REPORT

Project Title: Predicting Turbine Energy Yield(TEY) and gas emissions from gas turbines

Abstract: In Gas turbine based power plants the gas emissions (mostly CO and NOx) are exceedingly jeopardizing the environment. These emissions need to be closely controlled and monitored. In this paper, we introduce an Emissions control dataset collected over five years from a gas turbine for the predictive modeling of the CO and NOx emissions as well as the Energy Yield. We analyze the data using multiple machine learning models, and present useful insights about emission predictions. We use Regression models for the Gas emission predictions and further, use Clustering Machine Learning models to classify the Energy Yield Ranges. Furthermore, we analyze what the more accurate training models, amongst the ones used, for this set of data are.

Keywords: Regression, Clustering, CO emission, NOx emission, Gas Turbine

Introduction: In this project we will focus on how Gas turbine CO and NOx Emission impacts the energy yield of a Turbine using various machine learning models. We will use ambient features such as Turbine Inlet Temperature and Compressor Discharge pressure along with CO and NOx emission.

1. **Proposed Methodology**
   a. **Datasets**: Gas Turbine CO and NOx Emission Data Set

      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

Data Set Characteristics:	Multivariate
Number of Instances:	36733
Attribute Characteristics:	Real
Number of Attributes:	11
Associated Tasks:	Regression, Clustering
Missing Values:	N/A

Attribute Information:

1. Ambient temperature (AT) °C
2. Ambient pressure (AP) mbar
3. Ambient humidity (AH) (%)
4. Air filter difference pressure (AFDP) mbar
5. Gas turbine exhaust pressure (GTEP) mbar
6. Turbine inlet temperature (TIT) °C
7. Turbine after temperature (TAT) °C

8. Compressor discharge pressure (CDP) mbar
9. Turbine energy yield (TEY) MWH
10. Carbon monoxide (CO) mg/m3
11. Nitrogen oxides (NOx) mg/m3

Correlation matrix of various attributes:

Table 1.1

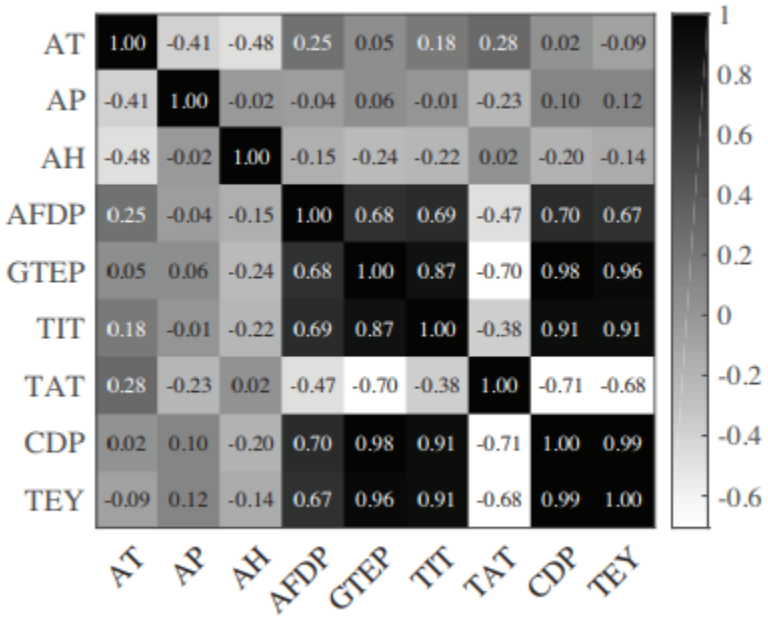


Table 1.2

Pairwise correlation between features and the two target variables

Feature	CO	NO_x
AT	–0.174	–0.558
AP	0.067	0.192
AH	0.107	0.165
AFDP	–0.448	–0.188
GTEP	–0.519	–0.202
TIT	–0.706	–0.214
TAT	0.058	–0.093
CDP	–0.551	–0.171
TEY	–0.570	–0.116

b. Pre processing:

i. For prediction of CO and NOx using Regression

1. For Univariate Regressions

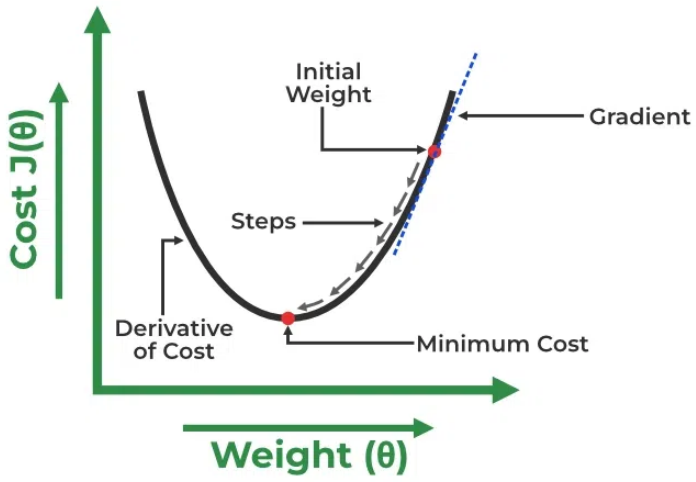
Independent variable = CO / NOx
Dependent variable = Turbine Inlet Temperature(TIT) / Atmospheric temperature(AT)

2. For Multivariate Regressions
   Independent variable = CO / NOx
   Dependent variable = [TIT,GTEP,CDP,TEY] / [AT,GTEP,TIT]

c. Regression analysis

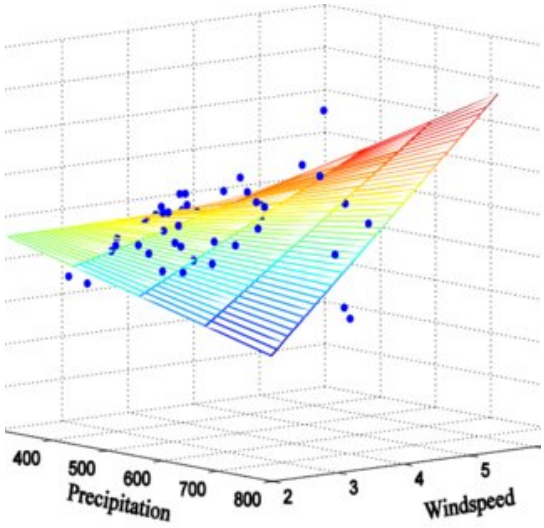
i. Linear regression(Gradient descent)

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.



ii. Multiple regression

Multiple linear regression, often known as multiple regression, is a statistical method that predicts the result of a response variable by combining numerous explanatory variables. Multiple regression is a variant of linear regression (ordinary least squares) in which just one explanatory variable is used. Example graph shown below:



iii. Bayesian Linear regression

In the Bayesian viewpoint, we formulate linear regression using probability distributions rather than point estimates. The response, y, is not estimated as a single value, but is assumed to be drawn from a probability distribution. The model for Bayesian Linear Regression with the response sampled from a normal distribution is:

$$y \sim N(\beta^T X, \sigma^2 I)$$

The aim of Bayesian Linear Regression is not to find the single "best" value of the model parameters, but rather to determine the posterior distribution for the model parameters. Not only is the response generated from a probability distribution, but the model parameters are assumed to come from a distribution as well. The posterior probability of the model parameters is conditional upon the training inputs and outputs:

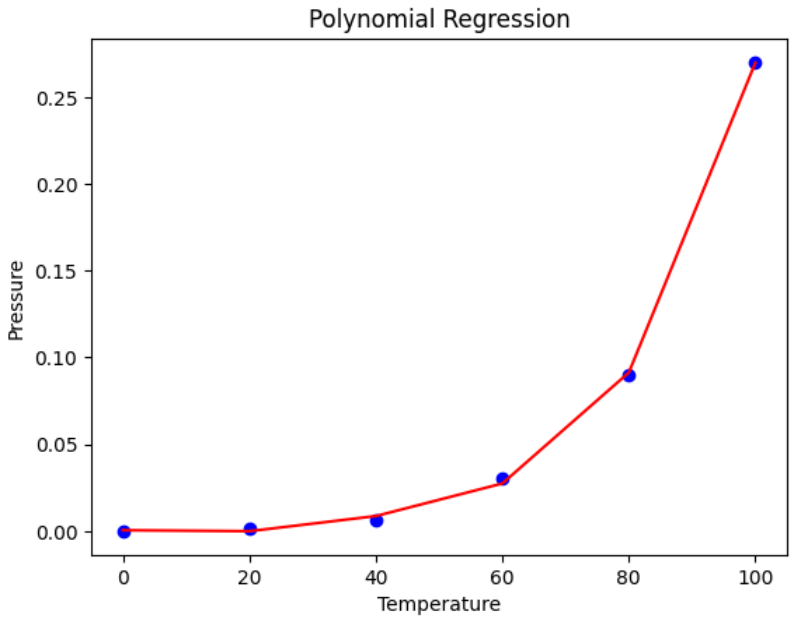
$$P(\beta|y,X) = \frac{P(y|\beta,X) * P(\beta|X)}{P(y|X)}$$

iv. Polynomial regression

There are some relationships that a researcher will hypothesize is curvilinear. Clearly, such types of cases will include a polynomial term. If we try to fit a linear model to curved data, a scatter plot of residuals (Y-axis) on the predictor (X-axis) will have patches of many positive residuals in the middle. Hence in such a situation, it is not appropriate.

An assumption in the usual multiple linear regression analysis is that all the independent variables are independent. In the polynomial regression model, this assumption is not satisfied.

Example graph shown below:



v. Robust regression

Linear regression fits a line or hyperplane that best describes the linear relationship between inputs and the target numeric value. If the data contains outlier values, the line can become biased, resulting in worse predictive performance. Robust regression refers to a suite of algorithms that are robust in the presence of outliers in training data.

Huber regression is a type of robust regression that is aware of the possibility of outliers in a dataset and assigns them less weight than other examples in the dataset. Example graph shown below:

d. Clustering Models

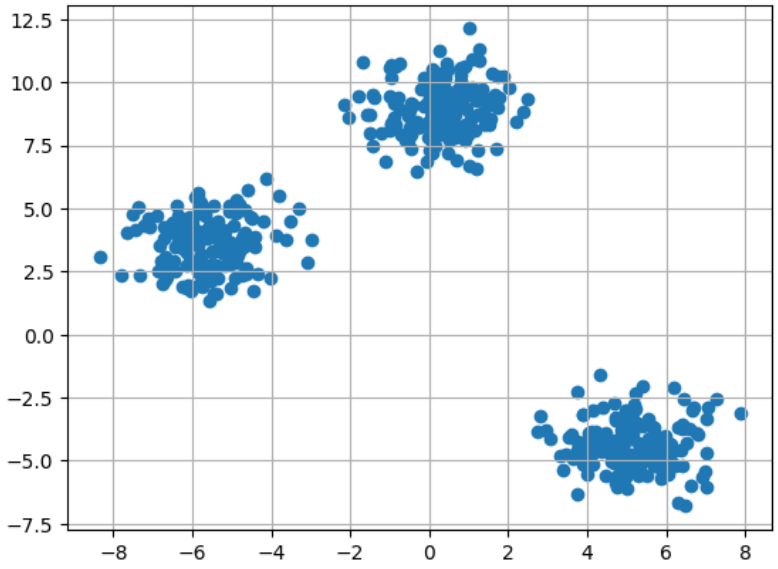
i. K-means clustering

The process of teaching a computer to use unlabeled, unclassified data and enabling the algorithm to act on that data without supervision is known as unsupervised machine learning. The machine's task in this scenario is to arrange unsorted data according to parallels, patterns, and variations without any prior data training.

In order to make the data points within each group more comparable to one another and distinct from the data points within the other groups, clustering divides the population or set of data points into a number of groups. In essence, it is a classification of things according to how similar and dissimilar they are to one another.

A data set of items with specific features and values for these features is provided to us (much like a vector). The assignment is to group those products into categories. We will employ the unsupervised learning algorithm K-means to do this. The number of groups or clusters we wish to divide our items into is indicated by the letter "K" in the algorithm's name.

