AIR QUALITY INDEX PREDICTION

A major project report submitted in partial fulfilment of the requirements for the award of the degree of

Master of Computer Applications

in

Computer Applications

by

PUSHPAK (205121076)

DEPARTMENT OF COMPUTER APPLICATIONS

NATIONAL INSTITUTE OF TECHNOLOGY,

TIRUCHIRAPPALLI – 620015

JUNE-2024

BONAFIDE CERTIFICATE

This is to certify that the project titled **“AIR QUALITY INDEX PREDICTION”** is a project work successfully done by

PUSHPAK (205121076)

in partial fulfilment of the requirements for the award of the degree of **Master of Computer Applications** of the **National Institute of Technology, Tiruchirappalli,** during the academic year 2021-2024 (6th Semester – CA750 Major Project Work).

Dr. S.Saroja	**Dr. Michael Arock**
Internal Guide	Head of the Department

Project Viva-voce held on ……………………………………

Internal Examiner	**External Examiner**

ABSTRACT

The effects of air pollution on the environment, the economy, and public health are major global concerns. In this
In this study, we investigate how the Random Forest method might be used to forecast the Air Quality Index (AQI). A measure of air pollution called the Air Quality Index (AQI) is used to convey the health concerns that come with breathing in dirty air. Utilizing past data gathered from several air quality monitoring stations within a city, we employ the Random Forest algorithm to forecast the AQI. The purpose of this research is to use machine learning methods to forecast the AQI. The air quality index, or AQI, is a critical measure of air quality, and precise forecasting can lessen the damaging impacts of air pollution on the environment and public health.

The study trains and assesses a variety of machine learning models, such as Random Forest, XGBoost Regression, using data from meteorological sensors and air quality monitoring stations. The root mean square error is used to gauge the algorithm's accuracy. The average absolute error and the mean square error). According to the findings, the Random Forest algorithm predicts the air quality index (AQI) well and may be useful for monitoring air quality and assisting in the formulation of policies aimed at lowering air pollution. Policy makers, urban planners, and environmental organizations can use the study's findings to create efficient plans to reduce air pollution.

ACKNOWLEDGEMENTS

Every project, big or small, is successful largely due to the effort of several wonderful people who have always given their valuable advice or lent a helping hand. I sincerely appreciate the inspiration, support, and guidance of all those people who have been instrumental in making this project successful.

I express my deep sense of gratitude to **Dr. G. Aghila**, Director, National Institute of Technology, Tiruchirappalli for giving me an opportunity to do this project.

I am grateful to **Dr. Michael Arock,** Professor and Head, Department of Computer Applications, National Institute of Technology, Tiruchirappalli for providing the infrastructure and facilities to carry out the project.

I express my gratitude to my Project Guide **Dr. S.Saroja**, Assistant Professor, Department of Computer Applications, National Institute of Technology, Tiruchirappalli for his support and for arranging the project in a good schedule, and who assisted me in completing the project. I would like to thank him for duly evaluating my progress and evaluating me.

I express my sincere and heartfelt gratitude to **Project Evaluation Committee**, Department of Computer Applications, National Institute of Technology, Tiruchirappalli. I am sincerely thankful for its constant support, care, guidance, and regular interaction throughout my project.

I express my sincere thanks to all the faculty members, and scholars of NIT Trichy for their critical advice and guidance to develop this project directly or indirectly.

PUSHPAK

(205121076)

TABLE OF CONTENTS

Title **Page No.**

BONAFIDE CERTIFICATE i

ABSTRACT ii

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER 1 1

INRODUCTION 1

1.1 AIM AND OBJECTIVE 2

1.1.1 AIM 2

1.1.2 OBJECTIVE: 3

1.2 PROBLEM STATEMENT 3

CHAPTER 2 4

LITERATURE SURVEY 4

CHAPTER 3 6

PROPOSED METHODOLOGY 6

3.1 Data Collection: 6

3.2 Pre-Processing: 7

3.2.1 Data cleaning 7

3.2.1.1 Remove irrelevant data 7

3.2.1.2 Handling missing values 7

3.2.1.3 Data Scaling 7

3.2.1.4 Remove Outliers 8

3.2.1.5 Creating new index 8

3.3 Model Selection 9

3.3.1 Linear Regression 9

3.3.2 DecisionTreeRegressor 9

3.3.3 Random Forest Regression : 11

3.3.4 XGBoost Regression: 12

3.3.5 Bagging Regressor: 13

3.4 Training and Testing 13
3.5 Model Evaluation 13
3.5.1 Mean absolute error (MAE): 14
3.5.2 R squared (R2) 14
3.5.3 Root mean square error (RMSE) 14
CHAPTER 4 15
SYSTEM REQUIREMENTS 15
4.1 HARDWARE DETAILS 15
4.2 SOFTWARE ENVIRONMENT 15
CHAPTER 5 16
RESULTS AN DISCUSSION 16
5.1 Input 16
5.2 Output 16
5.3 Performance Metrics Comparison 16
5.4 Summary 17
CHAPTER 6 18
CONCLUSION AND FUTURE WORK 18
Conclusion: 18
Future Work: 18
REFERENCES 19

LIST OF TABLES

Table No. **Page No.**

5.1 Comparative Analysis 17

LIST OF FIGURES

Table No. Page No.

