Biological Oxygen Demand (BOD) in river water

Name : Kalpana Rathod (Bach:DS22DES23)

2/2023

Table of Contents

1. [Introduction](#_TOC_250012) 3
2. [Client](#_TOC_250011) 3
3. [Dataset](#_TOC_250010) 3
4. [Data Wrangling](#_TOC_250009) 4
5. [Handling missing data](#_TOC_250008) **4**
6. [Handling inconsistent data](#_TOC_250007) **4**
7. [New Dataset](#_TOC_250006) 5

# Data Exploration………………………………………………………………………………...5

1. **Figures 6**
2. [Data Standardization 12](#_TOC_250003)
3. [Train and Test Sets 1](#_TOC_250001)2
4. Machine Learning 13
   1. [Regression **13**](#_TOC_250000)
5. Cross Validation 13
6. Scores 14
7. Plot model figure 15
8. Future Scope 16
9. Conclusion 16

# Introduction

# Build a model to predict biological oxygen demand in river water. Use cross-validation to determine model performance. Among the various modeling techniques applied to dataset, multiple linear regression (MLR) analysis is the most efficient way to figure out the relationship between the response variable and the predictive variables. This study emphasizes on establishment of multiple linear regression models to analyze Biochemical Oxygen Demand (BOD) removal efficiency for technologies, namely Dens deck, Extended Aeration and Activated Sludge Process. Assumptions of multiple linear regression like linear relationship, multivariate normality, multicollinearity and Homoscedasticity were examined.

# The data that verify the assumptions were analyzed with multiple linear regression. Time series plots indicate drastic decline in BOD removal efficiency in the month of Feb and March during the years 2012 and 2013. This study was significant as it gives the technology having the best-fit regression equation based upon multiple correlation coefficient (*R*), coefficient of determination (*R*2), standard error, residual and *F*-ratio value. Societal benefits include enhancement in the performance of sewage treatment plants.

# This dataset has data of the amount of biochemical oxygen demand, which is determined in 5 days ("BOD5" or "BOD"), in river water.

# Client

This analysis report can be an interest to any Real estate company, Real estate investors, Mortgage lenders and Home insurers. This report helps make decisions easy for the businesses and home seekers.

# Dataset

Dataset consists of train of **id and Target**. The dataset consists of 147 exploratory features with observations. The dataset is extracted from Kaggle <https://www.kaggle.com/datasets/vbmokin/prediction-bod-in-river-water>

The data set contains every minute detail of the house. Some of the major features in this data set are:

1. Id
2. Target
3. 2
4. 3
5. 4
6. 5
7. 6
8. 7

However, it is good idea to explore the data set from Kaggle to get good idea on the data

.

# Data Wrangling

Data Wrangling is an extremely I important step for any data analysis. It is very crucial for data to be organized. This process typically includes manually converting/mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

Data cleaning steps carried out in this project are:

1. Handling missing data
2. Handling inconsistent data in a few variables House Prices data set information:
3. <class 'pandas.core.frame.DataFrame'>
4. RangeIndex: 147 entries, 0 to 146
5. Data columns (total 9 columns):
6. # Column Non-Null Count Dtype
7. --- ------ -------------- -----
8. 0 Id 147 non-null int64
9. 1 target 147 non-null float64
10. 2 1 145 non-null float64
11. 3 2 145 non-null float64
12. 4 3 32 non-null float64
13. 5 4 31 non-null float64
14. 6 5 33 non-null float64
15. 7 6 37 non-null float64
16. 8 7 37 non-null float64
17. dtypes: float64(8), int64(1)
18. memory usage: 10.5 KB

The output above is produced from **info()** function. There are a few categorical and numerical variables with missing values

### Handling Missing Data:

* + **Categorical Data:** The categorical variables with missing values are ‘id’ and ‘taret’. Python provides many methods like fillna, forward/ backward filling, dropna etc. for handling missing data. I introduced another category called ‘**missing**’ to all the null values. This way I am retaining the original information of the data and not guessing anything.
  + **Numerical Data:** The most popular method to handle missing numerical data is **Mean Imputation**. I applied the same on my numerical data. Mean imputation is a method in which the missing value on a certain variable is replaced by the mean of the available cases. This is a reliable method for handling missing numerical data.

