

**Capstone Project Draft Report**

ALY6140 – Analytics System Technology Professor Joel Schwartz

2/16/22

By: Sneha Sharma Satheesha

NUID:002954432

# Introduction and Rationale

Air pollution in India is a serious health issue. Of the 30 most polluted cities in the world, 21 were in India in 2019. As per a study based on 2016 data, at least 140 million people in India breathe air that is 10 times or more over the WHO safe limit and 13 of the world's 20 cities with the highest annual levels of air pollution are in India.Data analytics provides a promising tool for solving air pollution management problems. The necessity for predicting Air Quality Index (AQI) incidentally is more suitable when it is known that data sets are available for free of charge and at a nominal cost from meteorological centers. An efficient tool once built for abstracting the large quantities of data sets that are available and deriving useful knowledge from them can help to detect the vulnerability of the exposed people in the survey areas. This will enable planners to utilize the information to improve the health conditions of these areas for effective air pollution mitigation and management. Hence, we are picking a dataset of the air particles from different cities in India.

Our primary goal for this project is to predict AQI, given the compositions of various gases in the environment at that point in time. Along with Air Quality Index, we also want to determine what gases have the highest effect on AQI.

## Questions to Investigate

1. What are all the features which affect the air quality index the most?
2. What are the features which affect the air quality index the least?

## Goals

Predict the Air Quality index given the quantities of various chemicals in the environment. Predict the bucket in which a particular air quality index falls given the quantities. Implement multiple models and find the most efficient ones.

# 2. Exploratory Data Analysis

## 2.1 Data Set

We have picked a dataset from Kaggle, and it has 16 columns and 29,531 rows. Since Kaggle is not the most reliable place to pick the dataset from this data needs to be cleaned to match the requirements.

1. PM2.5 refers to particles in the air that have diameters of less than 2.5 micrometers.
2. PM10 are particles with less than 10.
3. NO is nitrogen oxide.
4. NO2 is nitrogen dioxide
5. NOx is when nitrogen is combined with x number of oxygen atoms.
6. NH3 is ammonia
7. CO is carbon monoxide 8. SO2 is Sulphur dioxide and

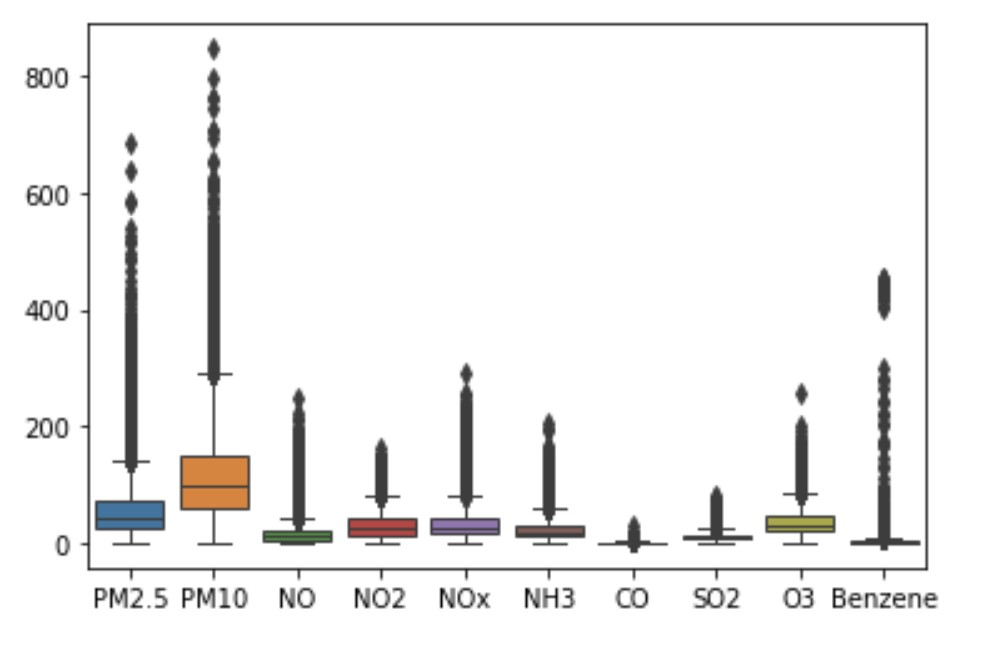
9. O3 is ozone.

These elements along with benzene, toluene, and xylene are what causes air pollution. In our project, we try to determine which of these are most significant to determine AQI (Air quality index) and what relationship this AQI has with these other particulars. We can then access what can be done to improve the air quality of any city.

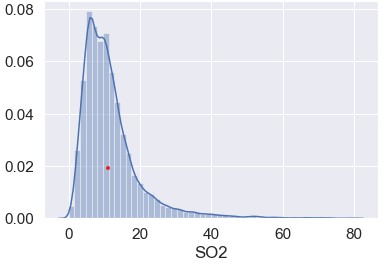
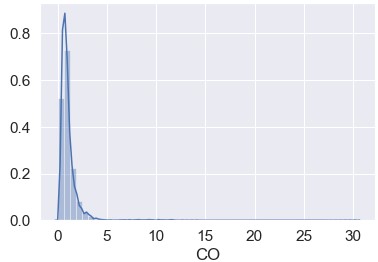
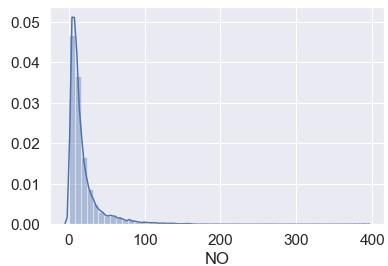
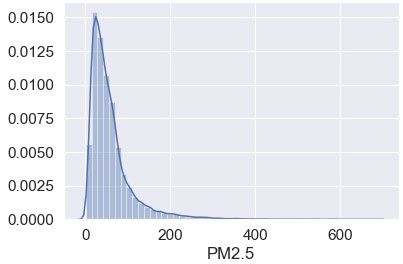
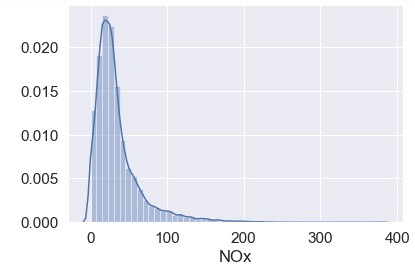
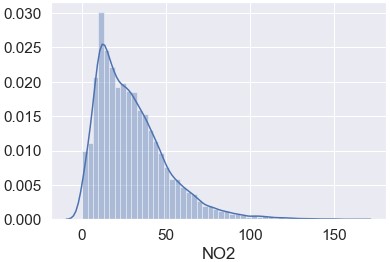
## 2.2 Initial Data Visualization

