INSTITUTE FOR ADVANCED
COMPUTING AND
SOFTWARE
DEVELOPMENT
AKURDI, PUNE

Documentation On

“Energy Prediction Model”

PG-DBDA MAR 2022

Submitted By:

Group No: 12
Ankita Jain 223309
Gauri Wale 223320

Mr. Prashant Karhale
Centre Coordinator

Mr. Akshay Tilekar
Project Guide

Contents

1. Introduction

1.1 PROBLEM STATEMENT

Energy Prediction Model

1.2 Abstract

Gas turbines are widely used in the energy production. In the present scenario, the quantity of the operating machines requires a special attention for prediction of power production in the energy marketing sector. Thus, the aim of this project is to support the sector by making the prediction of power production more computable. By using the data from an operating power plant, correlation and regression analysis are performed and model is developed to calculate energy yield. In this model, we observed the discrepancy between the predicted and actual energy yield was less than 0.2%.

1.3 Product Scope

The main use of this regression model is to predict the turbine energy yield of gas turbine. After giving input (AT, AP, AH, AFDP, GTEP, TIT, TAT, CDP, CO, NOX) to the model and pressing the Predict TEY Button, this model will predict and show the energy yield of gas turbine.

1.4 Aims & Objectives

The primary goal of this project is to extract quality patterns from the dataset of gas turbine power plant and use the trained models to predict turbine energy yield. Before giving data for prediction of energy yield the training of the models will be done using a bunch of different machine learning models and after the training the machine learning models willbe compared based on their mean absolute error, mead squared error, R^2 score and the best model will be selected which will then be used to make predictions for the turbine energy yield. The prediction which is given by our system will give the user the quite insight 9of energy yield of gas turbine.

2. Overall Description

2.1 Workflow of Project:

The diagram below shows the workflow of this project.

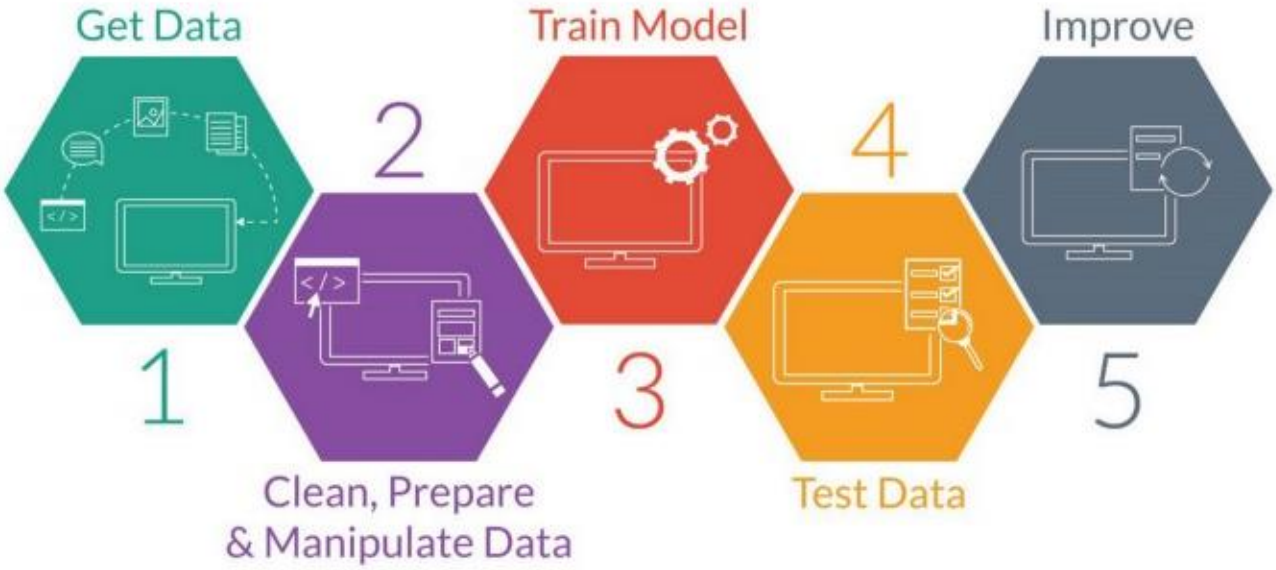


Figure 1.Workflow Diagram

2.2 Data Preprocessing and Cleaning:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data might be incomplete (lacking attribute values, lacking certain attributes of interest or containing only aggregate data), Noisy (containing errors or outliers) or inconsistent (containing discrepancies in codes or names).

2.2.1 Data Cleaning:

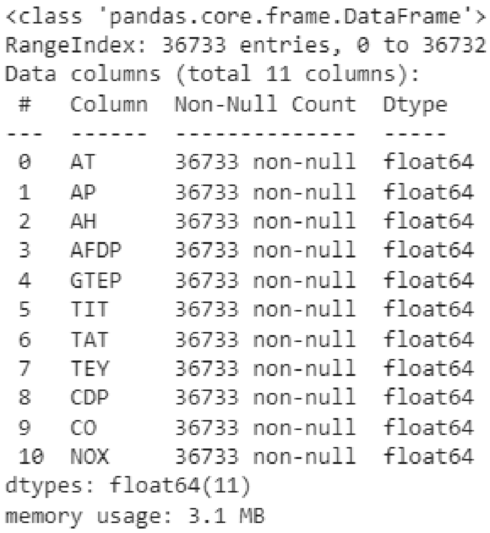
Data cleaning plays a significant part in building a model. As the accuracy of model depends on data so we checked if data have any incorrect, incomplete, irrelevant, duplicated or improperly formatted values in our dataset. But there were no inconsistent or missing values in our dataset.

