EDS 6340 – INTRODUCTION TO DATA SCIENCE

Spring 2023

Professor: Dr. Amaury Lendasse

Group 13

GAS TURBINE CO AND NOx EMISSION

Bairy Adwitha	2251808	abairy@cougarnet.uh.edu
Emmidi Madhukar	2151565	memmidi@cougarnet.uh.edu
Mohana Krishna Koripella	2252332	mkoripel@cougarnet.uh.edu
Lalichetti Aiswarya	2252761	alaliche@cougarnet.uh.edu

INTRODUCTION:

Gas Turbine CO and NOx Prediction is the process of predicting the emission of flue gases like CO and NOx from power plants based on various parameters like ambient temperature, ambient pressure, ambient Humidity, gas turbine exhaust pressure and other parameters. In gas turbine-based power plants, predictive emission monitoring systems are valuable tools for validating and supporting expensive continuous emission monitoring systems.

In this project, we used the predictive emission monitoring system (PEMS) dataset collected over five years from a gas turbine for the predictive modeling of the CO and NOx emissions. We evaluate the data using a contemporary machine learning methodology and give useful insights about emission projections

The fascinating points to choose this dataset point is, in many Industries gas turbine is largely used to generate electricity around the world. By developing the predictive model to predict the emission of CO and NOx we can identify which input parameters are causing for larger emission of flue gases and reduce the emission of CO and NOx emission. In recent years, the awareness of greenhouse gases has been raised and the emissions of the gas- turbine. So, to reduce the CO and NOx gas emissions this project study may be helpful.

<u>Key Factors about the dataset are as below:</u>

- Category: Regression, Clustering
- Number of Instances: 36733
- Number of Attributes: 11

  Monitoring the Gas flue emissions CO and NOx pollutants from a Gas Turbine. By using Machine Learning techniques, we predict the NOx and CO pollutants emitted during Combustion Operation in a Gas Turbine.

Table 1: Gas Turbine CO and NOx Emission Dataset description

Attribute Abbreviations	Name of Attribute	Unit	Min	Max	Mean
AT	Ambient temperature	C	6.23	37.10	17.71
AP	Ambient pressure	mbar	985.85	1036.56	1013.07
AH	Ambient humidity	%	24.08	100.20	77.87
AFDP	Air filter difference pressure	mbar	2.09	7.61	3.93
GTEP	Gas turbine exhaust pressure	mbar	17.70	40.72	25.56
TIT	Turbine inlet temperature	C	1000.85	1100.89	1081.43
TAT	Turbine after temperature	C	511.04	550.61	546.16
CDP	Compressor discharge pressure	mbar	9.85	15.16	12.06
TEY	Turbine energy yield	MWH	100.02	179.50	133.51
CO	Carbon monoxide	mg/m3	0.00	44.10	2.37
NOx	Nitrogen oxides	mg/m3	25.90	119.91	65.29

PRE-PROCESSING OF THE DATA:

In real life, data extracted from multiple sources of data, we cannot expect the data to be clean. To deal with the data and create models of the data, the data must be cleaned. This process of cleaning the data is known as data pre-processing. For our dataset, we are checking for missing values and finding outliers in the dataset. There were no missing and null values in the data but the data was very inconsistent due to the presence of outliers in the dataset. We use the interquartile range to remove the outliers of the data by replacing the extreme values of the data with the median values within the Interquartile range.