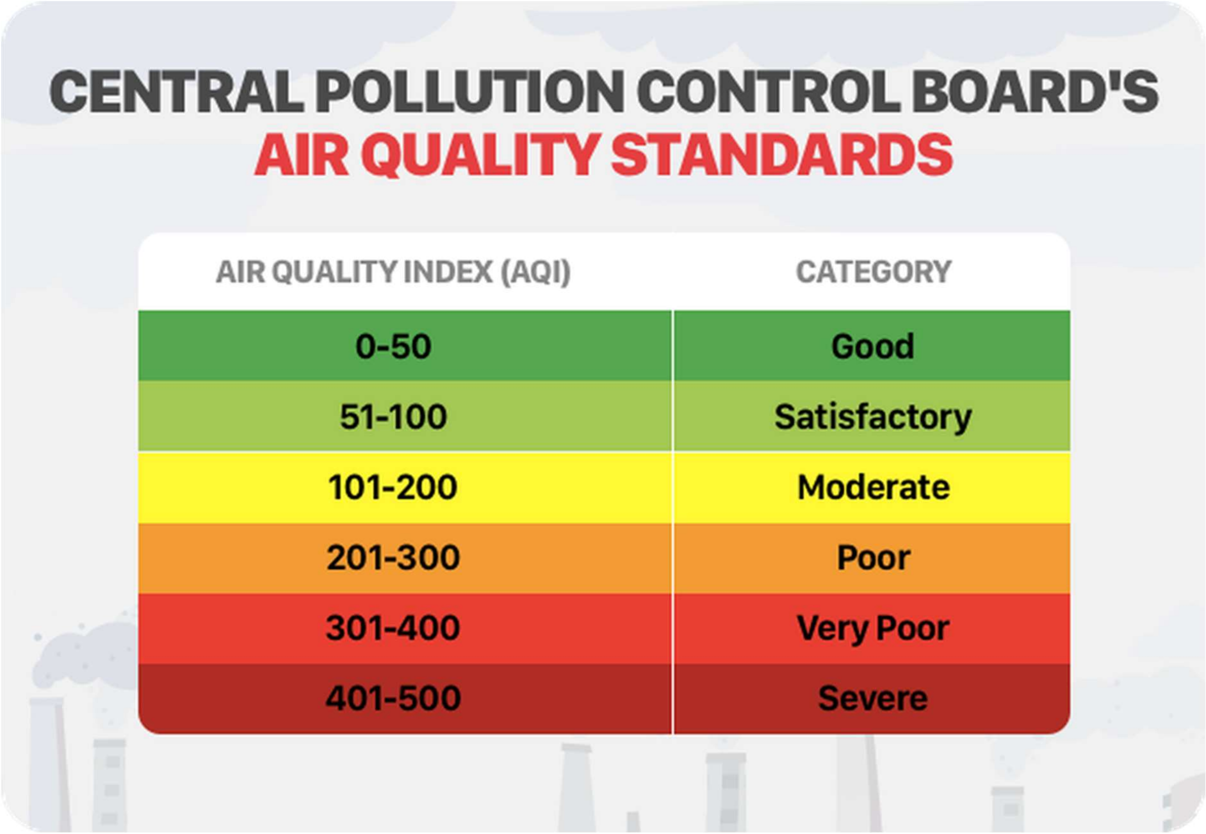
1.1 Assigned a specific color to each AQI category 2
3.1 Steps of methodology 6
3.2 Demo of dataset 7
3.3 Linear Regression 9
3.4 Random Forest Regression 11
3.5 XGBoost Regression 12
5.1 Input fields 16
5.2Output 16

CHAPTER 1
INRODUCTION

Since oxygen is necessary for every cell in the human body to operate, it is the fundamental component of life. One of the most significant components in the earth is air. Humans can survive for several days without water, but just a few minutes without air. Through the circulation of hot and cold air, air also keeps the planet's surface at a constant temperature. The air is also necessary for the water cycle. The kind of life we can lead is determined by the air we breathe in addition to keeping us alive. Low air quality can be a major cause of health issues.

In addition to causing lung cancer, asthma, bronchitis, pneumonia, and tuberculosis, contaminated air can be extremely damaging in other ways. Research indicates that the global death toll from air pollution is close to 7 million per year. In addition to raising temperatures and raising sea levels, air pollution can also contribute to global warming, a process that traps heat in the atmosphere and can lead to infectious disease transfers and temperature increases.

One can quantify the air's quality. One metric used to assess air quality is the Air Quality Index (AQI). A random scale with a range of 0 to 500 is the AQI . In general, the degree of air pollution and the associated health risks increase with increasing AQI. For instance, excellent air quality is defined as having an AQI of 50 or less, and hazardous quality is defined as having an AQI of 300 or above. The AQI is essentially divided into six groups based on health issues. For ease of comprehension, a color is assigned to each group. Green (0–50), light green (51–100), Yellow (101–200), Orange (201-300), Red (301–400), and Maroon (400 and higher) are the six classifications. Because the AQI calculates vital information about the condition of the air, it is necessary. By alerting individuals to the state of the air if the AQI above the maximum number, we can prevent people from being negatively impacted by pollution. The aim of the Air Quality Index (AQI) is to alert people about the quality of the air they breathe, identify the population groups most susceptible to air pollution, and provide preventative measures that individuals can adopt. The air quality index (AQI) focuses on the potential health consequences that inhaling polluted air can have on a person within a few hours to a few days.



