PROJECT-I REPORT

ON

Analysis of Gas Emissions from Gas Turbine Dataset

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING
(INFORMATION TECHNOLOGY)

Supervisor:
Dr. Veenu Mangat,
Associate Professor
I.T. , U.I.E.T ,Panjab University
Chandigarh

Submitted By:
Anshul Gupta
BE- IT Section-1
(2017-2021 batch)
UE178022

To

Department of Information Technology,
University Institute of Engineering and Technology,
Panjab University, Chandigarh

2020

DECLARATION BY THE CANDIDATE

I the undersigned solemnly declare that the project report is based on my own work carried out during the course of our study under the supervision of Dr. Veenu Mangat (Associate Professor, I.T., U.I.E.T, Panjab University, Chandigarh). I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that

I. The work contained in the report is original and has been done by me under the general supervision of my supervisor.
II. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.
III. I have followed the guidelines provided by the university in writing the report.
IV. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them in the text of the report and giving their details in the references.

Anshul Gupta (UE178022)

Signatures :

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my mentor Dr. Veenu Mangat (Associate professor-IT, UIET, Panjab University, Chandigarh) who supervised my project, guided me through the details and also helped in the successful completion of my project.

Table of Contents:

1. Introduction

Climate change is real and we need to act as soon as possible.

Executive summary of project

- **The purpose of this project is to analyze the gas emissions from the gas turbine dataset and find actionable insights, build regression model and perform clustering. The insights are needed so that the gas emissions can be monitored and then controlled.**
- **Data set used is sourced from UCI Machine learning repository and datasets consists ofof 36733 instances.**
- **Data preprocessing is done before the data modeling is done. Data visualizations are used to get various insights from the dataset. Two regression models are built and then two clustering models are built and evaluation of all of them is done so as to compare their performance.**
- **The analysis is motivated by the environmental problem of greenhouse gas emissions and presents various insights and models which can help tell us ways to tackle the problem of increasing greenhouse gases.**
- **Solution exists in the knowledge of when to trigger the maintenance of the industry equipment so that the flue gases (Carbon mono-oxide -CO and Nitrogen oxides- NOX here) emitted must be controlled.**

1.1 Problem setting:

Climate change is happening not only due to the pollution caused by industries and actions of humans but also due to the changing universal energies (shifting tectonic plates, change in orbit of earth, sun & moon) and due to various other reasons.

The main problem lays in the fact that in order to adapt with respect to the changing climate, we need to have extensive research capabilities so to as to study the climatic phenomenon's.

While it's true that it's next to impossible to control or predict the natural happenings. At the same time it becomes extremely necessary to not give up and to do what we can.

Thus, the analysis of gas emissions from industries is well needed so as to reduce the pollution and control the climate change.

1.2 Potential solution:

To control the flue gas emissions by the engine, we need to first monitor the emissions and then control them if they are above some threshold, by triggering

some form of control measure (maintenance trigger, self-clean of engine parts, etc).

So this analysis is motivated by the environmental problem of greenhouse gas emissions and presents various insights and models which can help tell us ways to tackle the problem of increasing greenhouse gases.

Solution exists in the knowledge of when to trigger the maintenance of the industry equipment so that the flue gases (Carbon mono-oxide and Nitrogen oxides here) emitted must be controlled.

1.3 Proposed approach:

In the analysis we try to use fundamental principles of science and mathematics in order to build model which predicts Turbine energy yield, CO and NOX emissions by using ambient features as the predictor features.

Then these three attributes are clustered so as to identify the data points which will trigger the maintenance of the gas turbine equipment.

1.4 Potential audience:

Researchers working on climate change might find an insight. Industrialists need to be reminded of the importance of our planet and the measures they need so

as to save it. The insights are not limited to single gas turbine industry but to many other similar industries such as other gas emitting power plants, vehicles, spacecrafts, air cleaning towers, etc.

AIM

The purpose of this project is to analyze the gas emissions from the gas turbine dataset and find actionable insights, build regression model and perform clustering. The insights are needed so that the gas emissions can be monitored and then controlled.

2. Technology used

2.1 Hardware

A personal computer with :

→ Intel core i5 7th generation CPU

→ NVidia 920MX 2GB GPU

→ 4GB Ram

2.2 Software

→ Code written in Jupyter notebook

→ Programming language – Python

→ Libraries used :

- Numpy
- Pandas
- Matplotlib
- Pyplot
- MiniSOM
- Seaborn
- Scikit-learn

3. Methodology

3.1 Data Source:

- Data set used is sourced from UCI Machine learning repository and it uploaded on 29th November'2019.
- While there were many similar datasets available online, this dataset was chosen due to the fact that it is relatively newer dataset and not much exploration has been done on this dataset.
- Being a relatively newer dataset, it provides information which is much more valuable since the equipment used in industries now is definitely different from the ones used earlier times due to rapid industrialization in 2000's.
- The number of instances in this dataset is also appropriate (not very small or very big), thus insights can be gathered from the data without the need of high computational power and generalized model can be made using the dataset.

3.2 Data description:

- 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.
- The instances do not have any labels specifying whether an instance is benign or malignant in the sense that benign instance will not trigger the maintenance of equipment or give a warning of increased emissions

Glimpse of the dataset

	Ambient temp	Ambient pressure	Ambient humidity	AFD pressure	GT exh pressure	Turbine inlet temp	Turbine after temp	CD pressure	Turbine energy yield	CO	NOX
0	4.5878	1018.7	83.675	3.5758	23.979	1086.2	549.83	134.67	11.898	0.32663	81.952
1	4.2932	1018.3	84.235	3.5709	23.951	1086.1	550.05	134.67	11.892	0.44784	82.377
2	3.9045	1018.4	84.858	3.5828	23.990	1086.5	550.19	135.10	12.042	0.45144	83.776
3	3.7436	1018.3	85.434	3.5808	23.911	1086.5	550.17	135.03	11.990	0.23107	82.505
4	3.7516	1017.8	85.182	3.5781	23.917	1085.9	550.00	134.67	11.910	0.26747	82.028
5	3.8858	1017.7	83.946	3.5824	23.903	1086.0	549.98	134.67	11.868	0.23473	81.748
6	3.6697	1018.0	84.114	3.5804	23.889	1085.9	550.04	134.68	11.877	0.44412	84.592
7	3.5892	1018.2	83.867	3.5777	23.876	1086.0	549.88	134.66	11.893	0.79996	84.193
8	3.7108	1018.5	84.948	3.6027	23.957	1086.3	549.98	134.65	11.870	0.68996	83.978
9	4.8281	1018.5	85.346	3.5158	23.422	1083.1	549.80	132.67	11.694	1.02810	82.654

Attribute Information (as uploaded on UCI ML website):

The explanations of sensor measurements and their brief statistics are given below.