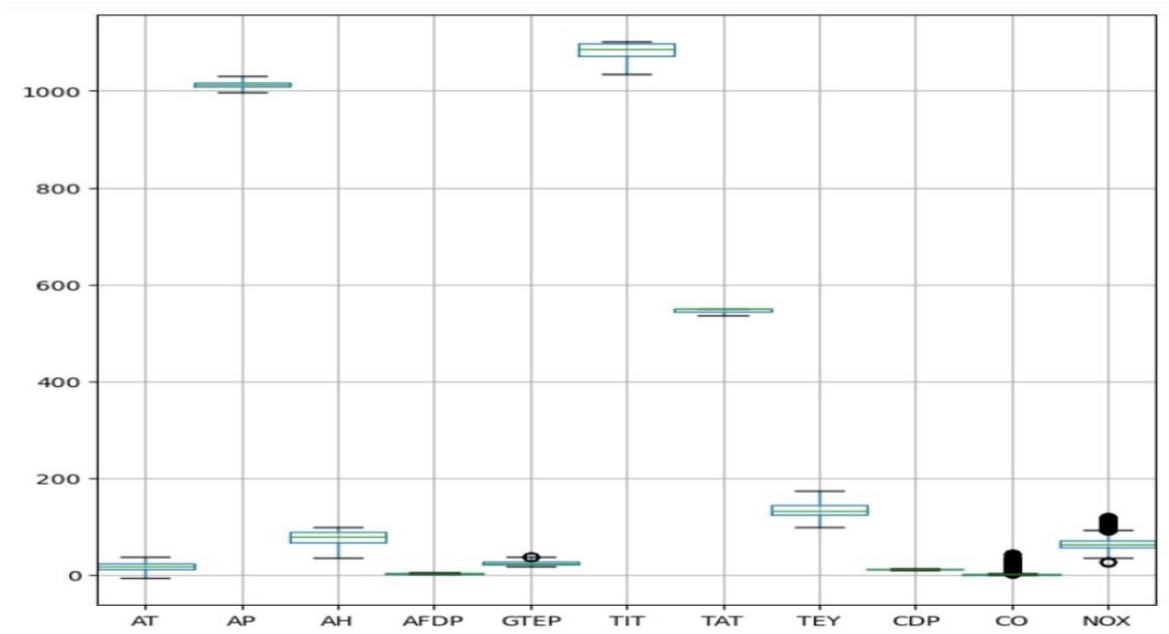

Fig 1: Outliers in Gas Turbine CO and NOx Emission Dataset description

EVALUATION OF SINGLE MODELS:

We performed different Machine learning models on our dataset, our approach is that first we split our data into training, validation, and testing datasets. Later, we implemented regression models on the training and validation set and calculated the performances of each model. For finding best parameters, we performed cross validation on the train set. It is observed that model performance improved after doing model structure selection.

Model 1: Linear Regression

A linear method for simulating the connection between a scalar response and one or more explanatory factors is linear regression. We fitted the training data to the model and predicted the outcome using the validation set. After calculating the performance using MAE and RSME metrics we performed GridSearchCV and Ridge regression to find the best parameters. Now, we tuned the model with the best parameters and performed the regression. The Performance for both CO and NOx is shown below.

Table 2: Performance for NOx and CO for linear Regression

Performance for NOx for linear Regression	Performance for CO for linear Regression
MAE on testset is: 5.409 MSE on testset is: 55.112 RMSE on the testset: 7.423	MAE on testset is: 0.855 MSE on testset is: 2.192 RMSE on the testset: 1.48

Naive baseline: It is taken as reference is taken a reference to compare results with all other models. The following is the result for our dataset:

- MAE for Naïve baseline for NOx is 8.82 and CO is 1.26
- Baseline Score for NOx is 63.8 and CO is 1.71

Model 2: K nearest neighbor Regression

KNN regression averages the data from the same neighborhood to approximately represent the relationship between independent variables and the continuous result. We implemented the KNeighborRegressor on the training and validation data. we performed GridSearchCV to find the best parameters.

Table 3: Performance for NOx and CO for K nearest neighbor Regression

Performance for NOx for KNN Regression	Performance for CO for KNN Regression
Mean Absolute Error: 3.146 Mean Squared Error: 23.739 RMSE : 4.872337603661056	Mean Absolute Error: 0.562 Mean Squared Error: 1.6795 RMSE : 1.2959787149884614

Model 3: Decision Tree Regression

A decision tree creates tree-like models for classification or regression. It progressively develops an associated decision tree while segmenting a dataset into smaller and smaller sections. We performed the decision tree regression same as the previous models. We have used Randomized search CV to find best parameters.

Table 4: Performance for NOx and CO for Decision Tree Regression

Performance for NOx for Decision Tree Regression	Performance for CO for Decision Tree Regression
MAE of decision tree modelis: 3.73807 MSE of decision tree modelis: 31.4135 RMSE of decision tree modelis: 5.6047	MAE on test set: 0.647 MSE on testset is: 1.811 RMSE on the Test set: 1.296

Model 4: Random Forest Regression

Random Forest Regression uses 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 the variance. We have used Randomized search CV to find best parameters.

