

Gas Turbine Emission Prediction

Submitted By: Lajpat Rai

Student ID: 501144138

Table of Contents

[Introduction 3](#_Toc107151146)

[Literature Review 4](#_Toc107151147)

[Methodology 5](#_Toc107151148)

[Data Description 6](#_Toc107151149)

[Exploratory Data Analysis 8](#_Toc107151150)

[Multivariate analysis of parameters 10](#_Toc107151151)

[Features Selection 13](#_Toc107151152)

[Filter Based Feature Selection Technique 13](#_Toc107151153)

[Feature Selection with the help of Correlation 13](#_Toc107151154)

[Feature Selection Based on Mutual Information Gain 15](#_Toc107151155)

[Wrapper Method Feature Selection Technique 17](#_Toc107151156)

[1. Step Forward Feature Selection 17](#_Toc107151157)

[2. Backward Elimination Feature Selection 19](#_Toc107151158)

[3. Standard Stepwise Feature Selection 21](#_Toc107151159)

[Embedded Method Feature Selection Technique 21](#_Toc107151160)

[LASSO 22](#_Toc107151161)

[Random Forest Regression 25](#_Toc107151162)

[References 26](#_Toc107151163)

# Introduction

A gas turbine is mainly used to generate electricity around the world. Due to rapid industrialization, humans have expelled copious amounts of CO and NOx emissions into the atmosphere causing environmental pollution and declined air quality. However, due to increased awareness of greenhouse gases and stringent environmental regulations, the emissions of the gas- turbines are controlled at certain limits according to the environmental regulations.

Generally, sensors are installed to measure the concentration of various gases continuously known as CEMS (Continuous Emissions Monitoring System). Nevertheless, the tendency to get erroneous data of gases emitted by gas turbine is high when sensors start to malfunction or fail to work. Also, continuous data can’t be captured in case sensors need to be replaced or overhauled. Furthermore, validity of gases’ concentration measured by CEMS is challenged in case of harsh environment emission sample collection system. In order to overcome these issues, Predictive Emission Monitoring Systems (PEMS) can deliver reliable real-time emission estimations exploiting regression models.

In addition to provide estimated emission data reducing initial investment, PEMS can be used to optimize plant production while minimizing plant emissions. The quantification of sensor failure is important to achieve reduced overhaul cost and predict the failure rate to optimize the production by reducing the unplanned downtime. There is no acceptable universal mathematical model that describes equipment wear or damage accurately because of complex nature of emission behavior coupled with multitude of operating parameters/conditions.

Combining both methods and using them simultaneously is an efficient way to address this limitation.

With technology advancements and a growing network of sensors that enable faster and higher frequency data gathering, there is a need to sustainably scale data management systems to the level required to handle volumes of data and faster processing response. The most comprehensive systems, however, are those that support efficient decision making. Accurate emission models of gas turbines can be derived using machine learning techniques that will help us to forecast the concentration of gases emitted from gas turbines allowing decision makers to take business decisions in performing a major overhaul.

|  |
| --- |
| The dataset contains 36733 instances of 11 sensor measures aggregated over one hour, from a gas turbine located in Turkey for the purpose of studying flue gas emissions, namely CO and NOx. |

Python will be used for applying various regression models to see the relationship or impact of gases emitted from gas turbine on CO and NOx emissions. Accuracy of models will be captured in order to see which model work best for estimating the emissions.

<http://archive.ics.uci.edu/ml/datasets.php?format=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table>

# Literature Review

An overview of the history of PEMS development as well as the regulatory framework can be found in [1]. There are two main approaches to model building for PEMS: analytical equations derived from the laws of thermodynamics, mass and energy balance, and data-driven models, including statistical and machine learning methods. The publication of the dataset in the open data repository of the University of California, Irvine, School of Information and Computer Science [2] in 2019 allowed researchers to start open discussion on the features of data and the quality of models derived from various machine learning methods to predict CO and NOx emissions. In [3], the dataset under study was presented for the first time and a description of its main statistical characteristics was given. The dataset was collected for 5 years which contained 36,733 instances of 11 sensor measures aggregated over one hour, including three external environmental parameters (air temperature (AT), air humidity (AH), atmosphere pressure (AP)), indicators of the gas turbine technological process and two target variables: the sensor measurements of emissions of carbon monoxide (CO) and the total nitrogen monoxide and nitrogen dioxide (NOx). The data were collected in an operating range between partial load (75%) and full load (100%).

At first, extreme learning machine classifiers (ELMs) for this problem were used to estimate NOx & CO emissions. The hyperparameters (the number of hidden nodes K and the regularization parameter C) and three fusion strategies were examined, and the best result has a coefficient of determination R2 = 0.56 and mean absolute error (MAE) 0.97 mg/m3 for CO prediction, and R2 = 0.67 and MAE = 4.57 mg/m3 for NOx prediction.

Subsequent articles by other researchers explored various machine learning techniques to produce better performing models. In [4], K-nearest-neighbor algorithm, based on the same dataset for predicting NOx emissions from the natural gas electrical generation turbines is proposed. In [5], the model of PEMS, using a gradient boosting machine learning method, is presented. The dataset from the continuous emission monitoring system (CEMS) with a sampling rate of 1 min for this research is not publicly available. In this study, the authors note that ANN-based models are treated as “black boxes” and regulators and decision makers without a statistical background often have difficulty understanding these models, which poses a significant challenge for a broader application of PEMSs. In [6], the M5P algorithm was used to predict CO emissions, and a binary decision tree with linear regression functions at the leaf nodes was built. The reported MAE predicting CO emissions ranged from 0.75 to 1.4 mg/m3 and that predicting NOx emissions ranged from 4.2 to 11 mg/m3. The advantage of the method is that the models are suitable for human interpretation.

In [7], the problem of choosing a feature normalization method and its impact on ANN-based models was investigated. Three datasets were examined, including CO and NOx emissions. Each dataset is specific, three methods of feature normalization showed similar results for the dataset under consideration. The authors calculated the proportional dispersion weights for each feature to improve the understanding of the features’ contribution to the model. The performance of the presented models is not perfect: average MAE = 0.73, R2 = 0.56 for CO and MAE = 5.4, R2 = 0.55 for NOx. In [3], the authors calculated the main statistical characteristics of the variables and found a strong correlation between the input variables, particularly between the compressor discharge pressure and the turbine energy yield (ρ(CDP, TEY) = 0.99), CDP and GTEP (ρ(CDP, GTEP) = 0.98), and also GTEP and TEY (ρ(GTEP, TEY) = 0.96). Thus, the problem of excluding variables containing redundant information is considered by most researchers. The idea of reducing the number of predictors using principal components analysis (PCA) was discussed in [4,6]. In [3], a canonical correlation analysis (CCA) was used — a method which uses two principal components to predict two explanatory variables.

In [8], the authors presented a new class of reliable-based multiple linear regression (MLR) models called Etemadi and evaluated the performance of the Etemadi model and classic MLR model using the same dataset to predict the hourly net energy yield (TEY) of the turbine with gas turbine parameters and the ambient variables as predictors. Given the high correlation between the predictors and the dependent variable, this dataset is not a valid example to support the conclusions of the article on the model for energy yield prediction.

It is worth noting that the problem of predicting power generation is vital and the Kalman filter has given good results for an open cycle gas turbine in [9] and for combined cycle gas turbine in [10]. The dataset under consideration requires a different method as emissions are more influenced by process parameters than temperature cycles.

The purpose of this project is to study the open dataset on CO and NOx emissions in order to choose suitable machine learning algorithms for emission predictions, selecting the features that are most important and investigate the quality of the resulting models as an explainable method of prediction.

# Methodology

This section describes the data exploration, inferences and machine learning approach that was applied and how they relate to the original business problem of gaining data insights specifically to identify attributes that are most important features and develop relation between them to predict the NOx & CO emissions. The methodology includes exploratory data analysis with the aid of histogram, boxplots, data visualization, regression analysis to investigate the influence of flue gas attributes as well as the choices and considerations within the methods.