.

# New Data Set

The data is now clean without any null/ inconsistent values. I transferred this data into a new csv file ‘**train.csv’**. I will use this data set for data exploration.

# Data Exploration

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a dataset. It is commonly conducted using visual analytics tools. Data Visualization is best way to explore the data because it allows users to quickly and simply view most of the relevant features of the dataset. By displaying data graphically scatter plots/ bar charts to name a few – users can identify variables that are likely to have interesting observations and if they are helpful for further in-depth analysis.

I used seaborn library provided by Python for my visualizations. I divided the data frame into numerical and categorical – containing quantitative and qualitative data respectively for the ease of analysis.

1. **Multicollinearity:** Multicollinearity exists when two or more of the predictors highly correlated, this might lead to an increase in the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. I used Heat map to find out highly correlated independent variables. From the graph, we can see that features like:

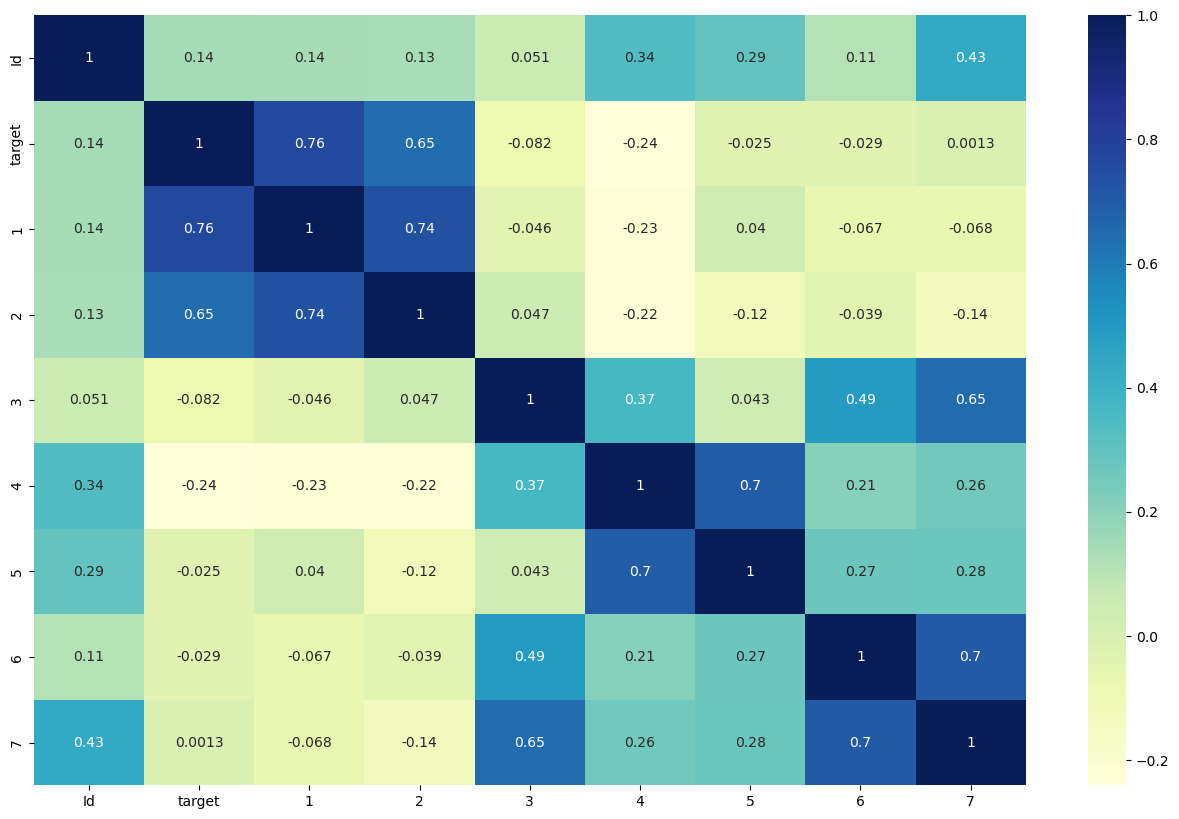
* 'id' and 'target',
* 'Total Basement square footage' and '1st floor square footage',
* 'Above grade (ground) area' and 'Total no. of rooms above grade(ground) are highly correlated with each other.

