PREDICTION & COMPARISON OF GROSS CALORIFIC VALUE OF INDIAN COAL USING MULTIPLE LINEAR REGRESSION (MLR) & RANDOM FOREST REGRESSION (RFR)

Thesis submitted to the Indian Institute of Technology, Kharagpur in partial fulfillment of the requirements for the degree of

Master of Technology

In the

Mining Engineering Department

By

ADARSH KUMAR
(18MI31001)

Under the guidance of

Prof. Rakesh Kumar

DEPARTMENT OF MINING ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR
KHARAGPUR – 721302 (INDIA)
May 2023.

Department of Mining Engineering
Indian Institute of Technology Kharagpur

CERTIFICATE

This is to certify that the project titled "Prediction and comparison of Gross Calorific Value of Indian Coal using Multiple Linear Regression (MLR) & Random Forest Regression (RFR) and comparing them" is a bonafide record of work carried out by **Adarsh Kumar (18MI31001)** under my supervision and guidance, as his Master of Technology Project, in the department of Mining Engineering during spring semester of the academic session 2022-2023.

****

Date:01/05/2023
Place: Kharagpur

Prof. Rakesh Kumar
Dept. of Mining Engineering,
IIT Kharagpur

ACKNOWLEDGEMENTS

I would like to express my heartiest gratitude to Prof. Rakesh Kumar for giving me an opportunity to work on a project under his able guidance. It was mostly due to his efforts that I have come to find this topic very interesting and have been able to learn how to apply theoretical knowledge practically. I have been fortunate that Prof. Rakesh Kumar has always ensured that I get to share my ideas and get appropriate knowledge, and discussions about the project. I would always be grateful to him for this.

I would also like to thank, Mr. Biswajit Halder, Fire and Explosion Lab, IIT Kharagpur, for his kind assistance in the laboratory.

Finally, I would like to thank my colleagues at the department who helped me at various stages of the project.

Place: Kharagpur
Dated:01/05/2023 Adarsh Kumar

ABSTRACT

One of the most important metrics for determining coal quality is the gross calorific value (GCV). As a result, good GCV prediction is one of the most important techniques to boost heating value and coal production. To make the task easier and lower the cost of analysis, multi-linear regression analysis and random forest regression analysis approaches have been introduced. The purpose of this study is to determine the applicability of these machine learning models in the context of Indian coals.

An attempt has been made in this paper to evaluate the applicability of proximate analysis data to develop a formula using multi-linear regression analysis and a model using random forest regression analysis to predict gross calorific value (GCV) and the results were further compared to achieve better model with a special focus on Indian coals. The models provided here were created using data from 117 coal samples, and its significance stems from the inclusion of all of the important GCV influencing variables, i.e., Moisture, Ash, Volatile Matter and Fixed Carbon content. The R-squared value for the complete data set using multi-linear regression was 0.996, which shows our model is reliable. For random forest regression, the model was trained on 81 coal samples and predicted the GCV of the remaining 36 coal samples. The R-squared value on the validation set was 0.948 which represents that the model was trained well. And the R-squared value for the entire data set i.e., 117 coal samples, using random forest was 0.983, which established reliability of our model.

List of Tables

List of Figures

Table of Content

Chapter 1: Introduction

1.0 Introduction

The global energy demand has recently increased, and fossil-based fuels such as natural gas, fuel oil, and coal are largely compensating. Coal is a critical energy source for many countries, and it is one of the fossil fuels that produces heat and electricity using various methods to meet our everyday needs. As a result, predicting coal quality is a critical undertaking that relies heavily on understanding of the material's chemical and physical properties.

Many enterprises in India are opting for coal-fired captive power plants due to a lack of continuous power supply and higher industrial pricing. By using a less expensive method, you can quickly test the quality of coal. A prerequisite is the ability to run the boilers efficiently. On the basis of proximate analysis data, an accurate prediction of coal's calorific value can be made. As a result, many research organizations are attempting it. As a result, several correlations have been created to forecast the future. Recently, proximate analysis data revealed the GCV of coals. Apart from reducing time and effort required experimentally determining GCV , the obtained correlations could be useful in performance modeling exercises of coal combustion, gasification, and pyrolysis processes. As a result, the accuracy of the prediction and versatility of any correlation like this is quite important.

It is usual practice to examine the quality of coals for power plant and industrial applications utilizing calorific value, proximate analysis, and ultimate analysis. Calorific value is the amount of heat released by complete combustion, and it is determined experimentally using a bomb calorimeter. This method of determination is costly and requires sophisticated equipment as well as trained chemists, whereas proximate analysis of coal requires only very expensive equipment and trained analysts, whereas moisture(M), ash(A), volatile matter (VM), and fixed carbon (FC) are easily determined using simple muffle chemistry.

In this study, Multiple Linear Regression and Random Forest Regression are utilized to separately create a model for the prediction of Gross Calorific Value (GCV). In Multiple Linear Regression(MLR), 117 data samples are used to develop the formula for GCV and in Random Forest Regression(RFR), 81 data samples are utilized to train the model and then we predict the GCV for remaining test data as well as entire data set. Multiple Linear Regression (MLR) and Random Forest Regression (RFR) are two widely used machine learning algorithms in industrial process research. Correlation analysis is carried out to

analyze the individual effect of moisture, volatile matter, ash and fixed carbon on the gross calorific value (GCV). Then, a comparison of the experimental value of GCV and its predicted values from the respective models is carried out to understand the variation.

1.1 Scope of Work

Presently, the way to determine the GCV of coal is the experimental way. But for small enterprises, buying and setting up the equipment to determine the GCV experimentally is very costly. In view of the above it is imperative that there is an urgent need to develop a simple but reliable model to predict the GCV of coals from various sources from the proximate analysis data involving all the major variables.

1.2 Objective

The objective of this project work is to predict the gross calorific value of Indian Coal using Multi - Linear Regression and Random Forest Regression algorithm using built-in methods in Python.

Keeping this in view, the objectives of the projects can be outlined as:

- To analyze the individual effects of moisture, volatile matter, ash and fixed carbon on the gross calorific value (GCV)
- To develop a model for prediction of Gross Calorific Value using the Multiple Linear Regression (MLR) and Random Forest Regression (RFR)
- To compare and analyze the experimental and predicted value of GCV obtained using both the models

1.3 Methodology

1.3.1 Proximate Analysis

Proximate analysis is an assessment of the moisture, volatile matter, fixed carbon, and ash content of a coal sample that is formally specified by a series of ASTM test procedures. These figures are derived from the mass loss experienced by a coal sample when heated to 900°C in a nitrogen environment, then kept at 900°C while the atmosphere is changed to air.
Proximate analysis results are usually expressed as a percentage of the air-dried material. On a 'dry' basis, the ash can be represented. Dry, dry ash-free, or dry mineral free are all terms used to describe volatile matter and fixed carbon. These statistics are derived from the air-dried' basis results.