The items will be divided up into k groups or clusters of resemblance by the algorithm. We will use the euclidean distance as a measurement to determine that similarity. Example graph shown below:



ii. Gaussian Mixture model clustering

In one dimension the probability density function of a Gaussian Distribution is given by:

$$G(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

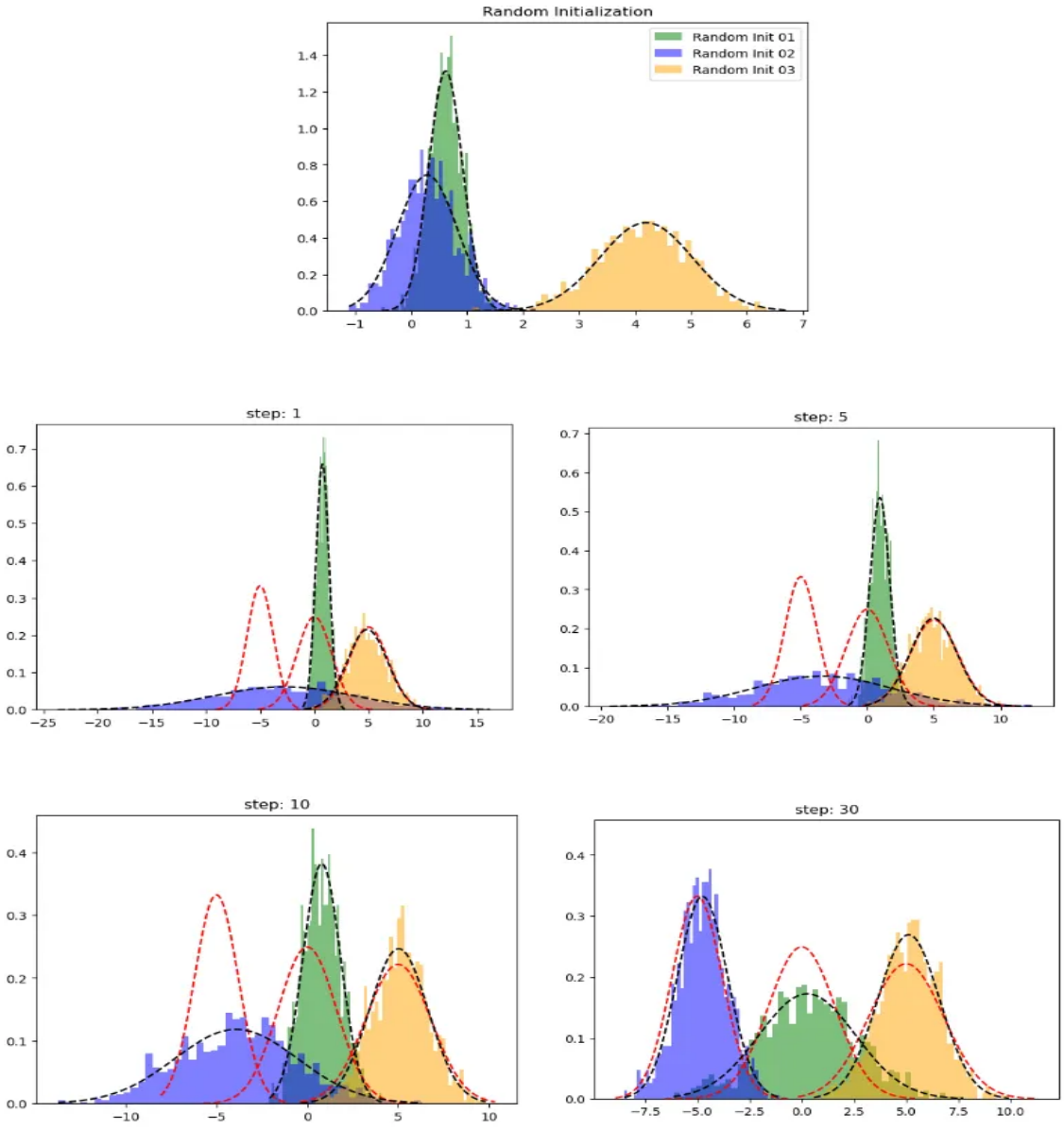
where μ and σ^2 are respectively the mean and variance of the distribution. For Multivariate (let us say d-variate) Gaussian Distribution, the probability density function is given by

$$G(X|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(X-\mu)^T\mathbf{\Sigma}^{-1}(X-\mu)\right)$$

Here μ is a d dimensional vector denoting the mean of the distribution and Σ is the d X d covariance matrix.

