# Air Quality Prediction Using Machine Learning

Hruthik Vinnakota, Maria Hovhannisyan, Sahanti Samarth Zade, Shubham Singh, and Utsav Rastogi

**Abstract**—Air pollution is a precarious environmental issue, specifically in a highly populated metropolitan city such as Beijing. With the increase in air pollution, countries have been taking precautions and developing an air quality index (AQI) to monitor air pollution, providing a warning system to the public. In this paper, we analyze air quality data collected from Beijing's air quality monitoring stations between March 2013 to February 2017. The study aims to examine the correlation between various pollutants and PM2.5 in Beijing. We also intend to identify the key factors that have the most significant impact on deteriorating air quality in the region. The high population density in Beijing has posed significant environmental challenges, necessitating the use of the Beijing Multi-Site Air-Quality Data Set to investigate air pollution in the area. The study found that PM10, CO, DEWP, SO2, NO2, and O3 were the most prevalent pollutants, and PM10 had the strongest correlation with the PM2.5. Furthermore, meteorological factors such as temperature, wind speed, and pressure were identified as additional AQI influencers. These features were fed to various regression models, with LGBRegressor achieving the best accuracy of 94%. The study will highlight the critical importance of closely monitoring air quality and identifying the key factors responsible for pollution in heavily populated urban areas such as Beijing.

**Index Terms**—ExtraTreesRegressor, RandomForestRegressor, KNeighborRegressor, XGBRegressor, GradientBoostingRegressor

## 1 Introduction

Air pollution is a developing environmental issue that poses a substantial risk to human health. With the rise in air pollution, developing countries have taken precautions, including the development of an air quality index (AQI) to monitor air pollution and provide a warning system to the public. In this backdrop, the purpose of this study is to examine the air quality data obtained from the Beijing Multi-Site Air-Quality Data Collection between March 2013 and February 2017.

The aim of this research is to define a relationship between a variety of pollutants and the PM2.5 levels of the city of Beijing in order to determine the important elements that are contributing to the region's deteriorating air quality. High population density in the city has created significant environmental concerns, making the use of Beijing Multi-Site Air-Quality Dataset to study the local air pollution levels. Commonly identified pollutants such as PM10, CO, DEWP, SO2, NO2, and O3, with PM2.5 having the greatest correlation with PM10. Meteorological elements such as temperature (TEMP), wind speed (WSPM), and pressure (PRES) are also regarded as AQI contributors.

The primary goal of this research paper is to build a machine learning model that is able to estimate the PM2.5 levels after learning from historical data on different input features like CO, SO2, etc. This study's dataset is a deep insight on how the air quality is measured using multiple metrics and atmospheric concentrations of different pollutants. This dataset has been collected by monitoring several air pollutants on a national scale, has been collected from 12 sites and linked with meteorological data from the nearest weather station. The effort is broken down into four stages: Data Collection, exploratory data analysis, data cleaning, data-preprocessing, modeling, and evaluation metrics.

The output of this research can aid policymakers and other stakeholders to create and fund impactful strategies in reducing air pollution and improving the air quality in the city. Resulting insights from this study will focus on the need of regular and careful monitoring of air quality and pinpointing on the primary sources of pollution in the populous areas like Beijing.

Additionally, machine learning algorithms can spot trends and offer insights by analyzing air quality data, including contaminants and other environmental factors, to help us better understand air quality levels. This study's potential impact could significantly improve the environment and people's health by dealing with important environmental and public health concerns. Using this approach to find solutions to critical problems such as air pollution, this study also aims to encourage professional and personal development in the machine learning sector and make a substantial contribution to the study of environmental degradation.

It's noteworthy that the problems associated with air pollution and its effects are complicated and multifaceted. A well thought and researched strategy is needed to successfully and completely address this issue. Using machine learning to predict the PM2.5 levels in the city solely based on the past data, , this research article intends to contribute to this effort.

### 1.1 Objective

The objective of this project is to analyze air quality data collected from the Beijing Multi-Site Air-Quality Data Set and identify the key factors that have the most significant impact on declining air quality in the region.

The project aims to develop a machine learning model that estimates PM2.5 levels using multiple input features, including meteorological data, air pollutant concentrations, and other environmental factors. The primary purpose of this model is to predict PM2.5 levels more accurately, providing public health officials and policymakers with valuable insights to make informed decisions about how to address air pollution in the region.

Ultimately, the project aims to contribute to the enhancement of the environment and public health by providing insights that improve our understanding of air quality levels, identify key factors driving air pollution, and inform evidence-based policies and interventions aimed at reducing air pollution and mitigating its associated health impacts. This will enhance air quality and public health by allowing authorities to better prioritize initiatives and develop more successful air pollution reduction plans.

# 2 Theoretical bases and literature review

## 2.1 Definition of the problem

The analysis of air quality data collected from the Beijing Multi-Site Air-Quality Data set between March 2013 to February 2017, aims to examine the correlation between various pollutants and the Air Quality Index (AQI) in Beijing and identify the key factors that have the most significant impact on deteriorating air quality in the region. The study aims to develop a machine learning model that estimates PM2.5 levels using AQI as a predictor variable to help policymakers and stakeholders develop effective strategies for mitigating air pollution and improving air quality in the region.