1)Variable (Abbr.) Unit

2)Ambient temperature (AT) C

3)Ambient pressure (AP) mbar

4)Ambient humidity (AH) (%)

5)Air filter difference pressure (AFDP) mbar

6)Gas turbine exhaust pressure (GTEP) mbar

7)Turbine inlet temperature (TIT) C

8)Turbine after temperature (TAT) C

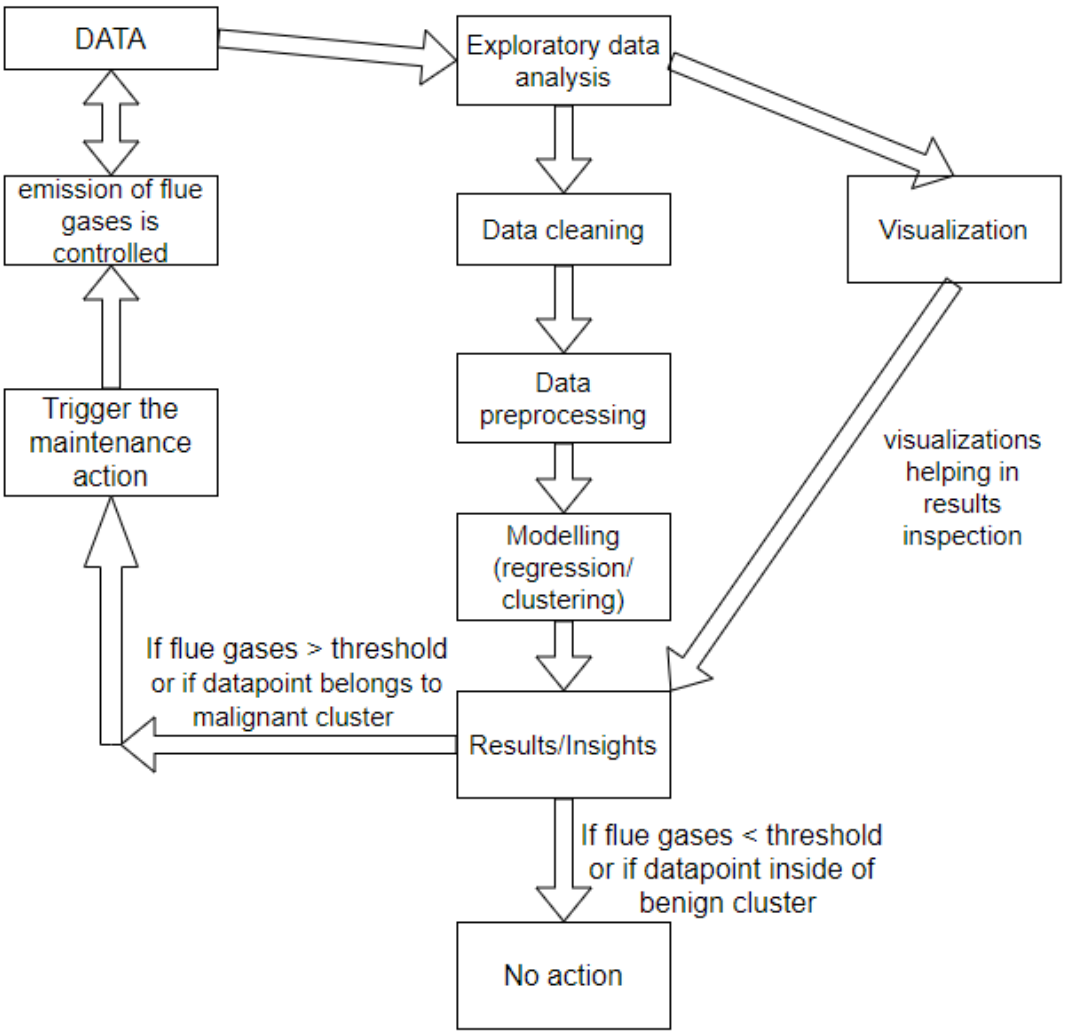
9)Compressor discharge pressure (CDP) mbar

10) Turbine energy yield (TEY) MWH

11) Carbon monoxide (CO) mg/m3

12) Nitrogen oxides (NOx) mg/m3

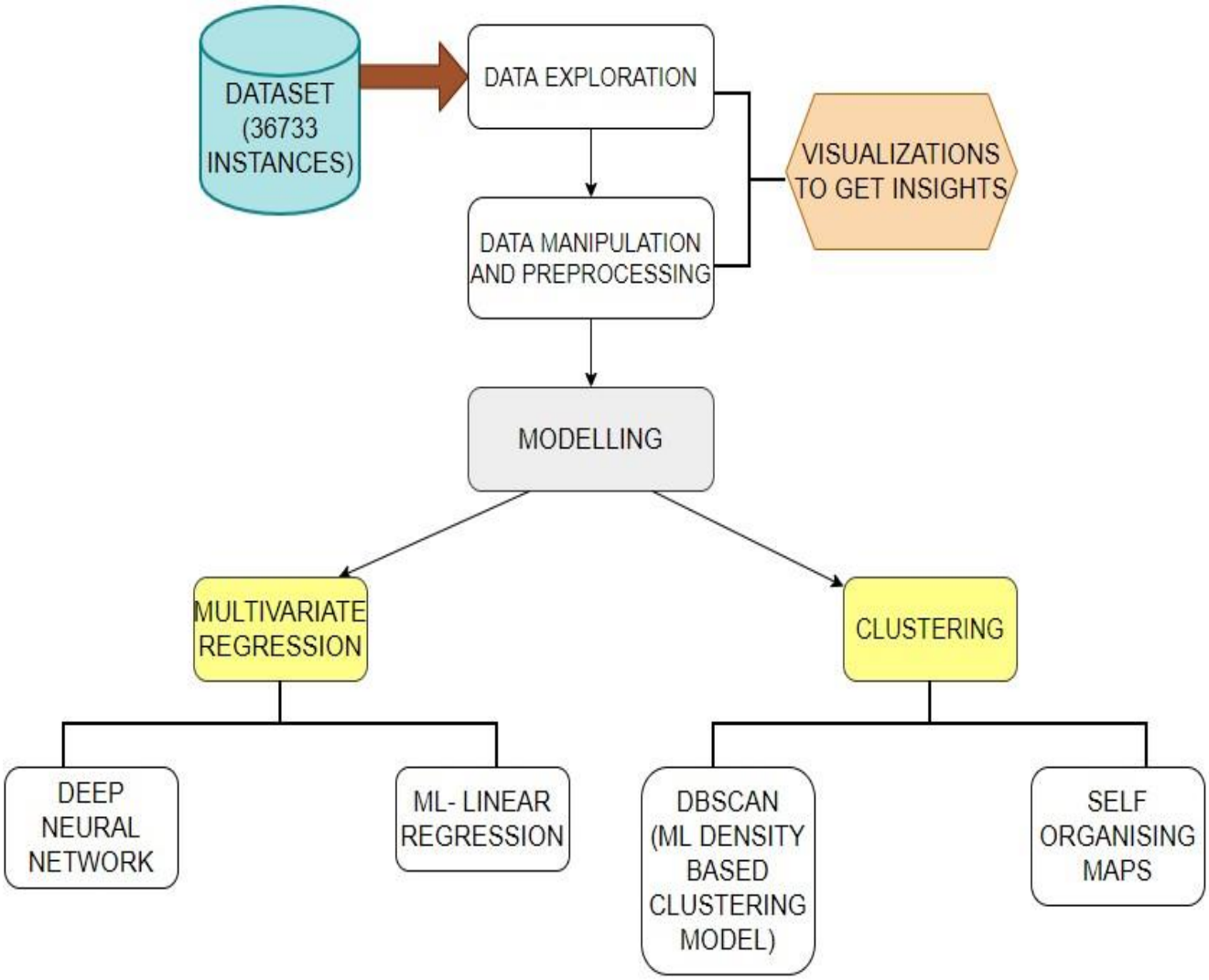
3.3 Flowchart of proposed solution



Here threshold is the difference between the predicted values by our model using the predictor variables and the actual values received by the sensors.

That means when the industry emission measure starts to become far from what it should have been then under specific conditions then maintenance should be triggered.

3.4 Steps and strategy used



Flowchart of Steps

1. **Data Exploration:** Exploring columns types, number of unique values in them, null value presence check, plotting distribution of column values, scatter plot of multiple features against each other.