Although, lot of work has already been done, this project study will try to select the most important features that have an impact on NOx & CO emissions and compare the results with and without features selection using various regression/classification machine learning models after checking accuracy of models.

Objective: To predict the gas emission CO & NOX from gas turbine

Data Collection

Data Cleaning

Exploratory Data Analysis

Data Preprocessing (Feature Extraction & Selection)

Algorithm Setup (with & without feature selection)

Model fitting

Prediction & Evaluation

Comparison of various ML models based on accuracy

# Data Description

The dataset is composed of hourly average sensor measurements of eleven variables (nine input and two target variables). There are a total of 36,733 instances collected over 5 years. The nine input measurements (independent variables) can be grouped into two as ambient variables (e.g., temperature, humidity, pressure) and process parameters (e.g., turbine energy yield, air filter difference pressure). The names, abbreviations and basic statistics of the variables used in the study are summarized in Table 1. Histograms for Carbon Monoxide (CO) and Nitrogen Oxides ( NOx) are given in Figure 1 suggesting that CO is right skewed and NOx is normally distributed. In Figure 2, the locations of sensors and sources of turbine parameters are shown on the illustration of the gas turbine.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Abbr.** | **Unit** | **Min** | **Max** | **Mean** |
| Ambient temperature | AT | ◦C | –6.23 | 37.1 | 17.71 |
| Ambient pressure | AP | mbar | 985.85 | 1036.56 | 1013.07 |
| Ambient humidity | AH | (%) | 24.08 | 100.2 | 77.87 |
| Air filter difference pressure | AFDP | mbar | 2.09 | 7.61 | 3.93 |
| Gas turbine exhaust pressure | GTEP | mbar | 17.7 | 40.72 | 25.56 |
| Turbine inlet temperature | TIT | ◦C | 1000.85 | 1100.89 | 1081.43 |
| Turbine after temperature | TAT | ◦C | 511.04 | 550.61 | 546.16 |
| Compressor discharge pressure | CDP | mbar | 9.85 | 15.16 | 12.06 |
| Turbine energy yield | TEY | MWH | 100.02 | 179.5 | 133.51 |
| Carbon monoxide | CO | mg/m3 | 0 | 44.1 | 2.37 |
| Nitrogen oxides | NOx | mg/m3 | 25.9 | 119.91 | 65.29 |

Table 1: Description of dataset

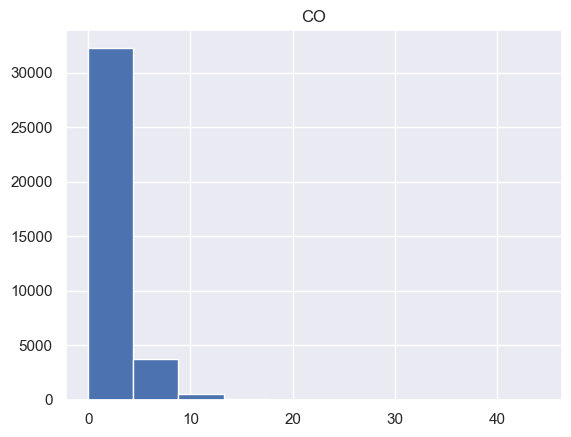
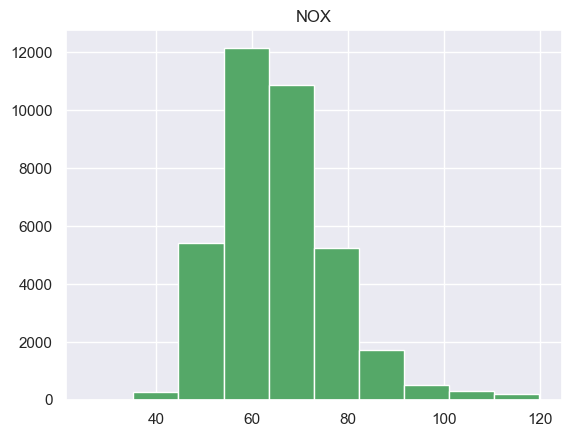
 

Fig 1: Histograms of CO & NOx

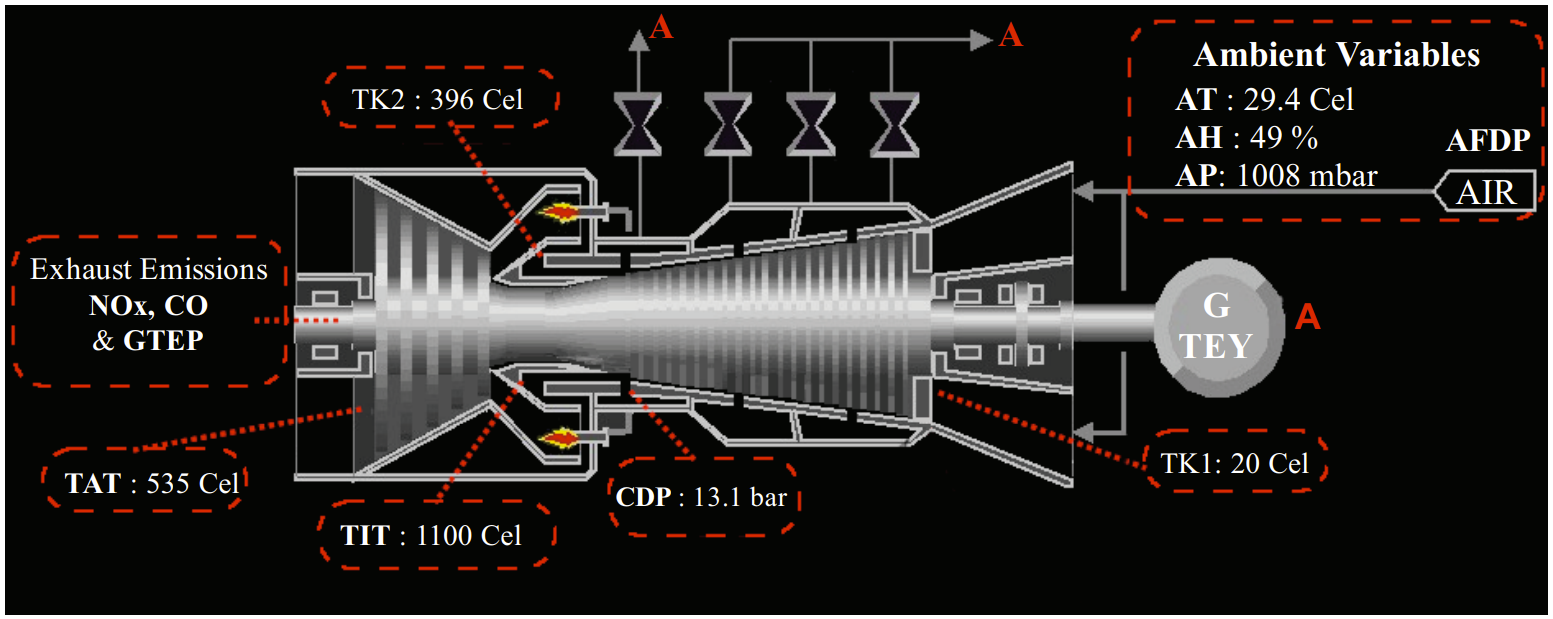
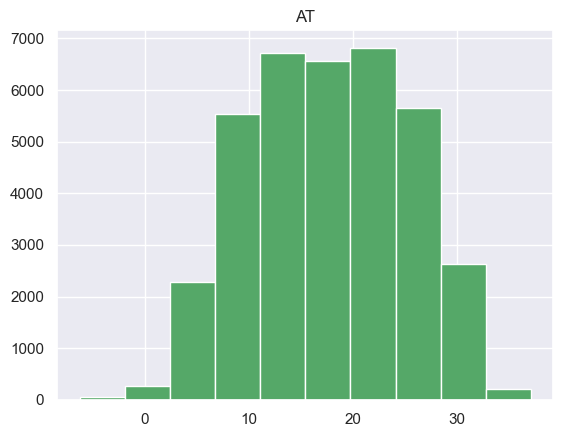
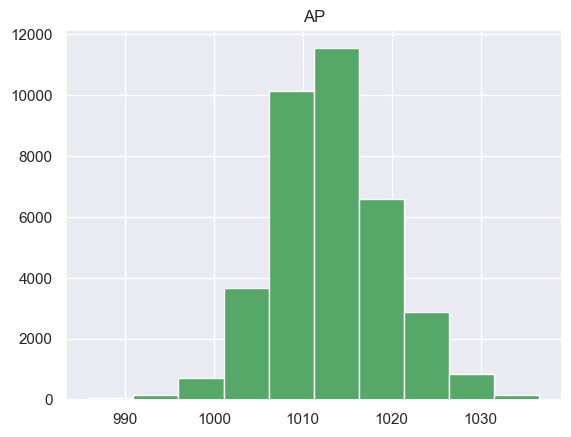
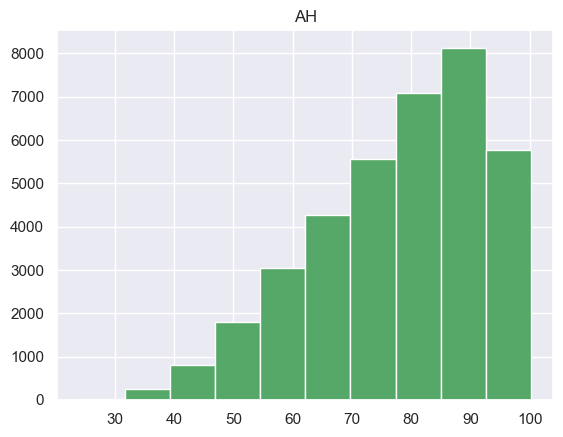
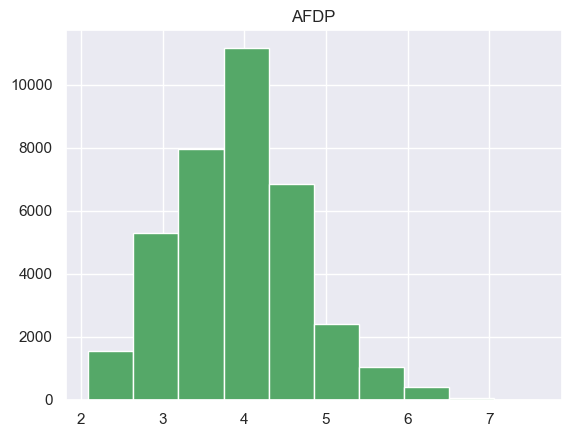