### Handling inconsistent data:

There are a few null values in the data set which are not actually nulls but are entered wrongly as nulls. Referring to the actual data set description file (data\_description.txt) from Kaggle, a few values were coded as ‘NA’ if a feature was not present in the house, but these NA values were entered as Nan in the .csv file. I decoded these misinterpreted values as ‘No feature\_name’ (feature\_name being name of the feature not present in the house).

1. **FIGURE**

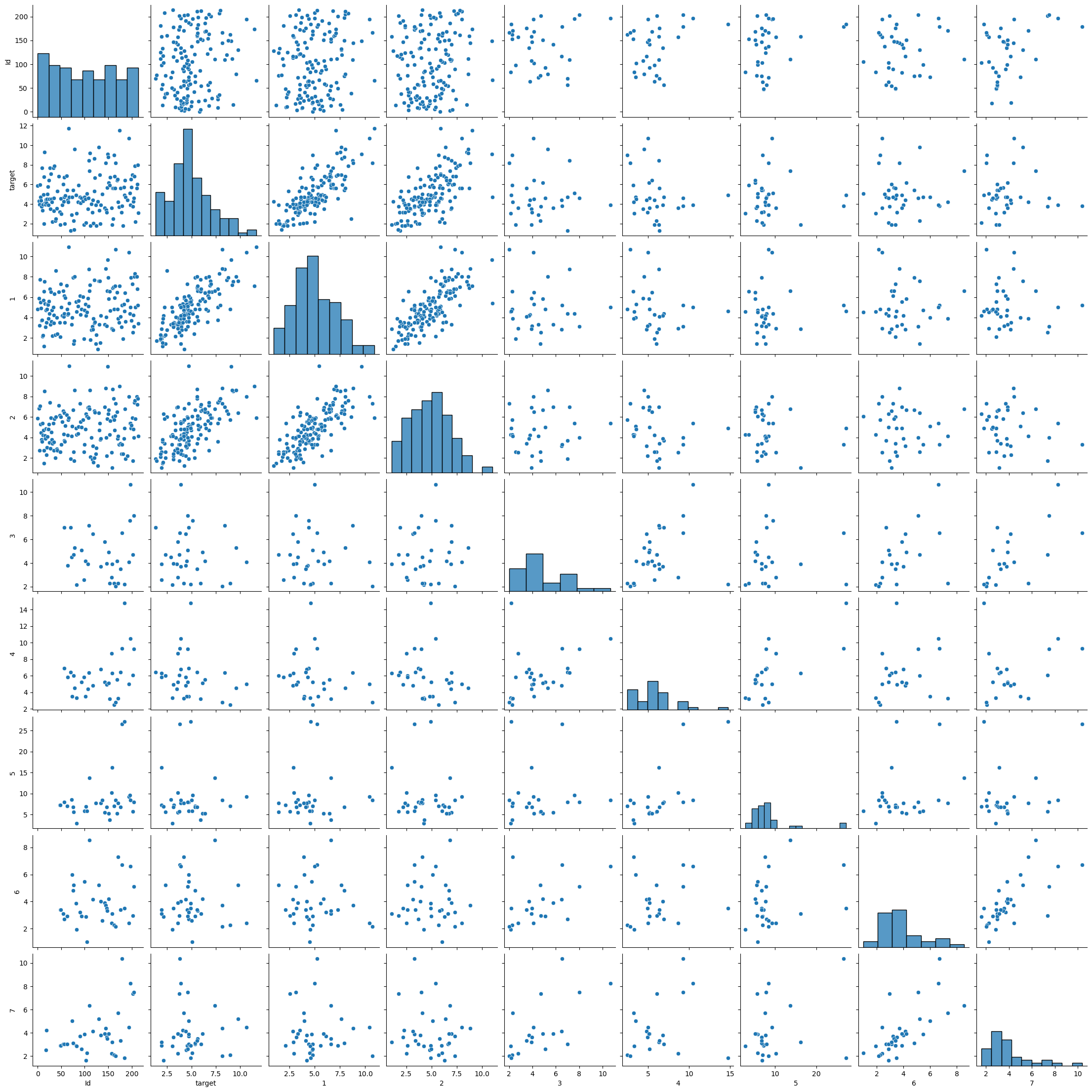
The issue with Multicollinearity can be addressed through Machine Learning algorithms such as Ridge and Lasso Regression.Other than that, the highly correlated independent variables with the target variable id, target Price are Overall Quality, Above Ground Living area and Garage cars