## 2.2 Theoretical background of the problem

Employing machine learning techniques for determining the air quality data from the Beijing Multi-Site Air-Quality Data Set. By examining the relationship between different contaminants and the Air Quality Index (AQI), the research's main goal is to pinpoint the main causes of Beijing's worsening air quality. The study proposes the development of a machine learning model that estimates PM2.5 levels utilizing AQI as a predictor variable. Furthermore, the research emphasizes the importance of close monitoring of air quality in densely populated urban areas and the identification of the key factors causing pollution to formulate effective strategies for mitigating air pollution and improving air quality. Theoretical concepts such as supervised learning, linear regression, decision tree regression, random forest regression, and support vector regression are employed in the research

## 2.3 Literature Review

In the context of the current situation being classified in China, Random Forest is considered one of the algorithms to predict the Air Quality Index in addition to SVM. While Support Vector Machine is better suited for classifying air quality levels. This study used machine learning to predict and classify air quality in Beijing in 2015, using data from China's air quality platform [1].

In similar research predicting air quality [2], this study looked at using machine learning to predict levels of sulfur dioxide, a harmful gas that can cause breathing problems. It found that SO2 concentration was highest in industrial areas and varied across different cities in India. The study also identified the need to include more parameters, such as atmospheric particulate matter, to improve air quality predictions in the future.

In the research [3] a brief introduction to air quality inspection machine learning techniques based on random forest algorithms to evaluate the Clean Air Action Plan in Beijing in 2013. The study concluded that the plan effectively reduced air pollution levels by implementing primary emission controls, leading to significant reductions in PM2.5, PM10, NO2, SO2, and CO levels from 2013 to 2017. The plan can be used as a successful example to develop air quality policies in other regions of China.

The paper Urban Air Quality Prediction Using Regression Analysis [4] focused on predicting air quality levels in urban areas based on past weather data using existing regression models in the sklearn library. The study is centered on New Delhi, one of the world's most polluted cities, and uses data from both air quality and meteorological sources to combine and analyze the impact of various factors on the Air Quality Index (AQI). The study embraced regression analysis and suggests that historic traffic data can be added to the weather data to improve accuracy. The study concludes that existing regression models have an accuracy of almost 85%, with Extra Trees obtaining the highest accuracy, making them useful as predictors of air quality.

This paper examines the use of machine learning (ML) algorithms in predicting air quality in Macao [5], which experiences high-pollution episodes that have adverse health effects. The study compares the performance of four ML algorithms, including random forest (RF), gradient boosting (GB), support vector regression (SVR), and multiple linear regression (MLR), in predicting the concentrations of particulate matter (PM10 and PM2.5) using meteorological and air quality data from 2013 to 2018 to train and validate on the air quality data from 2019 to 2021. The study finds that RF performs better than other methods, particularly during drastic changes such as the COVID-19 pandemic, due to its ability to handle non-linear relationships in the data. The paper provides valuable insights for researchers and policymakers to improve air quality forecasting to mitigate the adverse effects of air pollution on public health.

## 2.4 Solution for our problem

In this study, the proposed solution could significantly improve the environment and people's health by recognizing trends and providing insights that help better understand air quality levels. These findings can be used by regulators and other beneficiaries to create efficient plans for reducing air pollution and enhancing the region's air quality.

## 2.5 Why our solution is better

Specific statistics and preprocessing methods are used, and innovative feature engineering approaches are employed. The effectiveness of our model will also be assessed using unique performance evaluation measures, and various insights will be derived from examining the significance of the model's features. Finally, the study concentrates on the Beijing region with reliable data across several time frames.

# 3 Methodology

## 3.1 Data Collection

The dataset used contains the data on air contaminants that were logged hourly at 6 separate sites that are all subject to national air quality monitoring. The Beijing Municipal Environmental Monitoring Center collected this information, which is compared with weather data gathered from the closest weather station run by the China Meteorological Administration. The dataset consists of over 200,000 distinct rows and 17 separate informational columns.

The cities referred to in the dataset are Aotizhongxin, Changping, Dingling, Dongsi, Guanyuan, Gucheng. Overall the data was collected to monitor and analyze the air quality in Beijing and to provide information to policymakers and the general public to help them make informed decisions related to public health and environmental protection.

## 3.2 Exploratory Data Analysis

Exploratory data analysis (EDA) was performed based on the air quality data from six different monitoring stations in China. The CSV files were clubbed into one and were explored. Correlation between the target and feature variables was performed, which provided more clarity on the linearly related variables, along with its correlation heatmap was also generated as observed in Fig 1.