Table 5: Performance for NOx and CO for Random Forest Regression

Performance for NOx for Random Forest Regression	Performance for CO for Random Forest Regression
Test Mean Absolute Error: 3.499 Test Mean Squared Error: 28.055 RMSE on the Test set: 4.282	Test Mean Absolute Error: 0.5926 Test Mean Squared Error: 1.46274 RMSE on the Test set: 1.209441038

Model 5: Support Vector Machine – Linear

Support Vector Machine – Linear, Data is classified with the help of a hyperplane. It can be easily separated with a linear line. We have implemented SVR model and in the parameters, we assigned kernel as linear along with the other parameters

Table 6: Performance for NOx and CO for Support Vector Machine-Linear

Performance for NOx for Support vector Machine -Linear	Performance for CO for Support vector Machine- Linear
MAE on test set: 0.924 MSE on test set is: 2.624 RMSE on the test set: 1.62	MAE on test set: 0.88 MSE on test set is: 3.086 RMSE on the test set: 1.756

Model 6: Support Vector Machine – Non-Linear

Support Vector Machine Non-Linear, we use Kernels to make non-separable data into separable data. We map data into high dimensional space to classify. Kernel type is the most important parameter for SVR. It can be linear, polynomial or gaussian SVR. We have selected RBF (a gaussian type) kernel for Non-linear model.

Table 7: Performance for NOx and CO for Support Vector Machine-Non-Linear

Performance for NOx for Support vector Machine -Nonlinear	Performance for CO for Support vector Machine- Nonlinear
MAE on test set: 8.904 MSE on test set is: 126.909 RMSE on the test set: 11.265	MAE on test set: 0.931 MSE on test set is: 3.379 RMSE on the test set: 1.838

Model 7: Extreme Learning Machine

ELM is a type of machine learning model designed for fast and efficient learning with minimal tuning. It is based on the concept of random initialization of the input-to-hidden layer weights, which allows for a faster training process compared to other traditional neural network architectures.

Table 8: Performance for NOx and CO for Support Vector Machine-Linear

Performance for NOx for Extreme Learning Machine	Performance for CO for Extreme Learning Machine
mae_1: 4.3332753 mse_1: 36.108864 r_score_1: 0.727	mae_2: 0.9636312 mse_2: 2.0850870 r_score_2: 0.582

First Variable Selection:

Feature selection is primarily focused on removing non-informative or redundant predictors from the model. Many models, especially those based on regression slopes and intercepts, will estimate parameters for every term in the model. Because of this, the presence of non-informative variables can add uncertainty to the predictions and reduce the overall effectiveness of the model.

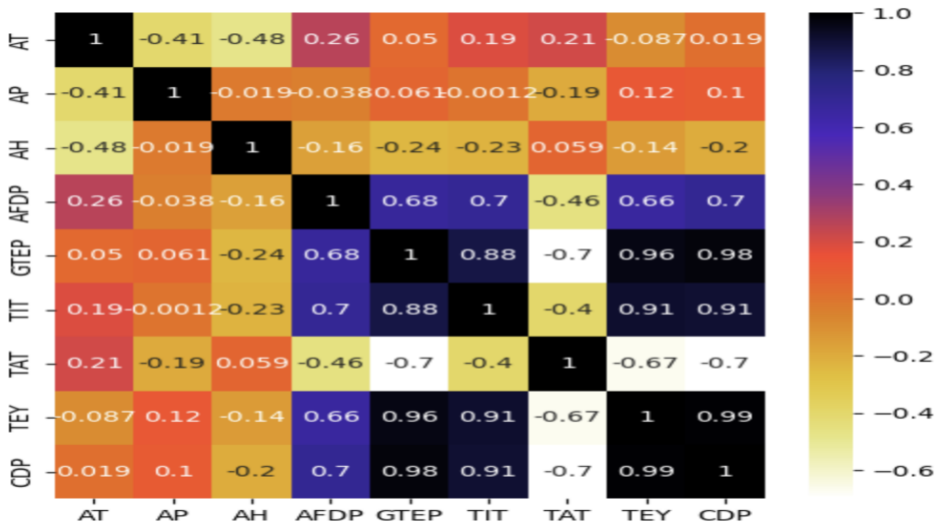

Fig 2: Correlation Gas Turbine CO and NOx Emission Dataset description

From the fig 2 we came to know Compressor discharge pressure (CDP), Turbine energy yield (TEY), Turbine inlet temperature (TIT) is highly correlated whose values are above 0.85, so we removed this feature and chosen model selection features are AT, AP, AH, AFDP, GTEP, TAT. On comparing best models with variable selection, we can conclude that random forest without variable selection has better performance for both NOx and CO target variables.