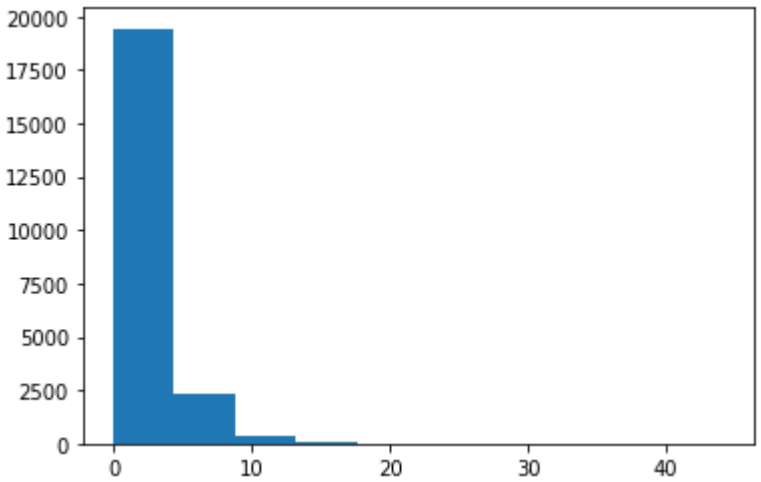


Fig 1. Distribution plot of feature 'CO'

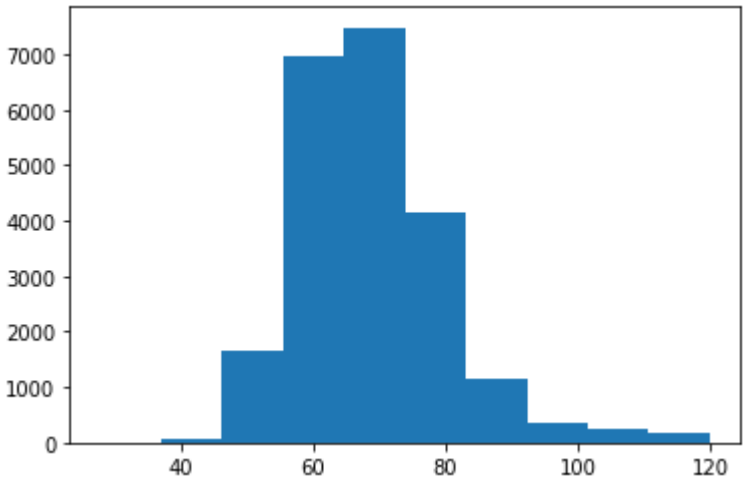


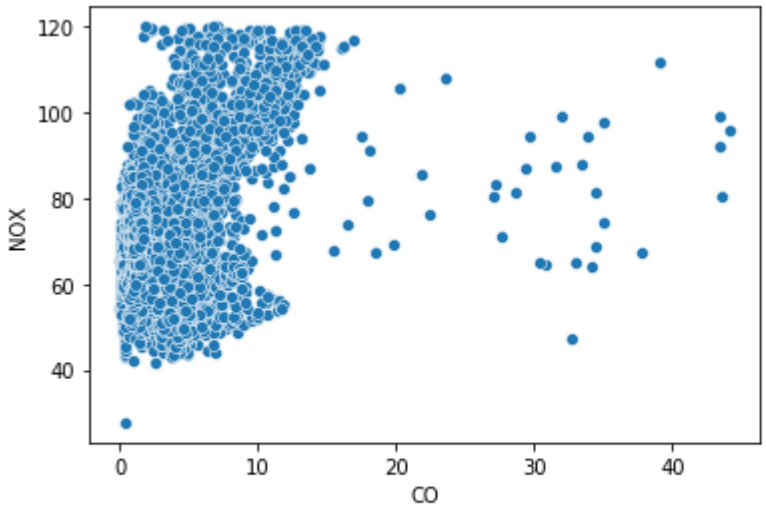
Fig 2. Distribution plot of feature 'NO'

	Ambient temp	Ambient pressure	Ambient humidity	AFD pressure	GT exh pressure	Turbine inlet temp	Turbine after temp	CD pressure	Turbine energy yield	CO	NOX
Ambient temp	1.000000	-0.406601	-0.476291	0.251974	0.045851	0.183706	0.281869	-0.091152	0.015287	-0.174326	-0.558174
Ambient pressure	-0.406601	1.000000	-0.015184	-0.040363	0.057533	-0.005390	-0.225601	0.118224	0.102636	0.067050	0.191938
Ambient humidity	-0.476291	-0.015184	1.000000	-0.147840	-0.235153	-0.221809	0.022965	-0.137360	-0.196275	0.106586	0.164617
AFD pressure	0.251974	-0.040363	-0.147840	1.000000	0.678485	0.691292	-0.466882	0.665483	0.702568	-0.448425	-0.188247
GT exh pressure	0.045851	0.057533	-0.235153	0.678485	1.000000	0.874234	-0.699703	0.964127	0.978470	-0.518909	-0.201630
Turbine inlet temp	0.183706	-0.005390	-0.221809	0.691292	0.874234	1.000000	-0.380862	0.910297	0.908469	-0.706275	-0.213865
Turbine after temp	0.281869	-0.225601	0.022965	-0.466882	-0.699703	-0.380862	1.000000	-0.682396	-0.706438	0.058353	-0.092791
CD pressure	-0.091152	0.118224	-0.137360	0.665483	0.964127	0.910297	-0.682396	1.000000	0.988778	-0.569813	-0.116127
Turbine energy yield	0.015287	0.102636	-0.196275	0.702568	0.978470	0.908469	-0.706438	0.988778	1.000000	-0.551027	-0.171256
CO	-0.174326	0.067050	0.106586	-0.448425	-0.518909	-0.706275	0.058353	-0.569813	-0.551027	1.000000	0.340606
NOX	-0.558174	0.191938	0.164617	-0.188247	-0.201630	-0.213865	-0.092791	-0.116127	-0.171256	0.340606	1.000000

❖ *Correlation table showing correlation among all features with each other*

❖ *As evident from above table , 'CD pressure' is highly correlated with 'GT exh pressure' and thus there is presence of Muliticollinearity ,a problem since the variables are considered to independent of each other*

2. **Various Visualizations** to get insights:



Here it seems like there is no direct relationship between emissions of NOX and CO.

Fig 3. Scatter plot (NOX vs CO)