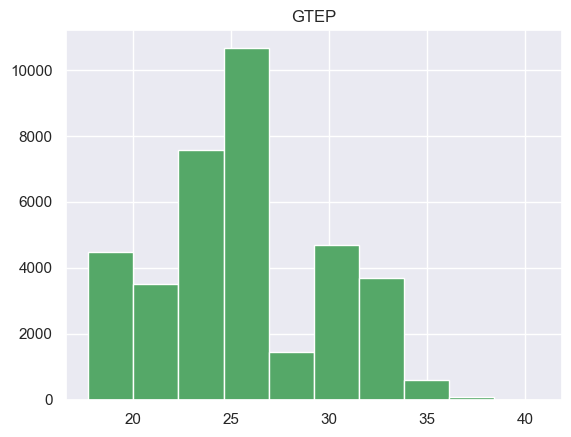
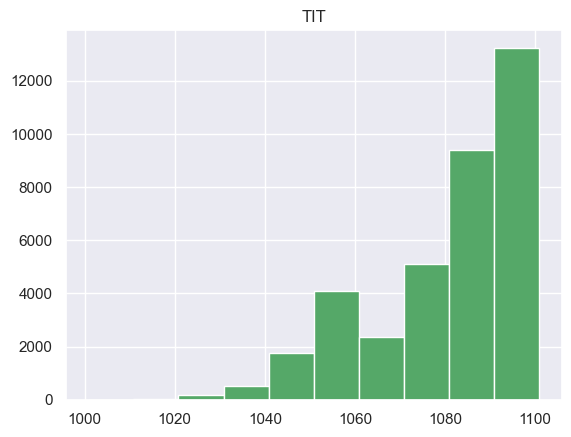
Fig 2. The sensor locations/parameter sources for all input and output variables used in the study. The parameters used are shown in dashed red rectangles, whereas the dashed white arrows show the sensor locations.

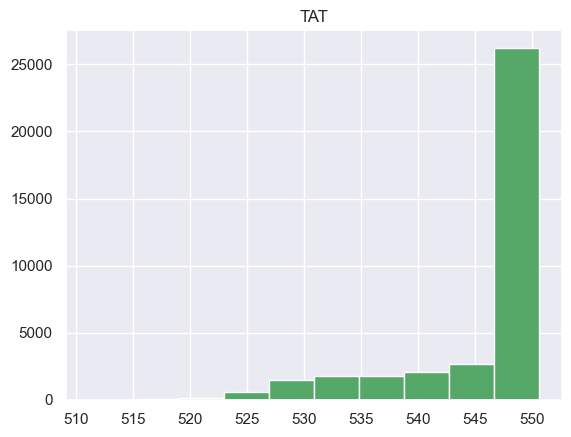
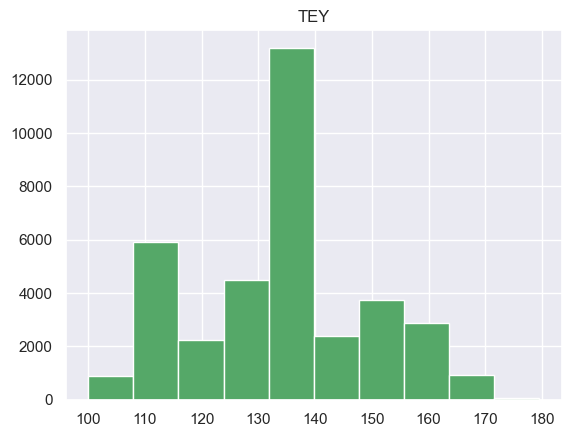
## Exploratory Data Analysis

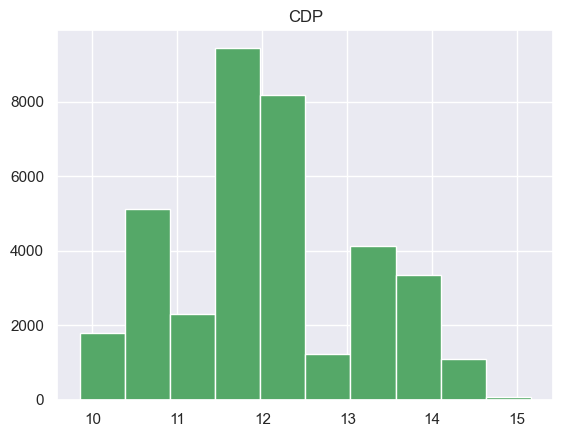
Exploratory data analysis was conducted utilizing histogram, boxplots & barplot. Dataset is clean and has no missing values. Few parameters are normal except TIT, TEY, TAT, GTEP, CO & AH. Furthermore, Standard deviation is small for all parameters hence most of data is close to mean. Data will be normalized before splitting it into training & testing data.