2.3 Exploratory Data Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Dataset Information:

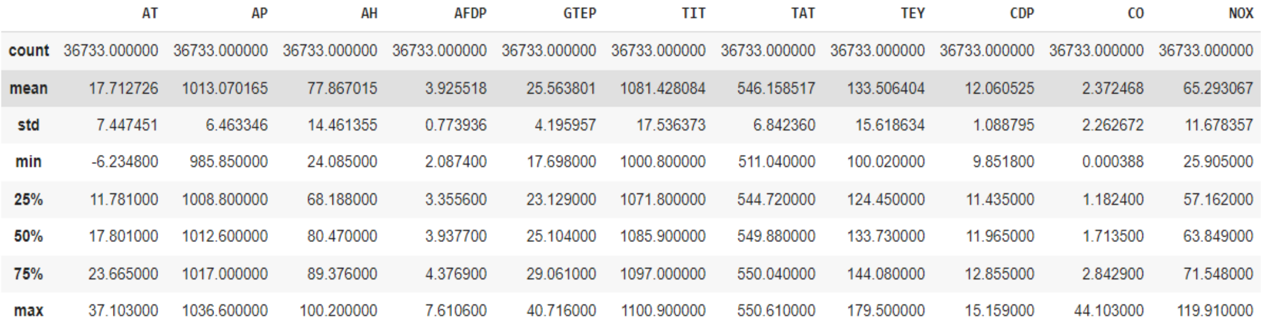
We can see that the features and its data type and the count of null is zero.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36733 entries, 0 to 36732
Data columns (total 11 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   AT      36733 non-null  float64
 1   AP      36733 non-null  float64
 2   AH      36733 non-null  float64
 3   AFDP    36733 non-null  float64
 4   GTEP    36733 non-null  float64
 5   TIT     36733 non-null  float64
 6   TAT     36733 non-null  float64
 7   TEY     36733 non-null  float64
 8   CDP     36733 non-null  float64
 9   CO      36733 non-null  float64
 10  NOX     36733 non-null  float64
dtypes: float64(11)
memory usage: 3.1 MB
```

Figure 2 Data Information

Summary of Statistics:



	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
count	36733.000000	36733.000000	36733.000000	36733.000000	36733.000000	36733.000000	36733.000000	36733.000000	36733.000000	36733.000000	36733.000000
mean	17.712726	1013.070165	77.867015	3.925518	25.563801	1081.428084	546.158517	133.506404	12.060525	2.372468	65.293067
std	7.447451	6.463346	14.461355	0.773936	4.195957	17.536373	6.842360	15.618634	1.088795	2.262672	11.678357
min	-6.234800	985.850000	24.085000	2.087400	17.698000	1000.800000	511.040000	100.020000	9.851800	0.000388	25.905000
25%	11.781000	1008.800000	68.188000	3.355600	23.129000	1071.800000	544.720000	124.450000	11.435000	1.182400	57.162000
50%	17.801000	1012.600000	80.470000	3.937700	25.104000	1085.900000	549.880000	133.730000	11.965000	1.713500	63.849000
75%	23.665000	1017.000000	89.376000	4.376900	29.061000	1097.000000	550.040000	144.080000	12.855000	2.842900	71.548000
max	37.103000	1036.600000	100.200000	7.610600	40.716000	1100.900000	550.610000	179.500000	15.159000	44.103000	119.910000

Figure 3 Statistical Summary

Pairwise relation between different variables:

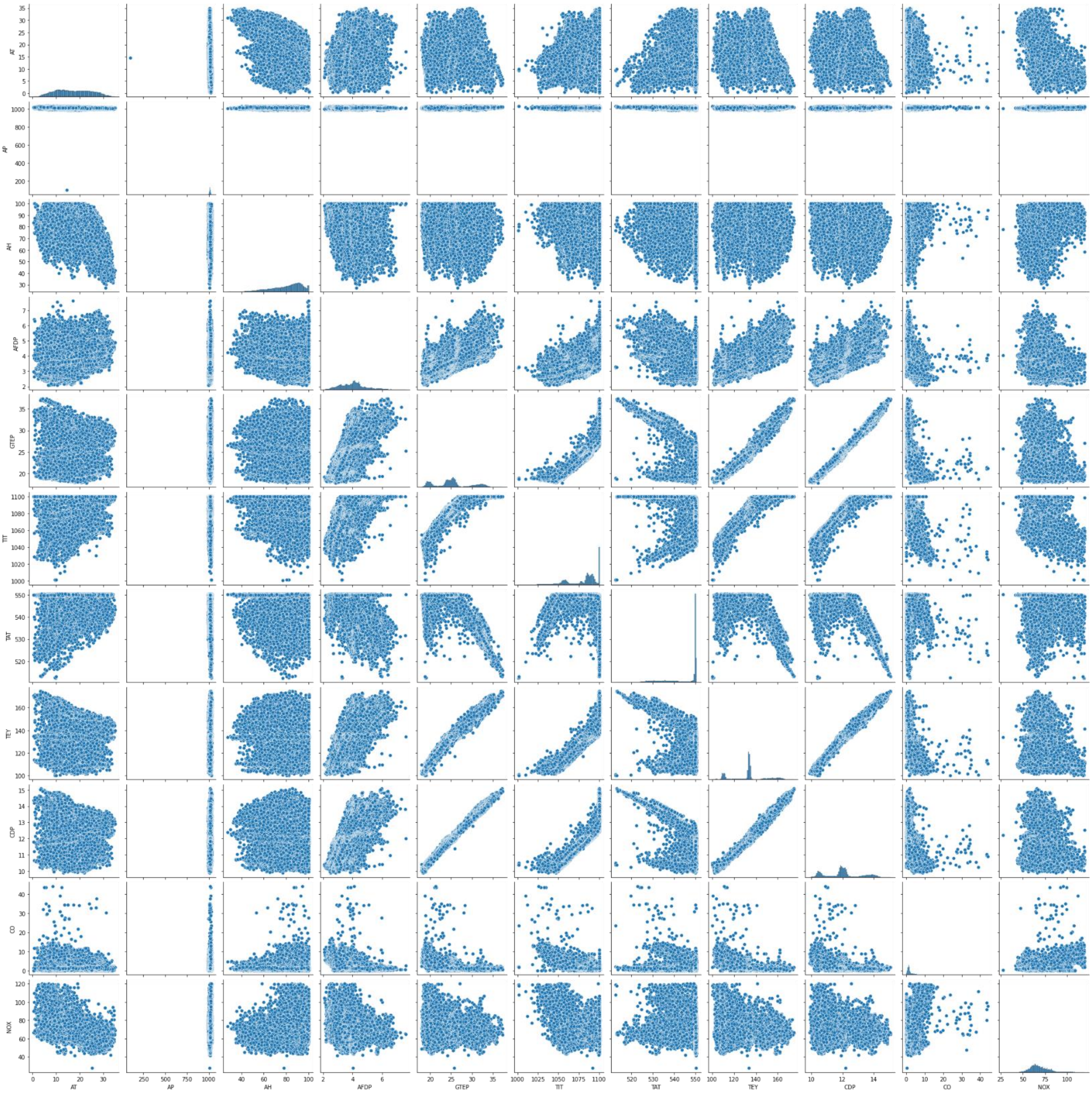
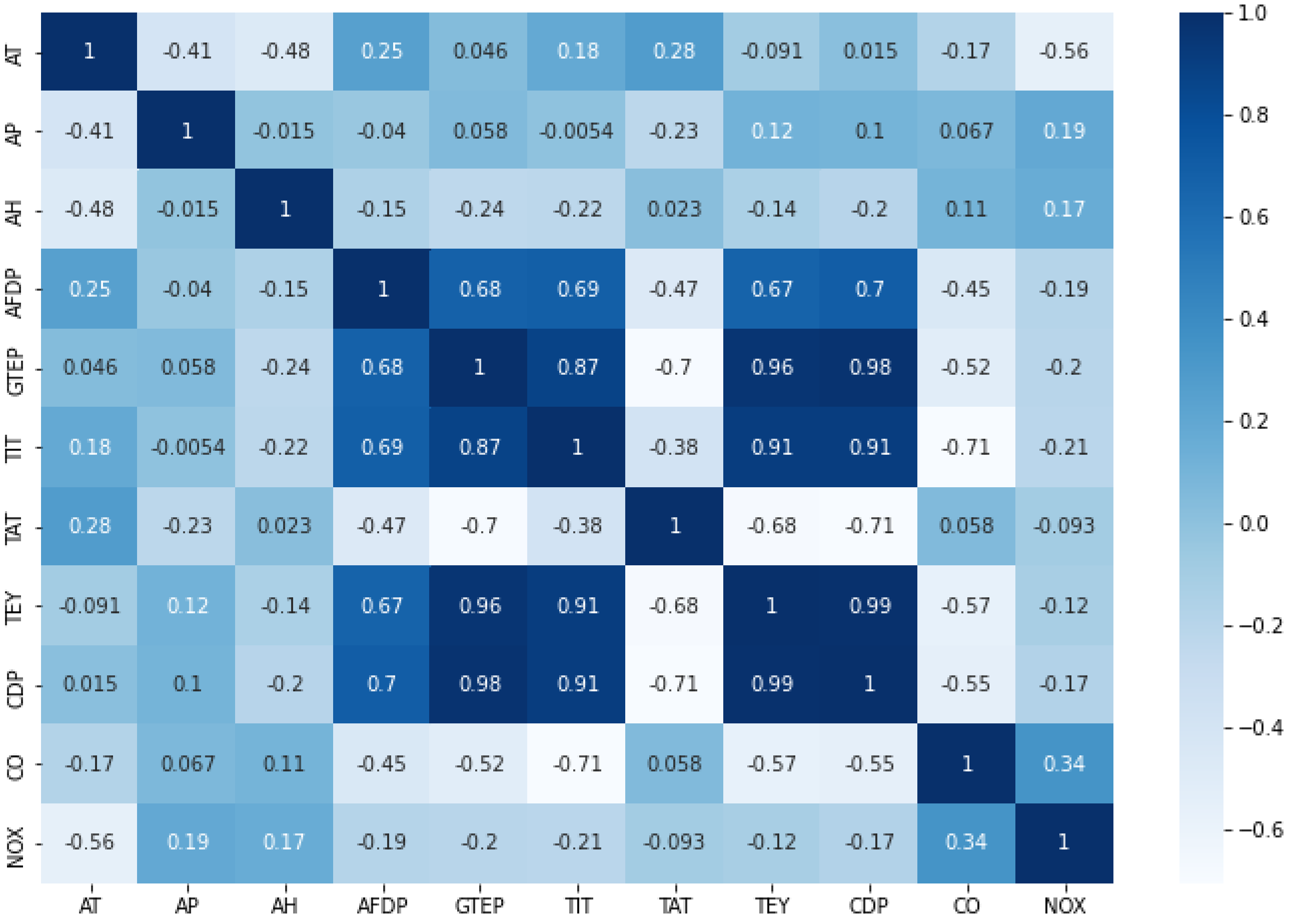


Figure 4 Multivariate Analysis (Pair Plot).

Observation:

Here we can see the relationship between each pairwise combination of variables.

Correlation:



	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
AT	1	-0.41	-0.48	0.25	0.046	0.18	0.28	-0.091	0.015	-0.17	-0.56
AP	-0.41	1	-0.015	-0.04	0.058	-0.0054	-0.23	0.12	0.1	0.067	0.19
AH	-0.48	-0.015	1	-0.15	-0.24	-0.22	0.023	-0.14	-0.2	0.11	0.17
AFDP	0.25	-0.04	-0.15	1	0.68	0.69	-0.47	0.67	0.7	-0.45	-0.19
GTEP	0.046	0.058	-0.24	0.68	1	0.87	-0.7	0.96	0.98	-0.52	-0.2
TIT	0.18	-0.0054	-0.22	0.69	0.87	1	-0.38	0.91	0.91	-0.71	-0.21
TAT	0.28	-0.23	0.023	-0.47	-0.7	-0.38	1	-0.68	-0.71	0.058	-0.093
TEY	-0.091	0.12	-0.14	0.67	0.96	0.91	-0.68	1	0.99	-0.57	-0.12
CDP	0.015	0.1	-0.2	0.7	0.98	0.91	-0.71	0.99	1	-0.55	-0.17
CO	-0.17	0.067	0.11	-0.45	-0.52	-0.71	0.058	-0.57	-0.55	1	0.34
NOX	-0.56	0.19	0.17	-0.19	-0.2	-0.21	-0.093	-0.12	-0.17	0.34	1

Figure 5 Heat Map

Observation:

Dark shades represents negative correlation while lighter shades represents positive correlation. Here we can observe the positive correlation between turbine energy yield(TEY) and compressor discharge pressure (CDP), Air filter difference pressure (AFDP), Gas turbine exhaust pressure (GTEP), Turbine inlet temperature (TIT).

Correlation of TEY with other features:

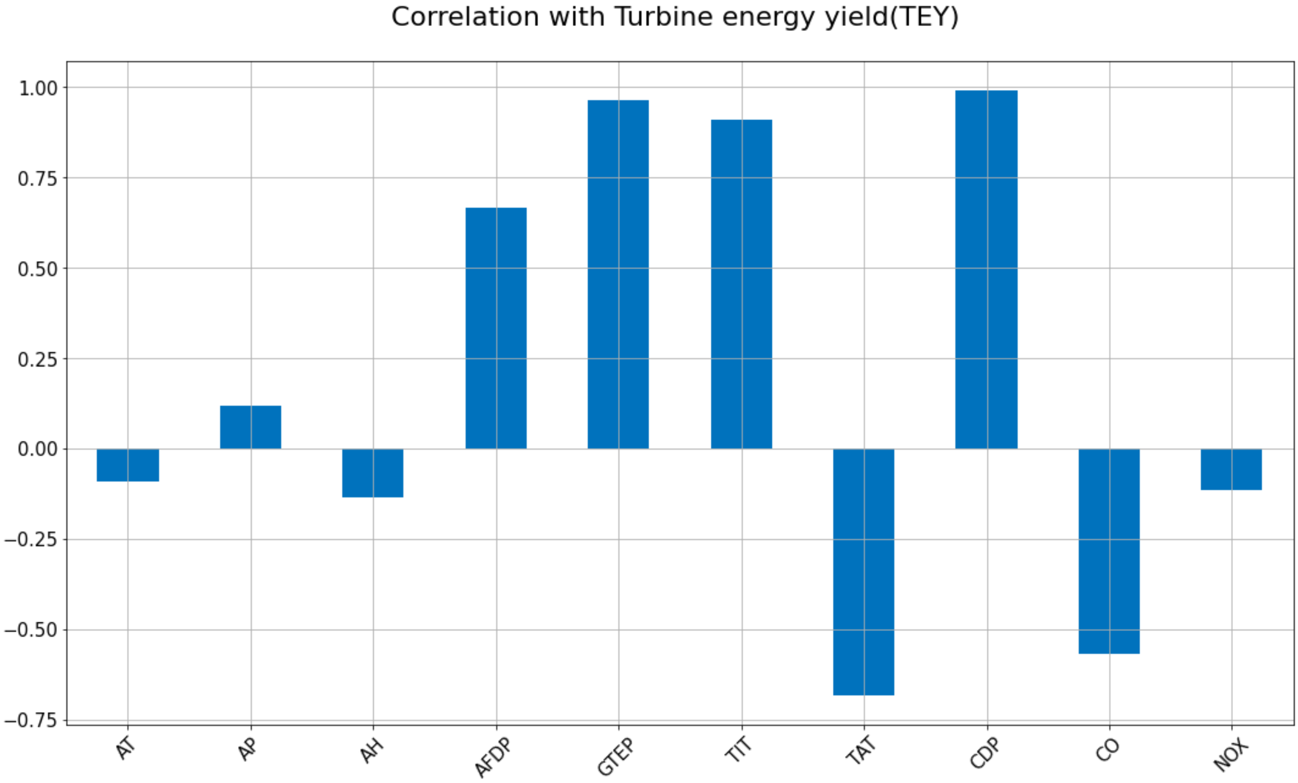
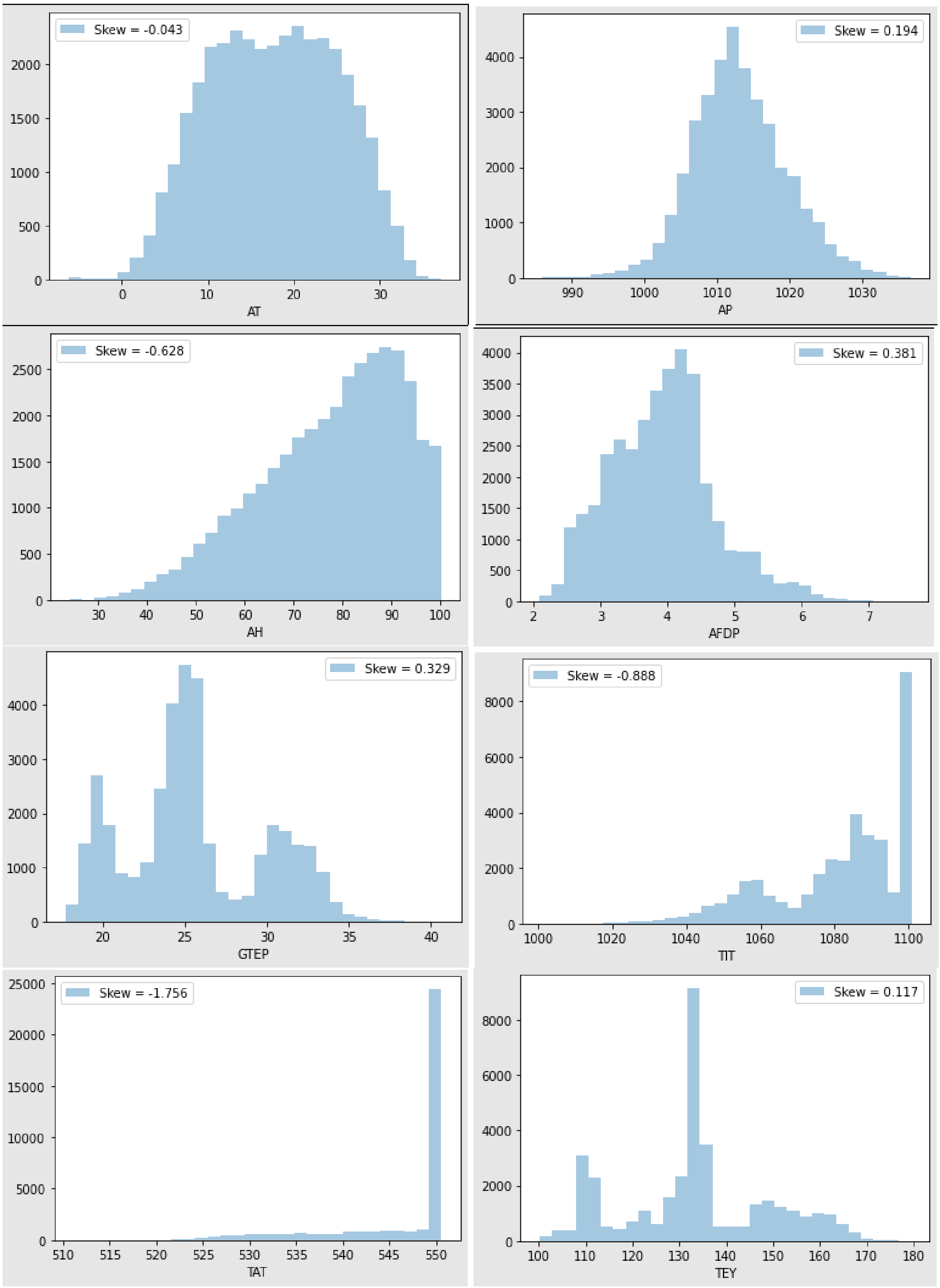


Figure 6 Correlation of TEY with other features.

Observation:

We can observe the positive correlation between Turbine energy yield (TEY), Compressor discharge pressure (CDP), Air filter difference pressure (AFDP), Ambient pressure (AP), Gas turbine exhaust pressure (GTEP), Turbine inlet temperature (TIT). From plot we can see there is negative correlation between Turbine energy yield (TEY), Ambient temperature (AT), Ambient humidity (AH), Turbine after temperature (TAT), Carbon monoxide (CO), Nitrogen oxides (NOx).

Skewness of features:



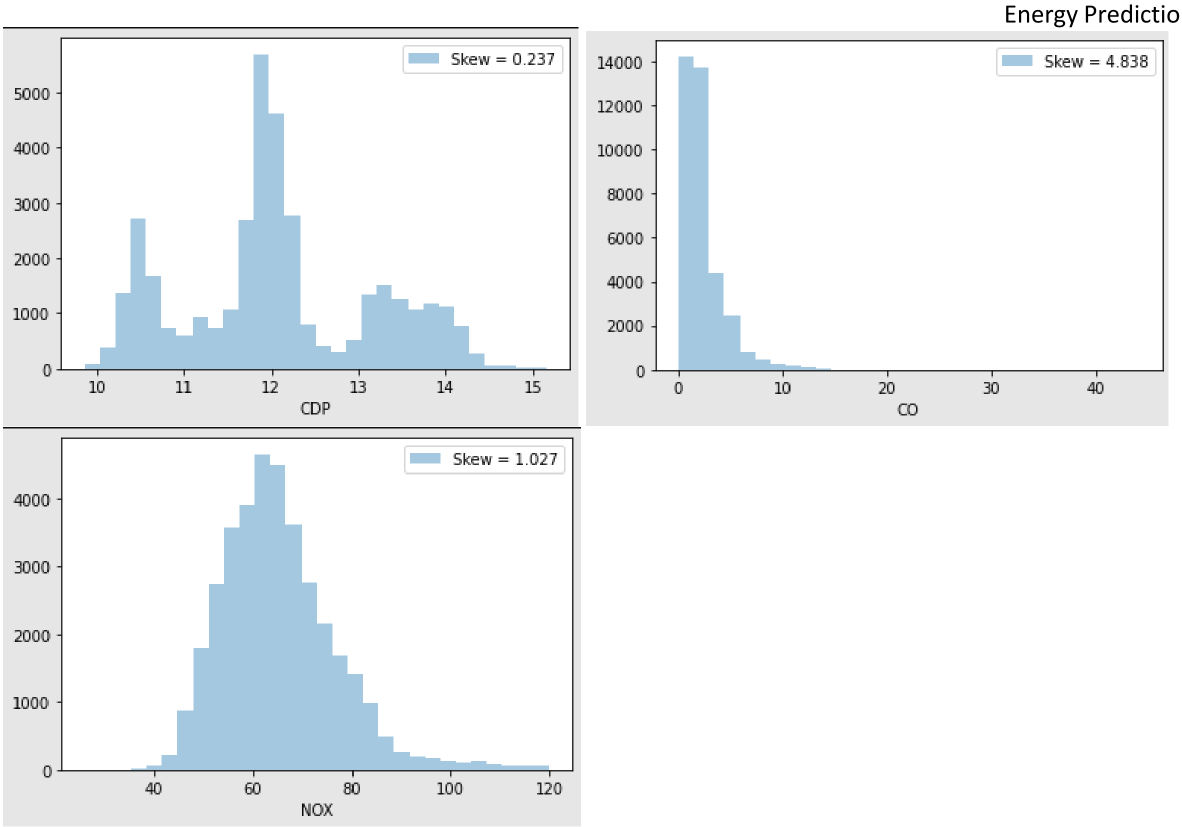


Figure 7 Skewness of all features

Observation:

The skewness in feature may violate model assumptions or may reduce the interpretation of feature importance. From above plots, we can observe that:

- In the features AT, AP, AFDP, GTEP, TEY and CDP, the data are fairly symmetrical as the skewness values are in between -0.5 and 0.5.
- TIT and AH are having moderate negative data skewed as skewness values are in between -1 and -0.5.
- Whereas CO and NOx data are highly positively skewed as skewness values are greater than 1.
- TAT feature’s skewness value is more than -1 so data is highly negatively skewed.

Feature Selection:

It is used to reduce the input variable to the model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for machine learning model based on the type of problem to be solved.

Benefits:

1. Reduces Overfitting:
   Less redundant data means less opportunity to make decisions based on noise.
2. Improves Accuracy:
   Less misleading data means modeling accuracy improves.
3. Reduces Training Time:
   Less data means that algorithms train faster.

For this we have used extra-tree-regressor from sklearn libraries. This class implements a Meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Property feature_importances_:

The higher, the more important the feature. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature.

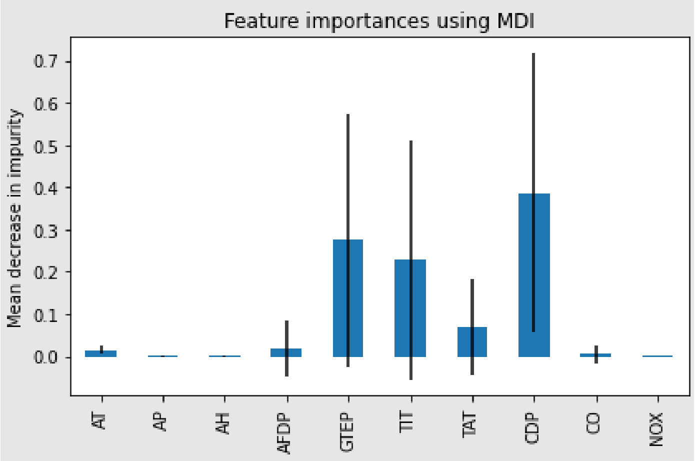


Figure 8 Feature importance

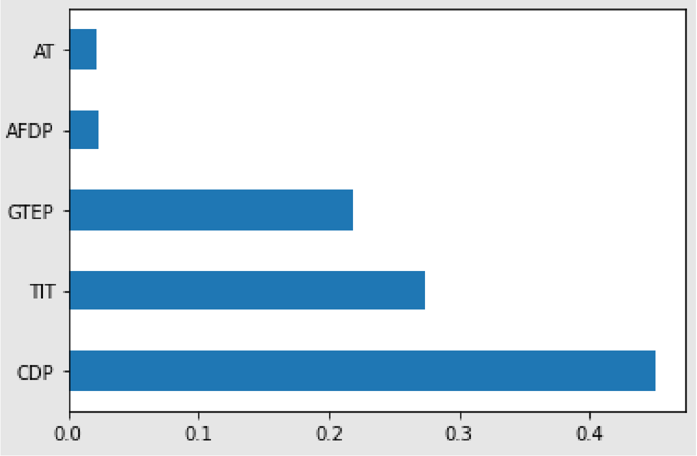


Figure 9 Top 5 important features

2.4 Model Building:

1. Train/Test split:

One important aspect of all machine learning models is to determine their accuracy.Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that modeland hence, find the accuracy of the model. A better option is to split our data into two parts: first one for training our machine learning model, and second one for testing our model.

- Split the dataset into two pieces: a training set and a testing set.
- Train the model on the training set.
- Test the model on the testing set, and evaluate how well our model did.

Advantages of train/test split:

- Model can be trained and tested on different data than the one used for training.
- Response values are known for the test dataset, hence predictions can be evaluated
- Testing accuracy is a better estimate than training accuracy of out-of-sampleperformance.

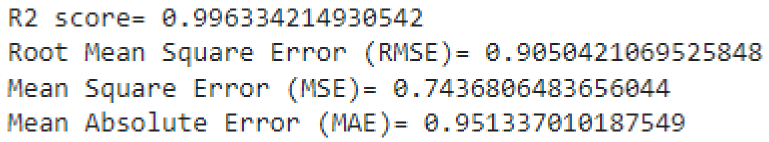
Machine learning consists of algorithms that can automate analytical model building. Using algorithms that iteratively learn from data, machine learning modelsfacilitate computers to find hidden insights from Big Data without being explicitly programmed where to look.

We have used the following algorithms to build predictive model.

2. Linear Regression:

It is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. The algorithm shows a linear relationship between a dependent (y) and one or more independent (X) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Model Results:



```
R2 score= 0.996334214930542
Root Mean Square Error (RMSE)= 0.9050421069525848
Mean Square Error (MSE)= 0.7436806483656044
Mean Absolute Error (MAE)= 0.951337010187549
```

Scatter plot:

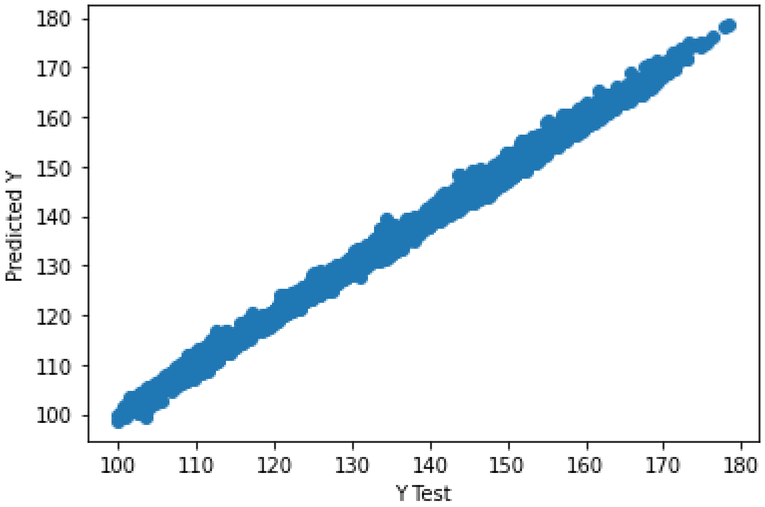
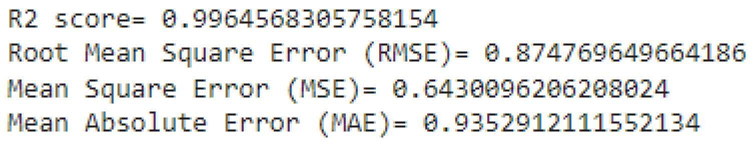


Figure 10 Scatter Plot (Y Test vs Predicted Y) For Linear Regression Model

3. Decision Tree:

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data). Each internal node of the tree corresponds to an attribute and each leaf node corresponds to a class label.

Model Results:



```
R2 score= 0.9964568305758154
Root Mean Square Error (RMSE)= 0.874769649664186
Mean Square Error (MSE)= 0.6430096206208024
Mean Absolute Error (MAE)= 0.9352912111552134
```

Scatter plot:

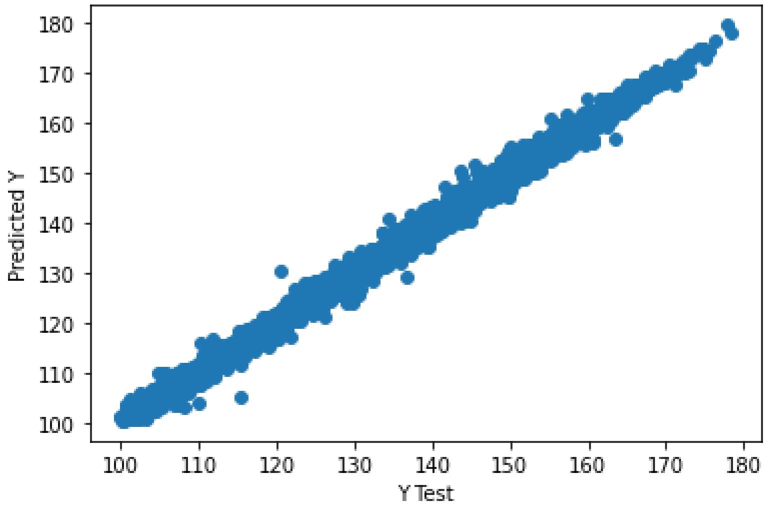


Figure 11 Scatter Plot (Y Test vs Predicted Y) For Decision Tree Model

4. Random Forest:

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called Aggregation.

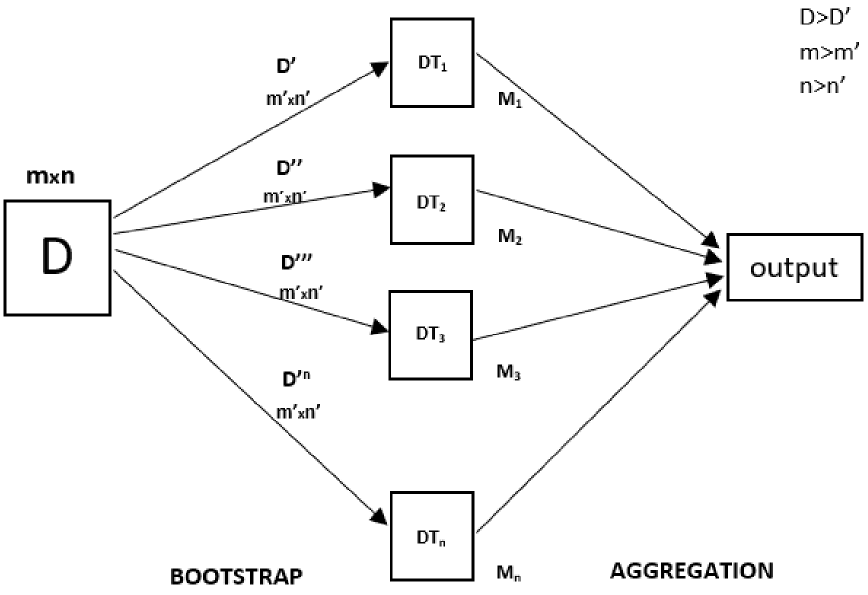


Figure 12 Random Forest

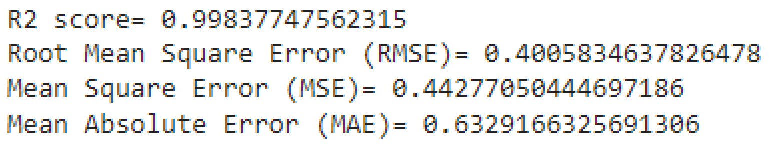
Hypertunning:

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process.

Grid search is arguably the most basic hyperparameter tuning method. With this technique, we simply build a model for each possible combination of all of the hyperparameter values provided, evaluating each model, and selecting the architecture which produces the best results.

RandomizedSearchCV randomly passes the set of hyperparameters and calculate the score and gives the best set of hyperparameters which gives the best score as an output.

Model Results:



```
R2 score= 0.99837747562315
Root Mean Square Error (RMSE)= 0.4005834637826478
Mean Square Error (MSE)= 0.44277050444697186
Mean Absolute Error (MAE)= 0.6329166325691306
```

Scatter plot:

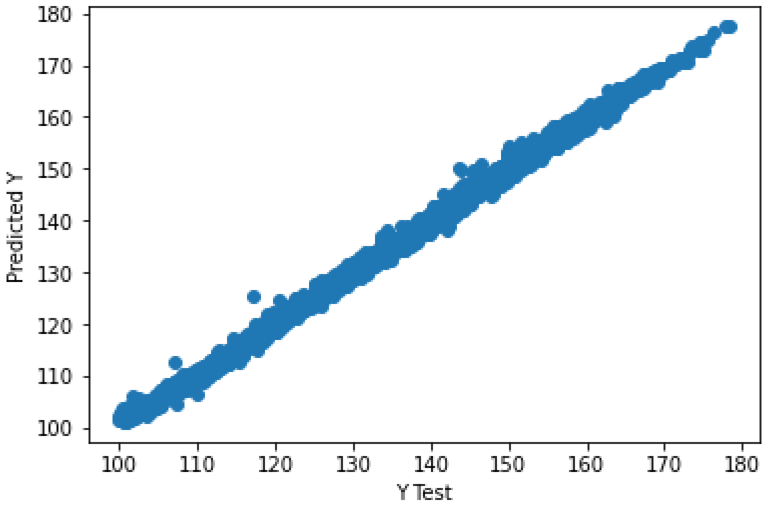
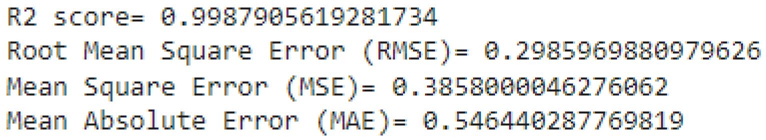


Figure 13 Scatter Plot (Y Test vs Predicted Y)
For Random Forest Model

5. XGBoost:

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is an optimized distributed gradient boosting library. In prediction problems involving unstructured data (text) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

Model Results:



```
R2 score= 0.9987905619281734
Root Mean Square Error (RMSE)= 0.2985969880979626
Mean Square Error (MSE)= 0.3858000046276062
Mean Absolute Error (MAE)= 0.546440287769819
```

Scatter plot:

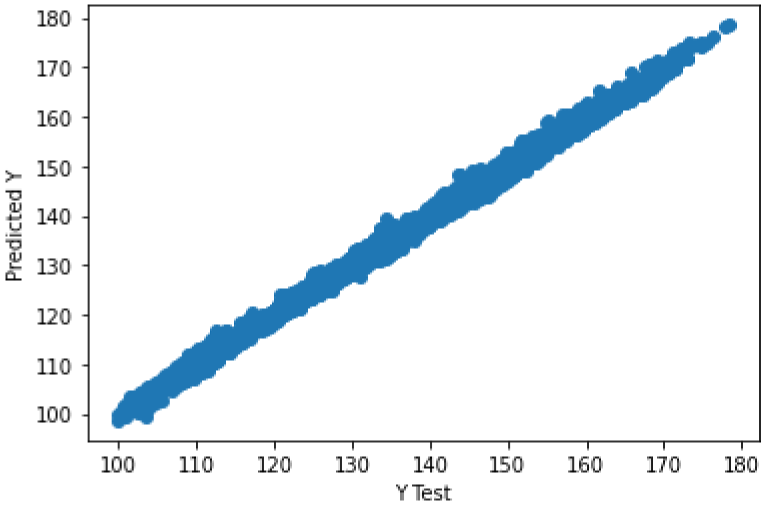


Figure 14 Scatter Plot (Y Test vs Predicted Y) For XGboost Model

6. Neural Network:

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Artificial neural networks (ANNs) are composed of a node layer, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

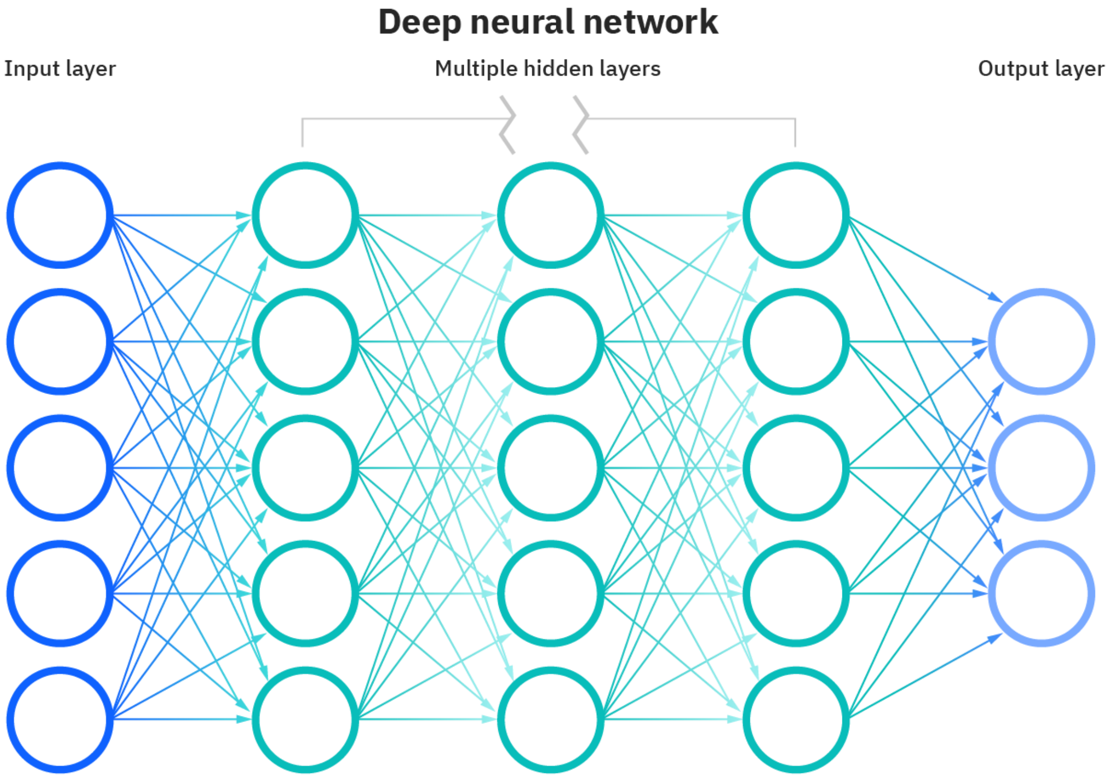
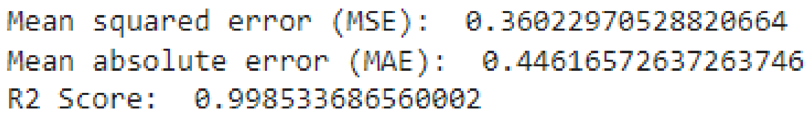


Figure 15 Deep Neural Network

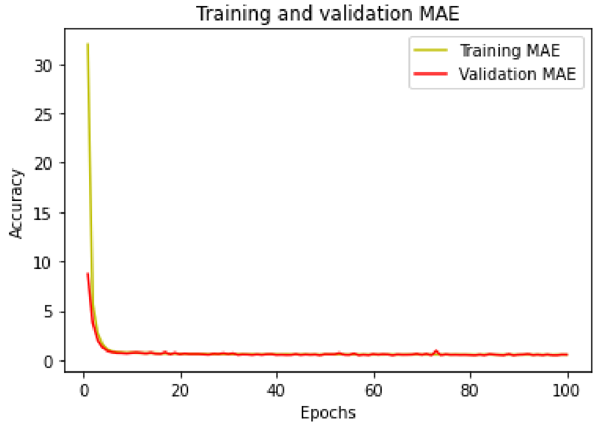
Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the most well-known neural networks is Google's search algorithm.

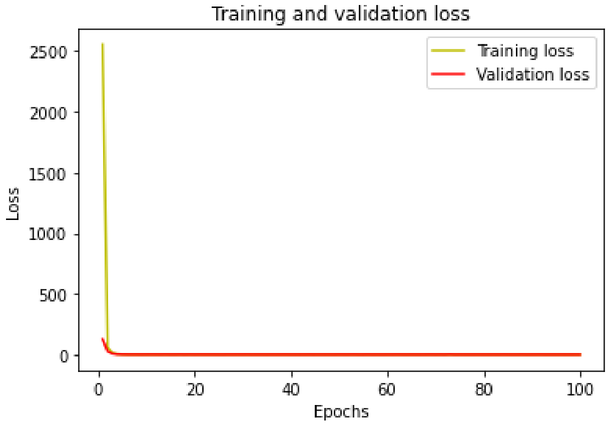
Model Results:



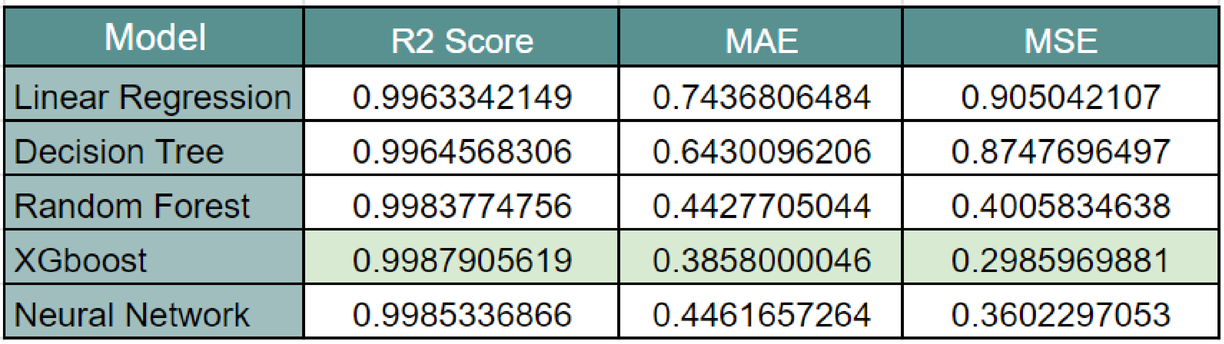
```
Mean squared error (MSE):  0.36022970528820664
Mean absolute error (MAE):  0.44616572637263746
R2 Score:  0.998533686560002
```

Plots:





All Model Result Comparison:



Model	R2 Score	MAE	MSE
Linear Regression	0.9963342149	0.7436806484	0.905042107
Decision Tree	0.9964568306	0.6430096206	0.8747696497
Random Forest	0.9983774756	0.4427705044	0.4005834638
XGboost	0.9987905619	0.3858000046	0.2985969881
Neural Network	0.9985336866	0.4461657264	0.3602297053

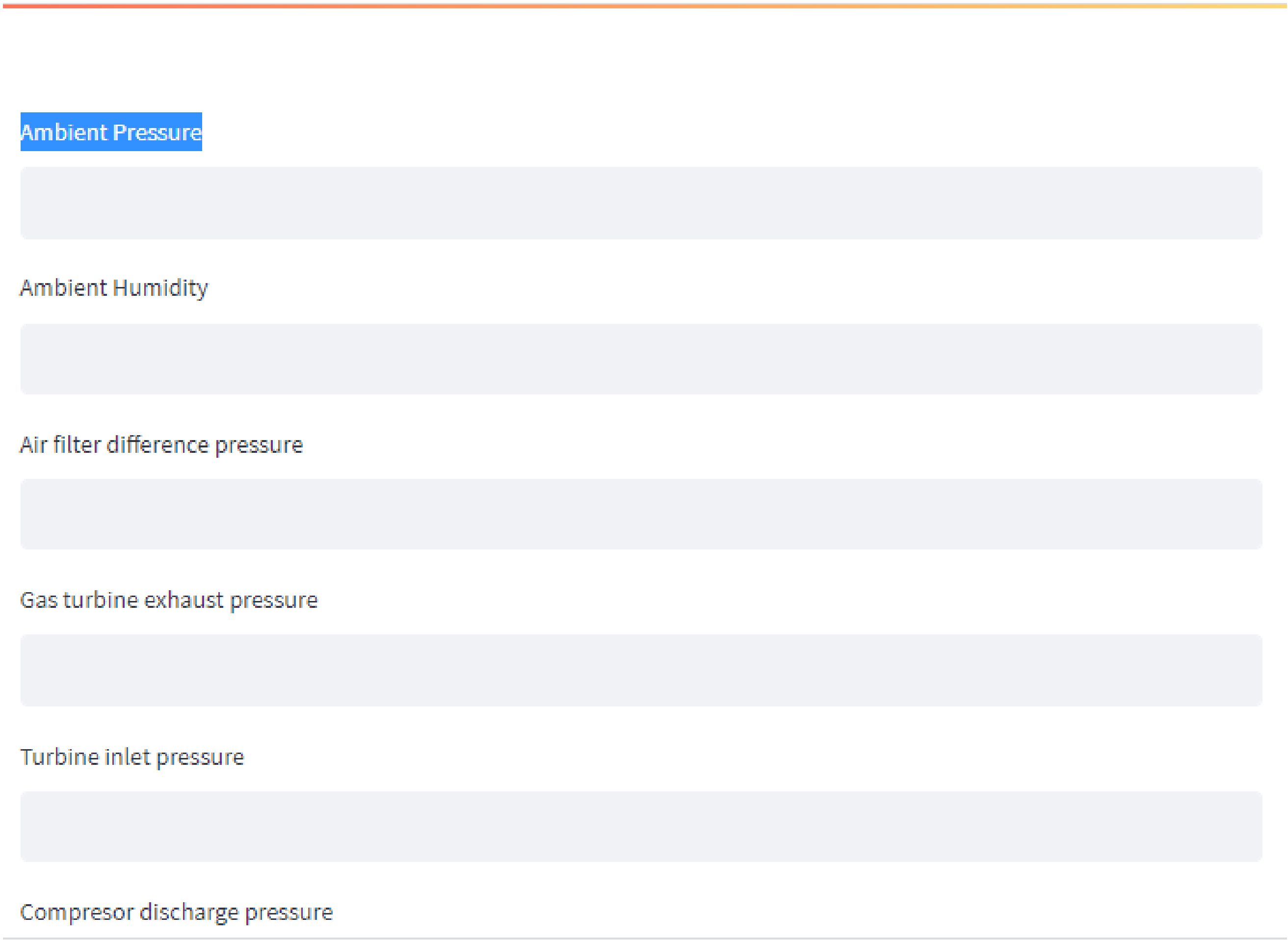
3. User Interface

After Training the models and finding out the best model to be Energy Prediction Model we can go ahead with building the user interface for our models. This will give us a clean and simple way of accessing our models and make the predictions depending on our inputs.

For building the user interface we adopted the **Streamlit** and **Flask Web Framework** which is very easy to use and robust.

Here are a few screenshots of the local and web app delivering the predictions for Turbine energy yield.

Local Application:



Ambient Pressure

Ambient Humidity

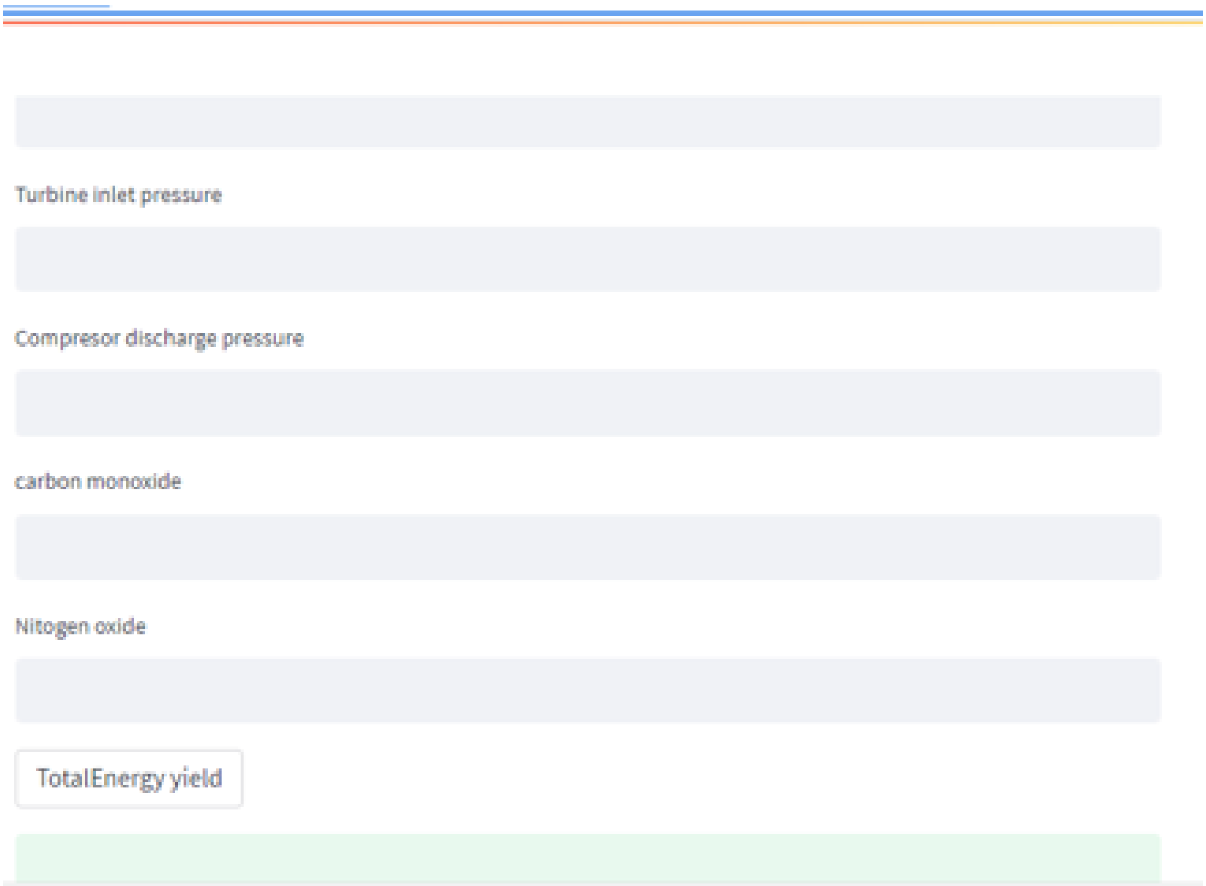
Air filter difference pressure

Gas turbine exhaust pressure

Turbine inlet pressure

Compresor discharge pressure

Figure 16 Energy Prediction Model Local application



Turbine inlet pressure

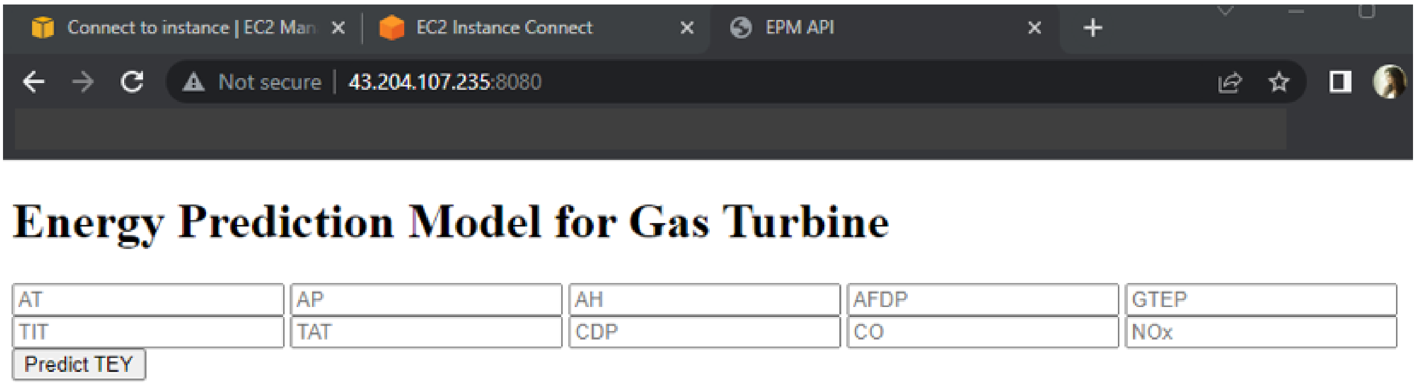
Compresor discharge pressure

carbon monoxide

Nitogen oxide

Figure 17 Energy Prediction Model Local application

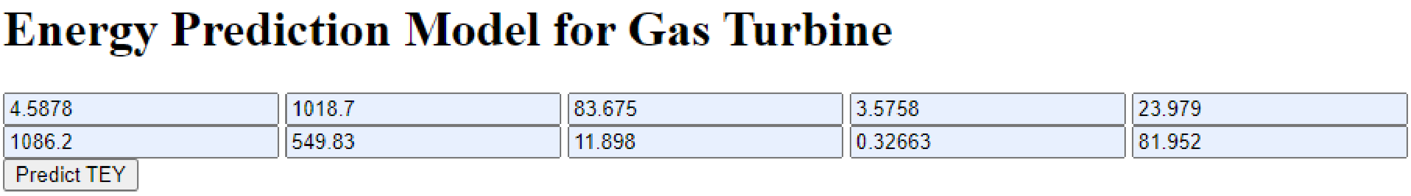
Web Application:



Energy Prediction Model for Gas Turbine

AT	AP	AH	AFDP	GTEP
TIT	TAT	CDP	CO	NOx

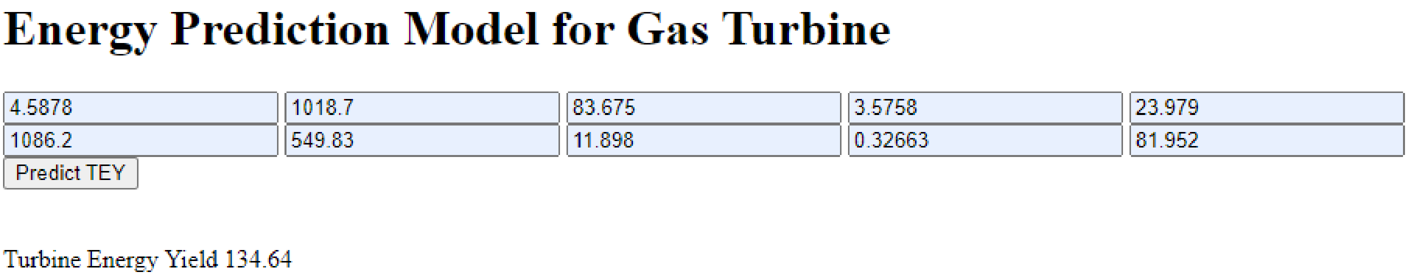
Figure 18 Web application interface



Energy Prediction Model for Gas Turbine