1.3.1.1. Determination of Moisture content

Coals are mined from the earth and contain a little quantity of moisture. When coal is heated to 1000°C, it loses weight due to drying, and this moisture content is termed as intrinsic moisture.

Method: The loss of mass from drying a known mass of material is calculated as moisture. The moisture content can be assessed by drying the coal in a single step at 108±2°C or by a two-stage procedure in which the coal is first air-dried under ambient circumstances and then dried in an oven at 108±2°C to remove any leftover moisture. The total moisture is estimated in the latter situation by adding the losses from air drying and oven drying. The minimum free-space oven method with a temperature of 200±5°C and a heating period of 4 hours is used to determine moisture in coke.

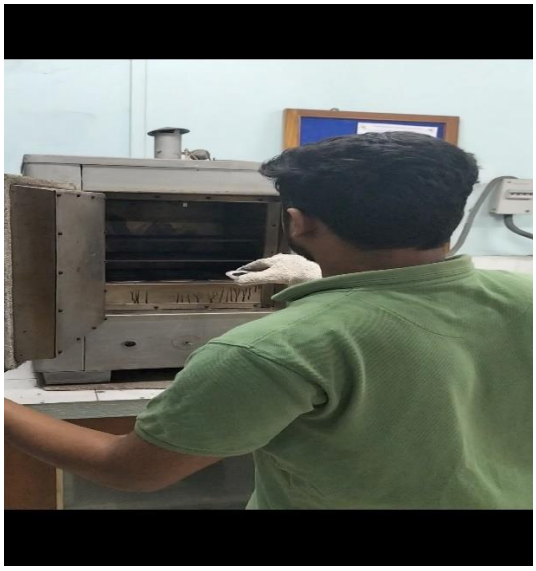

Image1: Removing sample from Oven
to determine moisture content

The calculation is given by the following formula,

$$\%\text{Moisture}=((Y-X)/(Y-Z))*100 \quad \textbf{(1)}$$

Where, X = weight of empty crucible, g.
Y = weight of crucible + coal sample before heating, g.
Z = weight of crucible + coal sample after heating, g.
Y - X = weight of coal sample, g.
Y - Z = weight of moisture, g.

1.3.1.2 Determination of Volatile Matter

Except for water, all of the components of coal that are freed when burnt at high temperatures in the absence of oxygen are classified as volatile matter. The

probability of coal spontaneous combustion is increased when volatile stuff is present. When coal is heated to a high temperature, the thermal decomposition of various coal elements occurs, resulting in a reduction in the mass utilized to determine volatile matter.

Method: The procedure entails heating a weighted quantity of air-dried coal or coke sample at 925°C for seven minutes without coming into touch with air. To ensure a non-oxidizing environment, 2–4 drops Benzene is added to the weighted material for testing coke.

Image2: Removing sample after heating from Muffle furnace

The calculation is given by the following formula,

$$\%\text{Volatile Matter} = ((Y - Z - M)/(Y-X))*100 \quad \textbf{(2)}$$

Where, X = weight of empty crucible, g
Y = weight of crucible + coal sample before heating, g
Z = weight of crucible + coal sample after heating, g
M = Moisture content, Y - X = weight of coal sample, g
Y - Z = weight of volatile matter + moisture,

1.3.1.3 Determination of Ash Content

Ash is the non-combustible residue formed from the inorganic or mineral components of the coal. Indian coals are of drift origin.

Method: Using a laboratory mechanical mixer, thoroughly mix the air-dried material, ground to pass through a 212-micron IS sieve, for one minute. Weigh an empty, clean, dry dish with a lid. Depending on the size of the dish, carefully weigh one to two grams of the material into it. Distribute the material evenly, with a distribution of no more than 015 g per cm3. At room temperature, place

the uncovered dish in the muffle furnace, raise the temperature to 500°C in 30 minutes, then to 825°C in another 30 to 60 minutes, and keep it there for 60 minutes.
The calculation is given by the following formula,

$$\%\text{Ash Content} = ((Z-X)/(Y-X))*100 \quad \textbf{(3)}$$

Where, X = weight of empty crucible in grams
Y = weight of coal sample + crucible in grams (Before heating)
Z = weight of coal sample + crucible in grams (After heating)
Y - X = weight of coal sample, g
Z - X = weight of ash, g

1.3.1.4. Determination of Fixed Carbon

It is determined by subtracting the sum of all the above three parameters from 100% and is given as Fixed Carbon,

$$\%FC = 100 - (M+V+A) \quad \textbf{(4)}$$

Where, M: Moisture content,
V: Volatile matter content, and
A: Ash content

1.3.2 Determination of Calorific Value

The calorific value (heating value) is a direct indication of the energy available for the production of steam and probably the most important parameter for determining the commercial usefulness of coal.
Calorific value is defined as the amount of calories generated when a unit amount of a substance is completely oxidized and is determined using the bomb calorimeter.
The **bomb calorimeter** is a type of constant-volume calorimeter used to measure the combustion heat of oxygen-burnable samples. Four critical parts are needed in every bomb calorimeter. The bomb calorimeter is a laboratory instrument used to measure the amount of a sample's combustion heat or heat power when excess oxygen combustion occurs. The purpose of this research is to determine the effect of using the bomb calorimeter on the ability of physics students to process science. Influences involve the efficacy of using the devices and learning how to develop the abilities of the scientific method of students before and after using materials. If the heat of the capacity calorimeter of to the calorimeter is known, then one determines the heat generated by only needing to note the change in the temperature process. Calorimetry is widely used in present-day laboratories.

Method: First of all, energy equivalent of calorimeter is determined with benzoic acid pellets as reference material. Then, samples of known weight are burned and the resultant temperature rise is measured and recorded. The amount of heat

obtained from each sample is then determined by multiplying the observed temperature rise by the energy equivalent of the calorimeter. Then, by dividing this value by the weight of the sample, we obtain the calorific value (heat of combustion) of the sample on a unit weight basis.

$$\Delta Hc = Cv.\Delta T$$

Where, ΔHc = Calorific Value;
Cv = Heat capacity of bomb;
ΔT = Change in temperature.