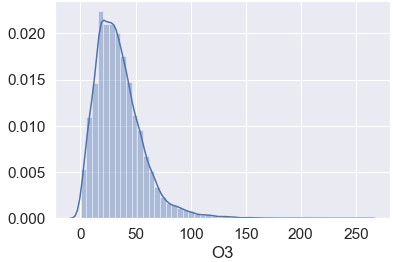
Below we have graphs depicting what the data looks before cleaning them. We can observe that all the graphs are right skewed and, in the boxplot they are after the Q3+1.5\*IQR.

***Graph 1***



### Graph 2





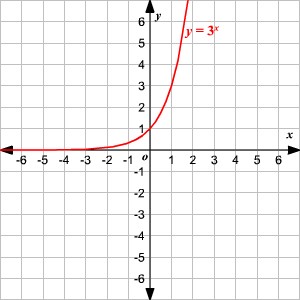
## 2.3 Data Cleaning

Data cleansing is the process of identifying corrupt, inaccurate, or irrelevant data. This is the most critical or important step in data science. Not cleaning the data properly could have dire consequences on the end results and the accuracy of the model. There are different ways in which we can clean the data:

1. We can remove the entire rows or columns
2. We can use different statistical methods to fill the missing data
   1. Mean
   2. Median
   3. Mode
   4. Imputation- infer them from the known part of the data

The steps followed for data cleansing is to first identify the critical data fields, collect the data, discard duplicate values, resolve empty values, standardize the cleaning process (which data is most often used, when it needs to be cleaned and who is responsible for the cleaning process).

We then review, adapt, and repeat the process.

Xylene and Toluene data were missing nearly 50% of the records, so we deleted those variables from the data frame for further analysis. The graph 1 in section 2.2 above depicts how the data looked before cleaning. Few of the variables like PM2.5, NO, and CO had less than 2% missing values. We decided to use imputation method here. We can observe in the graphs that the data is highly skewed, and the number of outliers is also high which will affect the mean value for these features and mode is generally recommended when we have string values. So, we use Median method to replace the missing values.

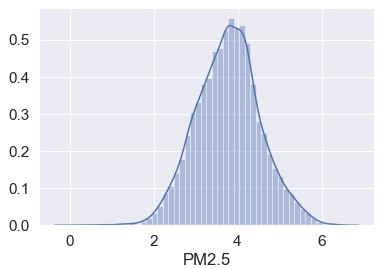
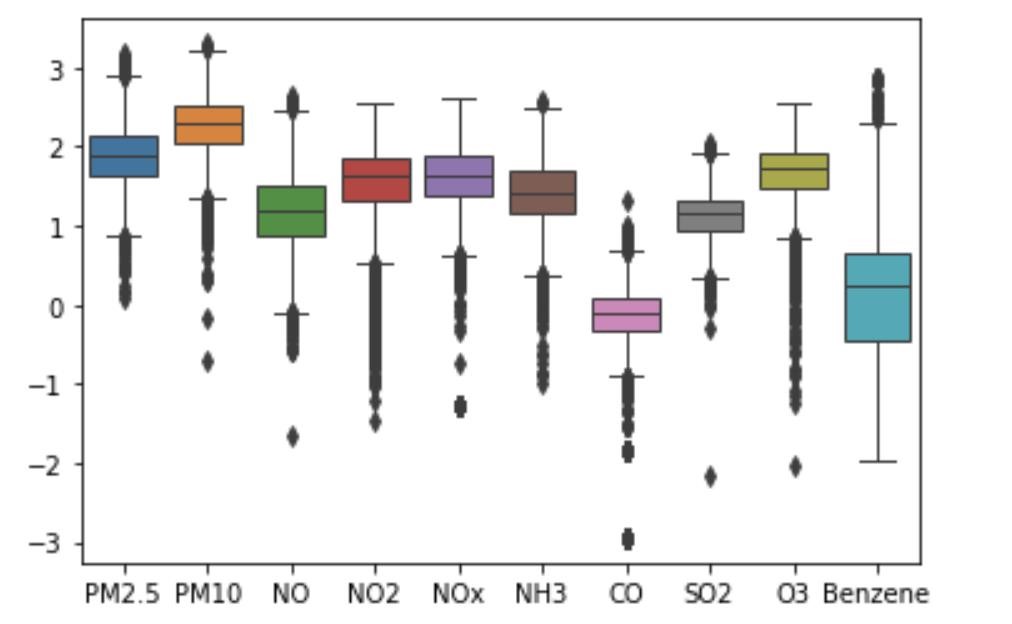
We still had 2 problems: Data is not uniformly distributed and has a very huge range and Large number of Outliers. To solve the above problems, we chose to transform data by applying the log of each value [1] Applying the log function reduces the range of data and give us uniform data and normally distributed graph. It can be explained better using the e graph. **Box plot of cleaned Data**

### Graph 3

we can observe that the graph for PM2.5 is now uniform.