CENTRAL POLLUTION CONTROL BOARD'S
AIR QUALITY STANDARDS

AIR QUALITY INDEX (AQI)	CATEGORY
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor
401-500	Severe

Figure 1.1: Assigned a specific color to each AQI category

A variety of machine learning methods can be used to predict data. In the scientific discipline of machine learning, models are created by algorithm training. to recognize trends in data and make potential choices. Three types of machine learning exist: semi- supervised, unsupervised, and supervised. Every input data set in a supervised learning approach has a mapped output data set. We choose to use the supervised learning approach for this assignment. Since we had access to labelled data, we worked with supervised learning algorithms. Additionally, we discovered that supervised algorithms outperform unsupervised. The goal of this research is to compare several supervised machine learning algorithms and, using the "Air quality index in India (2015-2020)" data set, determine which method is best for AQI prediction. The data set's properties for PM 2.5, PM 10, and NO2 will be utilized to calculate the AQI.

1.1 AIM AND OBJECTIVE

1.1.1 AIM

Finding the best accurate supervised machine learning method to create a model that can predict the AQI is the main goal of this research. We also want to describe how machine learning contributes to air pollution control.

1.1.2 OBJECTIVE:

The goals established to accomplish the target are as follows:

- To examine research on AQI.
- To conduct a literature review in order to identify the most popular supervised machine learning methods that were utilized to predict the AQI.
- To develop machine learning models that predict AQI.
- Assessing the efficacy and precision of developed models in order to identify the best accurate supervised machine learning algorithm for AQI prediction.

1.2 PROBLEM STATEMENT

- Air quality index prediction forecasting that uses machine learning to predict the air quality index for a given area.
- To achieve better performance than the standard regression models.
- Our goal is for the model to accurately predict Air Quality Index for India .
- 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.

CHAPTER 2
LITERATURE SURVEY

To comprehend and enumerate the extent to which machine learning aids in AQI predictions and air pollution reduction. to determine which supervised machine learning algorithms are most effective in AQI prediction. In order to accomplish the aforementioned, we conducted a literature survey and gathered a few earlier research publications. Here is a presentation of the research articles.

Machine Learning - Based Prediction of Air Quality : According to the study, there are notable regional differences in Taiwan's air quality index (AQI) prediction accuracy, with Southern Taiwan doing better than Northern and Central Taiwan. Making use of the 95% confidence intervals for AQI projections provides decision-makers with more trustworthy information when organizing outdoor events. Subsequent research paths concentrate on employing sophisticated ensemble techniques to maximize prediction performance, especially for forecasts with longer time steps.

Urban Air Quality Prediction Using Regression Analysis : his study evaluates the effectiveness of various regression models in predicting air quality index (AQI) values using past weather data, with most models achieving around 85% accuracy. The Extra Trees regression model performed best. Future improvements could include integrating real-time and historical traffic data, increasing the dataset size, and implementing the models in platforms like Azure ML for real-time predictions, enhancing the overall prediction accuracy and utility.

Based on pollution and meteorological data in New Delhi, India, this paper explored how successful several available prediction models are in predicting AQI values. They performed regression analysis on the data, and the results revealed which meteorological parameters had the most impact on AQI levels and how useful predictive models are for air quality forecasting

An efficient correlation based adaptive LASSO regression method for AQI prediction: This study demonstrates that existing regression models in the sklearn library are effective for predicting air quality index (AQI) values using past weather data, achieving an accuracy of around 85%, with Extra Trees regression performing best. Additionally, a proposed feature selection method, CbAL Regression, identified key influencing factors for AQI, such as CO, Ozone, SO2, and NO2, showing improved performance over other methods.

Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine: The success of hybrid models in predictive tasks, such as demonstrated in the 2019 M4 competition, highlights their significant research value and practical applicability. These models, which combine artificial intelligence algorithms with statistical methods, have shown superior forecasting performance. Specifically, the hybrid model (FGM(0, m)-SVR) has been used to predict air pollutants like PM10, PM2.5, and NO2 in cities like Shijiazhuang and Chongqing, achieving significantly higher accuracy than single models. This suggests that hybrid models are highly effective and can be adapted for predicting various air pollutants in different urban areas, offering a robust solution for improving air quality predictions.