Image3(a): Sample ready to be put in the bucket of the bomb calorimeter

Fig3(b): Filling in the oxygen for combustion of the sample

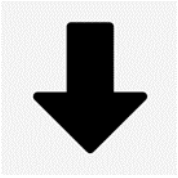

Image3(c): Putting the sample in bucket filled with distilled water

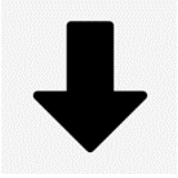

Image3(d): Taking the reading from the screen

Image3(e): Removing burnt sample from the bucket

1.3.3 Multivariable linear regression

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Multivariable linear regression model attempts to find the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. Formally, the model for multiple linear regression, given observation, is

$yi = \beta 0 + \beta 1xi1 + \beta 2xi2 + ... + \beta pxip + \epsilon$

where, for $i=n$ observations;

yi = dependent variable;

xi = explanatory variables;

$\beta 0$ = y-intercept (constant term);

βp = slope coefficients for each explanatory variable;

ϵ = the model's error term (also known as the residuals)

Assumptions for Multiple Linear Regression:

- A linear relationship should exist between the Target and predictor variables.
- The regression residuals must be normally distributed.
- MLR assumes little or no multicollinearity (correlation between the independent variable) in data.

1.3.4 Random Forest Regression

Random Forest Regression algorithms are a class of Machine Learning algorithms that use the combination of multiple random decision trees each trained on a subset of data. The use of multiple trees gives stability to the algorithm and reduces variance. The random forest regression algorithm is a commonly used model due to its ability to work well for large and most kinds of data.

The algorithm creates each tree from a different sample of input data. At each node, a different sample of features is selected for splitting and the trees run in parallel without any interaction. The predictions from each of the trees are then averaged to produce a single result which is the prediction of the Random Forest.

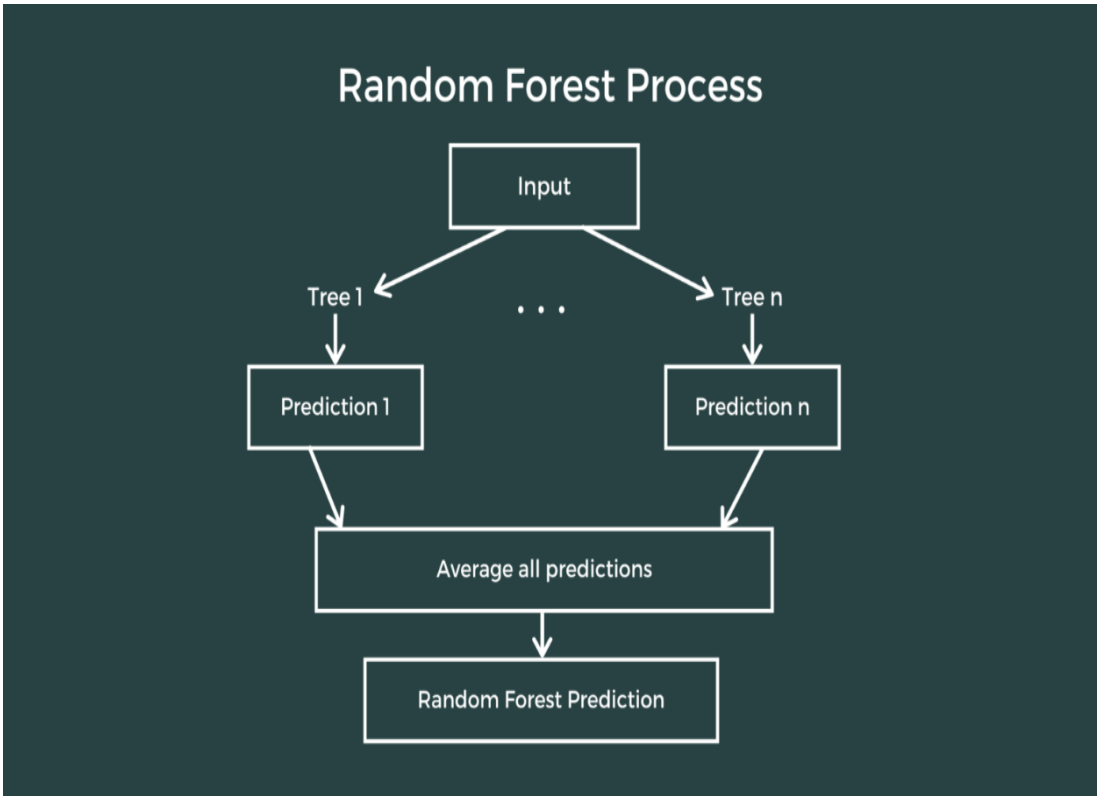

Figure 1: Random Forest Process

A random forest is a meta-estimator (i.e., it combines the result of multiple predictions), which aggregates many decision trees with some helpful modifications:

1. The number of features that can be split at each node is limited to some percentage of the total (which is known as the hyper-parameter). This limitation ensures that the ensemble model does not rely too heavily on any individual feature and makes fair use of all potentially predictive features.

2. Each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness that prevents overfitting.

Chapter 2: Literature Survey

2.1 Introduction

Researchers all over the world have been doing quite some research to determine the correlation between proximate analysis parameters and GCV of coal so that they can develop a machine learning model which is cost effective as well as time saving.

2.2 Theory

Yerel et al. (2013) studied the coal quality parameters such as ash content, calorific value and moisture content, 79 borehole samples were collected from western turkey. In their study, they predicted calorific values using linear regression analysis using both simple linear regression and multiple linear regression analysis and models developed for predictions. Linear regression was applied to determine the relationship between dependent variable calorific value and ash content, moisture as independent variables. The aim of regression analysis was to determine the values of parameters for a function that causes the function to best fit a set of data observations provided and it was concluded that calorific values can be estimated using a multiple linear regression model.