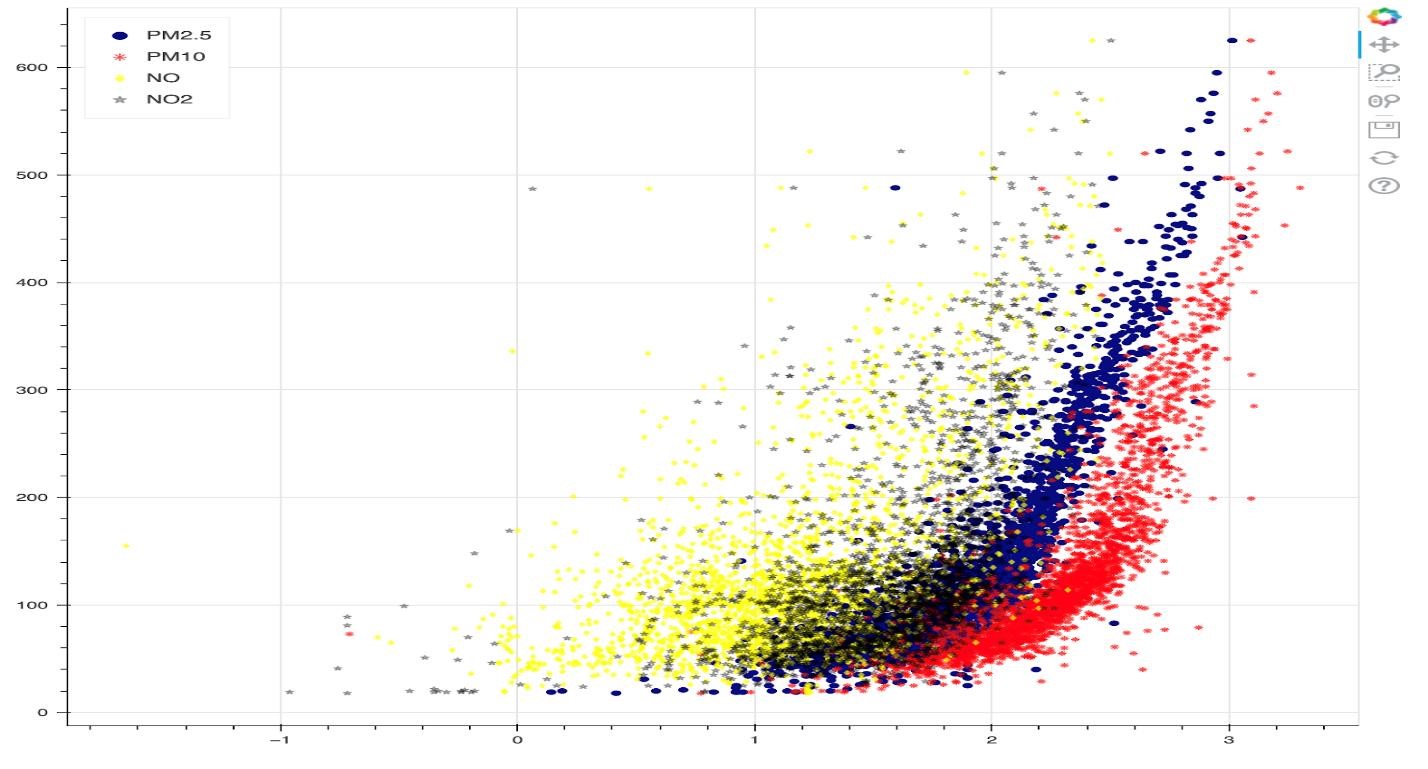
We apply the same log function to all the other features to