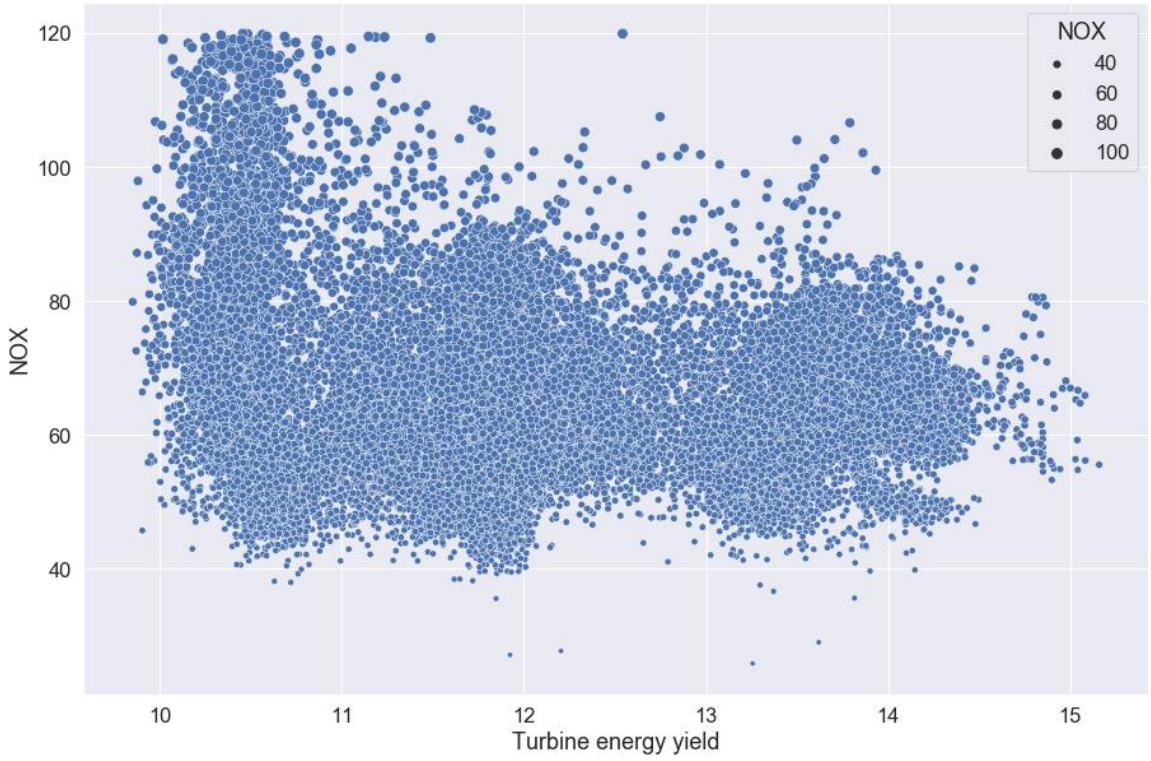


Fig 4. Scatter plot (NOX vs Turbine energy yield)

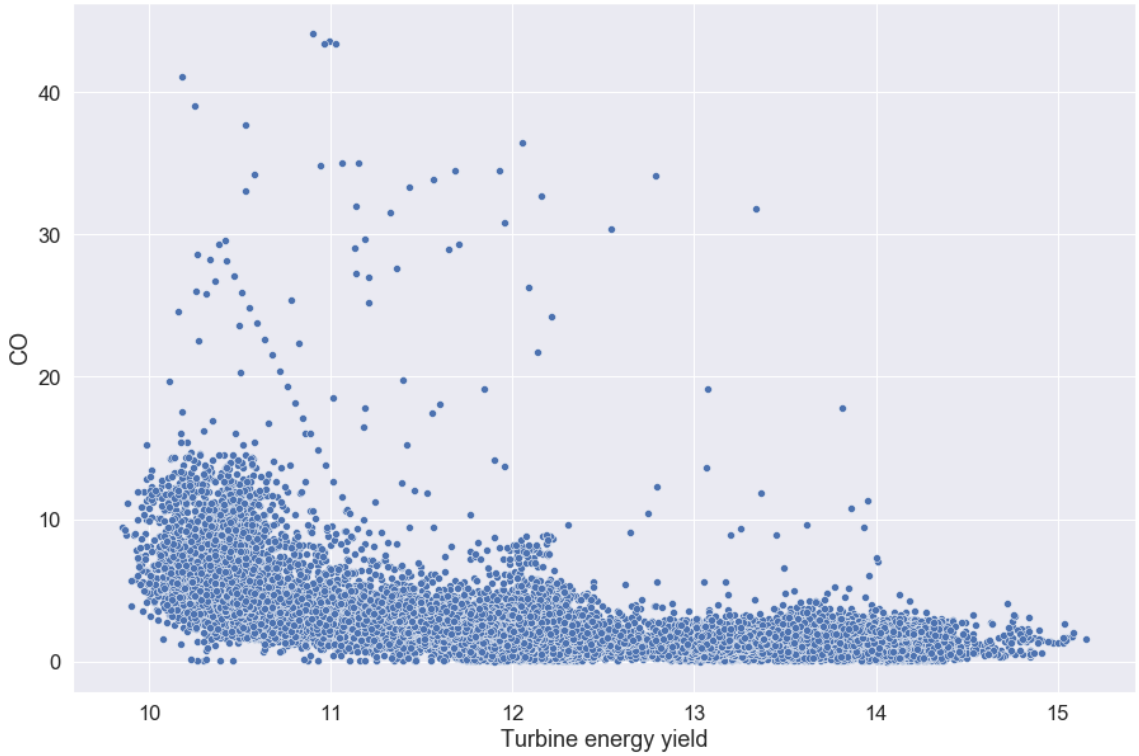


Fig 5. Scatter Plot (CO vs Turbine energy yield)

Insights from fig 5. :

- There is a very good insight from above plot, that with increase in Turbine energy yield ,the CO in the gas emissions reduces
- This means that when the energy yield is high there is complete combustion in gas turbine and the greenhouse gases (CO) will be become negligible , thus using the basic definition of combustion, we can progress towards our goal.
- The general equation for a complete combustion reaction is: Fuel + O2 → CO2 + H2O, but in case of

incomplete combustion equation the release of CO and NO increases.

- Also there are many seemingly outliers when the turbine energy yield is low (incomplete combustion), these outliers are the instances which are supposed to trigger the maintenance of gas turbine.

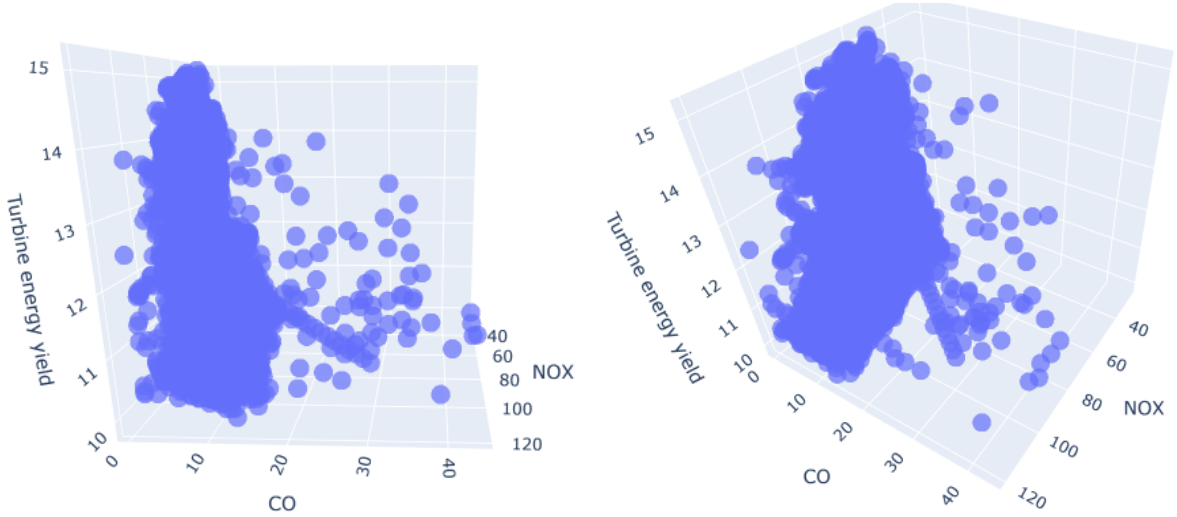


Fig 6. 3d Scatter plot (TEY vs CO vs NOX)

Insights from fig 6. :

- From above 3d plot we can see that the datapoints need to be clustered since we can see there is one single big arbitrary shaped cluster and the points outside of it are to be put in a different cluster.
- Density based clustering algorithms might work here.