Sharma et al. (2012) used standard sampling procedures to investigate coal samples taken from various north eastern Indian coalfields (Assam, Meghalaya, Nagaland, and Arunachal Pradesh). Using a proximate analyzer (TGA 701,leco,USA) and 144 DR sulphur determinator, a petrographic study, proximate analysis, and sulphur analysis were performed, and the percentage of oxygen was calculated by difference. Using an automatic bomb calorimeter, the calorific values of coal samples were determined (LECO AC-350). Multi-variable linear regression analysis was used to explore the association between gross calorific value (GCV) and macerals content of these coal samples. The statistical tool used to analyze the relationship between variables is regression analysis. North eastern coal samples had a high vitrinite concentration (80.07 percent), moderate to low liptinite (10.23 percent), and a low inertinite content, according to the mineral analysis (9.3 percent). The inter-correlation between GCV of coal and maceral analysis revealed that when interinite content in coal increases, GCV decreases, while higher vitrinite and liptinite content in coal can result in higher GCV.

Upadhyay (2014) investigated the physical and chemical properties of coal in Korba district for assessment of coal quality, in order to check its suitability for thermal power station, by collecting samples from Gerva coal mines. Three different coal samples were collected from different areas of Gerva coal mines and analyzed for ultimate, proximate and calorific value as per standard methods and From overall analysis and according to useful heat value (UHV) of coal samples, they were concluded that the grade of Gerva coal was "F" and very useful for coal based thermal power station.

Many researchers have utilized other machine learning and deep learning models like Artificial Neural Network (ANN) to train and test sample data for predicting GCV.

Like **Krishnaiah et al. (2012)** carried out the study for around 150 lab analysis data of coal and both proximate and ultimate information used to train and test the ANN model. Ultimate analysis is the process to know elemental composition of coal. The ultimate analysis is expensive, time taking and also cumbersome in nature but at the power plants only the gross level coal composition is estimated which is known as proximate analysis. The elemental compositions were estimated by using standard empirical formulae based on the gross level compositions of coal. Relationship between the elemental composition and gross level composition was 5 nonlinear. To achieve better performance of boilers and control on the boilers, accurate information of elemental composition is required. So they suggested a method to compute ultimate analysis by proximate analysis using an artificial neural network model (ANN). The prediction of ANN model and empirical models were compared and found that ANN prediction is better with lab data than the predictions of empirical mode.

Chapter 3: Work done

3.1 Introduction

A correlation study was done on the coal samples dataset provided to see how the moisture, ash, volatile matter and fixed carbon content affects the GCV of coal, using sklearn library in Python. After that, a multi-linear regression model was designed to predict the GCV of coal again using the sklearn library in Python. Then, a random forest regression model was utilized to train a data of 81 coal samples out of 117 and predict the GCV for the remaining 36 coal samples. The model was evaluated based on the R-squared metric.

3.2 Laboratory investigation/ Numerical Modeling/ Data from public domain

3.2.1 Experimental Data

The proximate analysis data was experimentally determined in Fire and Explosion Lab, Mining Engineering, IIT Kharagpur. The data consists of moisture, ash, volatile matter, fixed carbon contents and experimental gross calorific values of 117 coal samples.

Table1: Data of first 10 samples

Sl. No.	Moisture(%)	Ash(%)	Volatile Matter(%)	Fixed Carbon(%)	Experimental GCV
01	8.48	34.8	25.58	31.15	4082.1
02	5.63	38.16	25.68	30.53	4014.2
03	4.95	47.05	22.57	25.43	3334.9
04	5.24	31.95	28.15	34.66	4525.6
05	4.85	33.51	27.25	34.39	4428
06	4.26	41.29	24.99	29.47	3790.4
07	7.35	18.76	31.38	42.51	5463.6
08	5.46	31.15	26.03	37.36	4568.4
09	5.29	24.57	28.55	41.59	5319.6

10	5.3	25.54	27.58	41.58	5105.9

3.3 Work Methodology

3.3.1 Effects of proximate analysis parameters on GCV

It is well-known that ash, volatile matter, moisture and fixed carbon combinedly decide the nature of coal and more precisely its GCV as a fuel. Therefore, the results obtained were carefully examined for observing the impacts of ash, moisture, volatile matter and fixed carbon on the GCV of coal. It was observed that ash has negative effects on GCV whereas moisture, volatile matter and fixed carbon have positive effects on GCV. In other research papers, moisture has been seen to have negative effects on GCV. To show the impacts of ash and moisture contents of coals on GCV, the GCV values are plotted as a function of the variables.