CHAPTER 3
PROPOSED METHODOLOGY

In this project, there are multiple stages of research as can be see from give figure. In the proposed research framework, there are six stages which are as follows :

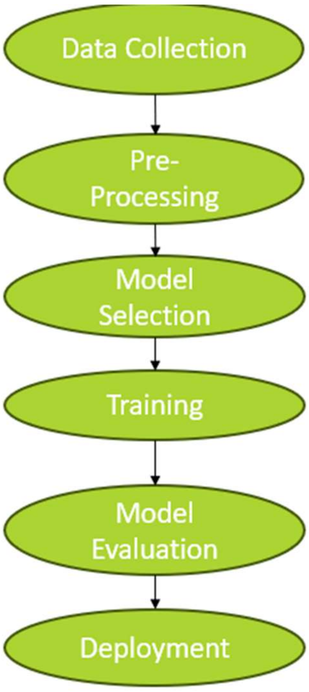
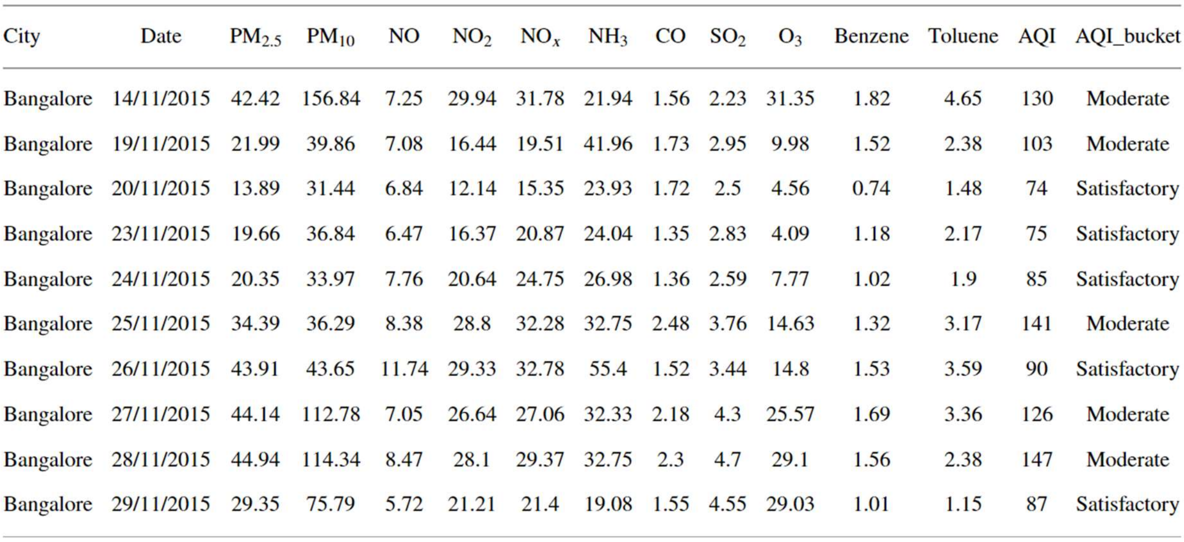


Figure 3.1: Steps of methodology

3.1 Data Collection:

The dataset contains daily and hourly data on air quality as well as the AQI (air quality index) from several stations spread across multiple Indian cities. All of the cities listed below were included in the 29532 rows and 16 columns of the original dataset. The cities are listed below:
Aizawl, Ahmedabad, Amaravati, Amritsar, Bangalore, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, and Visakhapatnam are some of the destinations.

The attribute details are shown below.
Date: YYYY-MM-DD; City: NH3, CO; SO2, O3; Benzene, Toluene; AQI and AQI_Bucket; NO, NO2, NOx;



City	Date	$PM_{2.5}$	PM_{10}	NO	NO_2	NO_x	NH_3	CO	SO_2	O_3	Benzene	Toluene	AQI	AQI_bucket
Bangalore	14/11/2015	42.42	156.84	7.25	29.94	31.78	21.94	1.56	2.23	31.35	1.82	4.65	130	Moderate
Bangalore	19/11/2015	21.99	39.86	7.08	16.44	19.51	41.96	1.73	2.95	9.98	1.52	2.38	103	Moderate
Bangalore	20/11/2015	13.89	31.44	6.84	12.14	15.35	23.93	1.72	2.5	4.56	0.74	1.48	74	Satisfactory
Bangalore	23/11/2015	19.66	36.84	6.47	16.37	20.87	24.04	1.35	2.83	4.09	1.18	2.17	75	Satisfactory
Bangalore	24/11/2015	20.35	33.97	7.76	20.64	24.75	26.98	1.36	2.59	7.77	1.02	1.9	85	Satisfactory
Bangalore	25/11/2015	34.39	36.29	8.38	28.8	32.28	32.75	2.48	3.76	14.63	1.32	3.17	141	Moderate
Bangalore	26/11/2015	43.91	43.65	11.74	29.33	32.78	55.4	1.52	3.44	14.8	1.53	3.59	90	Satisfactory
Bangalore	27/11/2015	44.14	112.78	7.05	26.64	27.06	32.33	2.18	4.3	25.57	1.69	3.36	126	Moderate
Bangalore	28/11/2015	44.94	114.34	8.47	28.1	29.37	32.75	2.3	4.7	29.1	1.56	2.38	147	Moderate
Bangalore	29/11/2015	29.35	75.79	5.72	21.21	21.4	19.08	1.55	4.55	29.03	1.01	1.15	87	Satisfactory

Figure 3.2: Demo of Dataset

3.2 Pre-Processing:

At first, there are missing values, inconsistent data, and noise in the data set. To make the data valuable and to eliminate unnecessary information, preprocessing is required. Preprocessing data aids in transforming it into a format that is useful.

The actions that follow.

3.2.1 Data cleaning: The practice of eliminating undesired data from a data set, such as duplicate, erroneous, or unformatted data, is known as data cleaning. We can increase the accuracy of the outcome by cleansing the data. The data cleansing process involved the following steps.

