TABLE OF CONTENTS

Introduction

An accurate predication on the Biochemical oxygen demand is important for what BOD is good for the location or any Particular station. Different station has different BOD and uses this data to find which Station gives accurate result.

2. **Dataset**

Dataset consist historical BOD for river water and different station. The dataset consist of 8 exploratory features and 147 observation. The dataset is extracted from kaggle https://www.kaggle.com/datasets/vbmokin/prediction-bod-in-river-water

The dataset contains BOD for 7 different stations . Some of the features in this dataset are:

1. Id
2. Target
3. Station1
4. Station2
5. Station3
6. Station4
7. Station5
8. Station6
9. Station7.

3. **Data Wrangling or Cleaning**

Data wrangling is very crucial steps for any analysis because it gives a better understanding of the data. Also gives flexibility for further analysis.

Data cleaning steps carried out in this project are:

a) Handling missing data

BOD in river water dataset information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 147 entries, 0 to 146
Data columns (total 9 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Id      147 non-null    int64
 1   target  147 non-null    float64
 2   1       145 non-null    float64
 3   2       145 non-null    float64
 4   3       32 non-null     float64
 5   4       31 non-null     float64
 6   5       33 non-null     float64
 7   6       37 non-null     float64
 8   7       37 non-null     float64
dtypes: float64(8), int64(1)
memory usage: 10.5 KB
```

The Output above is produced from Info () function.

a. Handling missing data:

I find some of the columns have lots of missing values that's why we remove those columns.