Wrapper method - Bi-directional elimination:

The wrapper method is the feature selection process in which it follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion. Bidirectional elimination is one of the wrapper methods which is essentially a forward selection procedure but with the possibility of deleting a selected variable at each stage, as in backward elimination, when there are correlations between variables. We performed bi-directional elimination method using python in-built function sffs () which is Selected Attributes using sequential feature selector (SFS). We fitted the feature dataset into sffs and found the important features that are shown.

By using above method selected features for gas turbine CO and NOx are AT, AH, GTEP, TIT, TAT. To choose the best model, we compared the best two models Random Forest Regressor and SVM linear Kernel models by applying bidirectional elimination wrapper method.

SFS feature parameters are as follows:

```
SequentialFeatureSelector(cv=0, estimator=LinearRegression(), k_features=(5, 5),
                          scoring='r2')
```

Table 9: Performance for NOx and CO for Support Vector Machine-Linear and Random Forest after bidirectional elimination feature selection

Model	NOx	CO
Random forest Regression	MAE on test set: 3.227 MSE on test set is: 23.26 RMSE on the test set: 4.822	MAE on test set: 0.555 MSE on test set is: 1.42 RMSE on the test set: 1.191
SVM Linear Kernel	MAE on test set: 6.945 MSE on test set is: 81.398 RMSE on the test set: 9.022	MAE on test set: 0.961 MSE on test set is: 3.026 RMSE on the test set: 1.739

From Table 9, we can clearly see that Random Forest Regression is the best model compared to SVM Linear Kernel on the selected features.

Clustering:

Clustering is a technique used to group similar data points together based on their attributes or characteristics. Thus, help identify patterns and structure in the data that can be used to inform feature selection and engineering.

Clustering can be useful in building a predictive model, particularly when dealing with large and complex datasets. In addition, clustering can be used for dimensionality reduction.

We also used approach is the silhouette coefficient, which measures how well each data point fits into its assigned cluster and can help determine the appropriate number of clusters. silhouette score ranges from -1 to 1, where a higher score indicates better clustering results: Results indicate optimal number of clusters for our model be 4.

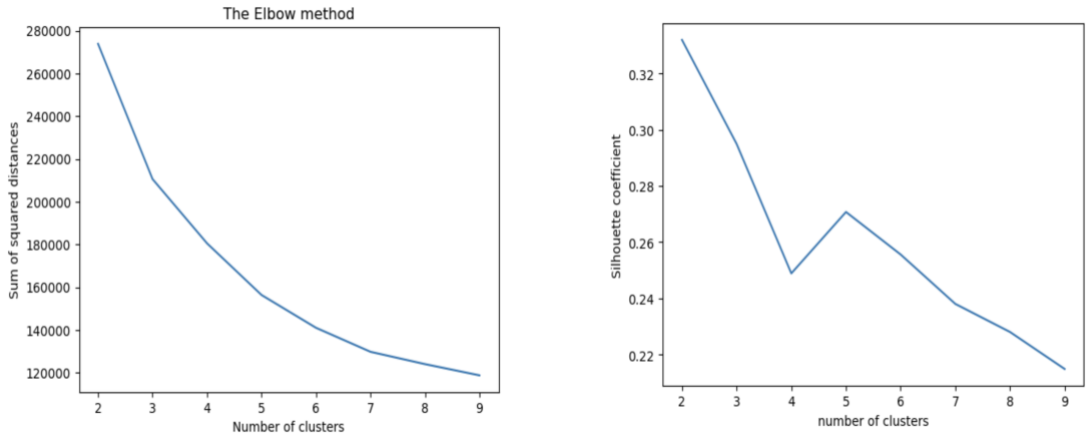

Fig 3: Elbow method and Silhouette analysis

Visualization:

Visualization can be a very useful tool for building a predictive model. Visualization allows you to gain insights into the relationships between different variables in the data, identify patterns, and detect outliers or anomalies. This information can be used to inform feature selection and engineering, as well as to diagnose issues with your data. It helps us identify areas where the model may be underfitting or overfitting, thus allowing in adjusting improve its performance. Principal Component Analysis (PCA) used to speed up model training for data visualization. It's used to reduce the number of components but still retain the information with less loss.