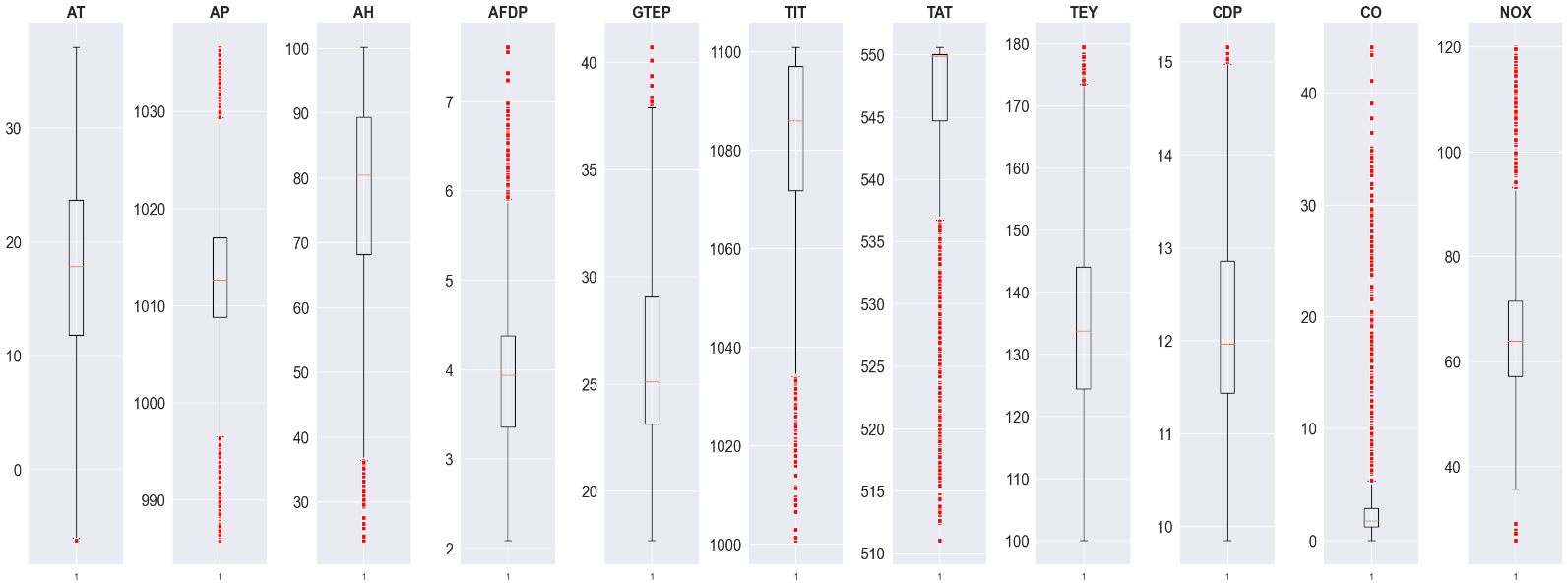
 





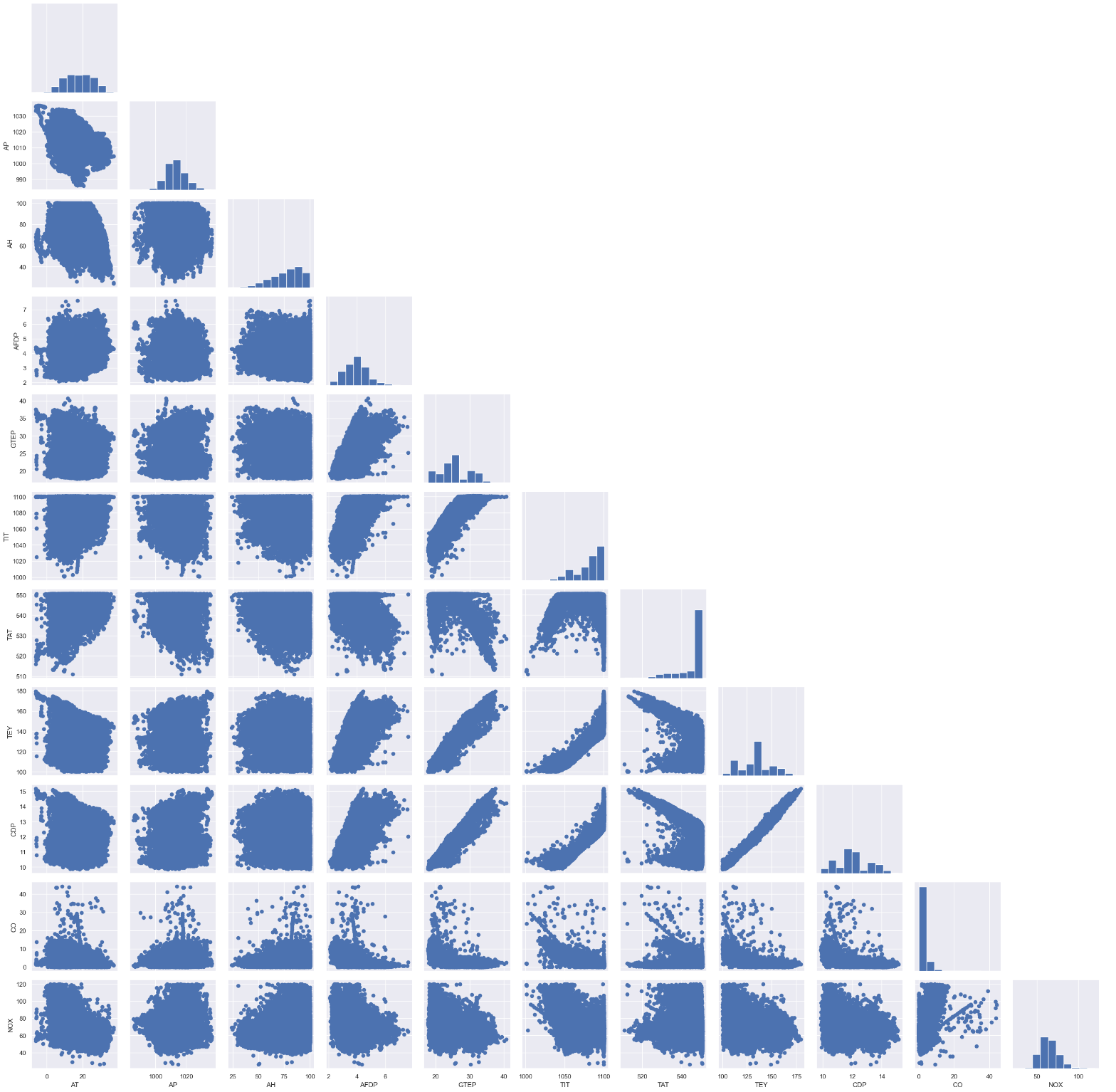
The boxplot shows that more input variables are outliers, so mean absolute error (MAE) will be used to evaluate the model in modeling. The MAE is not sensitive to the outliers.

## Multivariate analysis of parameters

Multivariate analysis is a set of techniques to analyze the multidimensional data, seeking patterns within data elements. Figure below illustrates the multivariable correlation between the variables using heatmap in numerical format. The red data points in the scatter plots indicate data readings with higher NOx values. As can be seen in Table 3, the higher values of NOx are usually clustered closer to lower temperatures and lower power production yields, which result lower turbine temperatures and pressures. Visual inspection of scatter plots and correlation values indicate the process variables are more correlated with each other and less with the weather condition parameters. Therefore, clustering parameters in a scientific and quantitative method can clarify the relationships in more details.

Github Link is https://github.com/Lajpat1985/CIND820-Project





# Features Selection

It is very important that the independent features to be highly related to the dependent one. But, this fails to happen in the real world and we may end up with a poor model. To help in improving the efficiency we need to perform feature selection.

Top reasons to use feature selection are:

* It enables the machine learning algorithm to train faster.
* It reduces the complexity of a model and makes it easier to interpret.
* It improves the accuracy of a model if the right subset is chosen.
* It reduces overfitting.

## Filter Based Feature Selection Technique

Filter methods are generally used as a preprocessing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable.