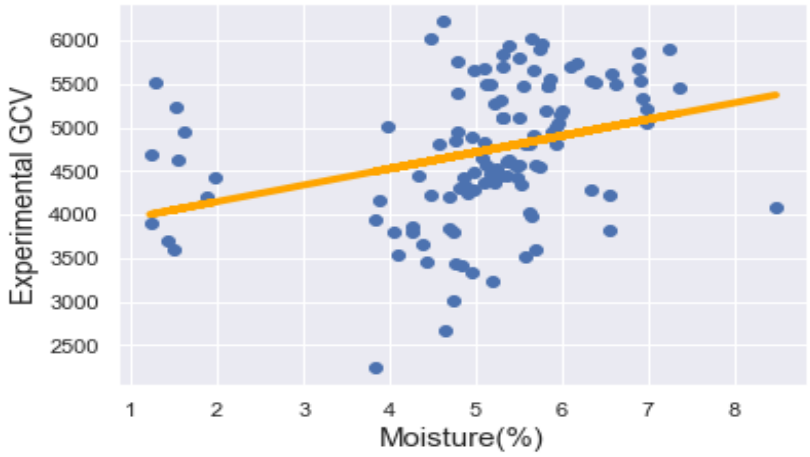

Figure 2: Effects of Moisture on GCV of Coal

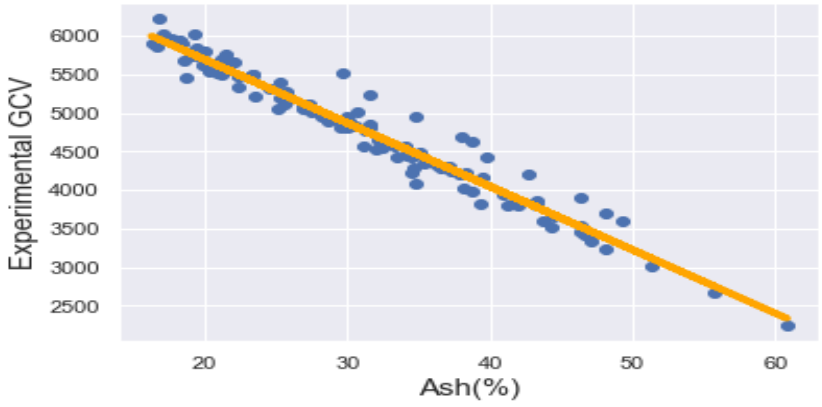

Figure 3: Effects of Ash on GCV of Coal

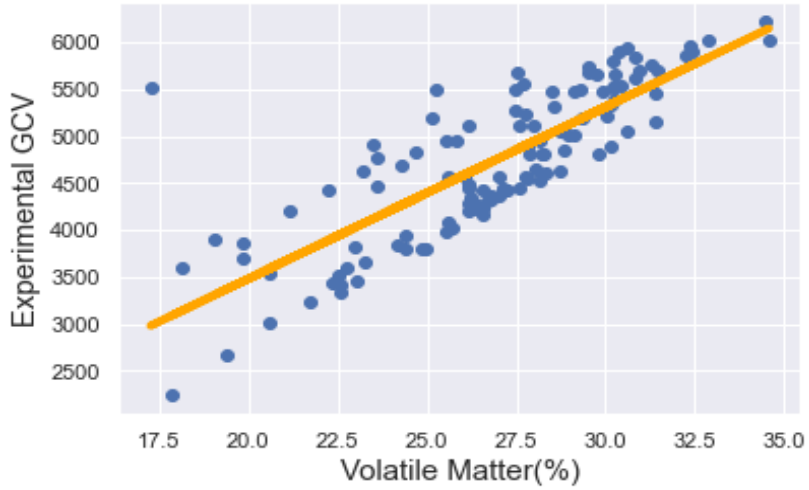

Figure 4: Effects of Volatile Matter on GCV

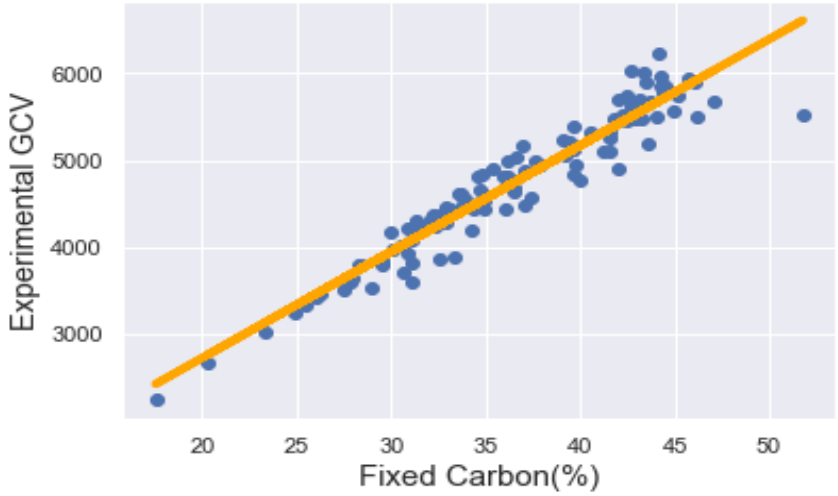

Figure 5: Effects of Fixed Carbon on GCV

3.3.2 Model Development

3.3.2.1 Model Development by Multi-Linear Regression Analysis

Using the 117 experimental data of proximate analysis and gross calorific value of coal, and utilizing the built-in functions in sklearn library of Python for linear regression, a formula for Gross Calorific Value (GCV) was developed:

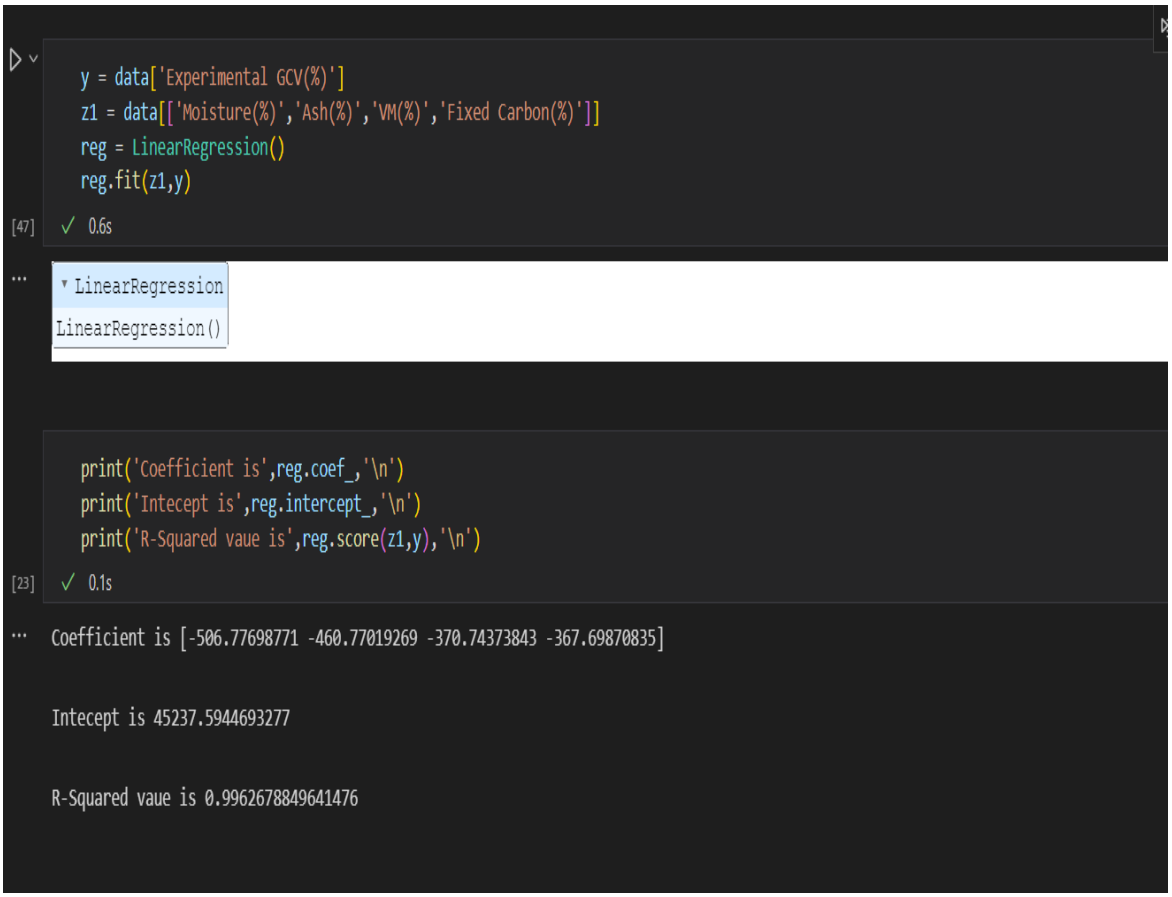

Fig 6: Code utilized for multi-linear regression

$$GC = 45237.594 - 506.777*M - 460.77*A - 370.744*V - 367.699*F \quad (5)$$

Where, **M** = Moisture (%), **A** = Ash (%), **V** = Volatile Matter (%), and **F** = Fixed Carbon (%).