#### Figure 7.1

.

**Plot Pairplot**



#### Figure 7.2

**Count plot for the dataset**

#### 

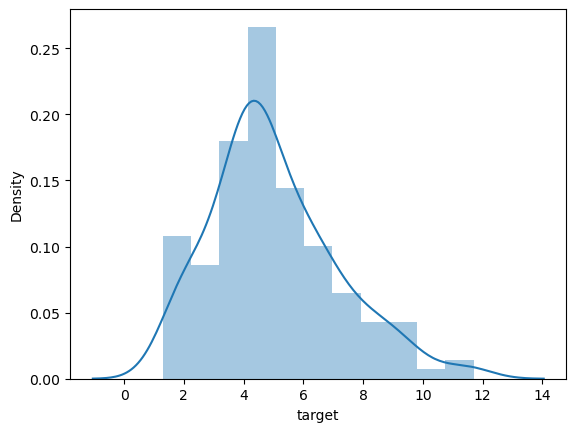
. **Figure 7.3**

### Plot hisplot

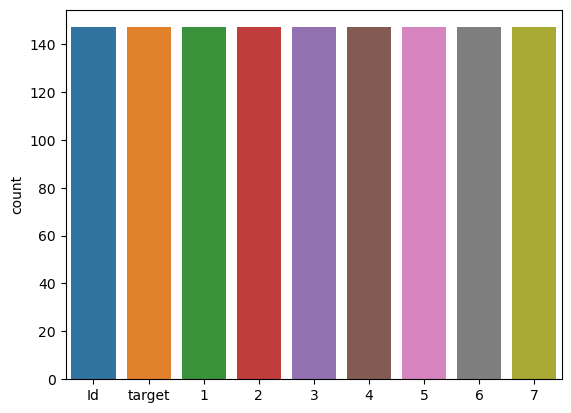
### 

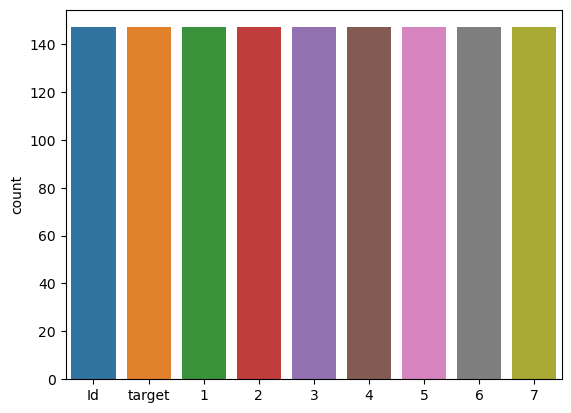
***Figure 7.4***

### Distplot



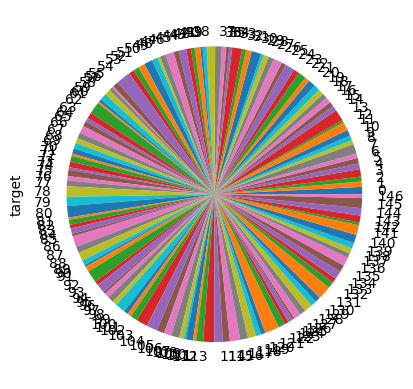
***Figure 7.5***

******



***Figure 7.6***

**Plot Pie chart**



***Figure 7.7***

.

# Data Standardization

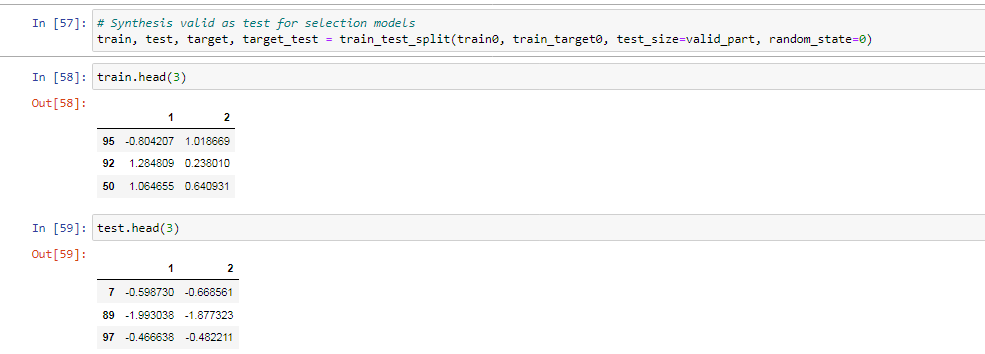
Before applying any Machine Learning Algorithms, it is extremely important to standardize the data. Data Standardization should be performed to make sure that all the **features are on the same scale so that they can be compared for analyzing results**. Data Standardization (or Z-score normalization) is the process where the features are rescaled so that they’ll have the properties of a standard normal distribution with μ=0 and σ=1, where μ is the mean (average) and σ is the standard deviation from the mean. I used functions from Scikit-learn library (a very useful Machine Learning library provided by Python) to standardize the data.

**9**. **Train and Test Sets**

Before applying ML algorithm, it is essential to split the data into train and test sets, so that there will be an untouched data set to assess the performance of the model. I split data the into train (70% of the entire data) and test (30% of the entire data).

Train – contains all the predictors of train data set target – the target variable in train set

test – all predictors in test set target\_test – target variable in test set Note: Target Variable – ‘target’

. 

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

# 10. Machine Learning

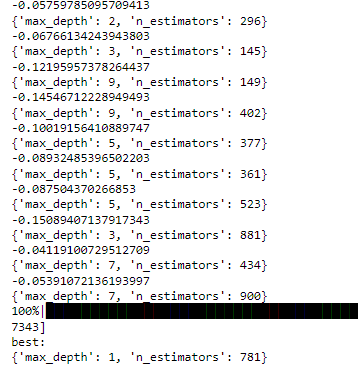
## a. Regression:

* I performed Multiple Linear Regression first and then moved to more advanced algorithms. Regression plot plotted between the actual and predicted prices produced a good fit of a line for the data. Models are applied in dataset below:
* Linear Regression
* Support Vector Machines and Linear SVR
* Stochastic Gradient Descent, GradientBoostingRegressor, RidgeCV, BaggingRegressor
* Decision Tree Regression, Random Forest, XGBRegressor, LGBM ,ExtraTreesRegressor
* MLPRegressor (Deep Learning)
* VotingRegressor

