

**MUMBAI REFINERY**

**SUMMER INTERNSHIP 2018**

**PREDICTING THE TOTAL STEAM USING MACHINE LEARNING**

**Submitted by**

**Daksha Singhal**

**Submitted to**

**Bharat Petroleum Corporation Limited, Mahul**

**Computer System and Services Department (CSS Dept.)**

CERTIFICATE

This is to certify that Daksha Singhal of Pes University, Bangalore has successfully completed his internship in Learning Center Bharat Petroleum Corporation Limited, Mumbai and has submitted his report on the topic “Predicting Total Steam Using Machine Learning”. He has worked on it during his internship period. The report submitted has been found to be satisfactory and as per the guidelines set by our company.

**Mr. VIMAL K. GULATI Mrs. DURGESH BHUTADA**

**CHIEF MANAGER MENTOR**

**(Training and Development) CHIEF MANAGER IS (R)**

**Computer System & Services**

Acknowledgement

We take great pleasure in submitting this project report carried out in the month of June-July 2018 at Bharat Petroleum and Corporation Limited, Mahul.

I am highly indebted to **Mr. Vimal K. Gulati** (Chief Manager) for allowing us to pursue our Summer Internship through their Technical Cell in this Organization.

I would like to express profound sense of gratitude to thank **Mrs.**

**Durgesh Bhutada** for her supervision and being a constant source of inspiration. I personally thank her for taking time off her busy schedule to make sure I get maximum from my Internship and sharing with me her knowledge and experience to the fullest.

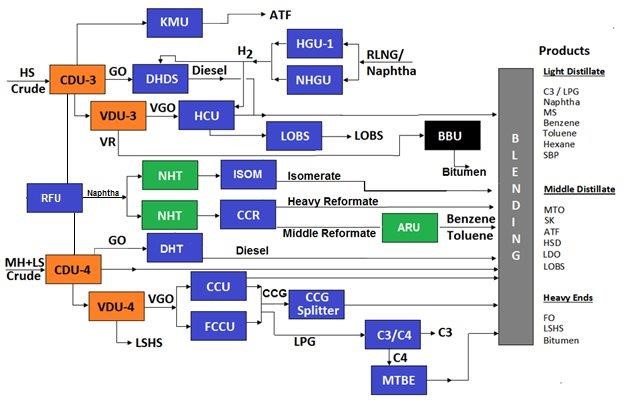
INDEX

1. About BPCL
2. BPCL Mumbai Refinery
3. Problem Statement
4. Project Requirement Specifics
5. Machine Learning
6. Dataset
7. Data Analysis
8. Data Cleaning
9. Data Preprocessing
10. Splitting The Data
11. Problem Approach
12. Target Prediction
13. Model Prediction
14. Shortcomings
15. Conclusion

About BPCL

* **Bharat Petroleum Corporation Limited** (**BPCL**) is an [Government of India](https://en.wikipedia.org/wiki/Government_of_India) controlled Maharashtra [oil](https://en.wikipedia.org/wiki/Oil) and [gas](https://en.wikipedia.org/wiki/Gas) company headquartered in Maharashtra. The Corporation operates two large refineries of the country located at [Mumbai](https://en.wikipedia.org/wiki/Mumbai) and Kochi with a capacity of 12 and 9.5 Million metric Tonnes per year.
* The company is ranked 358th on the [Fortune](https://en.wikipedia.org/wiki/Fortune_Global_500) list of the world's biggest corporations as of 2016.
* It is engaged in offering motor spirit (MS), high speed diesel (HSD) and liquefied petroleum gas (LPG). The Company is engaged in the business of refining of crude oil and marketing of petroleum products.
* It operates through two segments: Downstream petroleum, and Exploration and Production of Hydrocarbons (E&P). The Downstream petroleum segment includes refining and marketing of petroleum products.
* It is engaged in the production of liquid and gaseous fuels, illuminating oils, lubricating oils or greases or other products from crude petroleum or bituminous minerals.
* The Company also manufactures other petroleum products, including petroleum bitumen and other residues of petroleum oils or of oils obtained from bituminous minerals.
* Its marketing infrastructure includes network of installations, depots, retail outlets, aviation service stations and LPG distributors.

BPCL, MUMBAI REFINERY



PROBLEM STATEMENT

To predict the Total Steam of CDU-3 from historical data given by the department.

The prediction will depend on various parameters given in the dataset using machine learning.

PROJECT REQUIREMENT SPECIFICS

* Device:
* Processor: Intel Core i7
* Ram: 4 Gb
* OS: Windows 10
* Language: Python 3.6.1
* Editor used: Jupyter
* Compiler: Anaconda 3

MACHINE LEARNING

Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" with data, without being explicitly programmed.