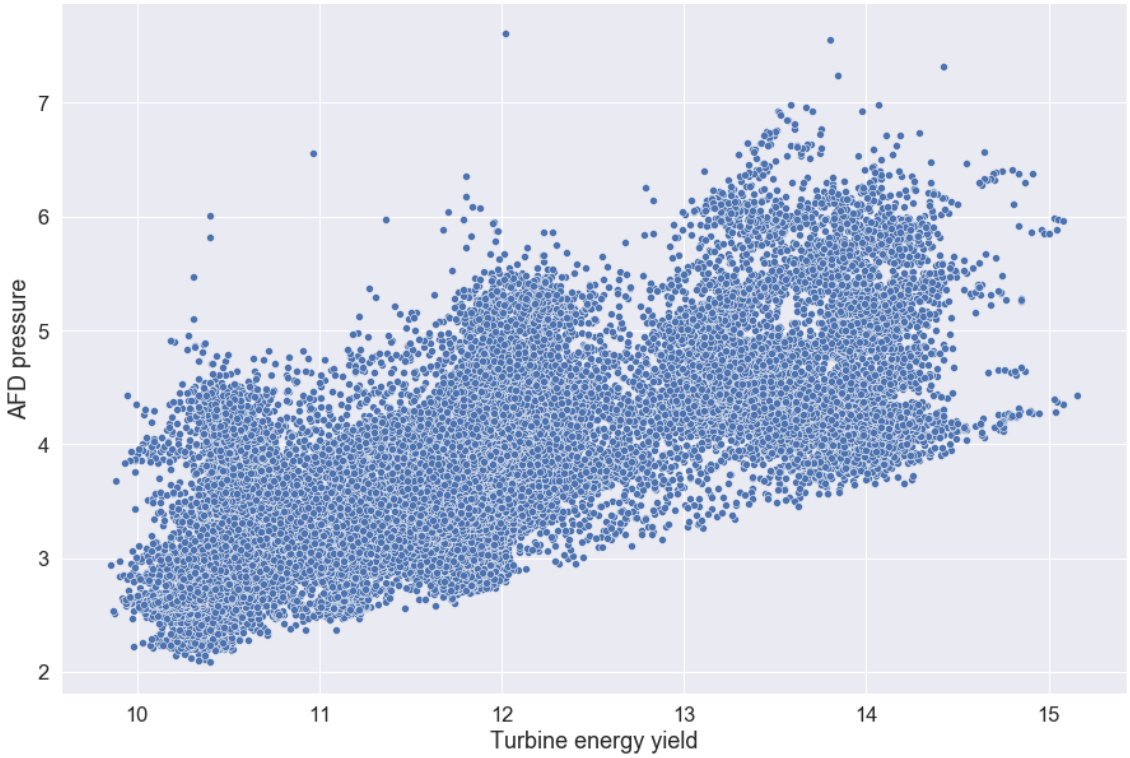


Fig 7. Scatter plot (AFD pressure vs Turbine energy yield)

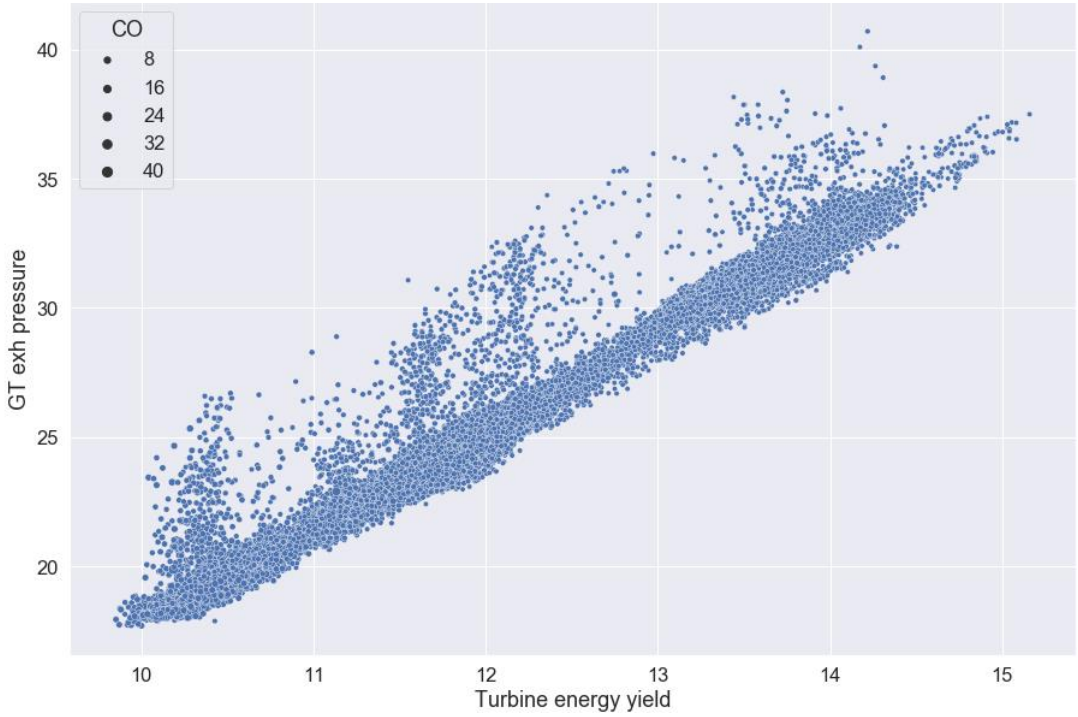


Fig 8. Scatter plot (GT pressure vs Turbine energy yield)

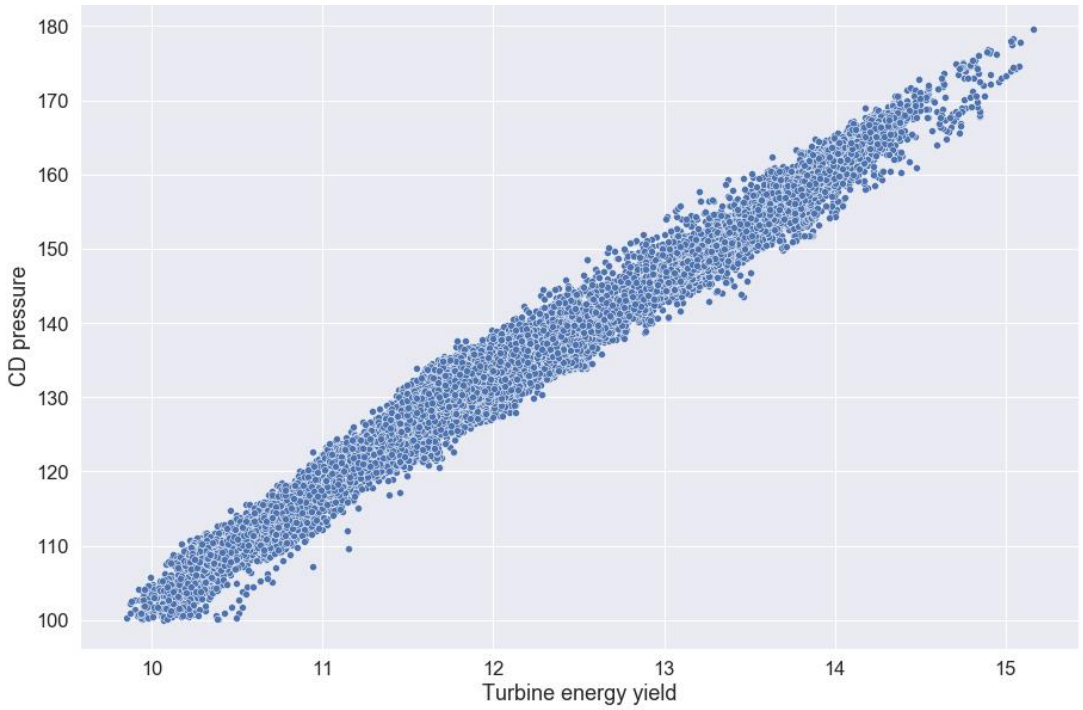


Fig 9. Scatter plot (CD pressure vs Turbine energy yield)

