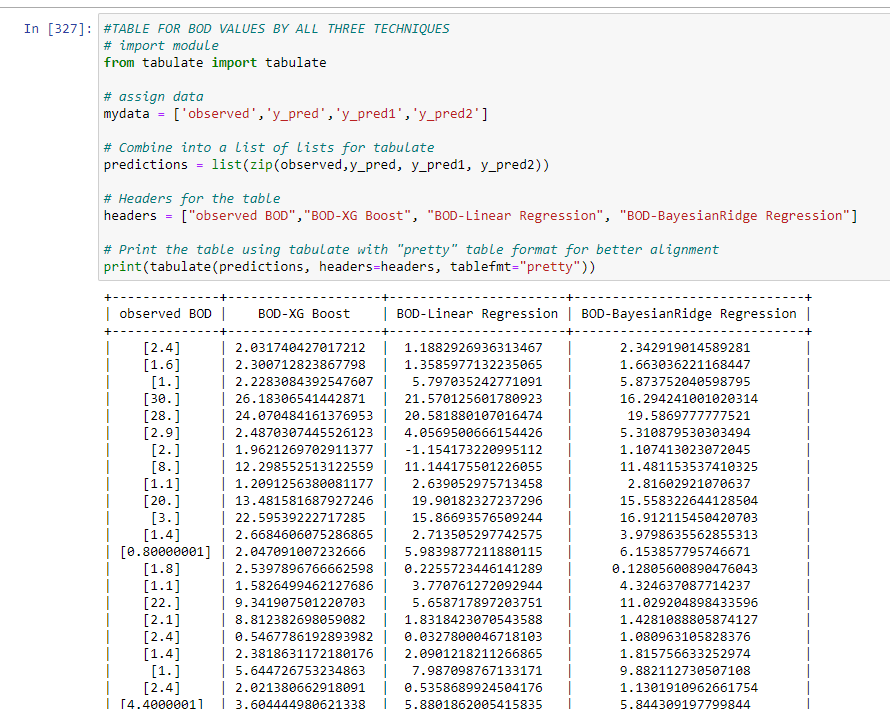


* *BOD Prediction using BayesianRidge Regression*

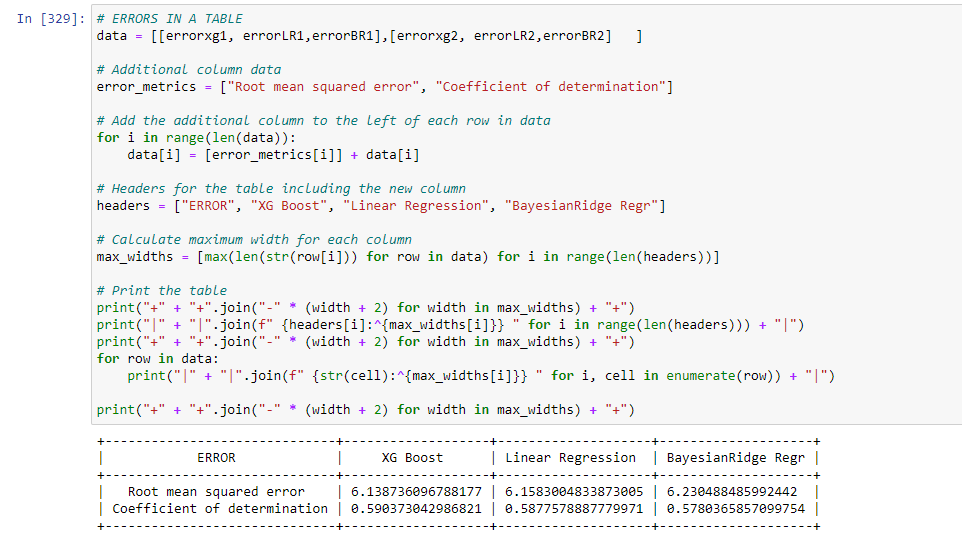
**



* *Predictions in Tabluar format*

**

* *Errors in table*

**

# Conclusion:

The present project’s aim is to predict BOD using Linear regression, XG Boost and BayesianRidge Regression. The exploratory data analysis also shows a higher correlation of BOD with DO and PH as input parameters.

* The error table shows that BOD prediction is done better by all the three techniques is almost same with an upper hand by XG Boost Regression (Coefficient of determination as 0.59) and a lower RMSE of 6.14.
* The Scatter plot with XG Boost shows a better scatter as compared to the same using other techniques.

# References:

* <https://www.w3schools.com/python/pandas/default.asp>
* <https://www.w3schools.com/python/numpy/numpy_random_seaborn.asp>
* <https://www.w3schools.com/python/matplotlib_pyplot.asp>