SUPERVISED UNSUPERVISED

MACHINE LEARNING MACHINE LEARNING

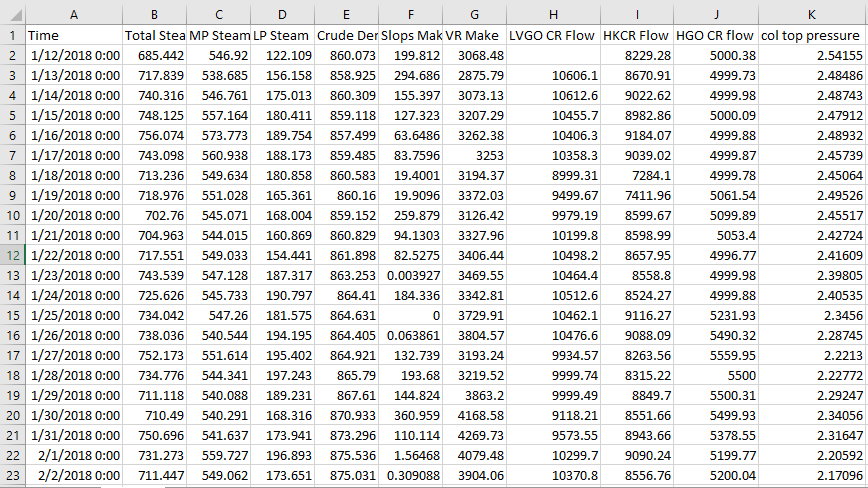
* SUPERVISED MACHINE LEARNING

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

* UNSUPERVISED MACHINE LEARNING

Unsupervised machine learning is the machine learning task of inferring a function that describes the structure of "unlabeled" data.

DATASET

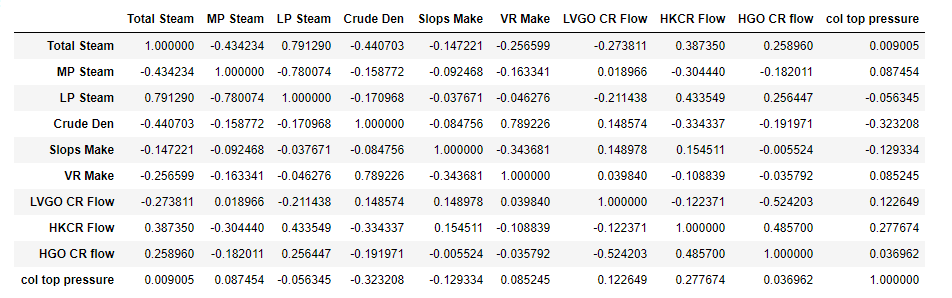


CDU-3 Steam Data 1

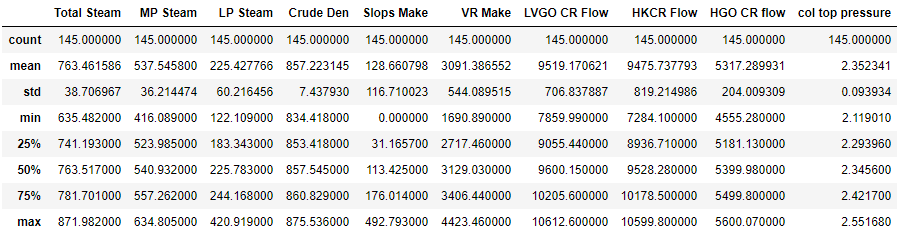
Source: computer center, BPCL, Mumbai refinery

DATA ANALYSIS

* Machine learning algorithms learns from data and it is critical that the right data is fed to the problem to be solved.
* Library: Pandas
* Functions: .corr() , .describe()
* DETERMINE CO-RELATION



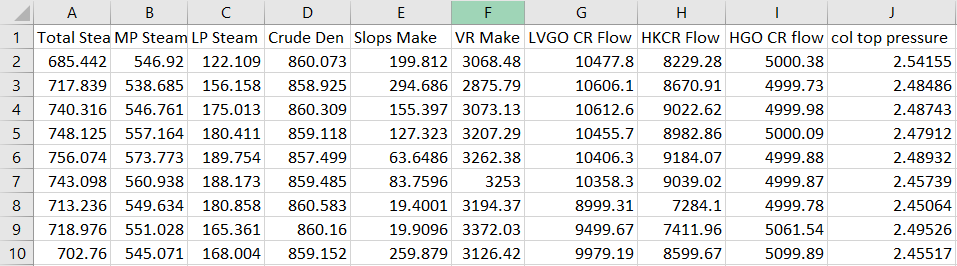
***Helps to determine how different features affect each other***

* STASISTICAL ANALYSIS

DATA CLEANING

* FEATURE SELECTION

Taking features necessary for making predictions and dropping the unnecessary ones.



**Refined dataset**

DATA PREPROCESSING

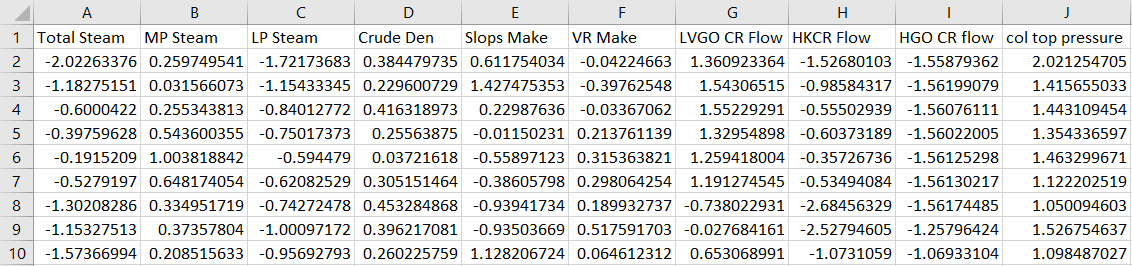
 Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean dataset. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

* SCALING OF DATA

**Data scaling** is a method used to standardize the range of independent variables or features of data. In [data processing](https://en.wikipedia.org/wiki/Data_processing), it is also known as data normalization and is generally performed during the data preprocessing step.

**Library**: Sklearn.preprocessing

**Function**: Standard Scaler: - It will transform the data such that the data will have a distribution with a mean value 0 and standard deviation of 1.



**Final data required for prediction**

SPLITTING THE DATA

* VARIABLE DISTRIBUTION
* Independent variable: All the features except ‘Total Steam’
* Dependent variable: ‘Total Steam’
* TEST/TRAIN SPLIT

To assess the model’s performance, we show it new data (data that it hasn’t seen before) for which we have labels.

This is usually done by splitting the labeled data we have into two parts.

* One part of the data is used to build our machine learning model, and is called the **training data***.*
* The rest of the data will be used to assess how well the model works; this is called the **test data**.

**Library**: sklearn.model\_selection

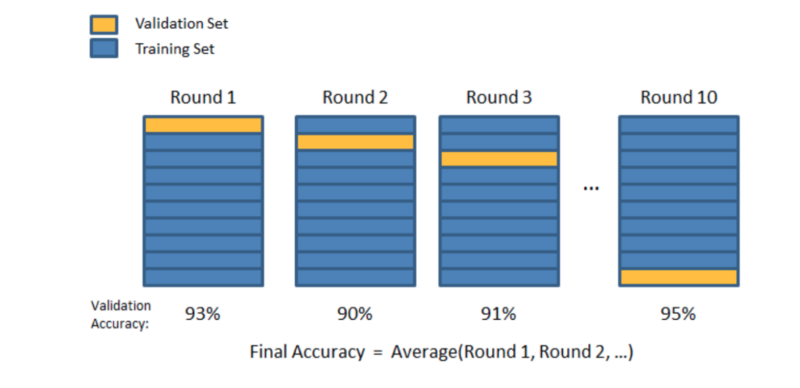
**Function**: train\_test\_split()

**Training Data**: 75%

**Testing Data**: 25%

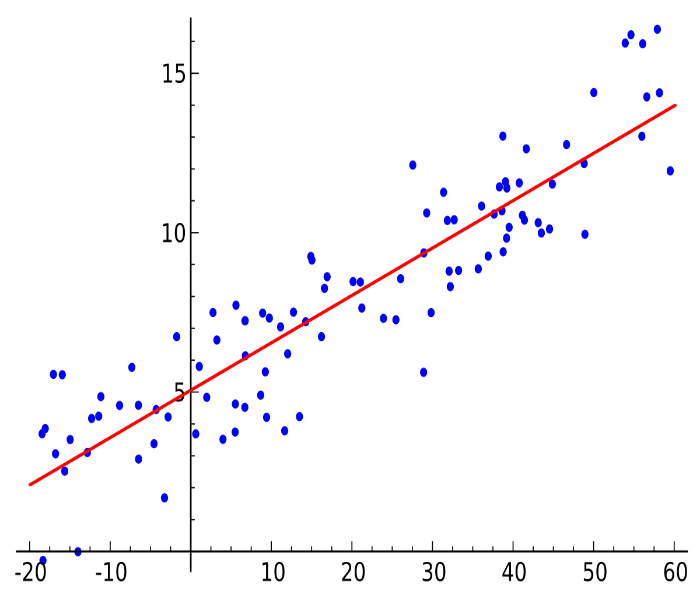
* K-FOLD CROSS VALIDATION

In K-Folds Cross Validation we split our data into k different subsets (or folds). We use k-1 subsets to train our data and leave the last subset (or the last fold) as test data. We then average the model against each of the folds and then finalize our model. After that we test it against the test set.

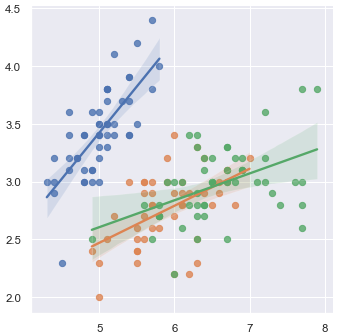


PROBLEM APPROACH

* The dataset is a supervised regressor dataset where the parameters and the corresponding singled valued output is given.
* Linear Regression is an approach to model the relationship between a dependent variable and its corresponding independent variable.
* If there is a single independent variable its simple linear regression.
* In case of multiple dependent variable, it’s called multiple linear regression. Linear regression model is fitted using the least square approach.
* If the goal is prediction or for casting or error reduction linear regression can be used to fit a predictive model to an observed dataset of values consisting of dependent and independent variables.
* After developing such a model if additional values of the independent variables are collected without the dependent variable the fitted can be used to make a prediction of the dependent variable.



Simple Linear Regression

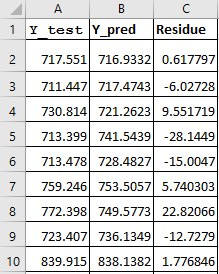


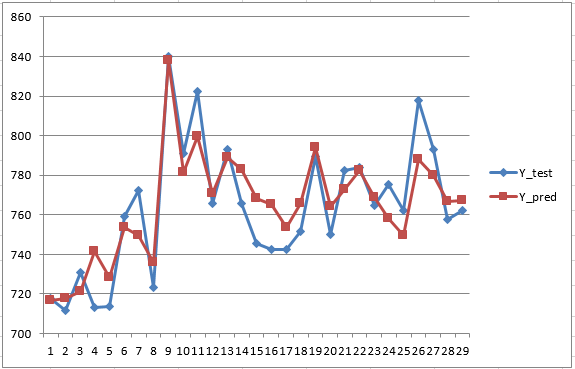
Multiple Linear Regression

* According to the best fit for the dataset **ridge regression model** was used to make predictions.
* In **Ridge Regression**, though, the coefficients (*w*) are chosen not only so that they predict well on the training data, but also to fit an additional constraint. The magnitude of coefficients should be as small as possible i.e. all entries of *w* should be close to zero.
* Intuitively, this means each feature should have as little effect on the outcome as possible (which translates to having a small slope), while still predicting well.
* This constraint is an example of what is called ***regularization*.** Regularization means explicitly restricting a model to avoid overfitting.

**Model Applied**: RIDGE REGRESSION

TARGET PREDICTION





MODEL PREDICTION

* MEAN ABSOLUTE ERROR: 12.434142

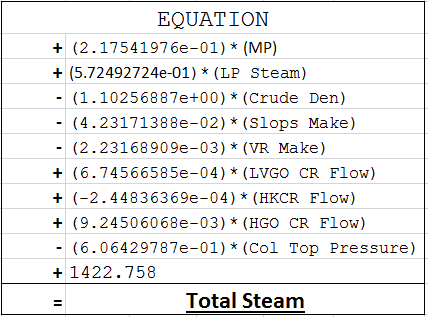
**Library**: sklearn.metrics

**Function**: mean\_absolute\_error()

* R2 SCORE: 82%

**Library**: sklearn.metrics

**Function**: r2\_score()



SHORTCOMINGS

* The r2 score was not improving more than 82%
* The mean absolute error should be between 0 and 1 but it was estimated to be roughly 11.
* Adding more independent variables may lead to skewing of the dataset which may result in a decrease of accuracy.

CONCLUSION

Through this project, I was successfully able to implement the regressor model to predict the Total Steam based on the underlying independent features as provided in the dataset with a satisfactory accuracy.

I was successfully also able to derive an equation with fitting coefficients for the required dataset.