Insights from figures – 7,8 & 9 :

- With increasing AFD pressure, GT exhaust pressure and CD pressure, the turbine energy yield also increases linearly, thought the increase is less steep in case of AFD pressure.
- Seems there is high correlation among turbine energy yield , CD pressure and GT exhaust pressure

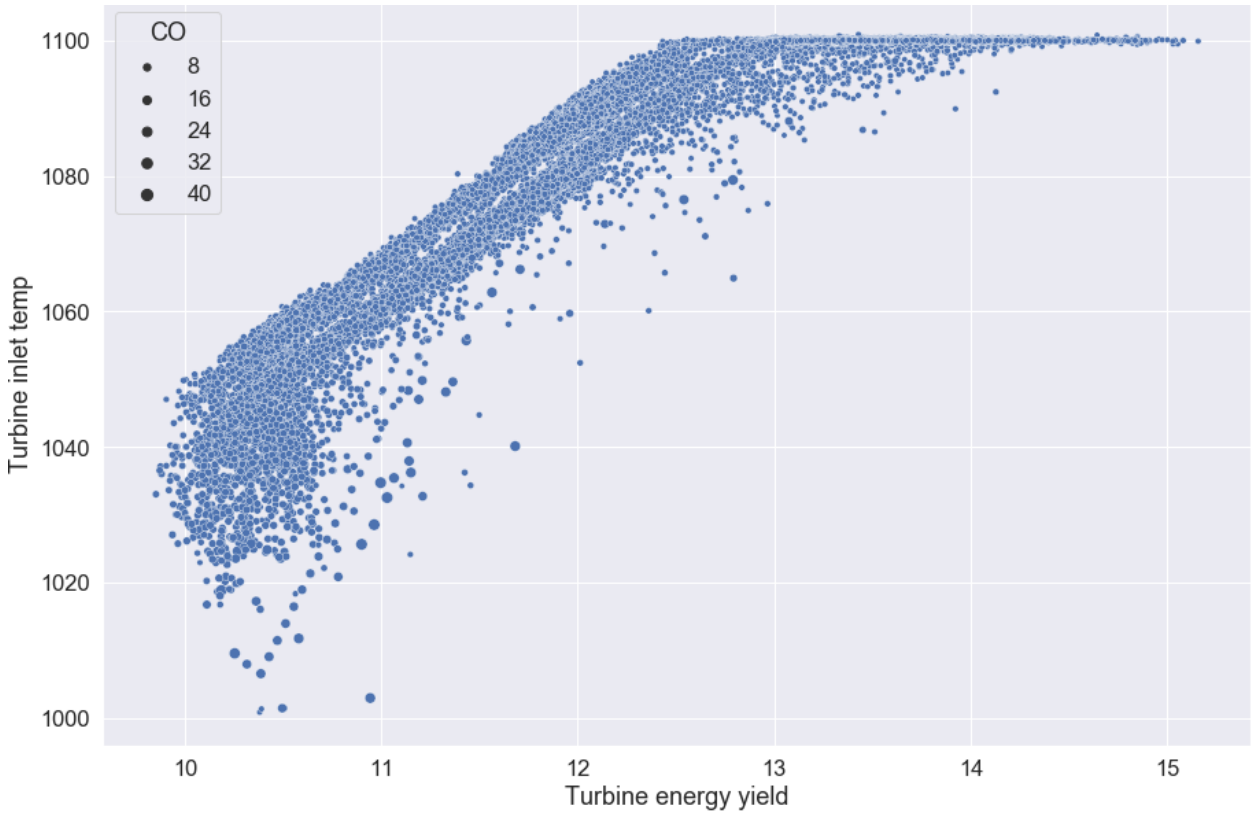


Fig 10. Scatter plot (Turbine inlet temp vs Turbine energy yield)

Insights from fig-10 :

- turbine inlet temp increases linearly with increase in TEY, upto a certain point (1100 C) and then the saturates at 1100 C , specifying the instances with low CO and NOX emissions (due to complete combustion).
- Turbine inlet temperature proves to be a strong indicator of whether the complete combustion is underway or not, thus one can get good idea of the emissions of CO and NOX that whether CO and NOX emissions are below a certain threshold by knowing the turbine inlet temperature.

3. **Modelling** process:

1. **Multivariate Regression analysis** :

Ambient features as predictor features and Turbine energy yield , CO and NOX as target features

- Neural network – Made pipeline consisting of Standard scaler which standardizes the dataset and then applies the neural network consisting of single hidden layer of 7 neurons.
- Chained linear regression - To exploit the correlation among the target variables we used chained regressor model which will basically chain model in this way:
  - *given X (predictor features), predict y1(first target feature)*
  - *given X and y1, predict y2(second target feature)*
  - *given X,y1 and y2, predict y3(third target feature)*

2. **Clustering**:

Turbine energy yield, CO and NOX are used as inputs

- DBSCAN - First the dataset is standardized using Standard scaler then it is clustered using density based clustering algorithm , evaluated using silhouette coefficient.

  Data is clustered by assuming :

  - First cluster should be the one with 'Not danger' values (High TEY, Low CO and Low NOX) and other cluster should consist of 'Danger' values (low TEY, high CO and high NOX)
  - Value of parameters of the algorithm are chosen by either heuristic methods or by hit and trial when we had enough assumptions

- Self-organising-maps(SOM)
  - The self-organizing maps differ from other artificial neural networks because they apply competitive learning as opposed to error correlated learning, which involves backpropagation and gradient descent.
  - In competitive learning, nodes compete for the right to respond to the input data subset. The training data usually has no labels and the map learns to differentiate and distinguish features based on similarities.

- The neurons basically try to become like the input and thus in this process it recognizes the boundaries present between the data