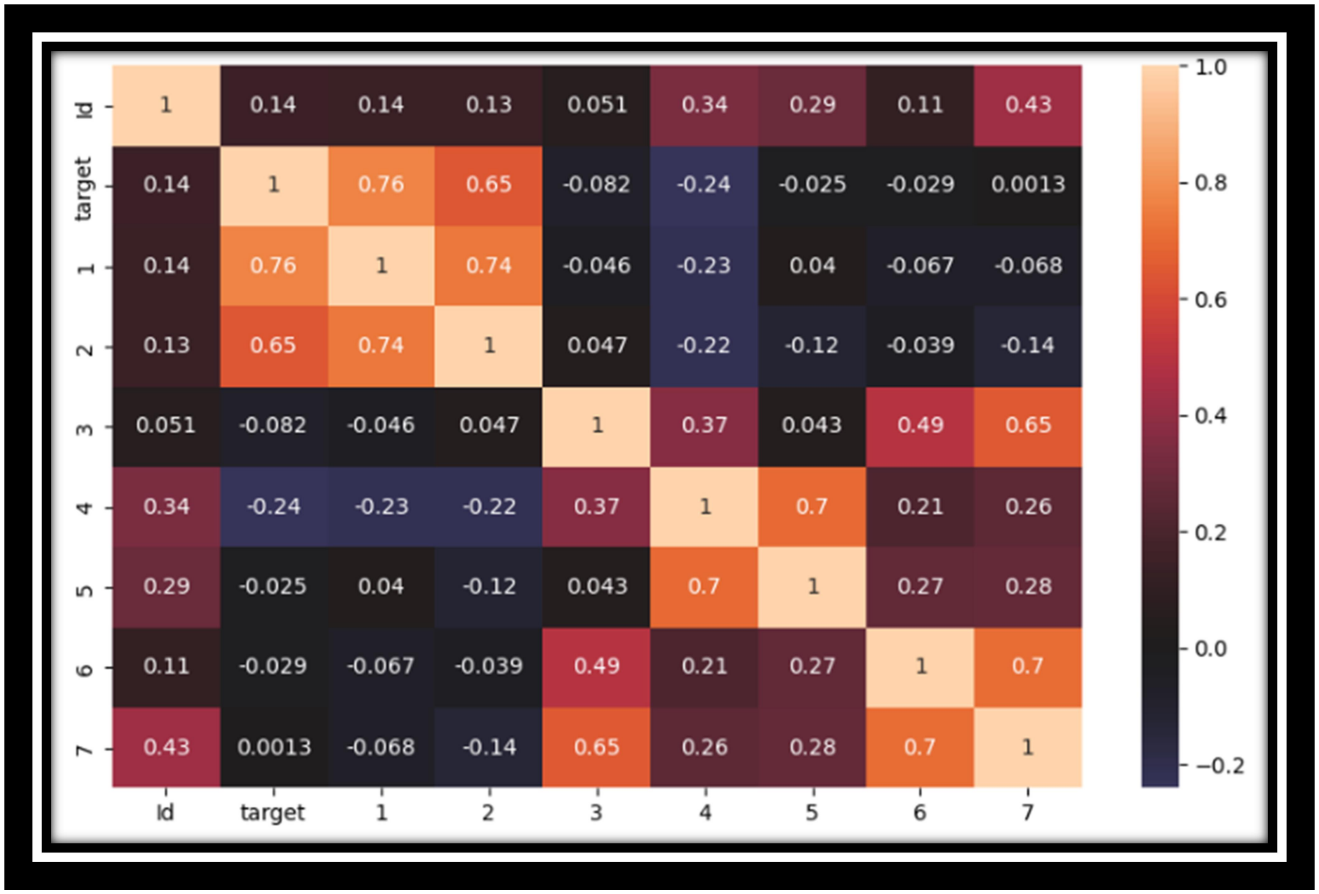
```
In [3]: train_data = train_data.drop(['3','4','5','6','7','Id'],axis=1)
```

4. Data Exploration

Data Exploration is the first step in data analysis and typically involves summarizing the main characteristics. It is commonly conducted using visual analytics tools. Data visualization is the best way to explore the data because it allows users to quickly and simply view most of the relevant features of the dataset.

I used sea-born library provided by python for my visualization.

a) Multicollinearity: I create a heat map for check the Multicollinearity. Heat map gives a power to visualize which features or columns are highly correlated, This might lean to an increase in the variance of the coefficient estimates and make the estimate very sensitive to minor changes in the model.



5. Data Standardization

Before applying any machine learning model. It is extremely important to summarize the data. Data standardization should be perform to make sure that all the **features are on the same scale so that they can be compared for analysing results.** I used function from Scikit-learn library (a very useful machine learning library provided by the python) to standardize the data.

6. Train and Test Sets

Before applying Ml algorithm , It is essential to split the data into train and test sets. So there will be untouched data set to access the performance of the model. I split the data into train (70% of the entire data) and test (30% of the entire data)

X_train – Contains all the predictors of train dataset

Y_train – The target variable in train set

X_test - all predictors in test set

Y_test - target variable in test set.