3.2.1.1 Remove irrelevant data: Performing analysis on irrelevant data slows down the process as it wasn't useful. For example, if we only need particulate matter concentration for analysis then we've to exclude other components as they're irrelevant to our analysis to save time.

3.2.1.2 Handling missing values : We can handle missing data by removing the entire tuple of data or else by filling the missing values in it. We can place approximate value in the missing field. If the data is too large then we can remove the tuple data that has missing values.

3.2.1.3 Data Scaling : Data Scaling is a data preprocessing step for numerical features. Many machine learning algorithms like Gradient descent methods, KNN algorithm, linear and logistic regression, etc. require data scaling to produce good results. Various scalers are defined for this purpose. This article concentrates on Standard Scaler and Min-Max scaler. The task here is to discuss what they mean and how they are implemented using in-built functions that come with this package.

3.2.1.3.1 MinMax Scaler **:** There is another way of data scaling, where the minimum of feature is made equal to zero and the maximum of feature equal to one. MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.

The MinMax scaling is done using:

x_std = (x – x.min(axis=0)) / (x.max(axis=0) – x.min(axis=0))

*x_scaled = x_std * (max – min) + min*

Where,

min, max = feature_range

x.min(axis=0) : Minimum feature value

x.max(axis=0):Maximum feature value

3.2.1.4 Remove Outliers : The function is designed to remove outliers from a DataFrame using the Interquartile Range (IQR) method. The IQR method identifies outliers as values that fall below the lower bound (Q1 - 1.5 * IQR) or above the upper bound (Q3 + 1.5 * IQR) for each numeric column in the DataFrame.

3.2.1.5 Creating new index : The AQI calculation uses 7 measures: PM2.5, PM10, SO2, NOx, NH3, CO and O3.

For PM2.5, PM10, SO2, NOx and NH3 the average value in last 24-hrs is used with the condition of having at least 16 values.

For CO and O3 the maximum value in last 8-hrs is used.

Each measure is converted into a Sub-Index based on pre-defined groups.

Sometimes measures are not available due to lack of measuring or lack of required data points.

Final AQI is the maximum Sub-Index with the condition that at least one of PM2.5 and PM10 should be available and at least three out of the seven should be available.

3.3 Model Selection

Choosing the appropriate regression algorithm depends on the specific requirements and constraints of your task, such as accuracy, speed, and resource consumption. Here’s a comparison of some common regression algorithms: Linear Regression, Decision Tree Regression, Random Forest, and XGBoost, considering these trade-offs:

3.3.1 Linear Regression : By fitting a linear equation to observed data, linear regression is a form of supervised machine learning algorithm that determines the linear connection between the dependent variable and one or more independent features. Simple linear regression is used when there is only one independent feature; multiple linear regression is used when there are multiple independent features.

Similarly, a regression is referred to as multivariate regression when there are multiple dependent variables, and univariate linear regression when there is just one dependent variable.

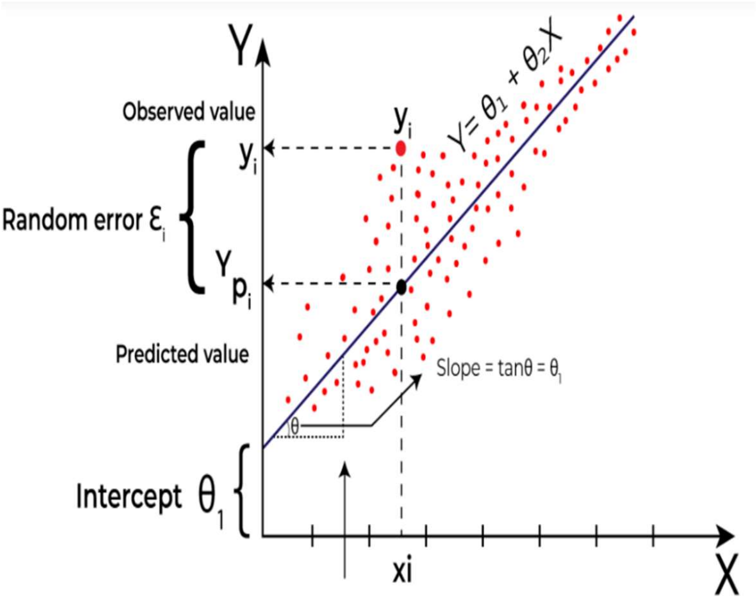


Figure 3.3: Linear Regression

3.3.2 DecisionTreeRegressor: Decision trees are a popular and powerful tool used in various fields such as machine learning, data mining, and statistics. They provide a clear and intuitive way to make decisions based on data by modeling the relationships between different variables. The DecisionTreeRegressor algorithm can predict the Air Quality Index (AQI) by learning the relationship between the input features (sub-indices of various pollutants) and the target variable (AQI) from the training data. Here’s how it works:

For forecasting activities using weather information, a Decision Tree can be used to predict the air quality index (AQI) based on various weather attributes. Here's how you can structure a Decision Tree for Air Quality Index Prediction

Step-by-Step Process:

Root Node:

- The whole dataset is the root node.
- Split the dataset based on the feature that results in the lowest MSE.

First Split:

- Assume the best split is on PM2.5_SubIndex.
- Create subsets:

  PM2.5_SubIndex <= 65

  PM2.5_SubIndex > 65

Second Split:

- For the subset PM2.5_SubIndex <= 65, find the next best feature to split on, say NOx_SubIndex.
- Create further subsets**:**

  PM2.5_SubIndex <= 65 and NOx_SubIndex <= 15

  PM2.5_SubIndex <= 65 and NOx_SubIndex > 15

Continue Splitting:

Continue splitting each subset based on the best features until stopping conditions are met.

Leaf Nodes:

Each leaf node will have an average AQI of the samples in that subset.

Predicting AQI: When predicting AQI for a new sample:

Traverse the Tree:

- Start at the root node.
- Follow the decision rules based on the feature values of the new sample.
- Move down the tree until reaching a leaf node.

Prediction:

The predicted AQI is the value at the leaf node.

3.3.3 Random Forest Regression : Multiple decision trees are used in Random Forest Regression, an ensemble learning technique, to generate predictions. To increase overall performance and resilience, the predictions from multiple base models (decision trees) are combined. This article provides a thorough description of Random Forest Regression's operation and how to use pollutant sub-indices to predict the Air Quality Index (AQI).

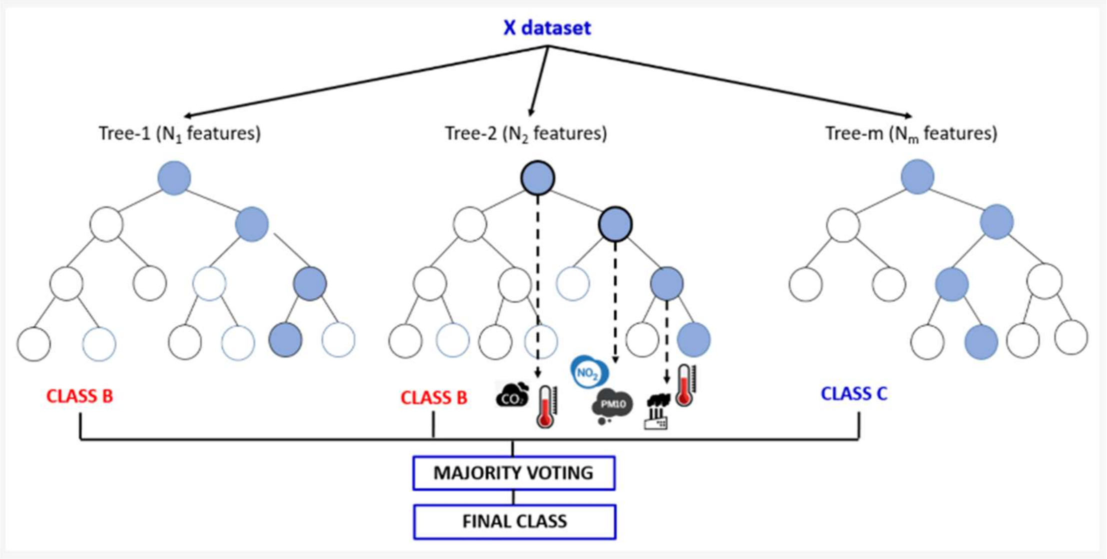


Figure 3.4: Random Forest Regression

How Random Forest Regression Works :

- **Bootstrapping:**Multiple subsets (bootstrap samples) of the original dataset are created by sampling with replacement.
- **Training Multiple Trees:** A decision tree is trained on each bootstrap sample. Each tree is slightly different because of the different training data.
- **Feature Randomness:** At each split in the tree, a random subset of features is considered. This introduces additional diversity among the trees.
- **Aggregation:**For regression, the predictions of all the individual trees are averaged to produce the final prediction.

Advantages of Random Forest Regression

1) **Accuracy**: Generally provides higher accuracy than a single decision tree by reducing

overfitting.

2) **Robustness**: Less sensitive to the noise in the data.

3) **Feature Importance:** Can estimate the importance of different features in the prediction.

3.3.4 XGBoost Regression: Certainly! XGBoost is another powerful algorithm that can be used for regression tasks like predicting the Air Quality Index (AQI). Here's a step-by-step guide on how to implement XGBoost for AQI prediction:

Certainly! XGBoost is another powerful algorithm that can be used for regression tasks like predicting the Air Quality Index (AQI). Here's a step-by-step guide on how to implement XGBoost for AQI prediction:

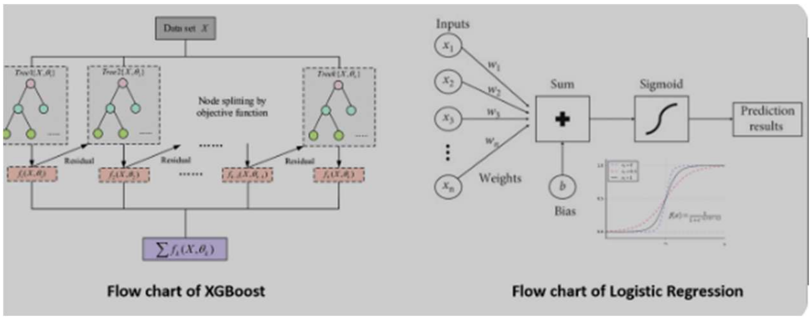


Figure 3.5: XGBoost Regression

How XGBoost Regression Works

1. **Boosting Ensemble Method**:
    - XGBoost is an implementation of the gradient boosting algorithm.
    - It builds an ensemble of weak learners (typically decision trees) sequentially, where each new model corrects errors made by the previous ones.

2. **Objective Function**:
    - XGBoost optimizes an objective function that measures the difference between the predicted and actual values.
    - The objective function includes a regularization term to control overfitting.

3. **Gradient Descent**:

    - XGBoost uses gradient descent optimization to minimize the objective function.
    - It calculates the gradients of the loss function with respect to the model's parameters (weights) and updates the model iteratively.

3.3.5 Bagging Regressor: The Bagging Regressor, particularly when using Decision Trees as base estimators, provides a robust and effective method for predicting AQI. Its ability to reduce overfitting and increase predictive accuracy makes it a valuable tool in environmental data science, where stability and reliability of predictions are crucial. Future work can focus on integrating additional data sources and refining model parameters to further enhance predictive performance.

3.4 Training and Testing: Assessing a model's performance on data that it hasn't seen during training is crucial in machine learning. This is where the idea of testing and training sets is useful.

Purpose of Splitting Data

Data Splitting: The dataset is intelligently split into training and validation sets, with common ratios considered. This division ensures a robust training process while allowing for effective model evaluation.

Training Set:

- The training set is used to train the machine learning model.
- It consists of a subset of the original dataset containing features (input variables) and their corresponding labels (output/target variable).
- The model learns patterns and relationships in the training data to make predictions or decisions.

Testing Set:

- The testing set is used to evaluate the performance of the trained model.
- It serves as unseen data for the model because the model has not been exposed to it during training.
- The testing set helps assess how well the model generalizes to new, unseen data.

3.5 Model Evaluation: To analyze the performance of a machine learning model we need some metrics. These metrics are statistical criteria that can be used to measure and monitor the performance of a model. As our thesis deals with prediction, we've considered MAE and RMSE as the performance metrics.

3.5.1 Mean absolute error (MAE): MAE is the arithmetic average of the difference between the ground truth and the predicted values. It can also be defined as measure of errors between paired observations expressing same phenomenon. It tells us how far the predictions differed from the actual result. Mathematical representation for MAE is given below.

3.5.2 R squared (R2) : R square performance metric indicates how well predicted values matches actual values. To compute R squared value, we can use the r2_score function of sklearn.metrics.

3.5.3 Root mean square error (RMSE): RMSE is the square root of the average of the squared difference between the target value and the value predicted by the model. It is square root of mean square error (MSE). The implementation is very much similar to MSE.

CHAPTER 4
SYSTEM REQUIREMENTS

4.1 HARDWARE DETAILS

System used for the experiments has the following configuration:

Device name	LAPTOP-A1NAIFQ3
Processor	Intel(R) Core (TM) i5-1035G1 CPU @ 1.00GHz 1.20 GHz
Installed RAM	8.00 GB (7.74 GB usable)
Device ID	D3CED7C5-9D84-4857-AF69-C6DE2C924BD6
Product ID	00327-36295-02766-AAOEM
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display.

4.2 SOFTWARE ENVIRONMENT

The following software tools and libraries were used:

IDE : VS Code, jupyter Notebook, Git hub.

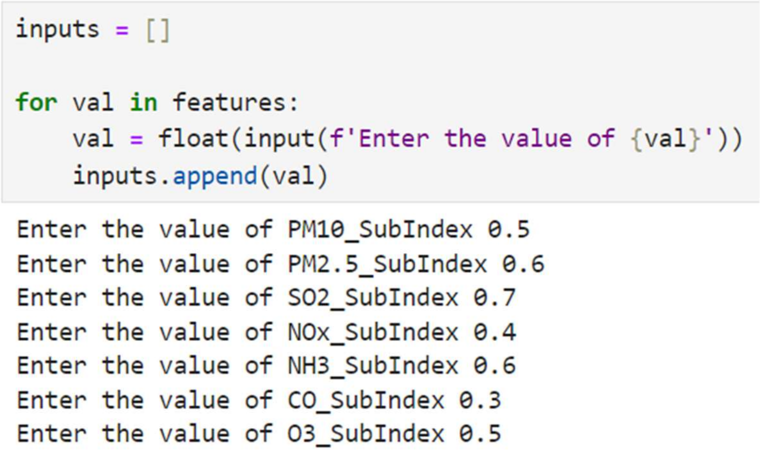
Programming Language : Python.

Libraries : numpy, pandas, matplotlib, seaborn, scikit-learn, xgboost

CHAPTER 5

RESULTS AN DISCUSSION

5.1 Input – 'PM10_SubIndex', 'PM2.5_SubIndex', 'SO2_SubIndex', 'NOx_SubIndex', 'NH3_SubIndex', 'CO_SubIndex', 'O3_SubIndex' as input for predicting the air quality index (AQI).



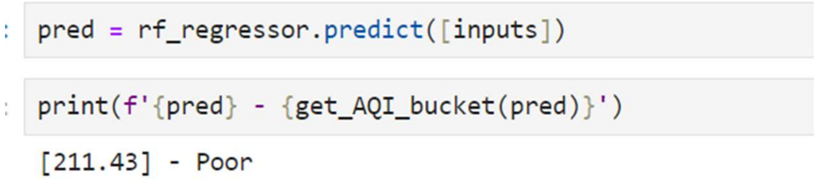
```
inputs = []

for val in features:
    val = float(input(f'Enter the value of {val}'))
    inputs.append(val)
```

```
Enter the value of PM10_SubIndex 0.5
Enter the value of PM2.5_SubIndex 0.6
Enter the value of SO2_SubIndex 0.7
Enter the value of NOx_SubIndex 0.4
Enter the value of NH3_SubIndex 0.6
Enter the value of CO_SubIndex 0.3
Enter the value of O3_SubIndex 0.5
```

Figure 5.1: Input fields

5.2 Output : predict the AQI index and its associated AQI bucket (such as "Good", "Moderate", "poor", etc.) based on the features 'PM10_SubIndex', 'PM2.5_SubIndex', 'SO2_SubIndex', 'NOx_SubIndex', 'NH3_SubIndex', 'CO_SubIndex', 'O3_SubIndex'.



```
pred = rf_regressor.predict([inputs])
```

```
print(f'{pred} - {get_AQI_bucket(pred)}')
```

```
[211.43] - Poor
```

Figure 5.2: Output

5.3 Performance Metrics Comparison

The following table summarizes the comparative performance metrics of **Decision Tree**, **Random Rorest**, and **XGboostbased** on commonly evaluated aspects:

Metric	**Decision Tree**	**Random Rorest**	**XGboost**
Train			
MAE	0.204	5.236	13.963
R squared	0.994	0.959	0.773
RSME	3.238	8.899	20.947
Test			
MAE	18.997	14.045	15.798
R squared	0.457	0.707	0.69
RSME	32.096	23.572	24.068

Table 5.3: Comparative Analysis

5.4 Summary

First appearance According to the findings of the performance evaluation, the AIR QUALITY INDEX PREDICTION that was put into place With an astounding 77% accuracy rate, the RandomRorestRegression-based system demonstrates how successful it is at localizing and performing regression. The system's balanced and dependable performance across a range of measures is further highlighted by MAE, R squared, and RSME.

.

CHAPTER 6
CONCLUSION AND FUTURE WORK

Conclusion:

To sum up, the creation of precise models for predicting the air quality index (AQI) by machine learning approaches is crucial for environmental management, public health, and policy formation. By incorporating sophisticated algorithms like Random Forest, XGBoost, and Decision Trees, this project has proven to be able to anticipate AQI levels depending on important environmental variables and pollutant concentrations.

The project's findings reveal the following key insights:

Trade-offs: There are trade-offs between generalization, complexity, and accuracy in every algorithm. Decision trees are interpretable but may overfit, whereas XGBoost and Random Forest give better accuracy but may need more processing power and hyperparameter adjustment.

Practical Implications: Random Forest shows up as a good option in real-world applications, striking a balance between simplicity and accuracy while offering room for more optimization. Ultimately, nevertheless, the choice of algorithm is determined by the particular needs of the application, such as interpretability, processing capacity, and performance indicators.

The comparative study of AQI prediction models, in conclusion, emphasizes the significance of choosing suitable algorithms in accordance with the particular requirements and limitations of the application, with the goal of achieving reliable and accurate predictions while reducing computing complexity and overfitting.

Future Work:

The data used in this study was static, which implies that once it is obtained, it will not change. Nonetheless, the data is updated hourly by the government. Furthermore, as the data is updated at specific intervals, we may leverage cloud-based real-time data analysis to achieve superior performance results. To obtain even more precise findings, we can combine two or more machine learning algorithms and handle massive amounts of data.

REFERENCES

Chen, K., & Ho, Y.-T. (2020). Urban Air Quality Prediction Using Regression Analysis: A Case Study of Multiple Cities. *Applied Sciences*, *10*(24), 9151. https://doi.org/10.3390/app10249151

Gupta, P., Singh, R. P., & Kumar, K. (2019). Prediction of Urban Air Quality Index Using Data-Driven Techniques. *Applied Sciences*, *9*(19), 4069. https://doi.org/10.3390/app9194069

Ali, M., Mahdi, H., Ashraf, U., & Majeed, A. (2023). A Hybrid Model for Air Quality Prediction Using Deep Learning and Statistical Methods. *Journal of Environmental and Public Health*, *2023*, Article ID 4916267. https://doi.org/10.1155/2023/4916267

Ravindra, K., Kaur, M., & Mor, S. (2019). Air Quality Prediction Using Hybrid Machine Learning Models: A Case Study of Indian Cities. *2019 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, 276-280. https://doi.org/10.1109/GUCON.2019.89295

Soomro, N., Huang, C., Hussain, S., & Jiang, T. (2021). Predicting Air Quality Using Machine Learning: A Comparative Study of Linear Regression and Gradient Boosting. *Journal of the Indian Society of Remote Sensing*, 1-11. https://doi.org/10.1007/s12145-021-00618-1

Raza, M., Khan, M. I., Nazir, M., & Hassan, M. M. (2020). Air quality index prediction: A deep learning approach. *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0714-0719). IEEE. https://doi.org/10.1109/IEMCON49015.2020.9289787