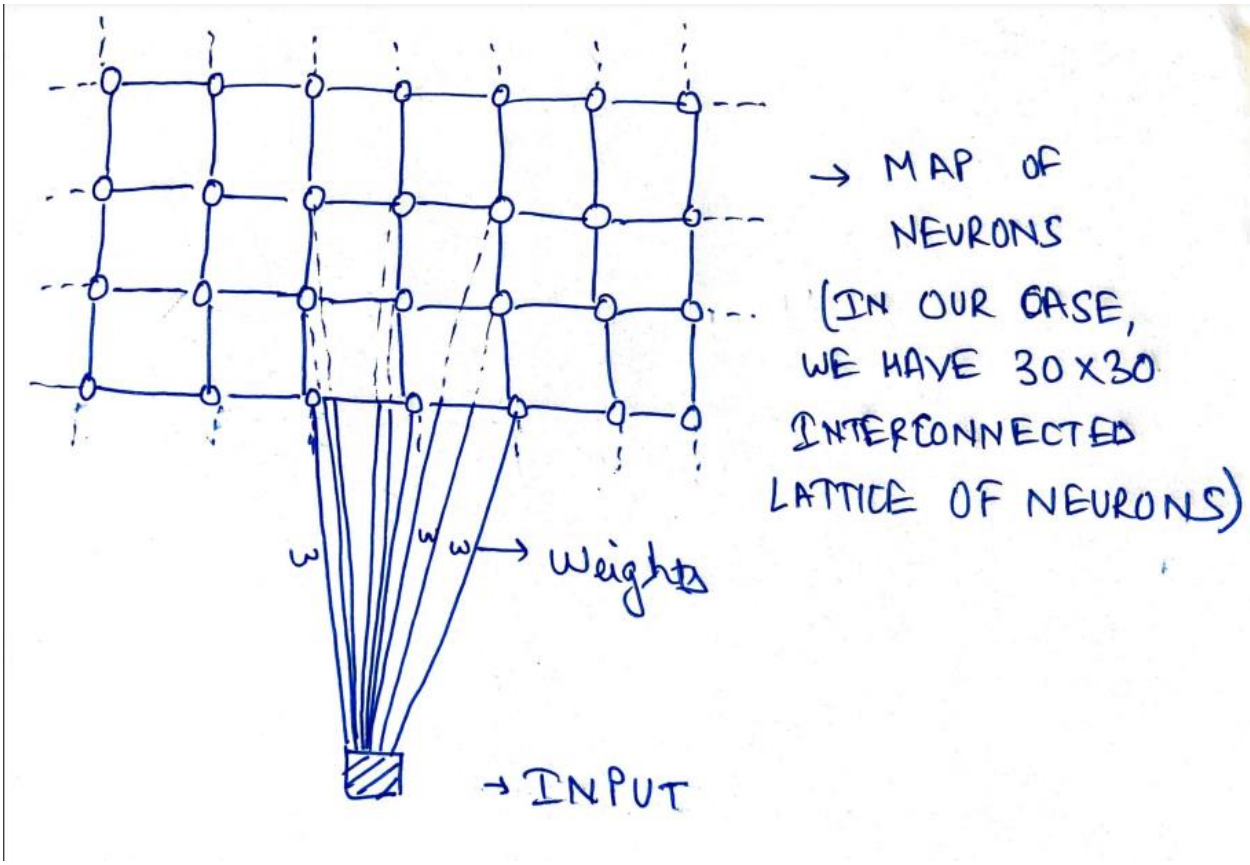


Fig 11. Diagram representing basic structure of SOM

SOM algorithm works in the following way:

- First the weights of neurons are initiated
- Then an input is selected at random
- The winning neuron is selected using the Euclidean distance (one with minimum).
- Then the neuron weights are updated
- Repeat the process from step 2 until training is done.

4. Results

Neural network – regression

- Cross validation score of pipeline with standardized data and neural network with 1 hidden layer gives us the estimate of models' performance on problem of unseen data.
- Thus we have generalized neural network model for multi-output regression
- We have the metrics as :
  - → Mean squared error ~ 11
  - → RMSE ~ 3.3 (Root mean squared error is the standard deviation of prediction errors. Prediction errors are a measure of how far from the regression line the datapoints are.)
  - → We have obtained low RMSE indicating that our model fits well and is able to predict the multi outputs also quite well

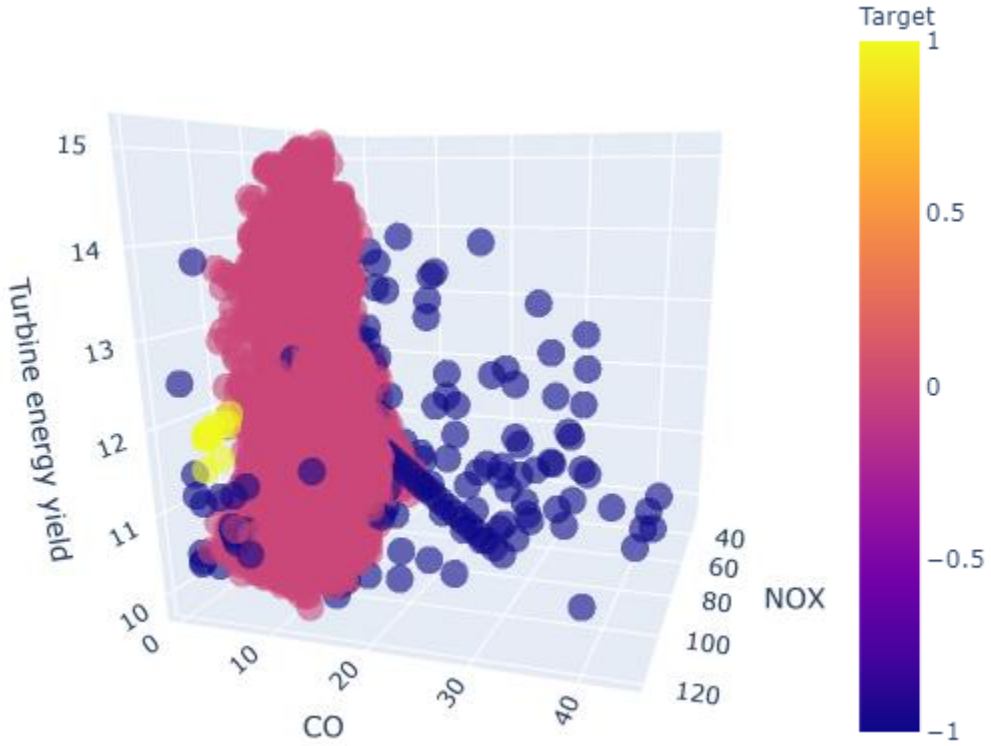
Chained Linear Regressor

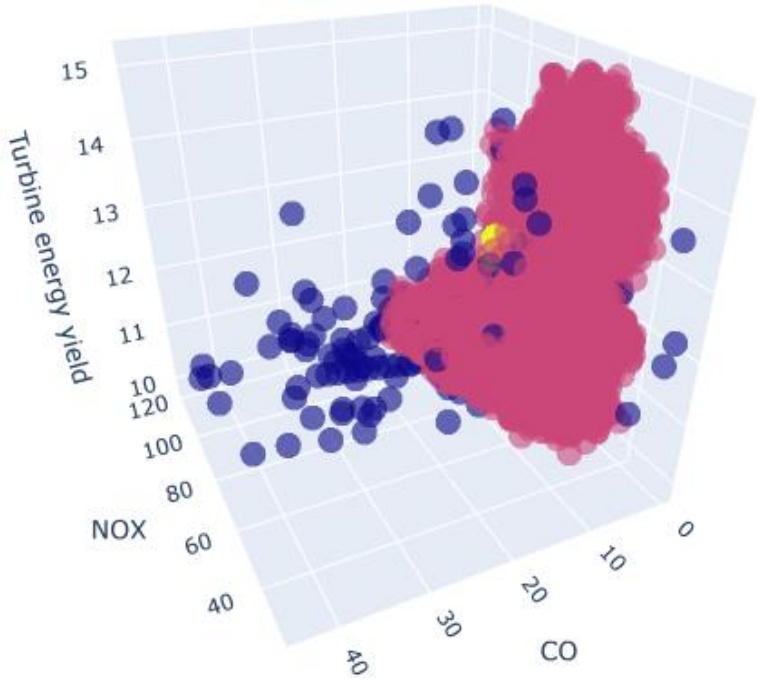
- We have the metrics as :
  - → the MSE (mean squared error) =0.4 which tells us that our model fits very well and
  - → Root mean squared error RMSE ~ 0.62
  - → Thus we can see that the chained regressor gives us much better fit when we need to predict multiple target features

Comparing the above two models, it is evident that the chained regressor here performs much better than the neural network

DBSCAN – (Density based spatial clustering of application with noise)