The R-squared value on the data set is 0.996.

The comparison of the predicted GCV by multi-linear regression model and that of the experimentally determined value of first ten samples has been presented in Table 2.

Table 2: Experimental GCV and Predicted GCV by Regression Model

Sl. No.	Experimental GCV	Predicted GCV	Difference
01	4082.1	3967.883314	114.2166862

02	4014.2	4054.908707	40.70870669
03	3334.9	3331.546484	3.353515607
04	4525.6	4679.601929	154.0019291
05	4428	4591.39147	163.3914696
06	3790.4	3952.556287	162.1562871
07	5463.6	5603.924191	140.324191
08	4568.4	4729.917359	161.5173589
09	5319.6	5358.297558	38.69755756
10	5105.9	5269.581114	163.6811141

3.3.2.1 Model Development by Random Forest Regression

Utilizing the built-in functions in sklearn library of Python for Random Forest Regressor, after splitting the data into test and train data with 30% test data, a model was developed to predict Gross Calorific Value (GCV). The methods are given in the figure on the next page:

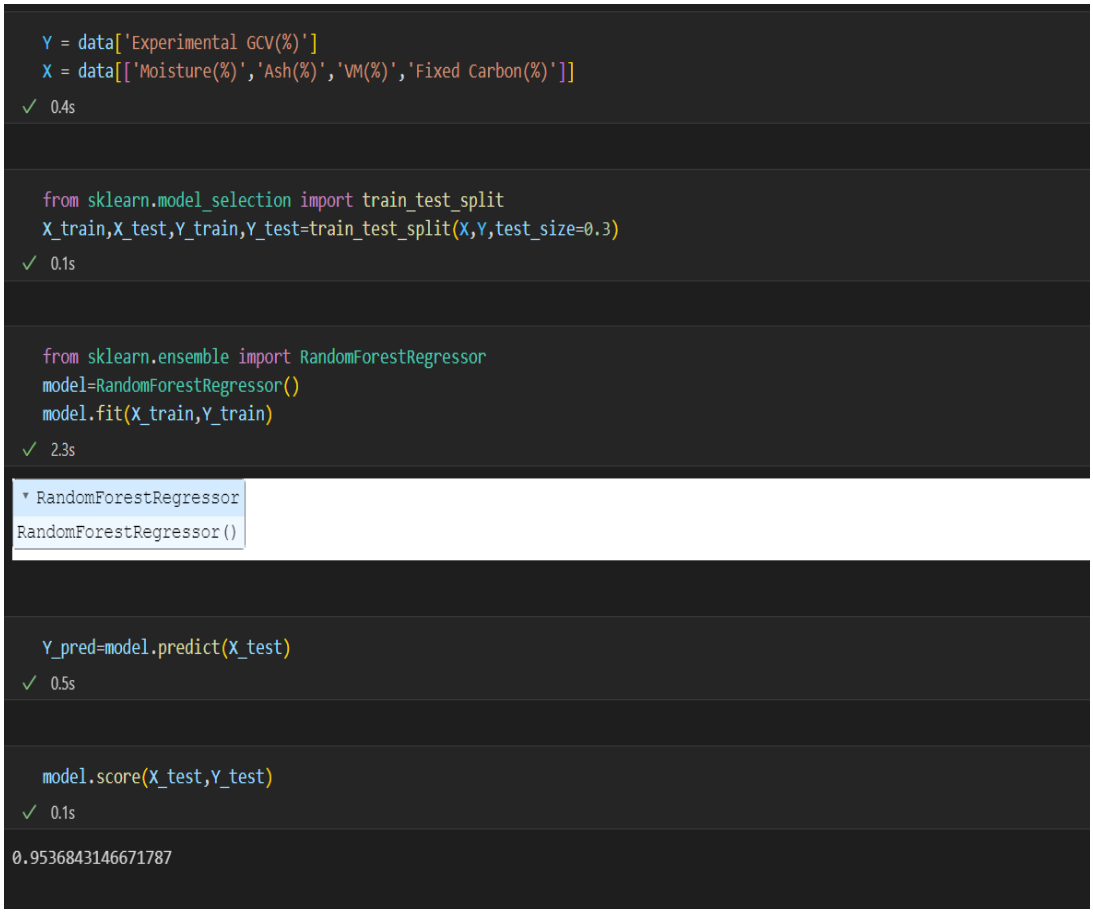

Fig 7: Code utilized for random forest regression

Then, predicted GCV of the 36-test data was found using the developed model. **The R-squared value on the data set is 0.954**.
The comparison of the predicted GCV by random forest regression model and that of the experimentally determined value of the test samples has been presented in Table 3.

Table 3: Experimental GCV and Predicted GCV and their difference of test samples using Random Forest regression

Sl.No.	Experimental GCV	Predicted GCV (Random Forest)	Difference
01	3011.1	3137.472	169.9211
02	4812.6	4818.197	7.428038
03	4365.3	4345.383	16.31968219

04	5706	5687.762	6.906752635
05	5488.5	5522.411	34.95873008
06	5564.7	5575.499	6.296899492
07	3532.7	3636.321	114.4330364
08	5845.2	5705.03	123.2910854
09	5498.2	5505.176	18.91278019
10	3603.1	3749.771	208.4997135
11	4949.3	4960.044	14.91071494
12	4416.6	4415.203	1.470545184
13	5198.8	5180.175	23.82544889
14	5185.5	5462.932	262.7237589
15	5693.5	5554.495	106.3119539
16	4812.9	4776.641	24.42691175
17	4280.3	4332.526	43.28366164
18	2660.7	2846.532	269.0264717
19	4606.1	4552.605	48.76186816
20	4549.7	4557.883	3.355665462
21	4446.1	4420.827	17.10542903
22	5528.1	5565.433	21.07325445
23	4558.9	4453.142	4.915912338
24	5466.9	5435.795	3.252362533
25	5110.9	5155.446	48.56851778
26	4825.6	4856.777	19.16985591
27	5514.9	5417.163	57.19194473

28	4688.3	4666.06	25.07628595
29	3892.5	3825.07	102.0178548
30	4941	4821.895	104.5090458
31	4629	4626.252	56.25141262
32	3698.3	3646.974	81.05394251
33	5225.8	4879.996	343.3094128
34	4194.9	4071.702	97.83068101
35	4434.4	4469.693	40.88414926
36	3596	3575.496	49.68595774

Chapter 4: Results, Discussion & Future Work

4.1 Introduction

1. To confirm the validity of the developed multi-linear regression model, the experimentally determined values of GCV of 117 coal compared with the predicted values using the model. The mean absolute percentage error (MAPE) between the experimental and the predicted data is found to be 0.68%. The error is quite low and therefore establishes the validity of the developed multi-linear regression model.