Here we used scree plot to plot the five principal components and variance graph to find the amount of variance for different number of components.

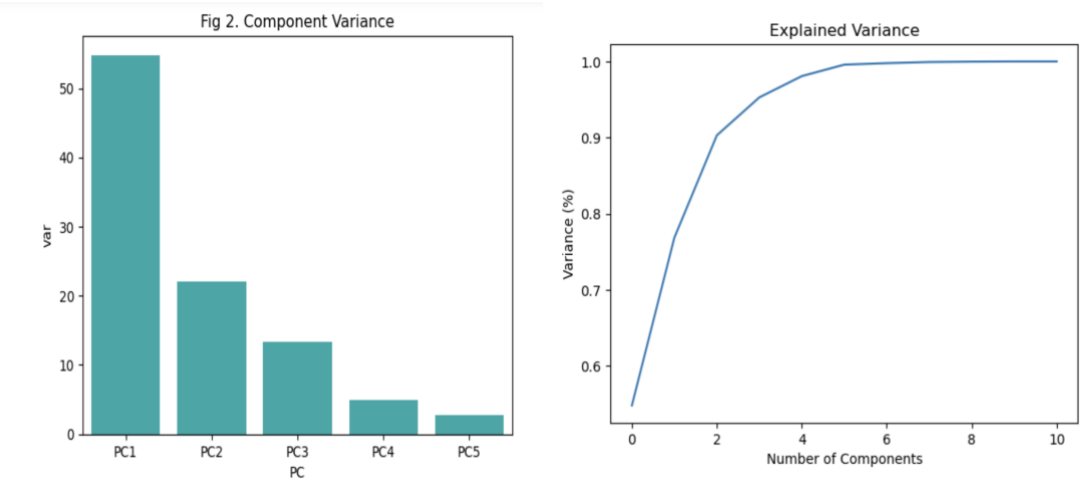

Fig 4: Scree plot and variance analysis

Ensemble Modelling:

Ensemble models are machine learning models that combine multiple individual models to improve their overall predictive performance. This can be done by either bagging, boosting or combining their outputs using a weighted scheme (e.g., stacking). Ensemble models are popular because they can mitigate the risk of overfitting, reduce bias and variance, and enhance the generalizability of the model.

We finally build an Ensemble model with combination of linear regression, Random Forests, Decision tree, KNN, SVM Linear, SVM Non-Linear Model. Based on the analysis, it can be concluded that the Random Forest model with Bidirectional Elimination (BDE) performed the best for our specific data set and problem.

On comparing the results of the ensemble with the results of single models- we found Random Forest with BDE has better performance with mean absolute error 1.224 for NOx, 0.204 for CO as target variable. The following graph shows the mean absolute errors for different machine learning algorithms.

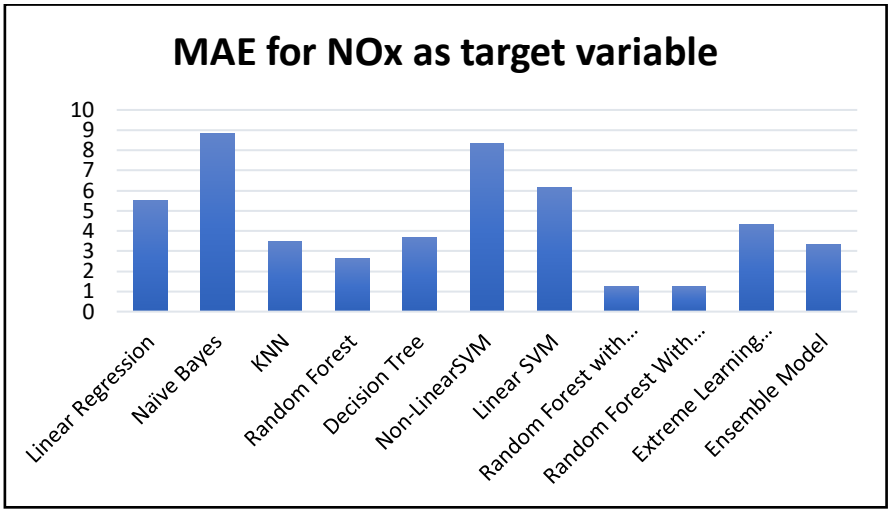

Fig 5: Model Comparison of MAE values for NOx target variable.