Set of all Features

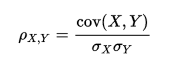
Selecting the best Subset

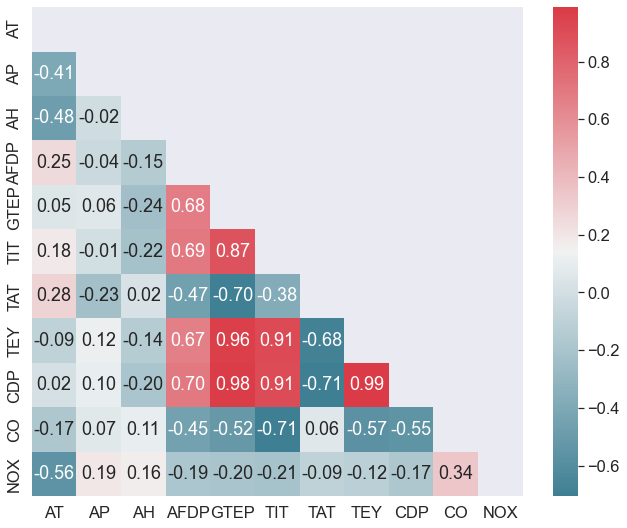
Learning Algorithm

Performance

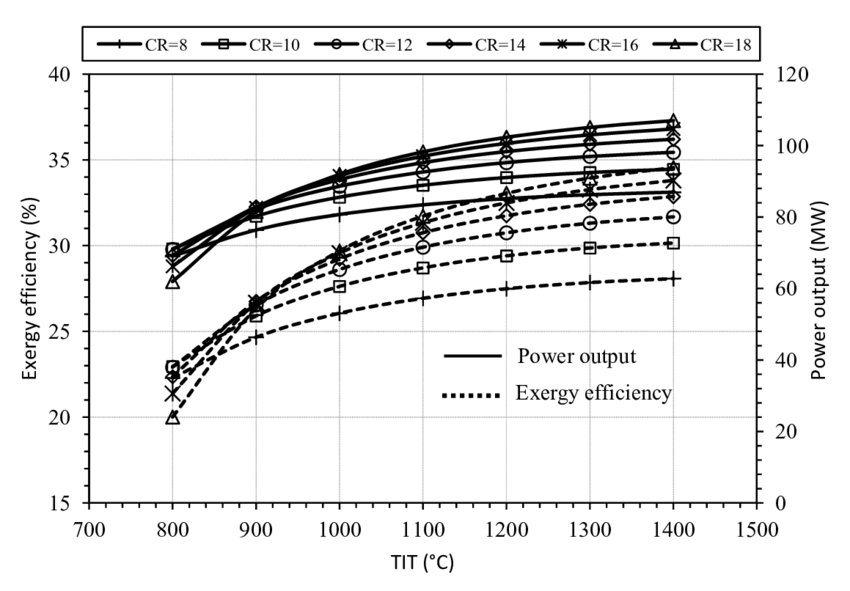
### Feature Selection with the help of Correlation

As input variables and output variables (NOx & CO) are in numerical data type so we will use Pearson’s correlation, which is used as a measure for quantifying linear dependence between two continuous variables X and Y. Its value varies from -1 to +1. Pearson’s correlation is given as:





The heatmap above shows the correlation between features and output variables (CO and NOx). It is easy to see that some features are negatively correlated each other. For example, the correlation between TIT and CO is -0.71, and it means that when the Turbine Inlet Temperature (TIT) decreases, the gas-turbine engine will produce more CO because a low TIT reduces the efficiency of the gas-turbine engine(look at the figure below).



Based on correlation matrix, top 5 features for output variable CO are TIT, TEY, CDP, GTEP & AFDP and top 5 features for output variable NOx are AT, TIT, GTEP, AP & AFDP.

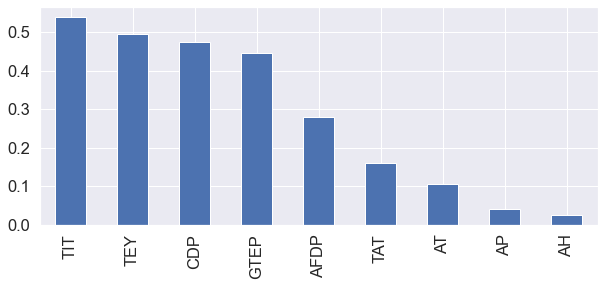
### Feature Selection Based on Mutual Information Gain

Mutual information(MI)between two random variables is a non-negative value, which measures the dependency between the variables .It is equal to zero if and only if two random variables are independent ,and higher values mean higher dependency.

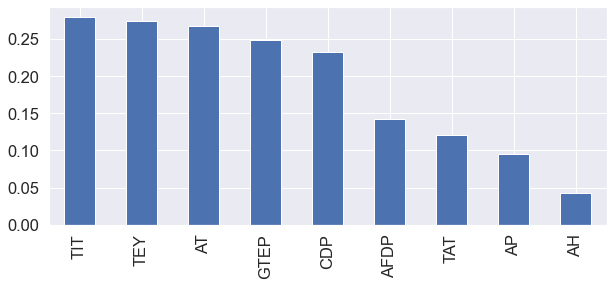
The mutual information between two random variables X and Y can be stated formally as follows:

I(X ; Y) = H(X) — H(X | Y)

Where I(X; Y) is the mutual information for X and Y, H(X) is the entropy for X, and H(X | Y) is the conditional entropy for X given Y. The result has the units of bits(zero to one).

From the graph above, we can infer that the TIT is having the highest mutual information gain(0.54) then TEY(0.49) followed by CDP(0.47), and so on. So TIT give 54% of the information about the target variable CO in this case. To select the top features we use another library called SelectKBest which picks up the top K features. There is also an option to pick up the top percentile of features. Here we select only the top 5 features.

['AFDP', 'GTEP', 'TIT', 'TEY', 'CDP']



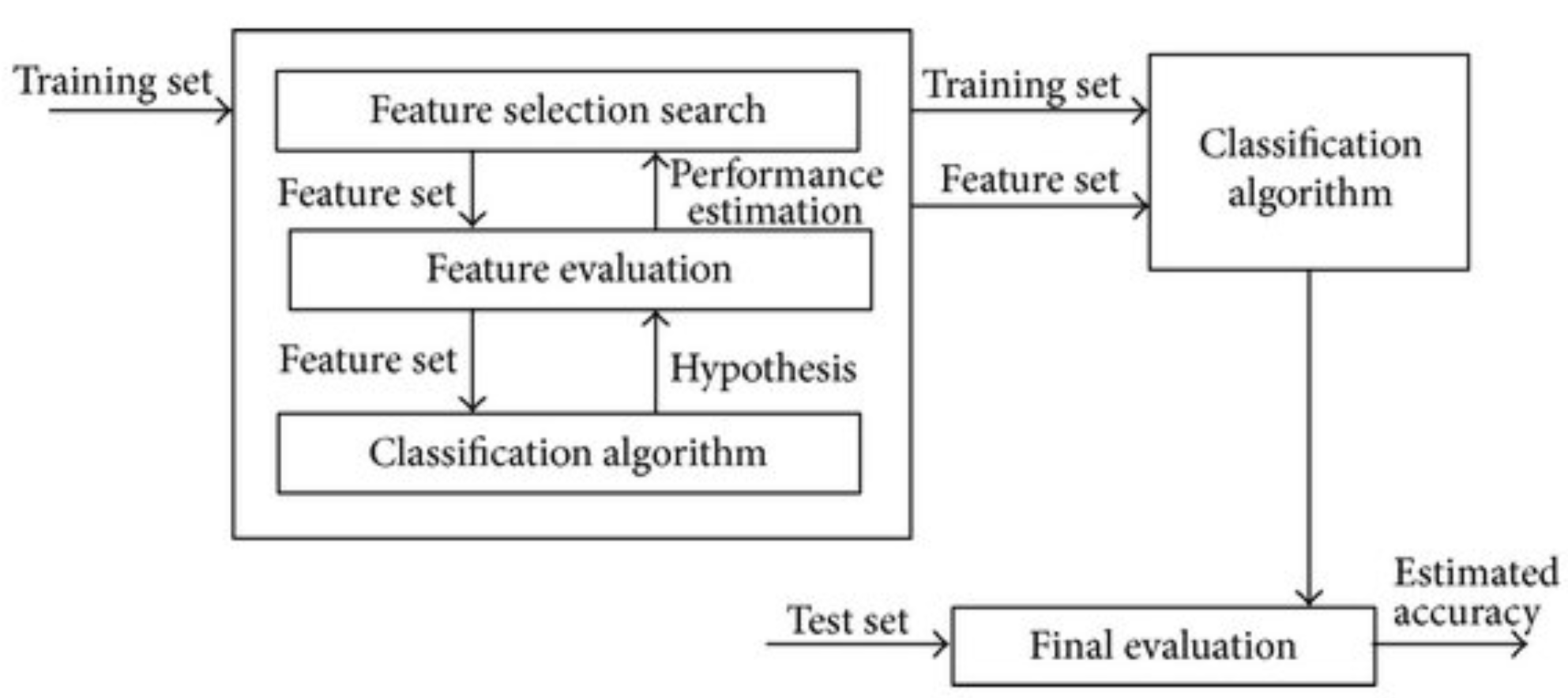
From the graph above, we can infer that the TIT is having the highest mutual information gain(0.28) then TEY(0.27) followed by AT(0.27), and so on. So TIT give 28% of the information about the target variable NOX in this case. To select the top features we use another library called SelectKBest which picks up the top K features. There is also an option to pick up the top percentile of features. Here we select only the top 5 features.