get normally distributed data

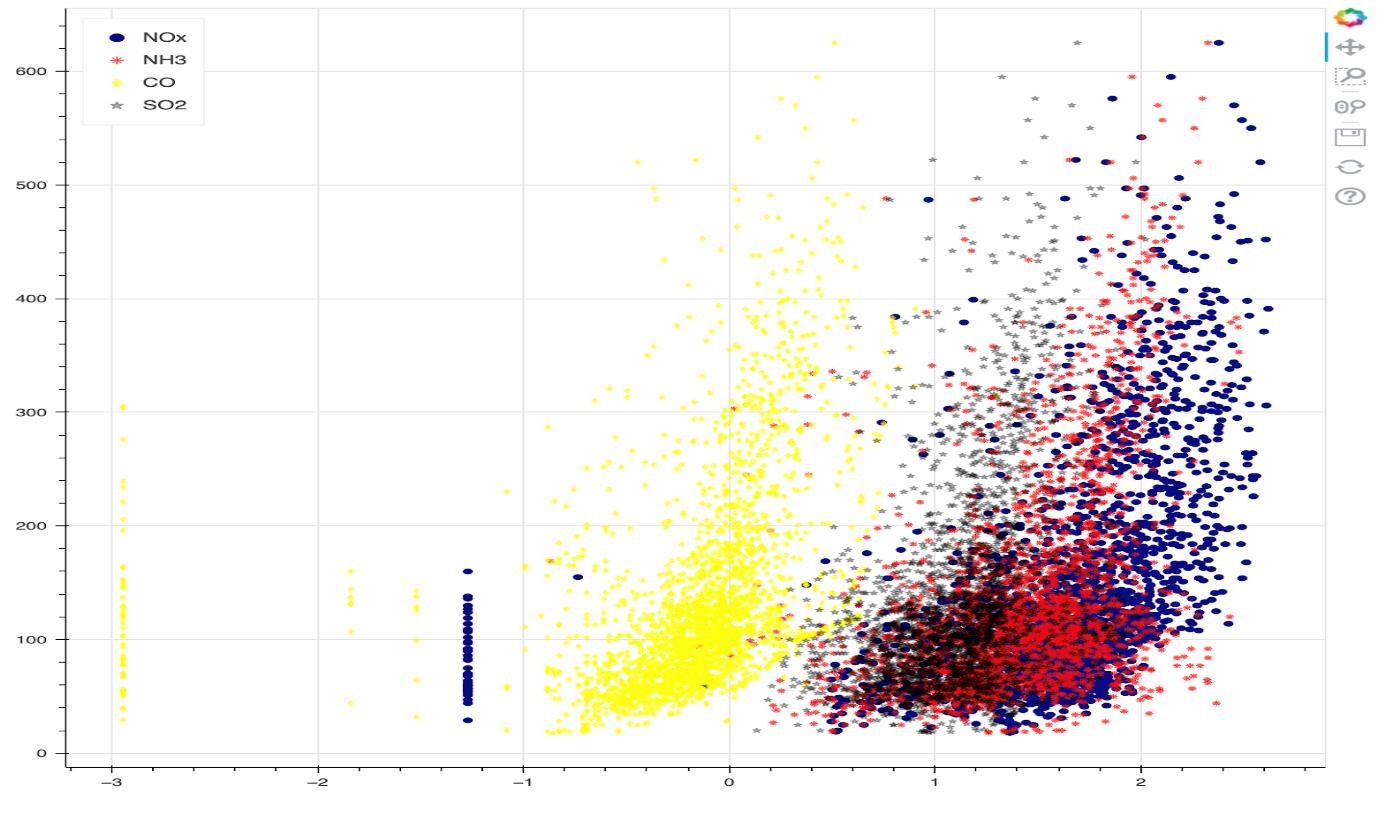


## 2.3 Further Visualizations of data distribution with respect to AQI

### Graph 4



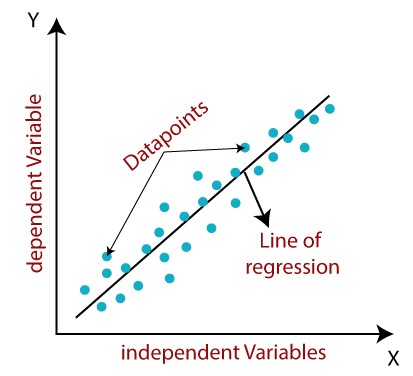
### Graph 5



# 3 Models Implemented

## Models for Predicting Air Quality index - Regression

1. **Linear Regression:**

It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables. The linear regression model provides a sloped straight line representing the relationship between the variables [2] There are 2 types of linear regression, simple and multiple linear regression. Simple for when we have a single independent variable and multiple for when we have multiple independent variables. These are the variables we use to predict our dependent variable which is prices in this case. The formula would be y=mx+c.

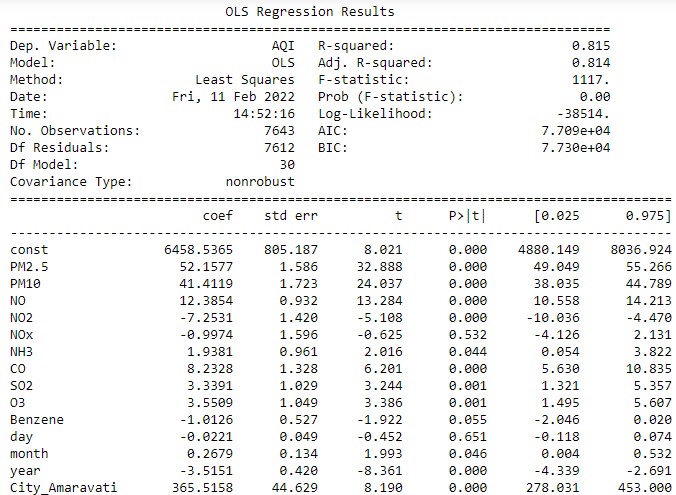
We can use sklearn or statsmodels library in python for our model.

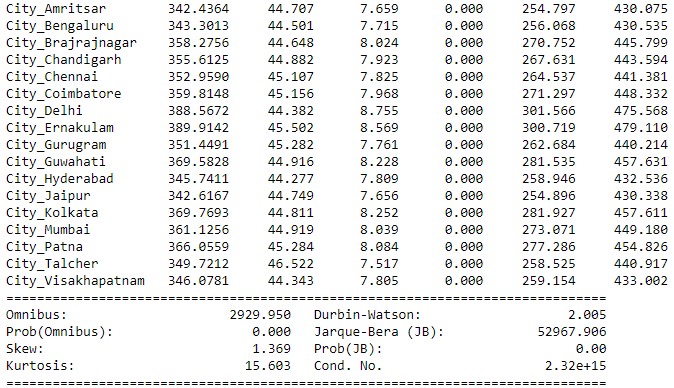
*Why?*

* 1. To find AQI
  2. To create a baseline model to compare with other models
  3. To get how significance and impact of the features

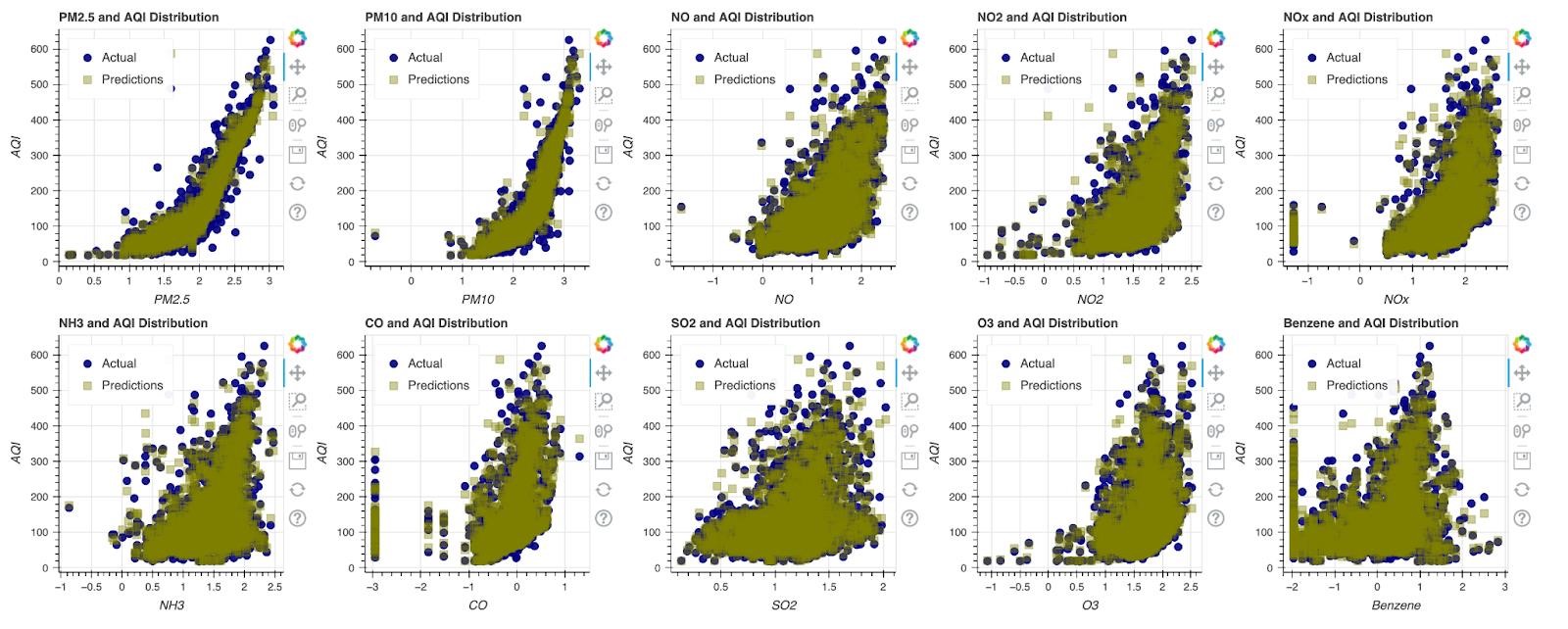
**Results:**

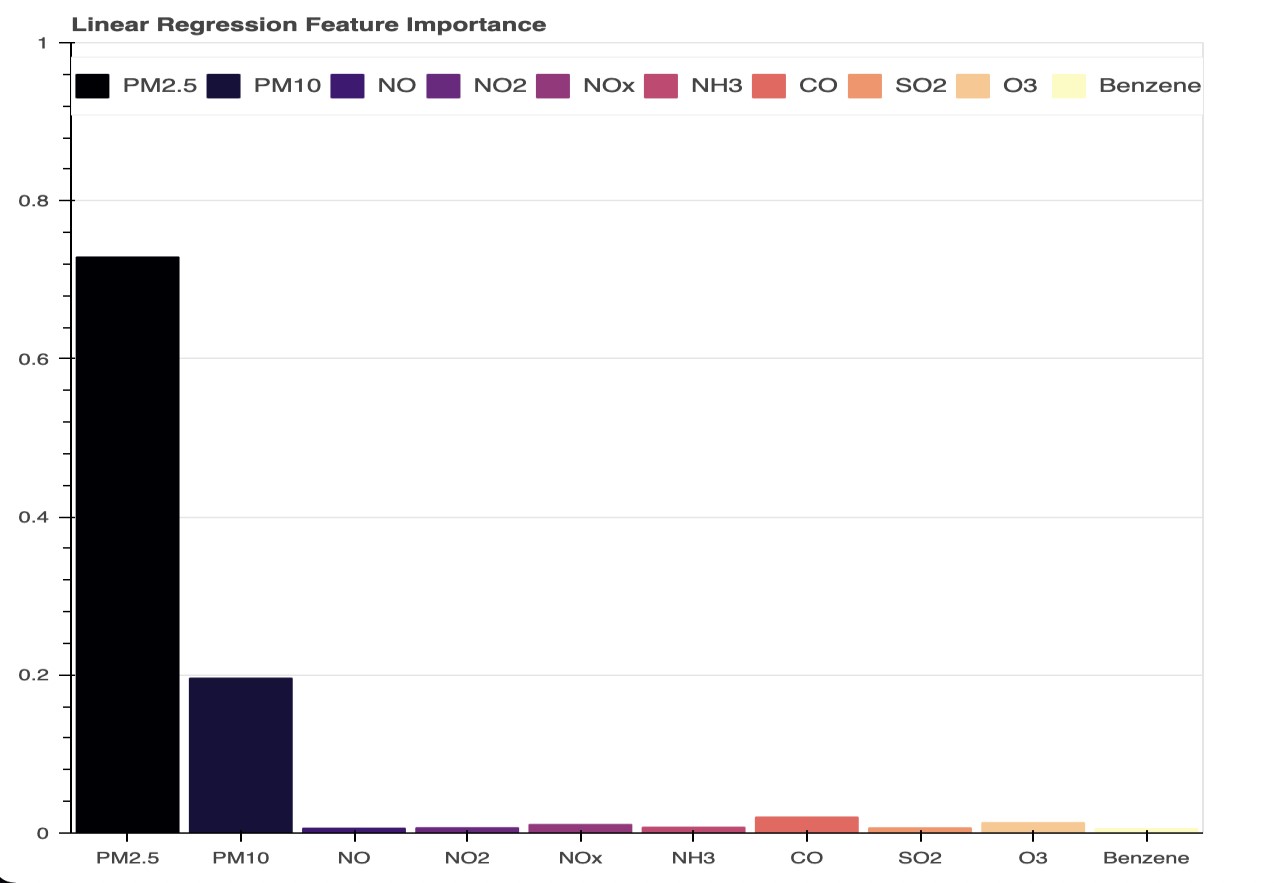
We can analyze which feature is most significant by checking the p>|t| values in the below result. The lower the p value, more is the significance of the feature. We can also check whether the feature has a positive or negative impact on the results based on the coefficient values.





*Accuracy: 90.9%*





1. **Random Forest Regression:**

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

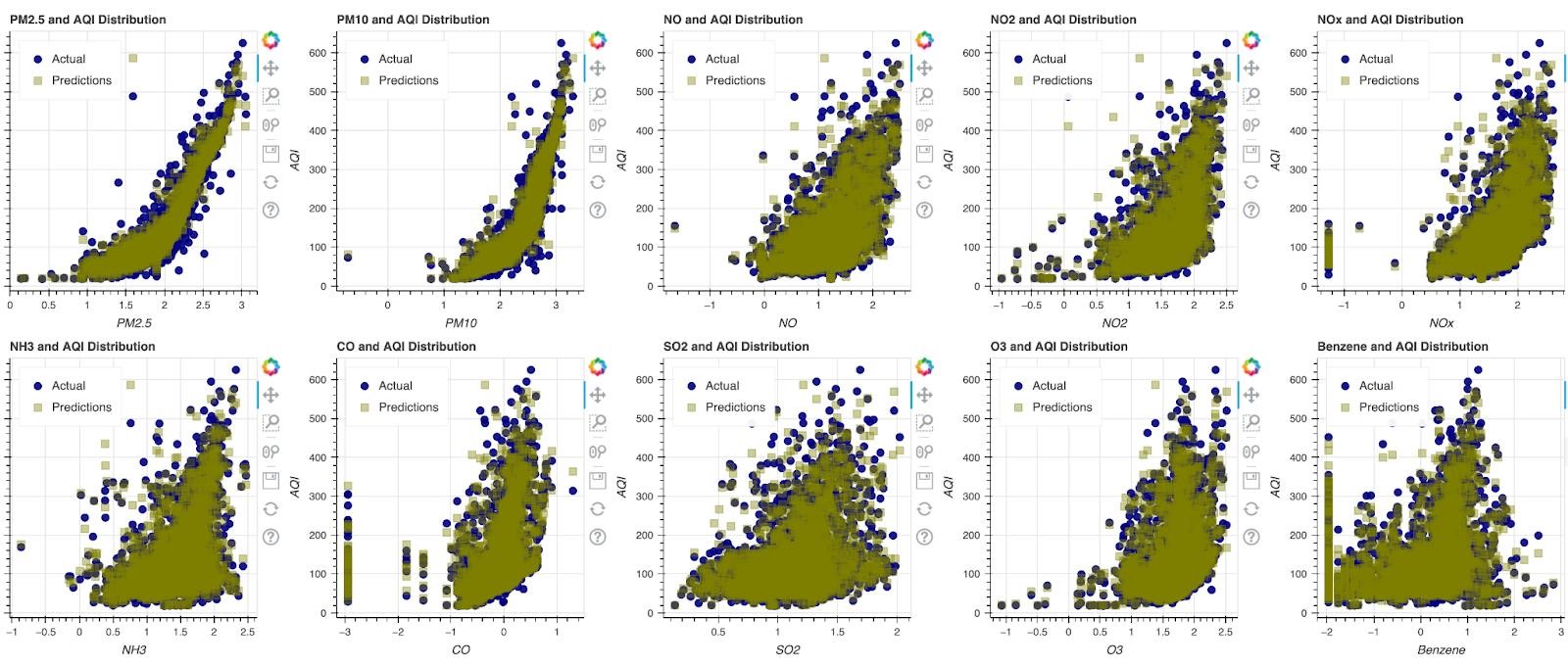
*Why?*

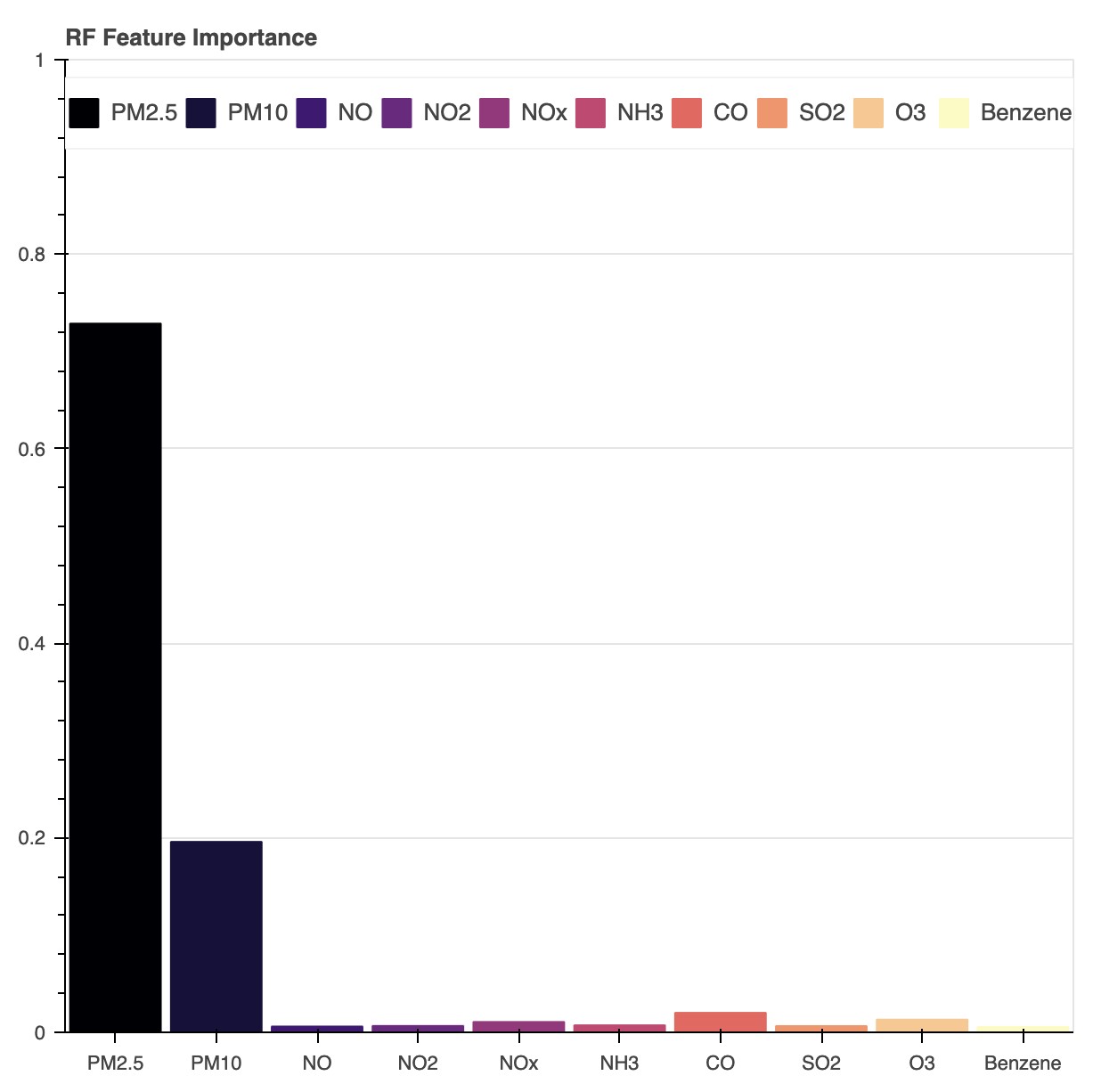
1. To get an accurate prediction of AQI
2. Analyze feature importance
3. It works great on nonlinear relationships
4. Deals with scattered data very well

***Results*** :

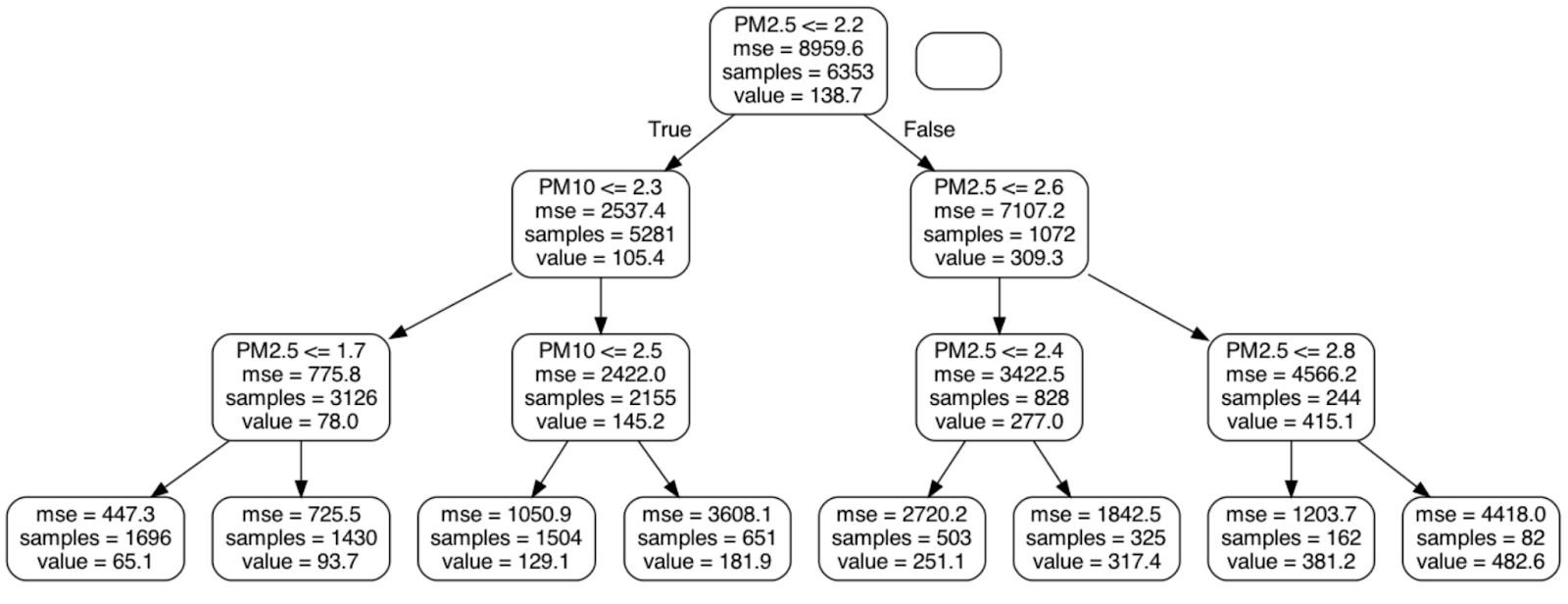
*Accuracy: 94.3%*

*Rf score: 0.94*



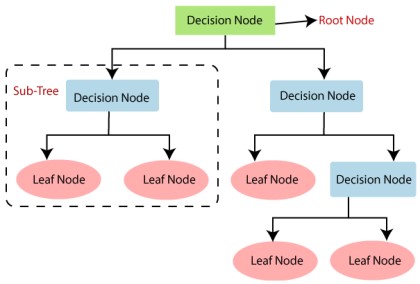


**Simple Visualization of Random Forest**



## Models for predicting Air Quality Index Bucket - Classification

1. **Decision Trees**

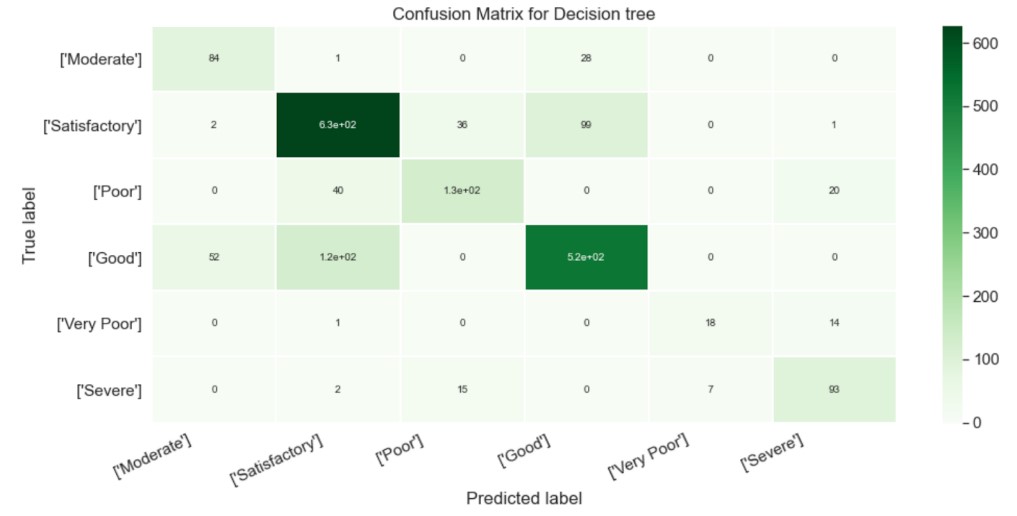
Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees. We are using the Entropy criteria to do this classification. The splits aim to maximize the decrease in entropy i.e., data impurity. We see that same as logistic regression due to the imbalance in the data, we have good negative predictions but our model is not good at predicting positive values. We now plot a feature importance graph, which is derived from which feature was repeated the most in the tree.

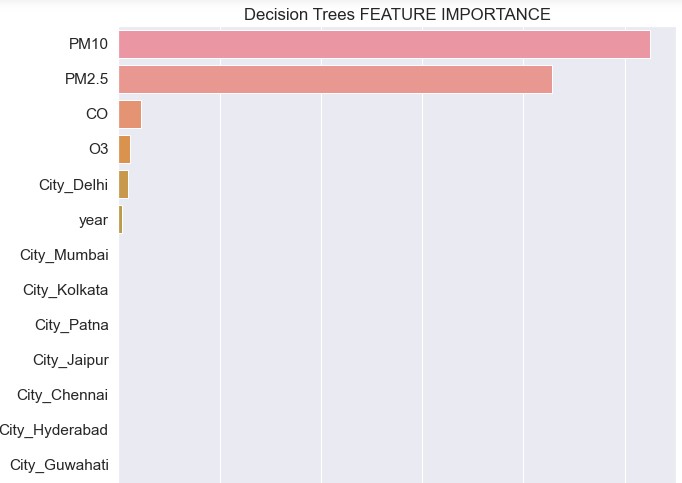
*Why?*

* 1. Decision Trees are one of the most effective classification algorithms
  2. It works well with Non-Linear data

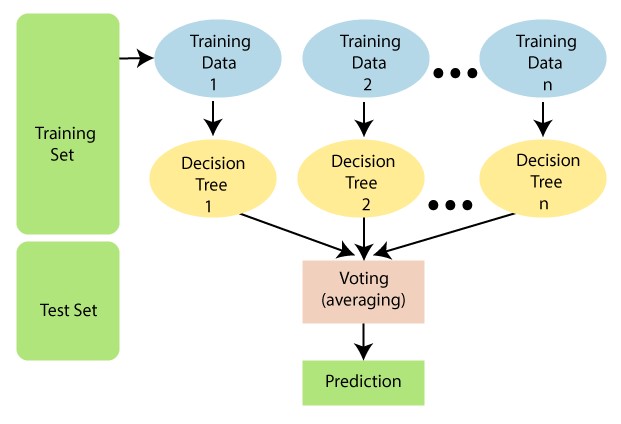
***Results:***

*Accuracy: 73%*





1. **Random Forest Classification:**

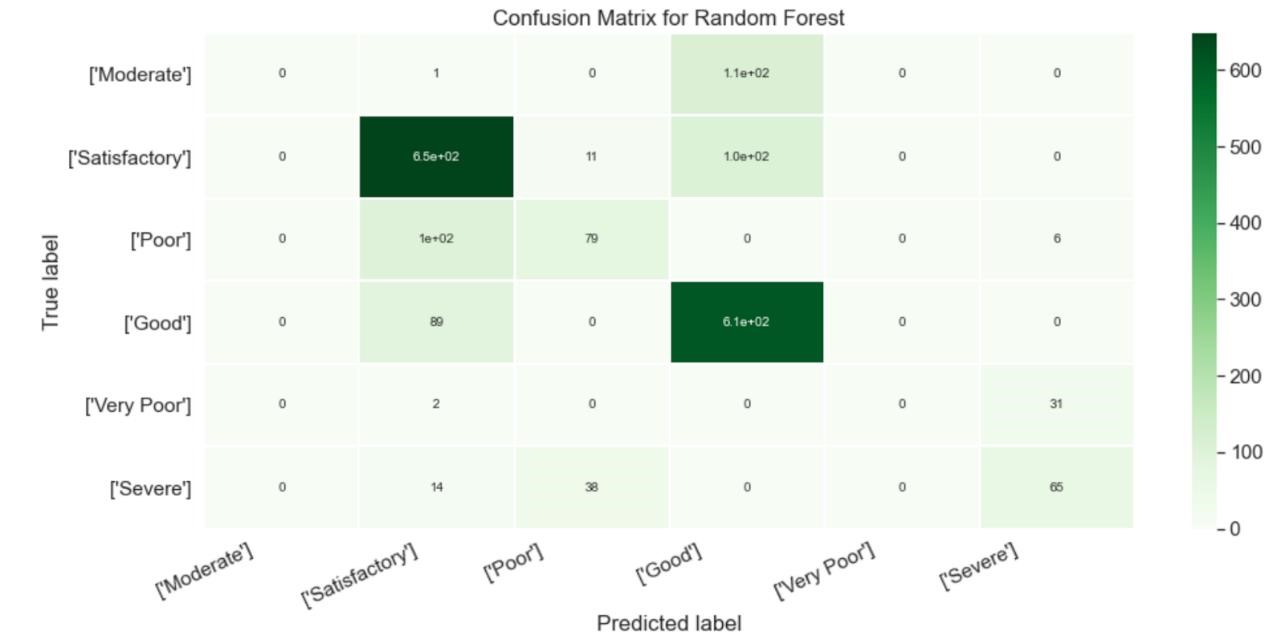
It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. We can give the maximum depth of the tree and how many trees the algorithm needs to create.

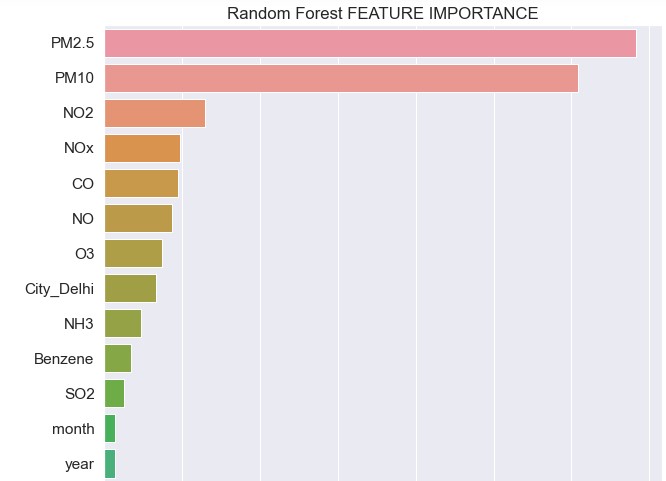
*Why?*

1. This model works well when data is very skewed
2. Suggested by Sciket Learn Roadmap for the data we have for this project
3. To understand the important features better

***Results:***

*Accuracy: 78%*





**Conclusion**

From the above analysis we can conclude the following:

1. Random Forest models are more accurate for both regression and classification compared to linear regression and decision trees comparatively.
2. We can observe that in most of these models, the most important/significant features are PM2.5, PM10, CO, O3 and NO2. We also see the city Delhi is also mentioned in few of the models.
3. We can observe from the linear regression results that PM2.5 and PM10 have the most positive coefficient value which proves that these particles are contributing most positively i.e., the AQI value are increasing as these particles increase in the air.
4. City Delhi also has the highest coefficient value which means AQI is high for this city. The air quality in Delhi, the capital territory of India, according to a WHO survey of 1,650 world cities, is the worst of any major city in the world. It also affects the districts around Delhi [3]
5. We can see from the linear regression results that NO2 coefficient has a negative value.

This means that NO2 is not impacting the air quality index as aggressively as the other particles. 6. We can see that day is not mentioned as important/ significant in any of the models. In the linear regression model results we can observe it has a very high p values which proves that this feature is the least important. What day of the month is the least important feature impacting the AQI.

**References**

1. Avoid Mistakes in Machine Learning Models with Skewed Count Data by [Mingjie Zhao](https://mingjie-zhao.medium.com/?source=post_page-----e3512b94d745-----------------------------------) dated September 14th 2021. Published in Towards DataScience 2. Linear Regression in machine learning published by Javapoint

3. Air Pollution in Delhi by Wikipedia.