#### 

#### 11. Cross Validation

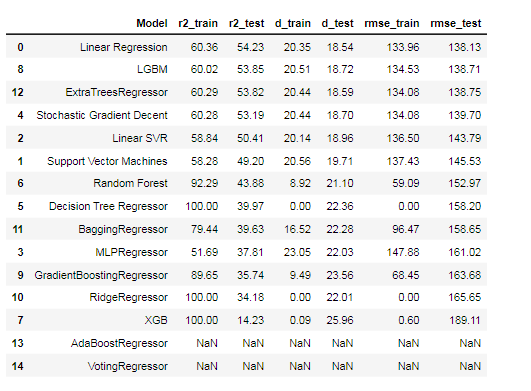
* When evaluating different hyper parameters for estimators, such as the alpha is this setting that must be manually set for an Ridge, there is still a risk of over fitting on the test set because the parameters can be tweaked until the estimator performs optimally. To solve this problem, yet another part of the dataset can be held out as a so-called “validation set”: training proceeds on the training set, after which evaluation is done on the validation set, and when the experiment seems to be successful, final evaluation can be done on the test set. GridSearchCV is used in Scikit-learn library to achieve this task.

****

Comparing the train and test scores (R2 and RMSE), Ridge regression with Cross Validation (Ridge\_GridSearchCV) seems to best suited for the data, because there is not much difference between the scores of train and test data sets. The regression and residual plots from Ridge regression using Cross Validation also seem to be a good fit for data.

# 12. Scores

With these many models applied on data how can we conclude the best model for data. For this I compared the R2 and RMSE scores produced by all the models.



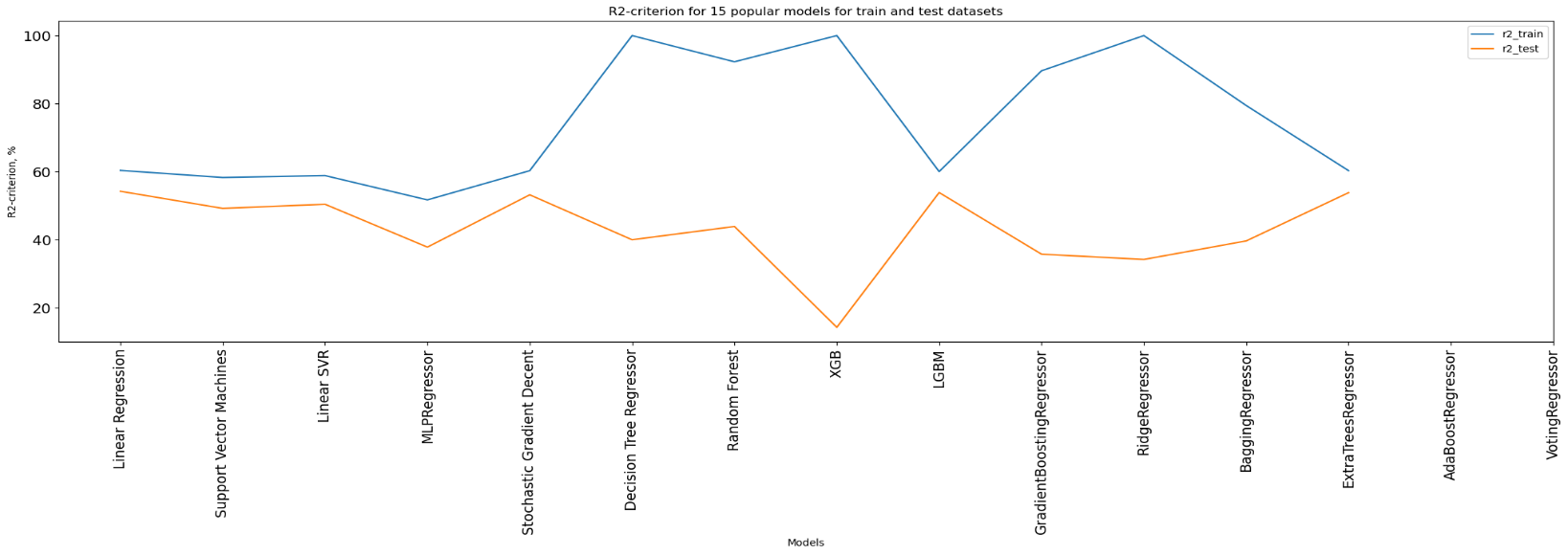


Fig13.1 : The 'RMSE for 15 popular models for train and test

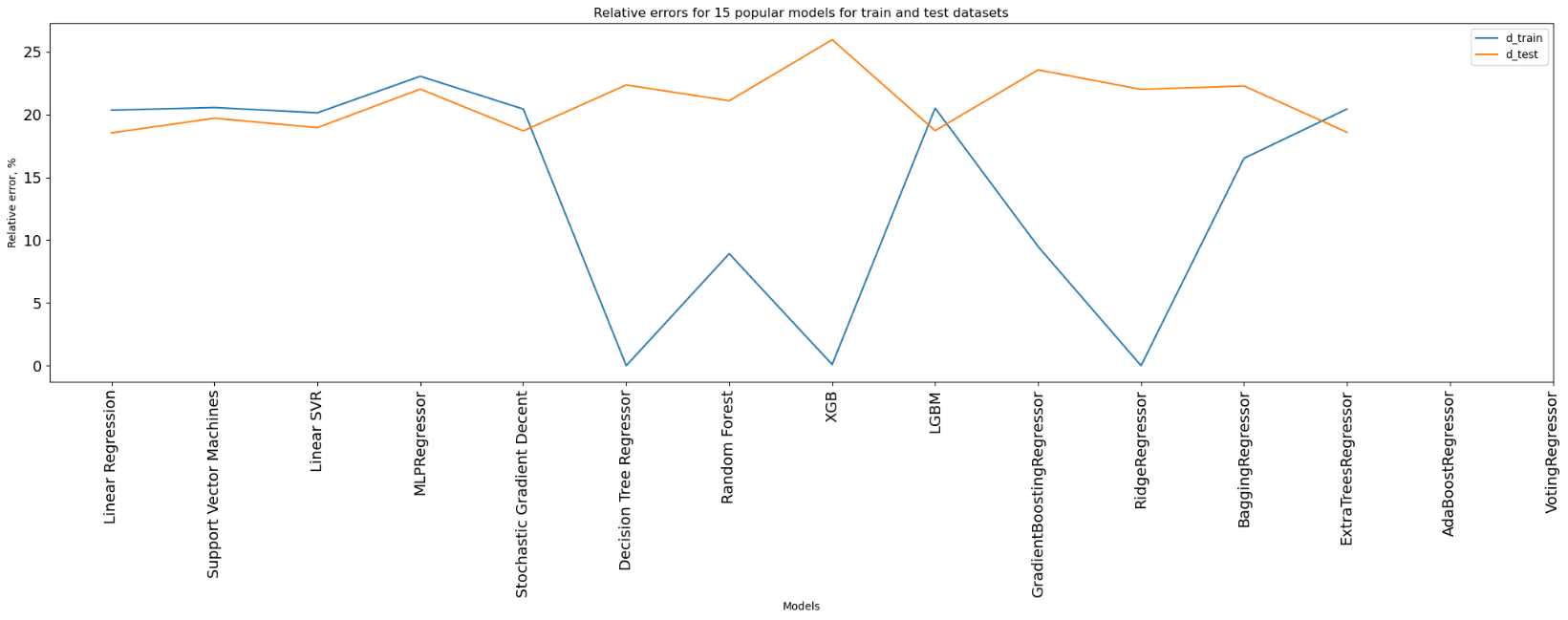


Fig13.2: The 'RMSE for 15 popular models for train and test datasets

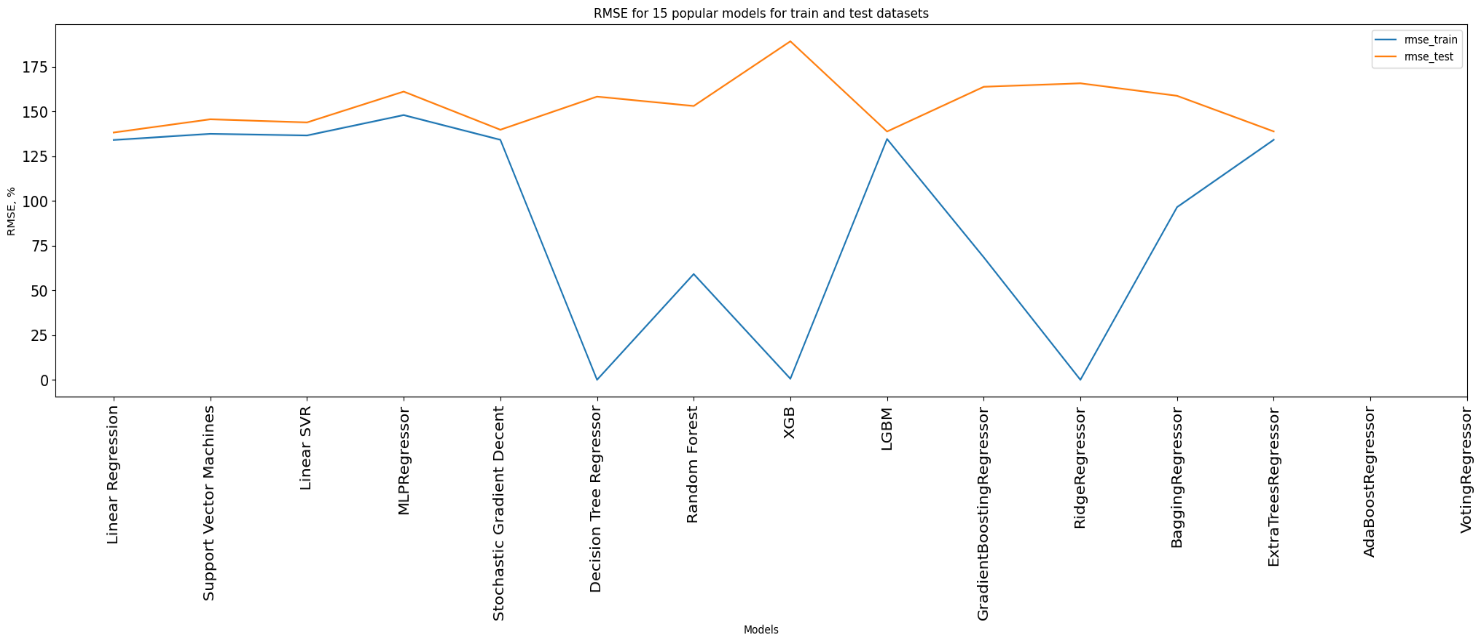


Fig13.3: 'RMSE for 15 popular models for train and test datasets

.

**13. Future Scope**

This research lays a foundation for further research in the prediction of dependent physical quantities that requires sensors. In addition to that, by predicting BOD and Ph of a water sample an intuition of the quality of that water sample can be achieved which will reduce the cost and efforts and hardware required to determine the quality of drinking water. This paper also proposed a way to predict the measurement of BOD hence saving the time and effort done to measure the amount of BOD using traditional methods.

**14. Conclusion**

* In this study, the usability of the regression analysis methods without measuring the daily BOD value was investigated. The results of the all regression models were compared with each other. The RMSE and MAE values were used to compare the performances of the methods. These two indices were used to determine errors and similarities according to the observed values. The MARS method applied in the testing set achieved improvements between 4% and 39% compared to the other models. As a result of the comparisons, it is understood that the MARS method can be used for the prediction of BOD. By means of the obtained model equation, BOD value can be reached with instantly measured parameters, without waiting for BOD test analysis results. Approximate BOD value can be obtained from this model without spending time and consuming material. It was observed that the established model of MARS method had estimates consistent with the measured values which showed the applicability of the model by giving close answers to the instantaneous changes. In addition to the methods used in this study, different meta-heuristic methods can be used and compared with the MARS method. Because the analysis of BOD is difficult, more successful results can be obtained through different models.