Air Quality Index Prediction

Mitrajeet Golsangi*, Divija Godse†, Vivek Ghuge‡,
Vishwajeet Haralkar§, Adityaraj Honraopatil¶ and Prof. Abha Marathe‖
dept. of Computer Science
Vishwakarma Institute of Technology
Pune, India
Email : *mitrajeet.golsangi20@vit.edu, †divija.godse20@vit.edu, ‡vivek.ghuge20@vit.edu,
§vishwajeet.haralkar20@vit.edu, ¶adityaraj.honraopatil20@vit.edu, ‖abha.marathe@vit.edu,

***Abstract*—The rapid pace of urbanization and industrialization in developed nations has been observed in the past few decades. Most developing countries are anxious about air pollution, which is one of the leading factors affecting environmental and public health. The introduction of harmful gases, droplets of liquid, solid particles creates a major threat to the quality of life in smart cities. To decrease air pollution, we need to use efficient air quality monitoring models to collect information about the fullness of air pollution and provide air testing pollution in each area. Therefore, air quality testing as well as forecasting has become an important research area nowadays. Quality of the air is influenced by the multiple elements that comprise the area, time, and uncertain flexibility.The purpose of this research paper is to research many different details as well as machine learning techniques that predict air quality.**

***Index Terms*—Air pollution, AQI, MLR, Random Forest, AQI prediction**

I. Introduction

Plants, animals, and even humans rely on air for their survival and existence on this planet. It is one of the most important natural resources available to life on this planet. Living organisms, therefore, depend on good quality air that is free of harmful gases for survival. But according to the data released by the World Health Organization (WHO), air pollution is responsible for around 1.3 billion deaths [1] worldwide each year. The decrease in air quality is just one of the adverse effects caused by pollution released into the atmosphere. Acid rain , global warming, aerosol formation and photochemical smog are also among those that have increased in the past several decades.

The Environmental Protection Agency (EPA) tracks pollutants that are commonly known, such as ground-level ozone (O_3), sulphur dioxide (SO_2), particulates matter (PM_{10} and $PM_{2.5}$), carbon monoxide (CO) , carbon dioxide (CO_2), Nitric oxide(NO), Nitric Dioxide(NO_2), Ammonia(NH_3), Benzene(C_6H_6), Toluene(C_7H_8) and Xylene(C_8H_10) [1]. A common index shows air quality in areas according to its composition through the Air Quality Index (AQI), which indicates if the air is polluted or clean at the moment or is expected to be in the future.

We propose machine learning models for the prediction of AQI concentrations in this paper.These data have been compiled from the website for the Central Pollution Control Board (CPCB), which is an official government agency. We investigated two machine learning models, the first one is Multiple Linear Regression and the second one is Random Forest to train our model. Furthermore this research demonstrates how the Air Quality Index would change for a city per day.Three common scale-dependent error indices are used to measure accuracy of MLR model : mean absolute error (MAE), root mean square error (RMSE), and R-squared error (R^2) and the accuracy of Random Forest measured using % variable explained. [2]

II. Literature Survey

The shift of a normal city to a smart city calls for the need of an immediate large scale urbanization. This urbanization further has led to an enormous increase in the pollution levels across different cities of the country. A large number of gases are let out which affect people's health in turn affecting the way of living. The gases mostly produced are ground-level ozone (O_3), sulphur dioxide (SO_2), carbon monoxide (CO), and carbon dioxide (CO_2) can be harmful for the various living organs. Based on the different levels of these gases, the Air Quality Index can be measured. The Air quality index indicates how clean or polluted the air in the surrounding is. The AQI is measured using various data processing techniques.

Most of the research papers simply use multiple regression models. [2] [1] These papers showcast the use of Random Forest, Gradient boosting regression, decision tree, KNN, ANN, SVM, AdaBoost for prediction. [2] [1] Multiple models predict values with different accuracies. Papers show that using the random forest algorithm gives the supreme accuracy for the AQI value prediction. [2] [1] Furthermore other researches also depict that the Gradient Boosting algorithm also gives more reliable values of predicted AQIs. [2] For finer data visualisation, different plots and graphs are plotted employing the processed regression models.

The papers have emphasized processing the regression model between 95% of the confidence interval according to the Central Limit Theorem. The prediction is carried out using the standard deviation(σ), z-score(z) and significant

values for the data set.

The formula for the prediction is:

$$PI = \frac{z\alpha}{2} \times \sigma$$

[1]

Another important factor highlighted by these papers was the importance of imputations. The MAE, RMSE, R^2 values show a significant change after they have been calculated post value imputations

III. Study Area and Methodology

A. Data Set Description

The data is compiled by the CPCB (Central Pollution Control Board) of India, which is the official governing body over the Pollution index in India.

The Dataset consists of the various pollutant concentrations collected daily from 9 cities all over India ranging from Jan 2015 to Jul 2020. The Dataset also calculates the AQI (Air Quality Index) for each day, and gives an AQI Bucket describing the Air quality categorically. The main pollutants are the particulate matters [3] given in the dataset which contain $PM_{2.5}$ and PM_{10}. The collected data is preprocessed to some extent but still needed certain amount of data cleaning to match the factors of data quality explained in [4].

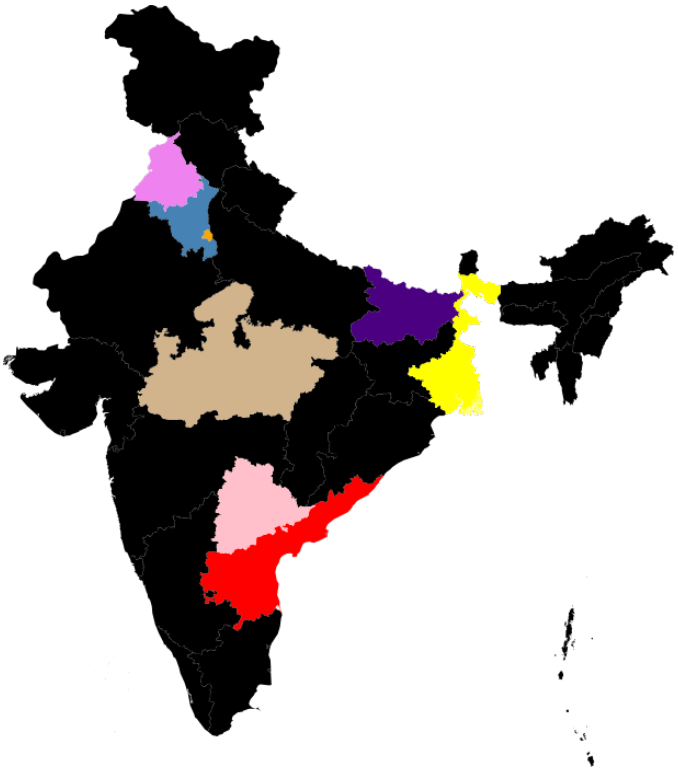

Fig.Data collection locations

The colored states are the locations from which data of certain cities was taken to create the dataet used in the given paper.

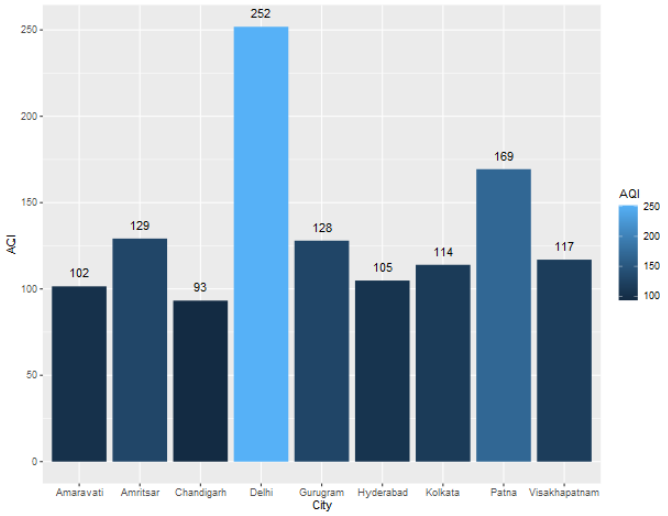

Fig.Average AQI per city

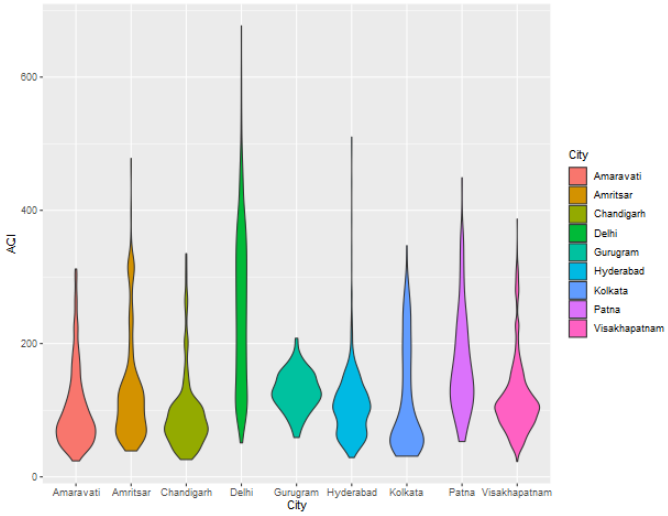

Fig.AQI frequency per city

B. Models

1) Multiple Linear Regression: Multiple linear regression is an extension of Linear regression method along with simple linear regression. This type of regression comes to aid when more than one predictors are to be considered. Multiple regression branches out as soon as a relationship between more then two variables are encountered. In this case, the best fit line is considered as the regression plot. According to the SSR(Sum of Squares due to regression) and SSE(sum of squares due to errors) the R^2(coefficient of determination) value is calculated and adjusted for the prediction value.

$$SSR = \sum(-y)^2$$

$$SSR = SST - SSE$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

2) Random Forest: It is the most popular ensemble learning algorithm of supervised learning that can be used for classification or regression problems. This classifier uses multiple subsets of the dataset to create the number of decision trees and combines them to improve predictive accuracy. The random forest does not use a single decision tree, but takes the predictions from each tree and reflects them in the final outcome based on the majority votes. [2] [5]

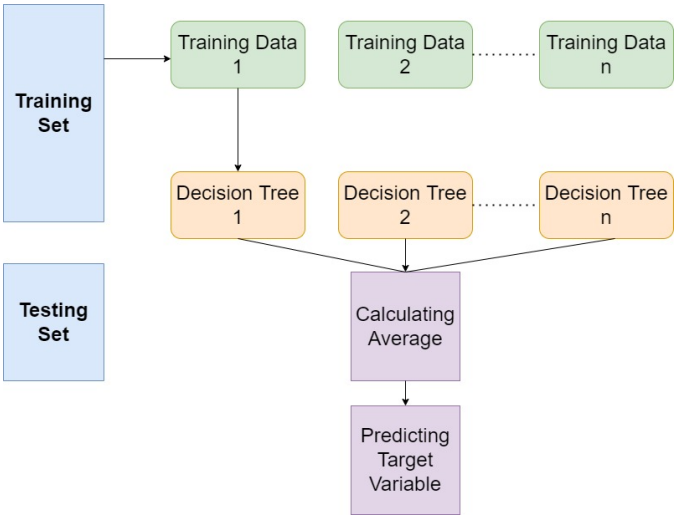

Fig.Workflow of Random Forest Algorithm

3) Performance Parameters: The regression models performance is measured using 3 main criterias

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{MSE}$$

Usually RMSE is used as the values of MSE can get really large making it difficult to interpret. The value of R^2 lies between the values of 0 and 1 giving a more interpretable meaning to it.

IV. Results and Discussion

We plotted the following graph based on our model to check how our model is working, and it is very easy to visualize the difference between actual value and predicted value. [6]

A. Multiple Linear Regression Without Imputation

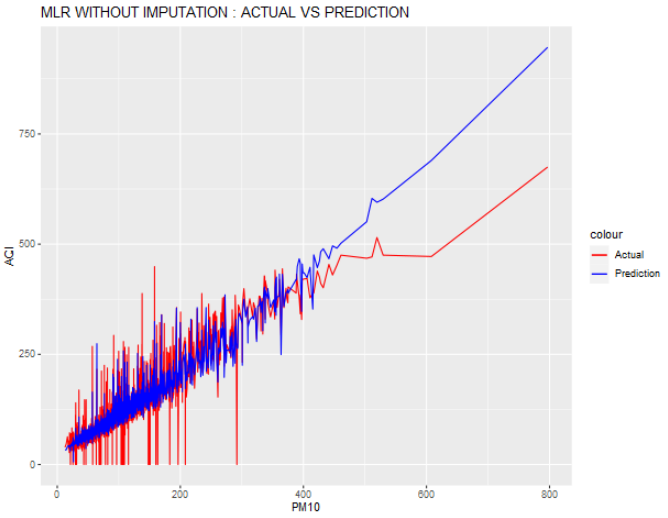

Fig.MLR Without Imputation

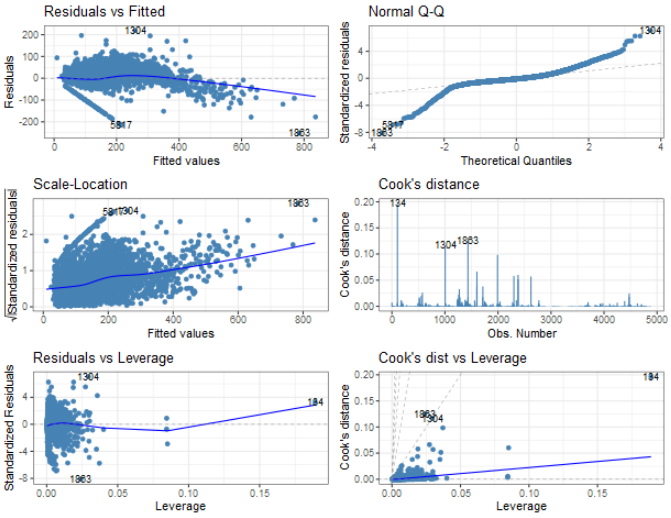

Fig.Performance Without Imputation

B. Multiple Linear Regression With Imputation

Fig.MLR With Imputation

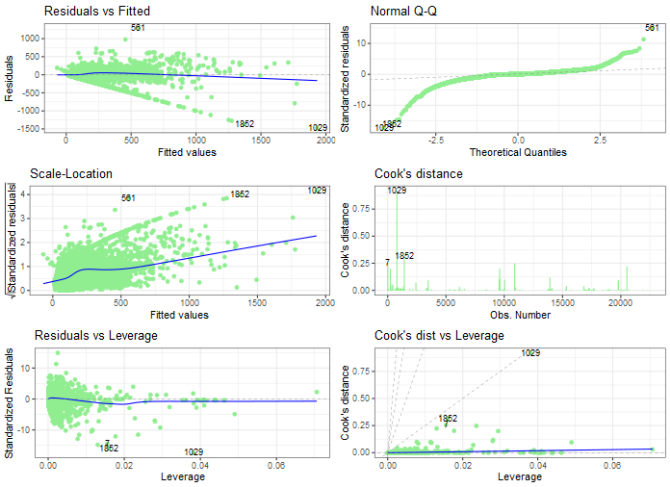

Fig.Performance With Imputation

C. Random Forest Without Imputation

Fig.RF without imputation

D. Random Forest With Imputation

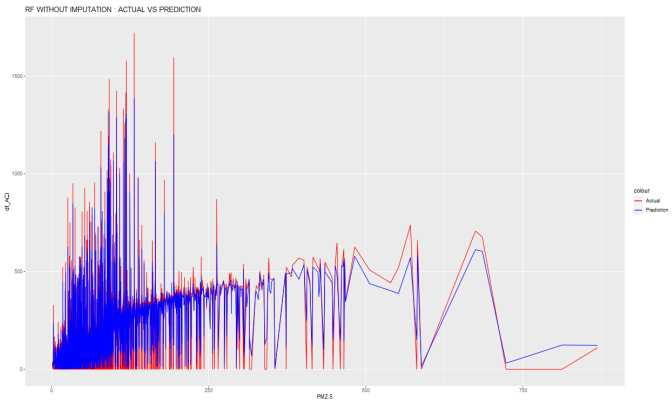

Fig.RF with imputation

E. Different Charts and their Significance

1) **Residuals vs fitted :** It is a frequently created plot. Residuals are plotted on the y axis and the true values are plotted on the x axis. It is very similar to scatterplot. It is used to check linearity and homoscedasticity.
2) **Scale - location :** Spread-location is another name for it. This graph indicates if residuals are distributed evenly across the predictor range. This is how we can test the equal variance assumption(homoscedasticity).
3) **Residuals vs leverage :** It is a plot between standardized residuals and leverage points of the points. The standardized difference between a predicted value for an observation and actual value of the observation is referred to as standardized residuals. The degree to which the coefficient in the regression model would vary if a specific observation was removed from the dataset is referred to as leverage.
4) **Normal Q-Q :** Normality of the data set can be checked using normal Q-Q plot. Scatterplot will be distributed along 45 degrees if normality exists in the data set.
5) **Cook's distance :** It measures the extent of change in model estimates when a certain observation is omitted. This plot measures the distance for each observation.
6) **Cook's distance vs leverage :** These plots are used to detect highly influential points of the data set. Cook's distance value above 1 indicates highly influential points.

F. Regression Results

Models	Imputation	R^2	$adjusted - R^2$	$RMSE$
MLR	No	0.8886	0.8883	39.392
MLR	Yes	0.6343	0.6341	86.576
RF	**No**	**96.96**	**–**	**16.17585**
RF	Yes	69.42	–	75.45717

V. Conclusion

Here, in this research paper we analyzed collected air pollution data from 9 major cities in India using the machine learning algorithm. Predicting the Air Quality Index(AQI) is based on the statistically calculated metrics, such as MSE, RMSE, and R^2. The results obtained demonstrate that the proposed model works better and also that the predicted values and actual values are very similar. Finally, as the conclusion of this research, we find that the Random Forest model without imputation of any values is better for forecasting air pollution in India.

References

[1] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, "Machine learning-based prediction of air quality," *MDPI*, 2020.

[2] Doreswamy, H. K. S1, Y. KM, and I. Gad, "Forecasting air pollution particulate matter (pm2.5) using machine learning regression models," *Science Direct*, 2020.

[3] Park, Seohui, M. Shin, J. Im, C.-K. Song, M. Choi, J. Kim, and S. L. et al., "Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over south korea."

[4] Z. Guan and R. O. Sinnott, "Prediction of air pollution through machine learning approaches on the cloud," *IEEE*, 2018.

[5] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, 2018.

[6] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Science Direct*, 2013.