```
In [10]: X_train, X_test, Y_train, Y_test= train_test_split(train_data,Y, test_size=0.3, random_state=0)

In [11]: print('X_train: ',X_train.shape)
         print('X_test: ',X_test.shape)
         print('Y_train: ',Y_train.shape)
         print('Y_test: ',Y_test.shape)

         X_train:  (100, 2)
         X_test:  (44, 2)
         Y_train:  (100,)
         Y_test:  (44,)
```

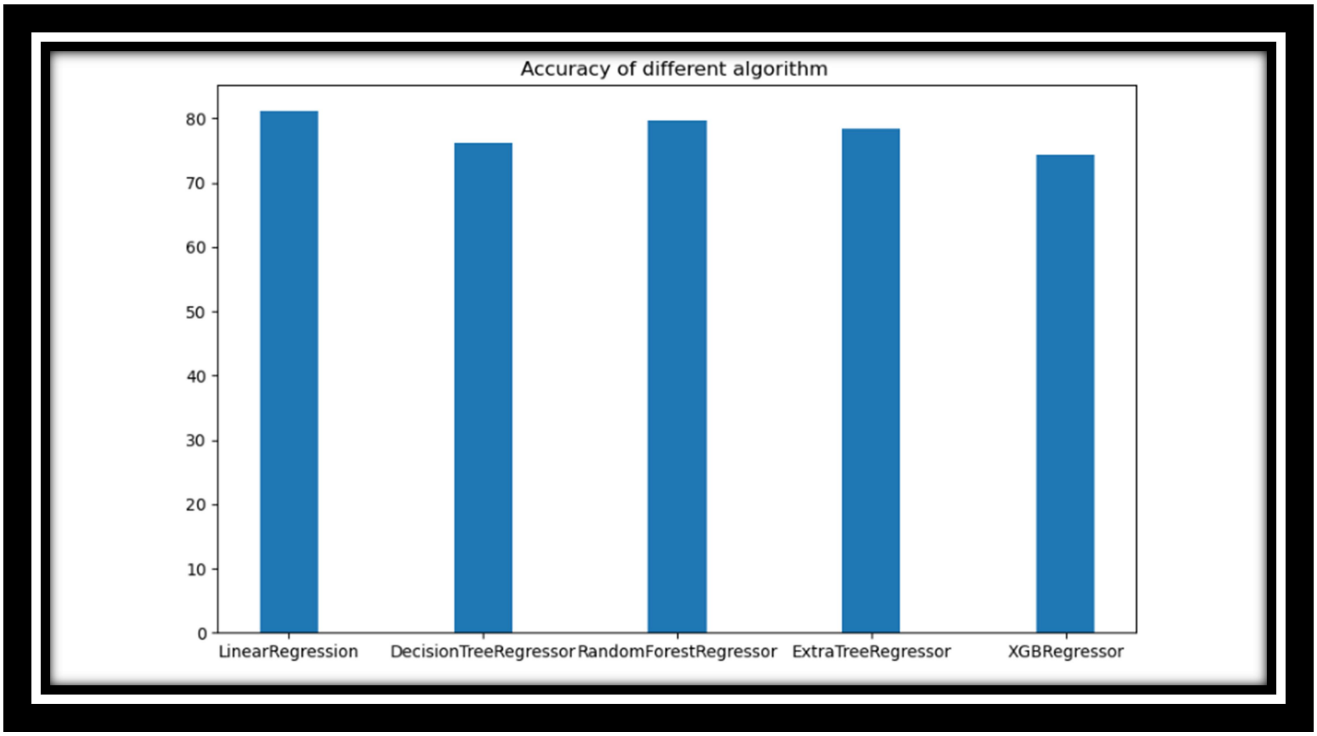
Notice that train data set is a matrix with all predicators and test data is a vector with only target variable.

7. <u>Machine Learning</u>

I used 5 algorithm to predict the better accuracy .

I. Linear Regression
II. Decision Tree Regressor
III. Random Forest Regressor
IV. Extra Tree Regressor
V. XG Boost Regressor

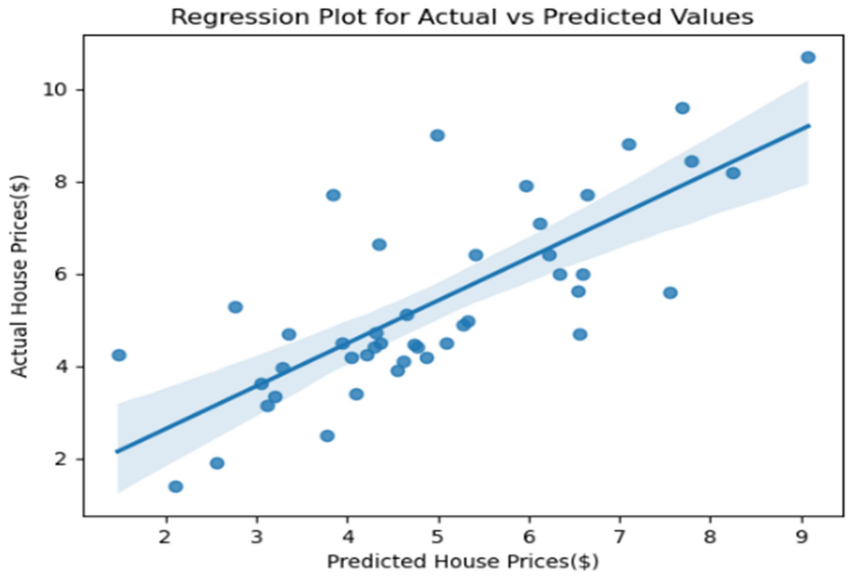
Here I show one graph that shows which model gives a highest accuracy.



8. Actual vs Predicted

For Linear Regression here I show a graph for understanding the variation of values;

```
Text(0.5, 1.0, 'Regression Plot for Actual vs Predicted Values')
```



9. Summary

Here we see some of the steps of data analysis that help to predict the better result. Linear Regression gives the highest accuracy in all of the models. Here we see station 1 and 2 are much closer to target BOD in the river .