GMM:

Determine the number of clusters for the provided dataset . Assume that there are 1000 data points and that there are only 2 categories. Set each cluster's mean, covariance, and weight parameters. Perform the following using the Expectation Maximization technique. Using the most recent estimate of the parameters, calculate the likelihood that each data point will belong to each distribution. To increase the predicted likelihood discovered in the E phase, modify the prior mean, covariance, and weight parameters. Up until the model converges, repeat these procedures. Example graph shown below:



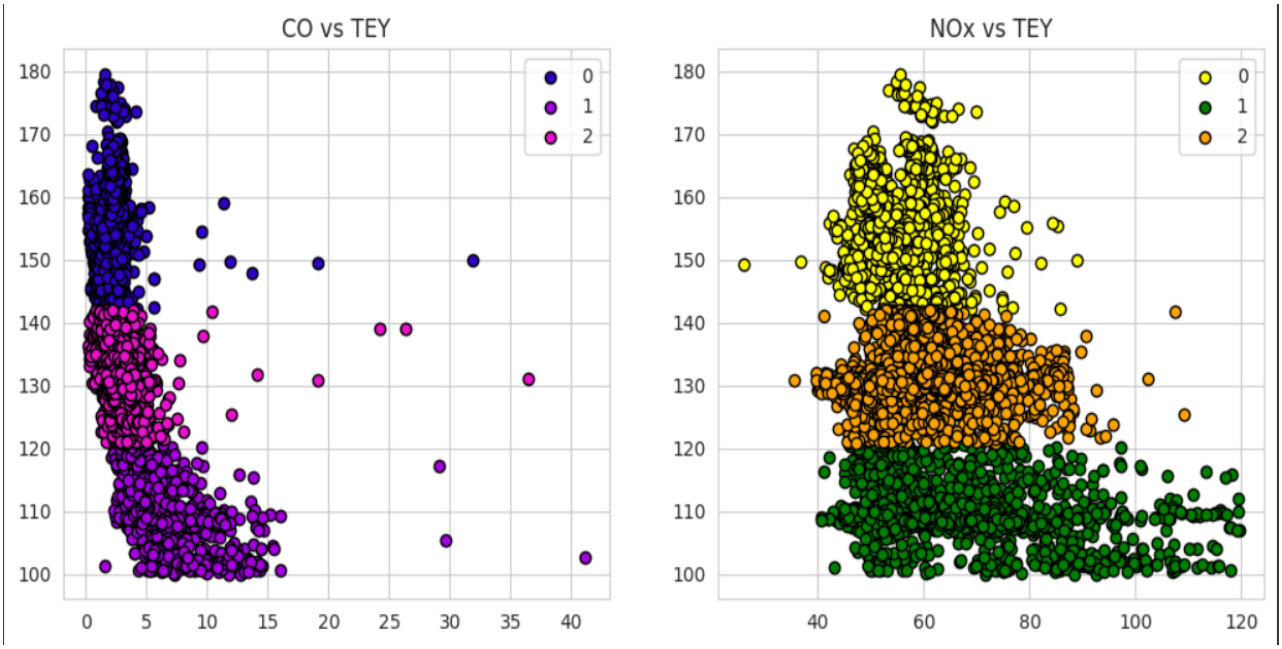
e. Comparison of different models

Table 2.1

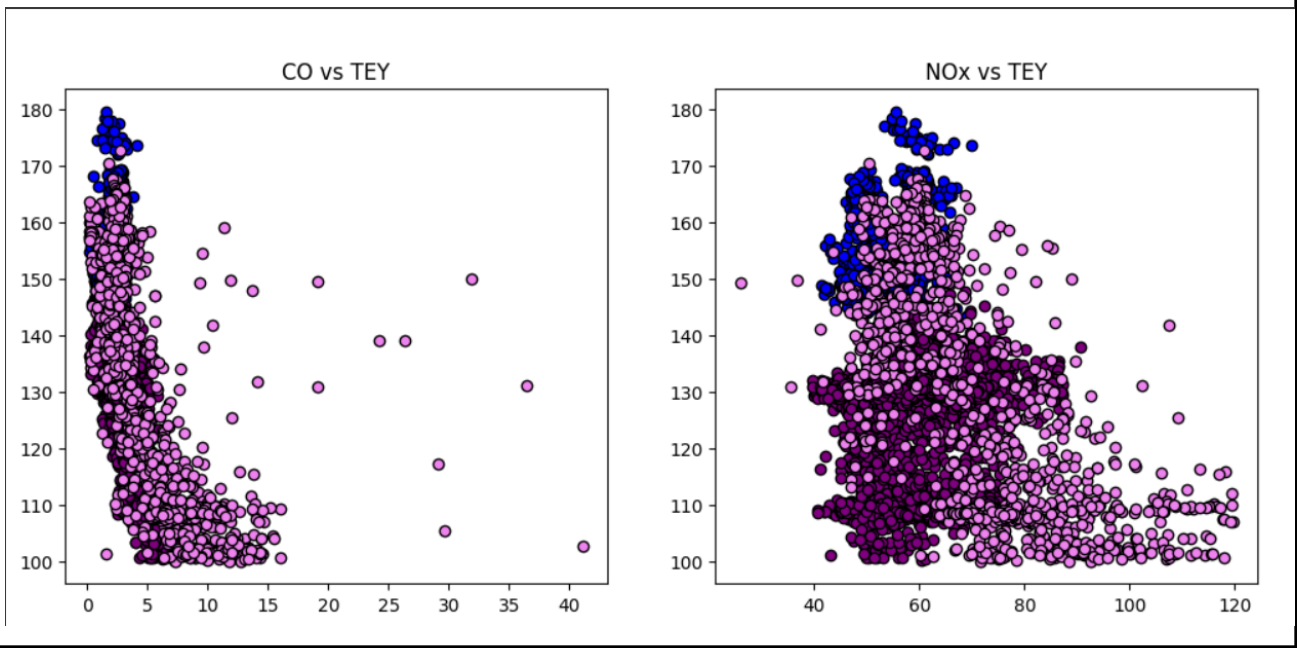
Sr. No.	Model	Mean Absolute Error	
		CO	NOx
1.	Univariate Linear Regression	0.826	6.379
2.	Multivariate Linear Regression	0.711	6.193
3.	Bayesian Linear Regression	0.735	5.971
4.	Polynomial Regression	0.520	3.334
5.	Huber Robust Regression	0.989	5.968

Table 2.2

K-Means Clustering



Gaussian Mixture Model



2. Result & Discussion

a. For the given dataset, the most accurate prediction model (for CO and NOx emissions) amongst the tried models is the Polynomial Regression model with minimum Mean Absolute error of 0.520 for CO and 3.334 for NOx.