['AT', 'GTEP', 'TIT', 'TEY', 'CDP']

## Wrapper Method Feature Selection Technique

Wrapper methods measure the “usefulness” of features based on the classifier performance. wrapper methods are essentially solving the “real” problem (optimizing the classifier performance), but they are also computationally more expensive compared to filter methods due to the repeated learning steps and cross-validation.

It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion. The evaluation criterion is simply the performance measure which depends on the type of problem. In this case, evaluation criterion can be p-values, R-squared or Adjusted R-squared.



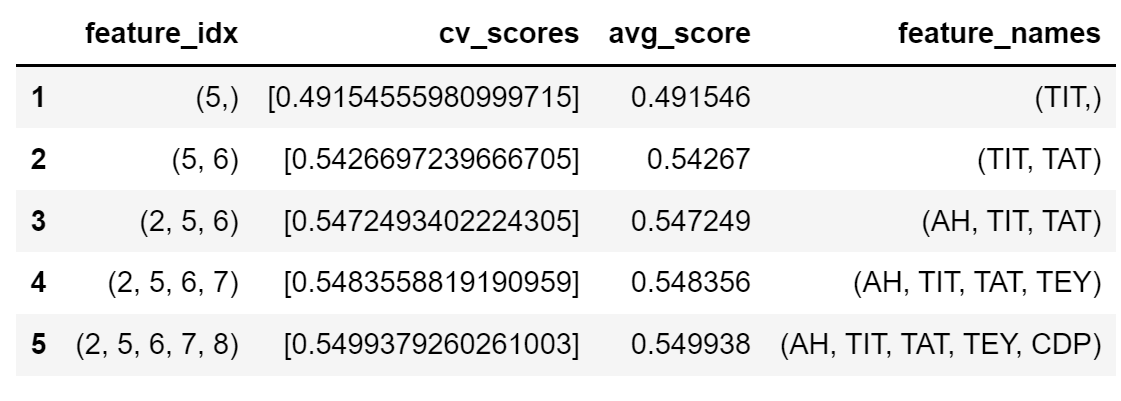
Wrapper methods for feature selection can be divided into three categories: **Step forward feature selection**, **Step backwards feature selection** and **Stepwise feature selection**.

### Step Forward Feature Selection

In the first phase of the step forward feature selection, the performance of the classifier is evaluated with respect to each feature. The feature that performs the best is selected out of all the features.

In the second step, the first feature is tried in combination with all the other features. The combination of two features that yield the best algorithm performance is selected. The process continues until the specified number of features are selected.

Table below is the output of independent top 5 features with respective R2 value for dependent variable CO.



Below is the plot of performance using SFS for dependent variable CO. It is evident that R2 reaches to maximum 0.55 after 5 – 6 top features as it is not changing afterwards.

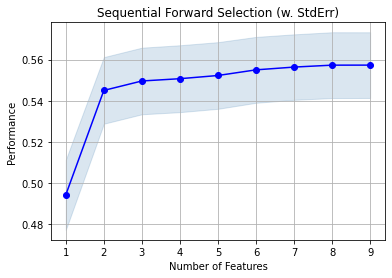
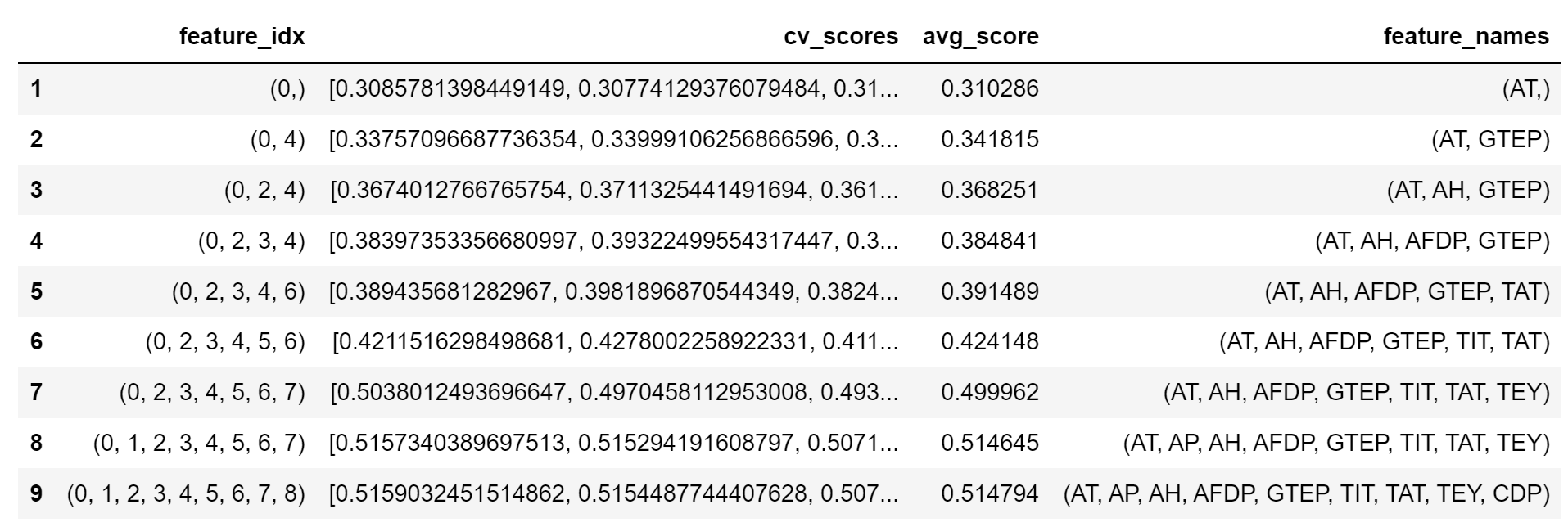
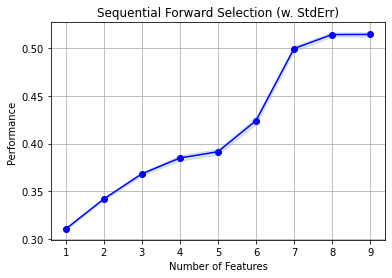


Table below is the output of independent top 5 features with respective R2 value for dependent variable NOX.



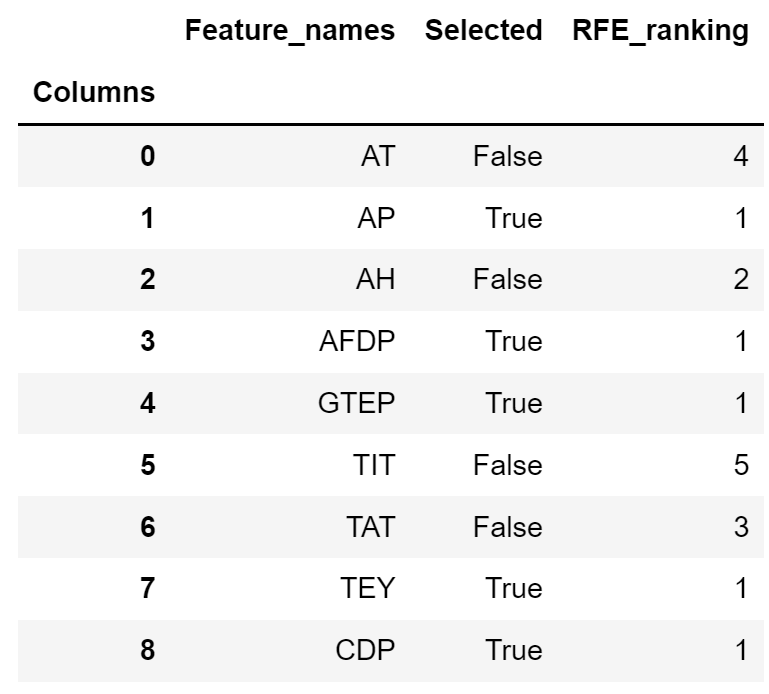
Below is the plot of performance using SFS for dependent variable NOX. It is evident that R2 reaches to maximum 0.52 after selecting all features.



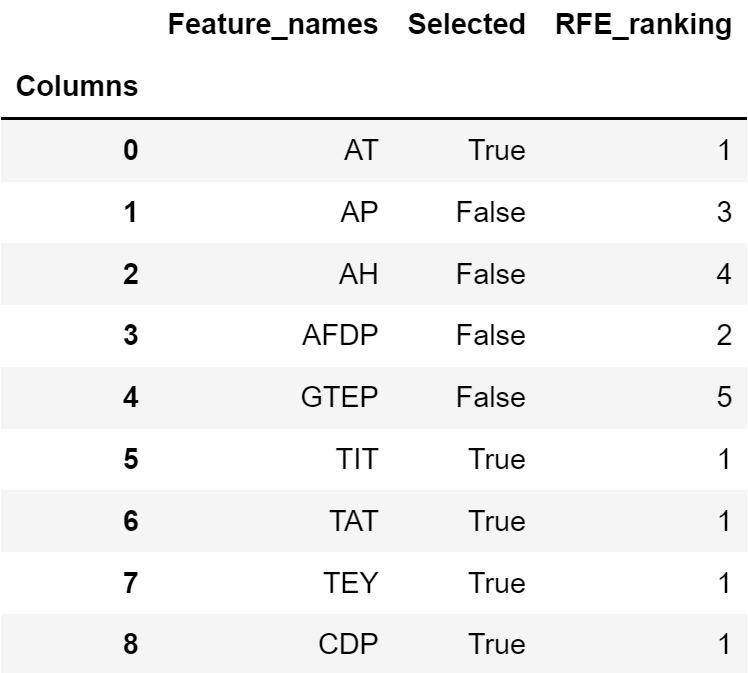
### Backward Elimination Feature Selection

In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

Below table suggests that 5 features that are most relevant using Recursive Feature Elimination for target variable CO. these 5 feature variables are [‘AP’, ‘AFDP’, ‘GTEP’, ‘TEY’, ‘CDP’].



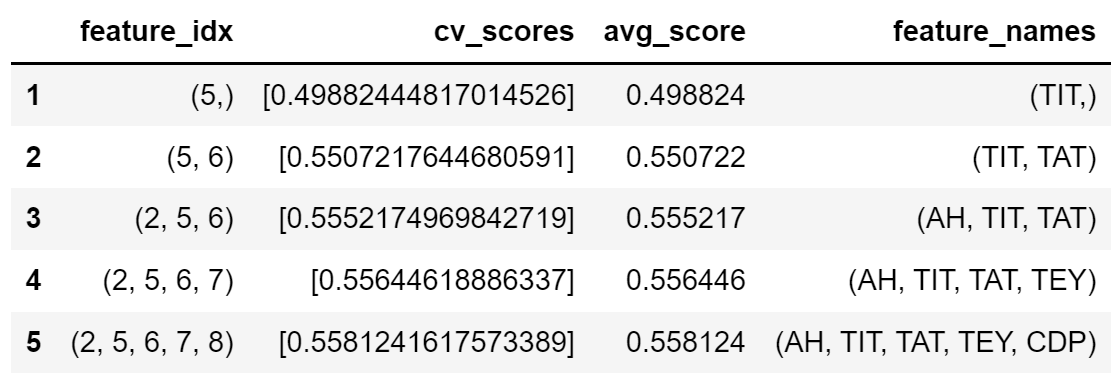
Below table suggests that 5 features that are most relevant using Recursive Feature Elimination for target variable NOX. these 5 feature variables are [‘AT’, ‘TIT’, ‘TAT’, ‘TEY’, ‘CDP’].



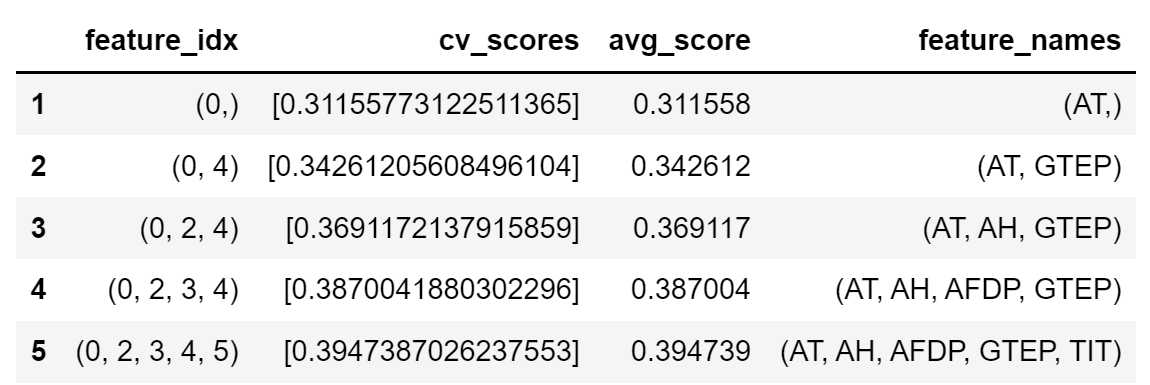
### Standard Stepwise Feature Selection

It is the most widely used Wrapper method for selecting features where at each step, features are added and removed on the basis of the improvement in R-square and thus acts as a combination of Forward Selection and Backward Elimination. As variables are constantly added and dropped considering the model’s accuracy the problem of multicollinearity is also resolved to a certain extent.

Below is the table using SFS for top 5 features that are highly correlated for target variable CO.



Below is the table using SFS for top 5 features that are highly correlated for target variable NOX.

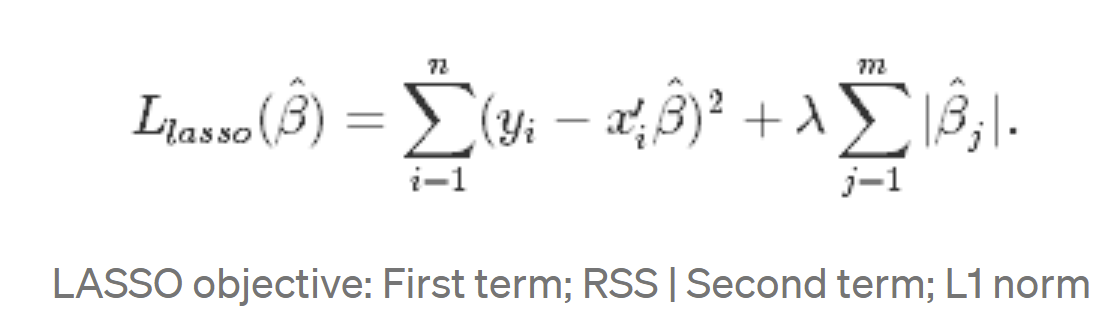


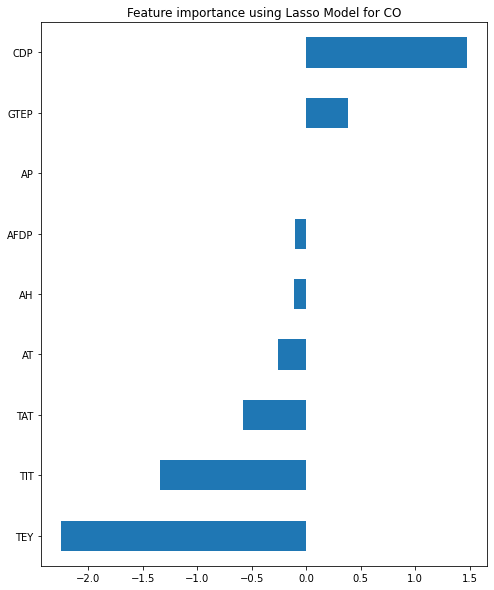
## Embedded Method Feature Selection Technique

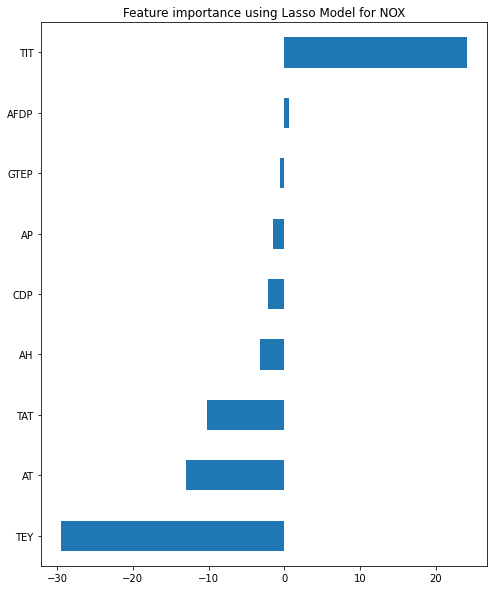
Embedded methods combine the advantageous aspects of both Filter and Wrapper methods. It **perform feature selection and training of the algorithm in parallel**. In other words, the feature selection process is an integral part of the classification/regressor model.

### LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) is a shrinkage method that performs both variable selection and L1 regularization at the same time. Regularization is a process that shrinks the coefficients (weights) towards zero that mean penalizing more complex models to avoid overfitting.







### Random Forest Regression

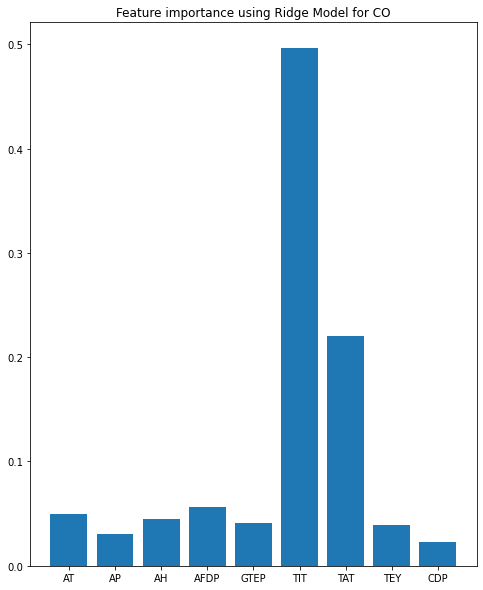
Random forest is an ensemble of decision trees. This is to say that many trees, constructed in a certain “random” way form a Random Forest.

* Each tree is created from a different sample of rows and at each node, a different sample of features is selected for splitting.
* Each of the trees makes its own individual prediction.
* These predictions are then averaged to produce a single result

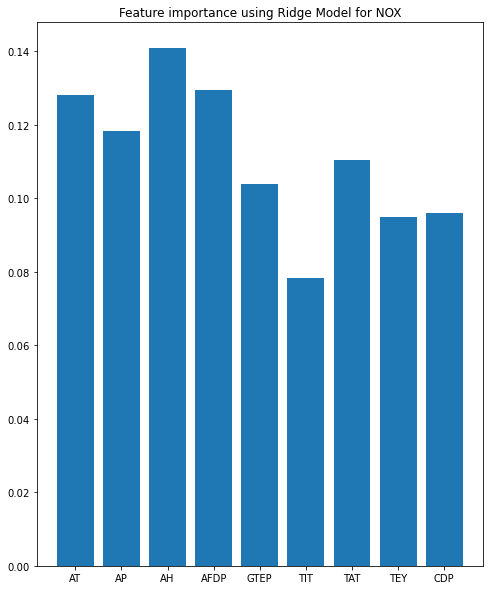
The averaging makes a Random Forest better than a single Decision Tree hence improves its accuracy and reduces overfitting.

A prediction from the Random Forest Regressor is an average of the predictions produced by the trees in the forest.

Below graph suggests that only 2 variables (TIT & TAT) are important to predict CO using Ridge Regression model.



Below is the graph that suggests almost all variables are equally important to predict the target variable NOX.



# References

1. Si, M.; Tarnoczi, T.J.; Wiens, B.M.; Du, K. Development of Predictive Emissions Monitoring System Using Open Source Machine Learning Library—Keras: A Case Study on a Cogeneration Unit. IEEE Access 2019, 7, 113463–113475. [CrossRef]

2. Dua, D.; Graff, C. UCI Machine Learning Repository. Available online: http://archive.ics.uci.edu/ml (accessed on 10 October 2021).

3. Kaya, H.; Tüfekci, P.; Uzun, E. Predicting CO and NOx emissions from gas turbines: Novel data and a benchmark PEMS. Turk. J. Electr. Eng. Comput. Sci. 2019, 27, 4783–4796. [CrossRef]

4. Rezazadeh, A. Environmental Pollution Prediction of NOx by Process Analysis and Predictive Modelling in Natural Gas Turbine Power Plants. Pollution 2021, 7, 481–494.

5. Si, M.; Du, K. Development of a predictive emissions model using a gradient boosting machine learning method. Environ. Technol. Innov. 2020, 20, 101028. [CrossRef]

6. Chawathe, S.S. Explainable Predictions of Industrial Emissions. In Proceedings of the 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 21–24 April 2021; pp. 1–7.

7. Nino-Adan, I.; Portillo, E.; Landa-Torres, I.; Manjarres, D. Normalization Influence on ANN-Based Models Performance: A New Proposal for Features’ Contribution Analysis. IEEE Access 2021, 9, 125462–125477. [CrossRef]

8. Etemadi, S.; Khashei, M. Etemadi multiple linear regression. Measurement 2021, 186, 110080. [CrossRef]

9. Manasis, C.; Assimakis, N.; Vikias, V.; Ktena, A.; Stamatelos, T. Power Generation Prediction of an Open Cycle Gas Turbine Using Kalman Filter. Energies 2020, 13, 6692. [CrossRef]

10. Giunta, G.; Vernazza, R.; Salerno, R.; Ceppi, A.; Ercolani, G.; Mancini, M. Hourly weather forecasts for gas turbine power generation. Meteorol. Z. 2017, 26, 307–317. [CrossRef]

11. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32. [CrossRef]

12. Abooali, D.; Khamehchi, E. Toward predictive models for estimation of bubble-point pressure and formation volume factor of crude oil using an intelligent approach. Braz. J. Chem. Eng. 2016, 33, 1083–1090. [CrossRef]

13. Khan, S.H.; Kumari, A.; Dixit, G.; Majumder, C.B.; Arora, A. Thermodynamic modeling and correlations of CH4 , C2H6 , CO2 , H2S, and N2 hydrates with cage occupancies. J. Petrol. Explor. Prod. Technol. 2020, 10, 3689–3709. [CrossRef]