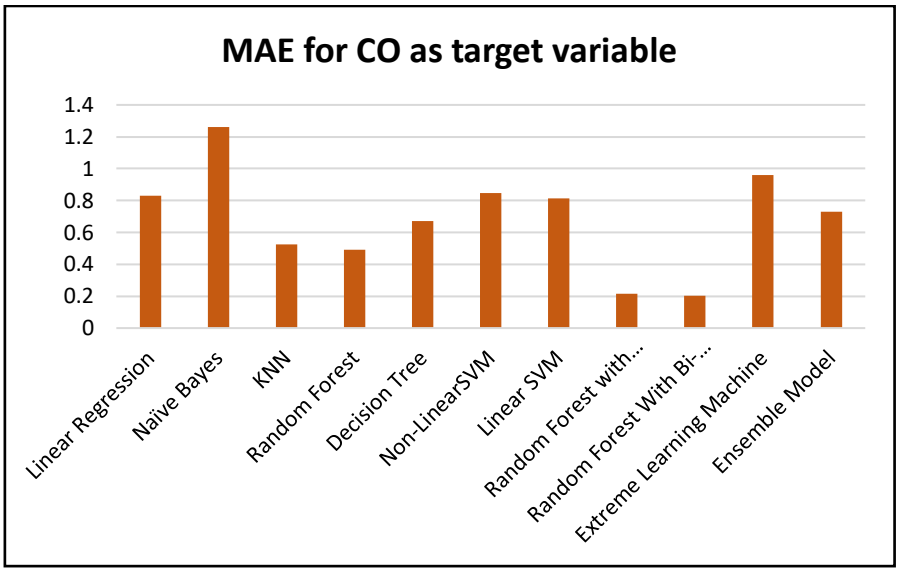

Fig 6: Model Comparison of MAE values for CO target variable.

General Discussion:

If we have a chance to start project again, we would have also analysed deep learning model performance and using different model structure selection parameters to understand the dataset more accurately. We would have focused more on developing a clear and comprehensive project plan. We should also have implemented a more effective communication strategy to ensure that all team members were aware of their roles and responsibilities.

If I had more time for the project, I would focus on improving our testing processes looked for a different strategy to run the models more quickly. Currently, we have a limited amount of time for performing, which increases the risk of errors and reduces the quality of our work.

We learnt many things from this project like we got the knowledge on how to pre-process the data to evaluate the model. We also got the knowledge on how to perform clustering for an unsupervised data model with the help of K-mean clustering. We might have tried reducing the project's execution time. Working on several other models might have further improved our coding abilities. I learned the importance of effective project management and the need for clear communication channels. In addition, I would spend more time on code reviews and debugging to ensure that our code is optimized and follows best practices.

Conclusion:

As part of this research, we worked on the dataset's preprocessing, including handling missin g values and outliers. We also investigated the five years Gas Turbine CO & NOx Emission dataset and assessed how well various machine learning models performed the task of predicting the output. After training and evaluating these models, we could state that input attributes Compressor discharge pressure (CDP), Turbine energy yield (TEY), Turbine inlet temperature (TIT) are strongly correlated with correlation values above 0.85. We trained and evaluated number of well-known models, including Naïve Bayes, Linear Regression, K-Nearest Neighbors, Decision Tree Random Forest, and Support Vector Machines Linear and Non-Linear Model, Extreme Learning Machine, Variable selection, Bidirectional Elimination, method and Ensemble modelling.

Our test findings demonstrated that the Random Forest model with the Bidirectional elimination method outperformed the other models with MAE of 1.224 for NOx and 0.204 for CO target variables. The results can also be used to guide the development and implementation of regulations and policies aimed at reducing emissions from gas turbines and other industrial sources.

References:

Data Set Link:

https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set

[1] Gomaa, M., El-Shahat, M., & El-Sayed, H. (2016). Modeling and prediction of gas turbine emissions using neural networks and support vector regression. Energy Conversion and Management, 108, 427-436.

[2] Ganesan, S., & Palanisamy, K. (2019). Modelling and analysis of gas turbine emission parameters using artificial neural networks. International Journal of Engineering and Advanced Technology (IJEAT), 9(3), 4086-4093.

[3] https://www.analyticsvidhya.com/blog/2022/01/machine-learning-algorithms/

[4] Liu, Z., Li, S., Li, Y., & Sun, Y. (2019). Prediction of NOx emissions from gas turbine