4.5878	1018.7	83.675	3.5758	23.979
1086.2	549.83	11.898	0.32663	81.952

Figure 19 Input to the Energy Prediction Model



Energy Prediction Model for Gas Turbine

4.5878	1018.7	83.675	3.5758	23.979
1086.2	549.83	11.898	0.32663	81.952

Turbine Energy Yield 134.64

Figure 20 Predicted output of Turbine Energy Yield

4. Requirements Specification

4.1 Hardware Requirement:

- 500 GB hard drive (Minimum requirement)
- 8 GB RAM (Minimum requirement)
- PC x64-bit CPU

4.2 Software Requirement:

- Windows/Mac/Linux
- Python-3.9.1
- VS Code/Anaconda/Spyder
- Python Extension for VS Code
- Libraries:
  - Numpy 1.18.2
  - Pandas 1.2.1
  - Matplotlib 3.3.3
  - Scikit-learn 0.24.1
  - Flask 1.1.2
- Any Modern Web Browser like Google Chrome
  - To access the web application written in Streamlit

5. Conclusion:

- In this project we have built a system for predicting value of Turbine Energy Yield with respect to all required features.
- For Model building we gathered the data from uci.com
- We cleaned the data and performed exploratory data analysis to understand the given data and relationship among those data.
- Built a different models on the cleaned dataset. Found out XGboost is the best model.
- Built a web application using the framework Flask to easily use the model and Streamlit used for local application.
- This will help decision makers to take their decisions about selection of the location of gas turbine power plants by considering ambient conditions and other key features.

6. Future Scope

On the basis of the work done in Energy Prediction Model, there is scope for future development which can be as follows:

- The project has achieved the objective of predicting the turbine energy yield but it can be modified to predict the CO and NOx emission released to the environment.
- By combing different models for predicting different features, the end user can have the choice of parameter to be predicted with available features.
- Different correlations between features can be observed by those different models and one can analyze the factors affecting on efficiency, emission rate.
- This further can be used to control the required related parameters to lower the emission or to achieve the required efficiency of any gas turbine power plant.

7. References

- Dataset - https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set#
- Novel data and a benchmark PEMS - https://journals.tubitak.gov.tr/elektrik/vol27/iss6/53/
- Technical Condition Estimation of the Gas Turbine Axial Compressor- https://iopscience.iop.org/article/10.1088/1755-1315/990/1/012037