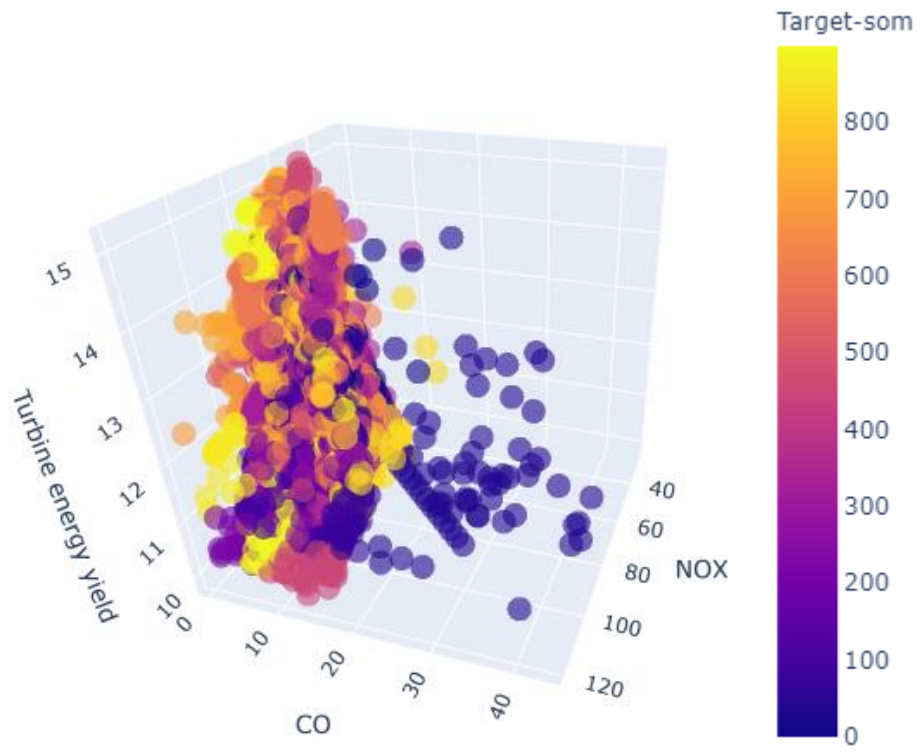
- Noise datapoints – 149
- Number of clusters – 2

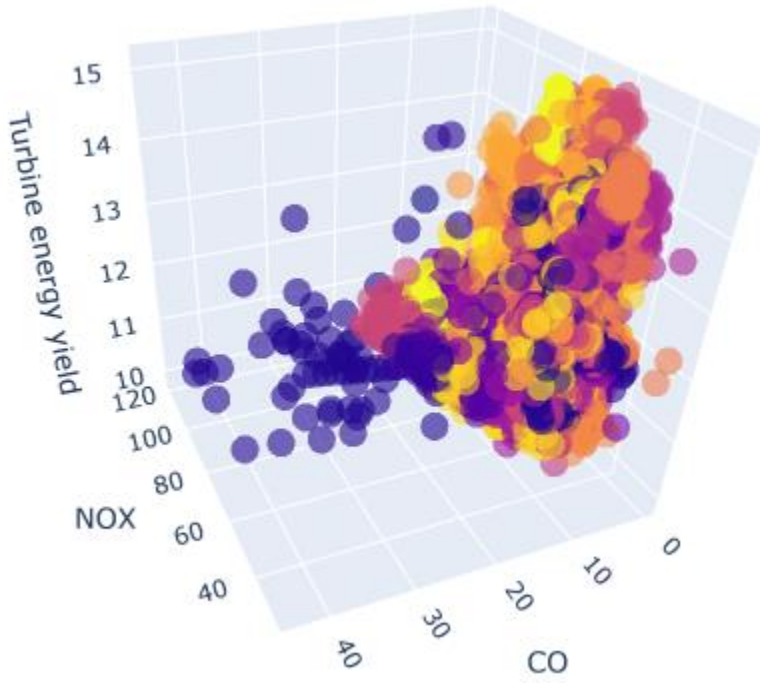




- By visual inspection , we can say that the goal we had in mind earlier (to cluster the datapoints into 2 clusters) is well completed.
- Since the true cluster labels are unknown, the model itself will be used to evaluate performance.
- Silhouette coefficient (SC) will be calculated and we evaluate using it as:
  - →SC is between -1 and 1
  - →if SC=1 , we have the perfectly defined dense clusters
  - →if SC=0, the datapoints are overlapping
  - →if SC= -1 , the datapoints are assigned to wrong cluster
- We get the silhouette coefficient = 0.560892044400304
- <u>We have got a decent silhouette score ,indicating that our DBSCAN model does not have overlapping clusters or mislabelled datapoints</u>

Self-Organizing maps (SOM)

- Time elapsed in the training process is : 0.58 seconds
- 

- 

- As you can see from above 3d plot that self-organizing maps do kind of soft clustering So basically we can

cluster the above ~ 900 clusters into 2 clusters based upon the maximum values of CO and NOX and then classify whether a datapoint is benign or malignant

- Silhouette coefficient = 0.655181249938469 5
- We can see that using the SOM , we get better defined clusters but the problem is that when we visually inspect SOM here don't seem to achieve our purpose very well because:
  - →SOM learn the similarities in input well but not the physical closeness in input data values.
  - →SOM are competitive neural networks and does not use the concept of backpropagation , thus the distribution of data will highly result of SOM.

5. Conclusion

- The basic cause of excessive gas emissions is understood without models, but the interactions are complex enough that models help in trying to fully understand all of the relationships between the components and help is controlling the emissions.
- Turbine inlet temperature and Turbine energy yield proved to be very beneficial attributes since they gives us very good idea whether emissions are excessive or not.
- Ambient features proved to be decent predictors of energy yield and flue gas emissions.
- Measures of CO and NOX alone do not give any idea whether emissions are excessive or not since they have no seemingly visible relationship between them.
- Chained regressor which exploits the correlation among the targets performs much

better when compared to neural network for multi output regression

- DBSCAN do better when the density related clustering is required but the SOM give us much better defined clusters
- With proper tuning of SOM model, there is decent chance that they might outperform DBSCAN here.
- The DBSCAN algorithm handles noise well and clusters few datapoints as noise
- SOM on the other hand are not sensitive to noise and thus cannot tell us which data points are noise.

6. Future Scope

- This type of project can be extended to other industries as well where different kind of emissions takes place.
- Automation of maintenance trigger system so that emissions are controlled.
- Determining which part can be updated in the machine to reduce the pollution by determining the factors affecting the emissions most.
- Tuning all models to better predict the targets so that the emissions can be monitored more closely.
- Only when we are able to monitor,control and predict the emissions closely, we can control the flue gases in the environment.

7. BIBLIOGRAPHY

- https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set
- https://journals.tubitak.gov.tr/elektrik/issues/elk-19-27-6/elk-27-6-54-1807-87.pdf
- https://towardsdatascience.com/build-your-own-neural-network-from-scratch-with-python-dbe0282bd9e3
- https://thecodacus.com/2017/08/14/neural-network-scratch-python-no-libraries/
- https://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/
- https://machinelearningmastery.com/deep-learning-models-for-multi-output-regression/
- https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/
- https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html
- https://stackoverflow.com/questions/12893492/choosing-eps-and-minpts-for-dbscan-r
- https://towardsdatascience.com/explaining-dbscan-clustering-18eaf5c83b31
- Machine Intelligence - Lecture 7 (Clustering, k-means, SOM) University of Waterloo : https://www.youtube.com/watch?v=lFbxTlD5R98
- SELF ORGANISING MAPS: INTRODUCTION: https://www.youtube.com/watch?v=0qtvb_Nx2tA

- https://stackoverflow.com/questions/43160240/how-to-plot-a-k-distance-graph-in-python
- https://medium.com/machine-learning-researcher/self-organizing-map-som-c296561e2117
- som documentation - https://github.com/JustGlowing/minisom/blob/master/examples/Clustering.ipynb

THANK YOU