b. For the given dataset, the most accurate Clustering model (for classification of Turbine Energy Yield into ranges) amongst the tried models is the K-Means model with more definite clusters as seen in table 2.2

3. Conclusion & Future Work

The CO and NOx emissions from such power plants are deadly for the environment albeit the product being important. In order to combat these rising pollution levels, this study can be used to determine the gas levels due to a plant and in turn to minimize said gas levels.

We have concluded a way to predict the gas levels based on the ambient features of the power plant/environment as well as a path to classify the Energy yield on the levels of these hazardous gasses.

4. References

a. Modern Regression Methods - Thomas P. Ryan
b. Data Preprocessing for Supervised Learning - S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas
c. Multiple Linear Regression - Mark Tranmer, Jen Murphy, Mark Elliot, Maria Pampaka
d. Introduction to Bayesian Linear Regression - Will Koehrsen
e. Geeksforgeeks.com
f. Understanding logistic regression
g. sklearn.cluster.KMeans
h. K-Means Clustering Algorithm - Javatpoint
i. Gaussian Mixture Model Clearly Explained
j. An Overview on Clustering Methods - T. Soni Madhulatha
k. Applied Regression: An Introduction - Colin Lewis-Beck, Michael Lewis-Beck
l. Robust Regression for Machine Learning in Python
m. Research Gate
n. Regression Analysis
o. Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS