Prediction of Air Quality Index using Random Forest Algorithm

Dipak Gaikar [1], Ujjwal Patel[2], Om Vispute[3],Sagar Singh[4], Takshil Sanghvi[5]

[1] Asst. Professor, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India
[2] B.E. student, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India
[3] B.E. student, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India
[4] B.E. student, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India
[5] B.E. student, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India

---***---

Abstract - Air pollution is a growing concern worldwide, and it has serious implications on human health, the environment, and the economy. In this project, we explore the prediction of Air Quality Index (AQI) using the Random Forest algorithm. AQI is a measure of air pollution that is used to communicate the health risks associated with breathing polluted air. We use historical data collected from various air quality monitoring stations in a city and apply the Random Forest algorithm to predict AQI. This study aims to predict the AQI using machine learning algorithms. The AQI is a crucial indicator of air quality, and accurate forecasting can help mitigate the negative effects of air pollution on human health and the environment. The study utilizes data from air quality monitoring stations and meteorological sensors to train and evaluate various machine learning models, including Random Forest, Support Vector Regression, and Artificial Neural Networks. The accuracy of the algorithm is measured using the root mean square error . The mean square error and the mean absolute erro). The results indicate that the Random Forest algorithm performs well in predicting AQI and has the potential to be used as a tool to monitor air quality and help in making decisions to reduce air pollution. The findings of this study can be used by policy makers, city planners, and environmental agencies to design effective strategies to combat air pollution.

***Keywords*: Prediction, Machine Learning, Random Forest, *Air Quality, P.M 2.5 , Root mean squared error(RMSE), Mean Squared error(MSE),mean absolute error (MAE)*.**

1. INTRODUCTION

Air pollution is a pervasive problem that affects millions of people worldwide, resulting in adverse health outcomes, environmental degradation, and economic losses. The World Health Organization (WHO) estimates that air pollution causes around 7 million premature deaths annually, making it one of the leading global health risks (WHO, 2021). Air Quality Index (AQI) is a measure of air pollution that provides information on the air quality status and associated health risks. AQI is a numerical value ranging from 0 to 500, and it is calculated based on the levels of major air pollutants such as particulate matter (PM), ozone (O3), nitrogen dioxide (NO2), and sulfur dioxide (SO2).

Various approaches have been developed to monitor and manage air quality, including regulatory policies, emission controls, and air quality forecasting. Air quality forecasting aims to predict future AQI levels using statistical and machine learning models based on historical data and meteorological factors. Machine learning techniques such as Linear Regression, support vector regression (SVR), and decision trees have been applied to air quality forecasting . Random Forest (RF) is a powerful machine learning algorithm that has been used for AQI prediction in recent studies.

2. OBJECTIVE

- Air quality forecasting that uses machine learning to predict the air quality index for a given region.
- To achieve better performance than the standard regression models.
- Our goal is for the model to accurately predict Air Quality Index for India as a whole.
- By forecasting Air Quality Index, we can track the main pollutants causing pollutants and the locations across India that are severely affected by pollutants.
- By creating a easily operated graphical user interface we will help the user to keep a track of the air quality index and its attribute on a single screen.

3. PROPOSED SYSTEM

AQI is an important environmental indicator that is used to inform public health and policy decisions. The proposed System using an Enhanced approach using ANN (Artificial Neural Network) is tested using the dataset of list 5 years (2013-2018). The results are compared with previous methods results. These methods are Random Forest, Linear regression, XG boost , K Nearest Neighbour Regression, ANN .The proposed enhanced method for AQI advantages over this methods. When compared to various other methods, our model gave the most precise forecasts. This technique makes it simple and accurate for meteorologists to forecast the weather and the AQI in the future. Fine material (P.M 2.5) may be significant because, once its level in the air is somewhat high, it poses a serious threat to people's health. Small airborne particles ,known as PM 2.5,reduce visibility and high levels make the air look like fog .

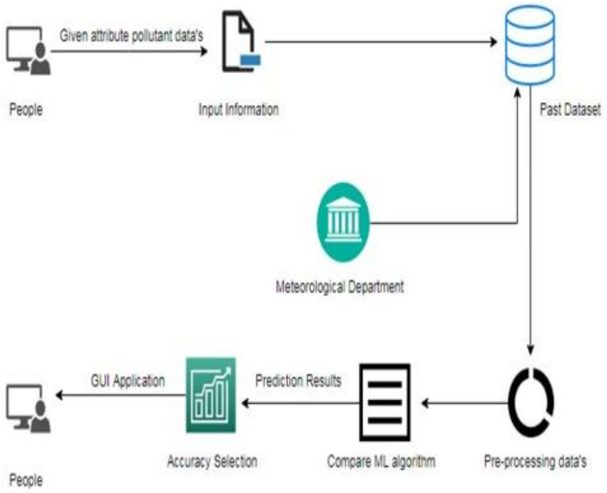


Fig -1 Proposed System Model

4. METHODOLOGY

4.1 Data sources

Collect data on various air quality parameters such as particulate matter (PM10, PM2.5), sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), etc. for a given location at different times. This data can be obtained of India from local environmental agencies or online sources.

T	TM	Tm	SLP	H	VV	V	VM	PM 2.5
16.9	25.1	6.6	1021.3	65	1.1	2	7.6	284.7958
15.5	24.1	7.7	1021	71	1.1	3.5	11.1	219.7208
14.9	22.8	8	1018.4	73	1.1	5.9	13	182.1875
18.3	24.7	11.5	1018.1	85	0.5	1.1	7.6	154.0375

Table–1 Sample Data

4.2 Preprocessing of data

Clean the data and remove any missing or inconsistent values. There are various techniques which are used in data preprocessing i.e data cleaning , data integration & data transformation, data reduction, data encoding. The overall goal aim of data preprocessing is to insure that the data is ready for analysis or machine learning and that it will produce accurate and meaningful results.

4.3 Feature Selection

Select the relevant features from the dataset that can impact air quality. This can be done using statistical techniques or domain knowledge. There are several techniques for feature selection, such as filter methods, wrapper methods, and embedding methods. Filter methods involve evaluating the relevance of each feature based on some statistical measure, such as correlation or mutual information, and selecting the top-ranked features. Wrapper methods involve selecting features based on the performance of a machine learning algorithm, such as decision trees or SVM, with a particular subset of features. Embedded methods involve incorporating feature selection into the learning algorithm itself, such as with regularization techniques like Lasso or Ridge regression.

4.4 Train-Test Split

Train-test split is a technique used in machine learning to evaluate the performance of a model on unseen data. The process involves splitting a dataset into two parts: a training set and a testing set. The training set is used to train the model, and the testing set is used to evaluate the model's performance. The goal is to train a model that can generalize well to new, unseen data. The splitting of the dataset can be done randomly or using a specific technique such as stratified sampling, where the split is done in a way that preserves the proportion of classes or values in the original dataset.

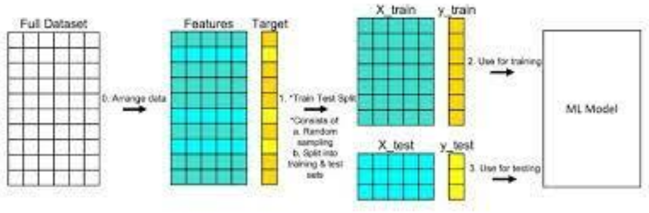


Fig -2 Training and splitting data

4.5 Model Selection

Build a random forest model using the training data. Random forest is an ensemble method that combines multiple decision trees and reduces overfitting. It belongs to the ensemble learning family of algorithms, which combines multiple models to make better predictions than any individual model. The basic idea behind the random forest algorithm is to build a collection of decision trees and combine their outputs to make a final prediction. Each decision tree in the forest is trained on a different subset of the original data and a random subset of the features. By creating different trees based on different data subsets and features, random forest reduces the risk of overfitting and improves the accuracy and stability of the predictions. When making a prediction, each decision tree in the forest predicts the outcome independently, and the final prediction is made by combining the outputs of all the trees. In classification tasks, the prediction is typically based on a majority vote of the trees, while in regression tasks, the prediction is typically based on the average of the outputs of the trees.

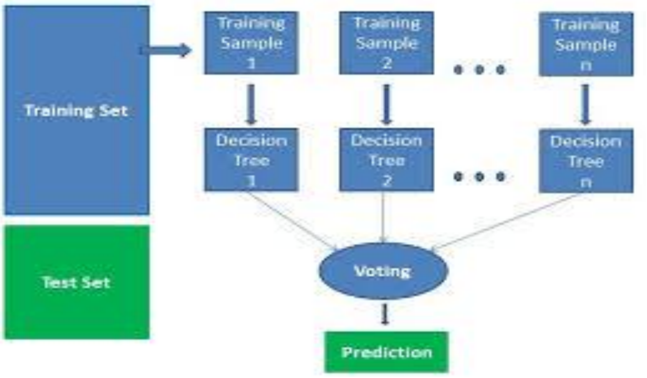


Fig -3 Selection of Model

4.6 Hyperparameter Tuning:

Hyperparameter tuning is an essential step in optimizing the performance of a Random Forest model for air quality index prediction. Here are the steps you can follow for hyperparameter tuning in Random Forest:

1)Split the data: Divide your dataset into a training set and a validation set. You can use a 70-30 split or a 80-20 split, depending on the size of your dataset.

2)Define hyperparameters: Select the hyperparameters to tune. In Random Forest, some of the hyperparameters that can be tuned include the number of trees in the forest, the depth of each tree, the minimum number of samples required to split an internal node, and the maximum number of features to consider when looking for the best split.

3)Choose a metric: Select a performance metric that you want to optimize. For air quality index prediction, you can use metrics like mean squared error (MSE), mean absolute error (MAE), or R-squared (R2).

4)Grid search: Use a grid search to try out all possible combinations of hyperparameters. Grid search is a technique that allows you to define a big variety of utility value for every hyperparameter and then conductes the evaluation for the model for all possible combinations of these values.

5) Cross-validation: Perform k-fold cross-validation on each combination of hyperparameters to get a more accurate estimate of the model's performance. Cross-validation helps to reduce the risk of overfitting and provides a more reliable estimate of the model's performance.

6)Evaluate performance: After completing the grid search and cross-validation, select the hyperparameters that give the best performance on the validation set.

7)Test on new data: Finally, test the model with the selected hyperparameters on a new test dataset to evaluate its performance in real-world scenarios.

4.7 Model Evaluation

Random forest is a popular machine learning algorithm used for regression and classification tasks. It is widely used for air quality index prediction due to its ability to handle non-linear relationships between the input variables and the target variable. However, it is important to evaluate the performance of the Random Forest model to ensure its accuracy and reliability. some commonly used evaluation metrics for a Random Forest model:

1)Mean Squared Error (MSE): MSE measures the mean squared difference between the predicted and actual values. Lower values of MSE indicate better performance of the model.

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

2) Mean Absolute Error(MAE): MAE measures the average absolute difference between the predicted and actual AQI values.

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i - \boldsymbol{x}|$$

3) Root Mean Squared Error (RMSE): RMSE measures the average squared difference between the predicted and actual AQI values, and it takes the square root of the result.

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

4)Rsquared(R^2): R-squared is a degree of the way properly the version suits the data.. It measures the proportion of the variance in the AQI values that can be explained by the model. R-squared values range from 0 to 1, with a value of 1 indicating a perfect fit.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

5. ARCHITECTURE

The figure below shows the system configuration of the proposed system. To train the model first the dataset is preprocessed. After pre-processing feature extraction is done for the dataset from which we get training data. These Training [illegible] science model. Next, you'll finally check the PM2.5 pollutant range predictions to predict whether the air quality levels are [illegible] model. Otherwise, , you will have to redeploy the model and dataset.

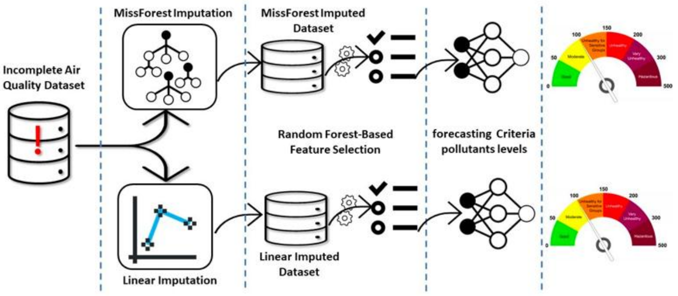

MissForest Imputation
MissForest Imputed Dataset
Incomplete Air Quality Dataset
Random Forest-Based Feature Selection
forecasting Criteria pollutants levels
Linear Imputation
Linear Imputed Dataset
0
50
100
150
200
300
500

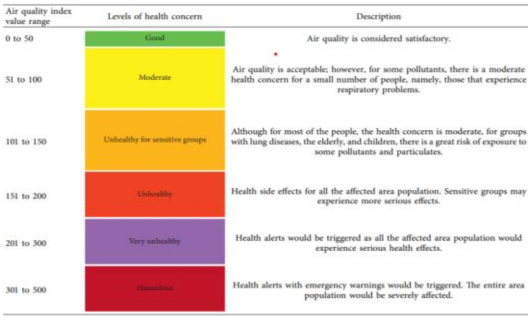
Fig -4 System Architecture

6. RESULT

In this project , we have shown how using Random Forest Algorithm we have obtained precise and accurate results for Air Quality Index . we have used parameters such as MAE,MSE and RMSE.

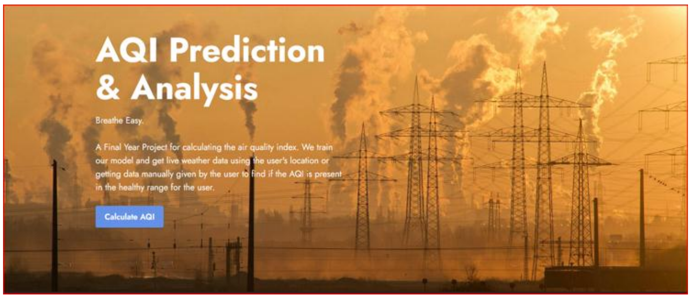
MAE: 36.326655063 86365	MSE: 2704.4949219 76799	RMSE: 52.0047586474 23785

The below representation shows us the categorical division by Environmental Protection Agency(EPA) for AQI. Here using a Graphical User Interface(GUI),We have established our results in the most simplest form using random forest algorithm with the best accuraty we could have achieved. The User Interface shows various fields which helps us to find Air Quality Index based on the data feeded in it.



Air quality index value range	Levels of health concern	Description
0 to 50	Good	Air quality is considered satisfactory.
51 to 100	Moderate	Air quality is acceptable; however, for some pollutants, there is a moderate health concern for a small number of people, namely, those that experience respiratory problems.
101 to 150	Unhealthy for sensitive groups	Although for most of the people, the health concern is moderate, for groups with lung diseases, the elderly, and children, there is a great risk of exposure to some pollutants and particulates.
151 to 200	Unhealthy	Health side effects for all the affected area population. Sensitive groups may experience more serious effects.
201 to 300	Very unhealthy	Health alerts would be triggered as all the affected area population would experience serious health effects.
301 to 500	Hazardous	Health alerts with emergency warnings would be triggered. The entire area population would be severely affected.

Fig -5 Category division for AQI

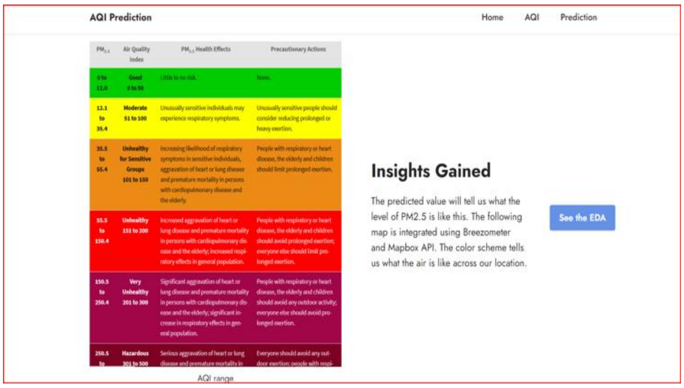


AQI Prediction & Analysis

Breathe Easy.

A Final Year Project for calculating the air quality index. We train our model and get live weather data using the user's location or getting data manually given by the user to find if the AQI is present in the healthy range for the user.

Fig -6 GUI for the Output



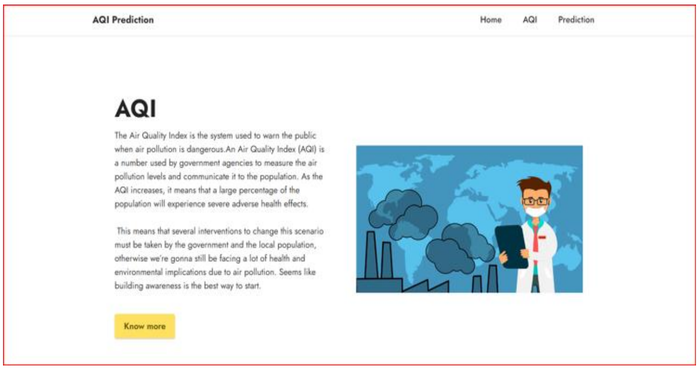
PM$_{2.5}$	Air Quality Index	PM$_{2.5}$ Health Effects	Precautionary Actions
0 to 12.0	Good 0 to 50	Little to no risk.	None.
12.1 to 35.4	Moderate 51 to 100	Unusually sensitive individuals may experience respiratory symptoms.	Unusually sensitive people should consider reducing prolonged or heavy exertion.
35.5 to 55.4	Unhealthy for Sensitive Groups 101 to 150	Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.	People with respiratory or heart disease, the elderly and children should limit prolonged exertion.
55.5 to 150.4	Unhealthy 151 to 200	Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid prolonged exertion; everyone else should limit prolonged exertion.
150.5 to 250.4	Very Unhealthy 201 to 300	Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid any outdoor activity; everyone else should avoid pro longed exertion.
250.5 to	Hazardous 301 to 500	Serious aggravation of heart or lung disease and premature mortality in	Everyone should avoid any outdoor exertion; people with respi-

AQI range

Insights Gained

The predicted value will tell us what the level of PM2.5 is like this. The following map is integrated using Breezometer and Mapbox API. The color scheme tells us what the air is like across our location.

Fig -7 GUI Information in the Output

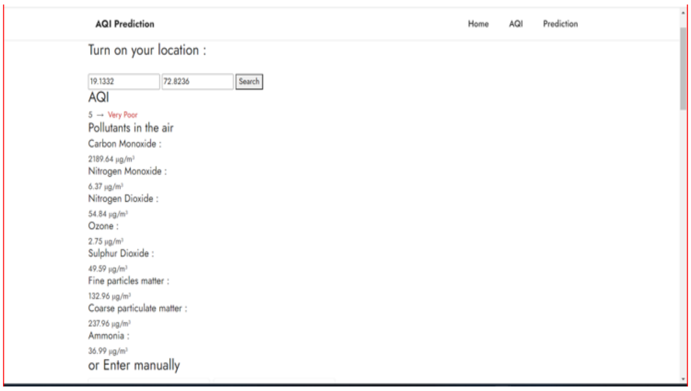


AQI

The Air Quality Index is the system used to warn the public when air pollution is dangerous.An Air Quality Index (AQI) is a number used by government agencies to measure the air pollution levels and communicate it to the population. As the AQI increases, it means that a large percentage of the population will experience severe adverse health effects.

This means that several interventions to change this scenario must be taken by the government and the local population, otherwise we're gonna still be facing a lot of health and environmental implications due to air pollution. Seems like building awareness is the best way to start.

Fig -8 GUI Information in the Output



Turn on your location :

19.1332 72.8236

AQI

5 → Very Poor

Pollutants in the air

Carbon Monoxide :

2189.64 μg/m³

Nitrogen Monoxide :

6.37 μg/m³

Nitrogen Dioxide :

54.84 μg/m³

Ozone :

2.75 μg/m³

Sulphur Dioxide :

49.59 μg/m³

Fine particles matter :

132.96 μg/m³

Coarse particulate matter :

237.96 μg/m³

Ammonia :

36.99 μg/m³

or Enter manually

Fig -9 AQI Prediction in GUI

7. CONCLUSIONS & FUTURE SCOPE

In conclusion, random forest is a powerful machine learning algorithm that can be used for air quality index prediction. It is a popular method for its ability to handle complex, high-dimensional datasets and to identify important features for prediction. By using random forest to analyze various air quality parameters, such as temperature, humidity, and particulate matter concentrations, it is possible to accurately predict the air quality index at a given location and time. However, it is important to note that prediction accuracy can be affected by the quality and quantity of data used to train the model, as well as other external factors such as weather conditions and human activity.

8. ACKNOWLEDGEMENT

This project was conducted and all the evaluations were implemented under the guidance of Prof Dipak Gaikar , Department of Computer Engineering at MCT'S Rajiv Gandhi Institute of Technology, Mumbai, India.

9. REFERENCE

[1] Dragomir, Elia Georgiana. "Air quality index prediction using K-nearest neighbor technique no. 1 (2010): 103-108.

[2] Carbajal-Hernández, José Juan "Assessment and prediction of air quality using fuzzy logic and autoregressive models." Atmospheric Environment 60 (2012): 37-50.

[3] Kumar, Anikender and P. Goyal, " Forcasting of daily air quality index in Delhi", Science of th Total Environment 409, no. 24(2011): 5517- 5523.

[4] Singh Kunwar Petal. "Linear and nonlinear modelling approaches for urban air quality prediction, " Science of the Total Environment 426(2012):244-255.

[5] Sivacoumar R, et al, " Air pollution modelling for an industrial complex and model performance evaluation ", Environmental Pollution 111.3 (2001) : 471-477

[6] Gokhale sharad and Namita Raokhande, "Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period", Science of the total environment 394.1(2008): 9- 24.

[7] Bhanarkar, A. D., et al, "Assessment of contribution of SO2 and NO2 from different sources in Jamshedpur region, India, "Atmospheric Environment 39.40(2005):7745- India." Atmospheric Environment 39.40 (2005): 7745-7760.

[8] Singh Kunwar P., Shikha Gupta and Premanjali Rai, " Identifying pollution sources and prediction urban air quality using ensemble learning methods", Atmospheric environment80 (2013): 426-437.

[9] Wang Jun, and Sundar A. Christopher, "Intercomparison between satellite derived aerosol optical thickness and PM2. 5 Mass: Impliances for air quality studies",Geophysical research letters30.21(2003).

[10] Sharma M E A McBean and U.Ghosh, "Prediction of atmospheric sulphate deposition at sensitive receptors in northern India", Atmospheric Environment 29.16(1995): 2157- 2162

[11] T. Madan, S. Sagar, and D. Virmani, "Air quality prediction using machine learning algorithms –a review," in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 140–145.

[12] C. Li, Y. Li, and Y. Bao, "Research on air quality prediction based on machine learning," in 2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), 2021, pp. 77–81.