2. Similarly, for checking the validity of developed random forest regression model the experimentally determined values of GCV of 36 coal samples were kept separately and compared with the predicted values using the random forest regression model. The mean absolute percentage error (MAPE) between the experimental and the predicted data of the Validation set is found to be 1.73% ,which is quite low and establishes the validity of the random forest model. The mean absolute percentage error (MAPE) between the experimental and the predicted data for the entire data set is 1.3%.

4.2 Results

1. Figure 8 shows the plot of the predicted GCV obtained from multi-linear regression vs experimental GCV. The R-squared value for the complete data set is 0.996, which shows our model is reliable.

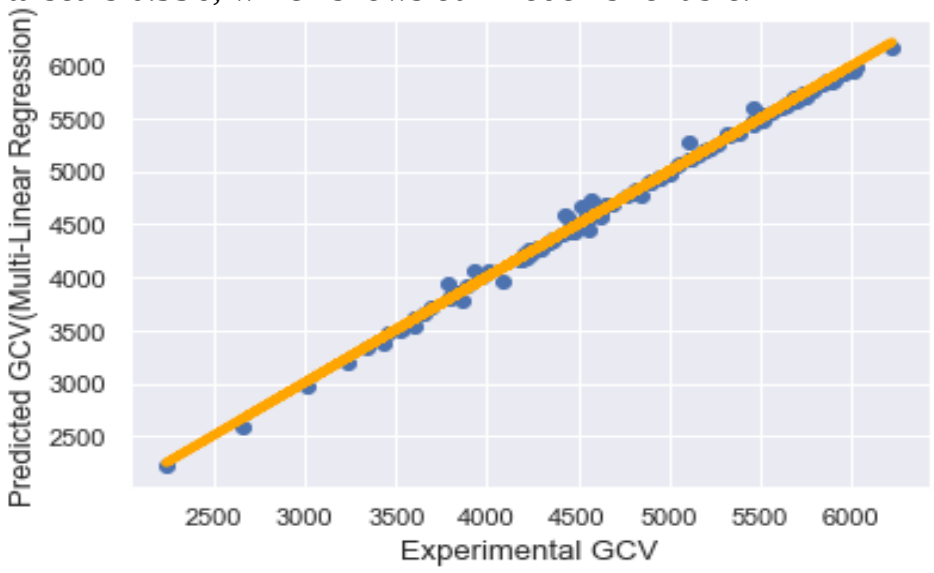

Figure 8: Correlation btw predicted GCV (Multi-Linear Regression) & experimental GCV

2. Figure 9 shows the plot of the predicted GCV obtained from random forest vs experimental GCV. The R-squared value for the validation set is 0.948, and the R-squared value for the entire data set i.e., 117 coal samples, is 0.983, which shows our model is reliable.

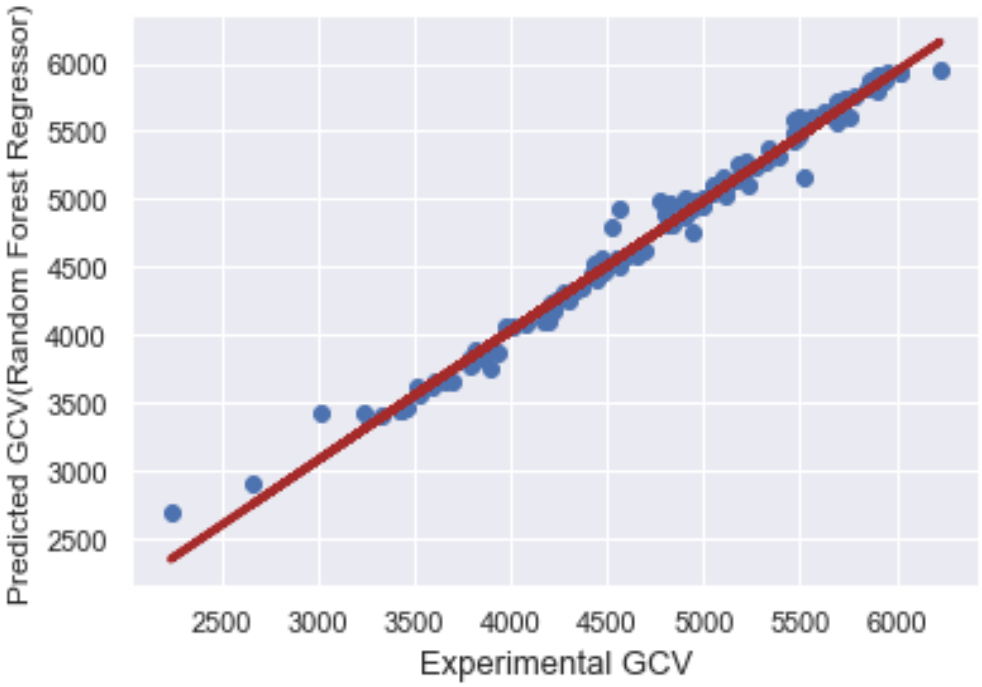

Figure 9: Correlation btw predicted GCV (Random Forest Regression) & experimental GCV

4.3 Future Work

1. In future, we can see if we can add more parameters to our models such as from Ultimate analysis data, it would train the model better and thus, will help the model to predict the GCV with precision.

2. We can utilize larger data sets in future for larger coal samples. We will try to get more proximate analysis data from various coal fields so that we can develop a more robust model.

3. Here, we've utilized only two machine learning algorithms. In future, we can apply deep learning models like ANN (Artificial Neural Network) to obtain better results for a larger sample of data.

4. A better outcome can be obtained which can predict the GCV of a vast variety of coals from different coalfields as a result of utilizing better models for prediction.

Chapter 6: Conclusions

1. On comparing the mean absolute percentage errors (MAPE) between experimental and predicted values of GCV of the two models for entire data set, the MAPE-score for multi-linear regression model is lesser i.e., 0.68% (<1.3% (MAPE-score for random prediction)). Thus, multi-linear regression model gives minimum error and is more accurate as compared to random forest regression.

2. Figure 10 provides visualization for comparison of Experimental GCV, Predicted GCV from Muti-linear regression, and Predicted GCV from random forest for first 12 samples. We can also develop an empirical relation between the three for our convenience as:

   GCV (experimental) = **0.786 * GCV (pred (**muti-linear regression**)) + 0.223*GCV (pred (**random forest regressor**)) - 43.54**

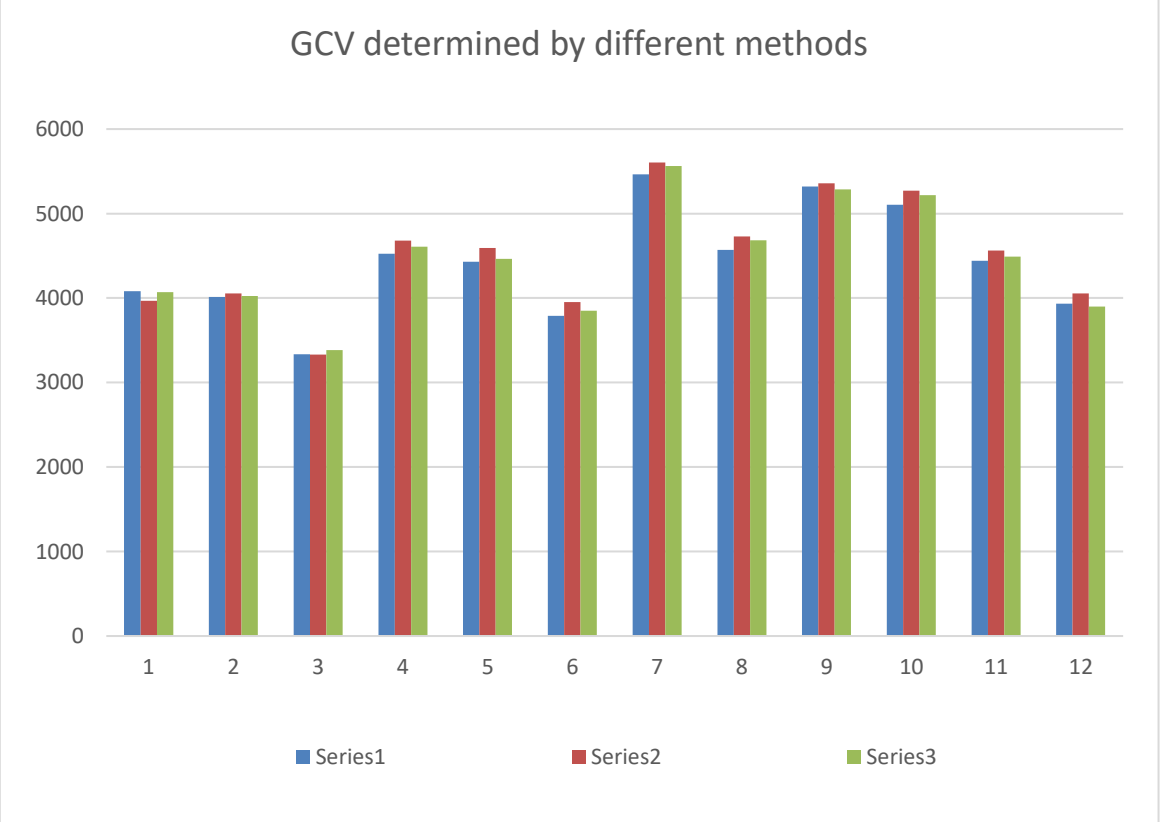

Figure 10: Comparison of GCV determined by different methods using column chart

Where, Series1 = Experimental GCV, Series2 = Predicted GCV (Multi-Linear Regression), and Series3 = Predicted GCV (Random Forest Regression).

3. Any industry where coal is utilized for heating applications, determinations of calorific value, proximate analysis and ultimate analysis are common practice to assess the quality of coals In India, due to non-availability of consistent power supply and higher industrial tariffs many industries are opting for coal-fired captive power plants. Quick assessment of coal quality by cheaper means to run the boilers efficiently is a pre-requisite, prediction of calorific value of coal based on proximate analysis data can be carried out by these two models of Machine Learning. Regression analysis results showed that the multiple regression model is seen as the best model. The determination R2 of the multiple regression model is 99.6%. This value is good and identifies the valid model. This result reveals the usefulness of a multiple linear regression model in the prediction of the gross calorific value.

Chapter 7: References

1. Donahue J., Rais A., Proximate analysis of coal, Journal of chemical education, vol. 86 no.2,2009, pp. 222 – 224

2. Yerel S., Ersen T., Prediction of the calorific value of coal deposit using linear Regression analysis, Energy sources, part(A), 35, 2013.pp. 976 – 980

3. Upadhyay M., Assessment of coal properties in Korba district , Indian Journal of Pharmaceutical Science and Research ,vol.4 Issue 2,2014,pp. 116 – 118

4. Sharma A., Saikia K., Baruah P., Maceral contents of tertiary Indian Coals and their relationship with calorific values, International Journal of innovative Research and Development, vol.1 Issue 7(special Issue), 2012, pp. 196 – 203

5. Mandavgade N.K., Jaju S.B., Lakhe R.R., Determination of Uncertainty in Gross Calorific Value of Coal Using Bomb Calorimeter, International Journal of Measurement Technologies and Instrumentation Engineering, 1(4), 2011,45 – 52

6. Matthesius G.A., Morris R.M. ,Desai M.J., Prediction Of The Volatile Matter In Coal From Ultimate And Proximate Analyses , Journal Of The South African Institut5 Of Mining And Metallurgy, Vol. 87,1987,pp. 157 – 161

7. Krishnaiah J., Lawrence A., Dhanuskodi K., Artificial Networks Model for Predicting Ultimate Analysis using Proximate Analysis of Coal, International Journal of computer applications, 2012, pp. 9 – 13