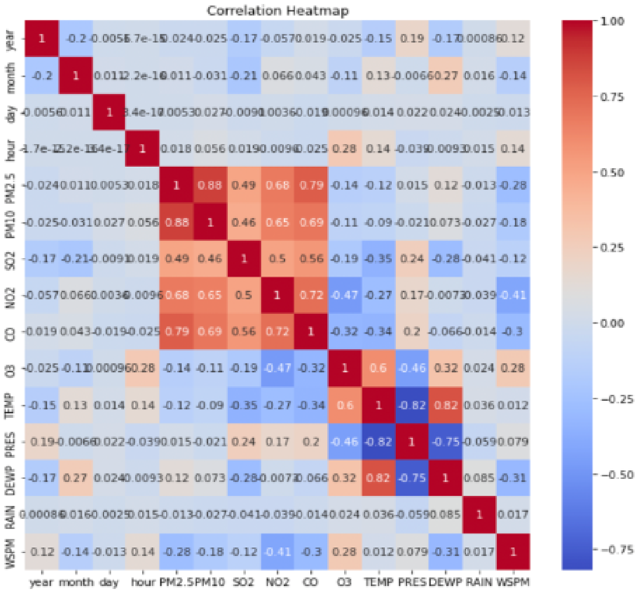


Fig. 1. Correlation Heatmap of features

By heat map, it was observed that the target variable "PM2.5" has a high correlation with columns "CO", "NO2", and "PM10. Hence, correlated heatmaps were plotted only for those.

Furthermore, data was examined to understand the average PM2.5 concentration of the cities in various months of the year, as seen in Fig 2. It was noted that among all months, December had the maximum concentration of PM 2.5 for most of the cities. In the paper [6] it was observed: "Notable seasonal variations of PM2.5-polluted days were observed, especially for the megacities in east-central China, resulting in frequent heavy pollution episodes occurring during the winter".

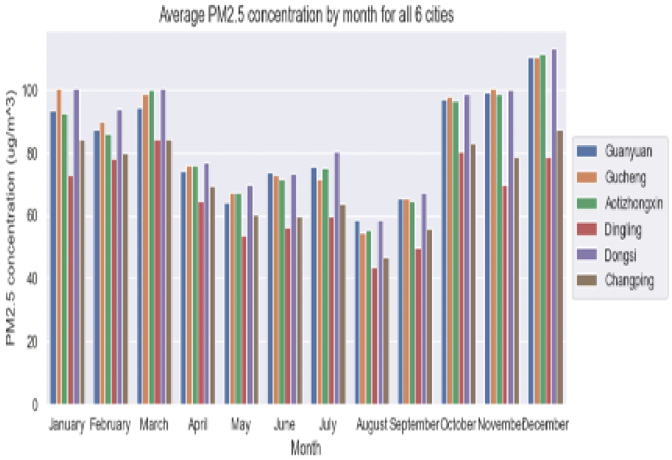


Fig.2. Average PM2.5 concentration by months

Another graph created as seen in Fig 3. displayed the average hourly recorded PM2.5 concentration in the air for all 6 cities, grouped by wind direction. Each line on the graph represents one of the 6 cities, and the x-axis represents the wind direction. The y-axis represents the average PM2.5 concentration (in micrograms per cubic meter) for each wind direction. The legend on the right side of the graph shows which line corresponds to which city.

From the graph, it was concluded that there was a variation in the average hourly PM2.5 concentration by wind direction for all 6 cities. It was also observed that for each city, the concentration of PM2.5 varies depending on the wind direction. For example, in Aotizhongxin, the PM2.5 concentration is highest when the wind direction is from the north, while in Dingling, the concentration is highest when the wind direction is from the southeast. Therefore, we can infer that wind direction is an important factor in determining the level of PM2.5 concentration in the air for these cities.

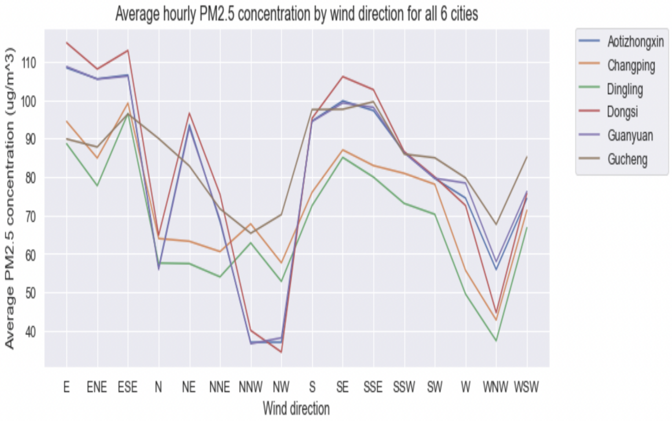


Fig. 3. Average hourly PM2.5 concentration by wind direction

Moreover, a histogram as seen in Fig. 4. was created to show how the values of PM2.5 are distributed over a dataset. The data has been categorized, and each category's bar height indicates how frequently observations fall into that category. Higher levels have a distinct color than lower ones, and different colors are used to depict the various PM2.5 level ranges. PM2.5 is on X-axis, while the frequency of observation is on the Y-axis. This method helped to portray PM2.5 values of the dataset by creating a histogram along with giving proper labels to x and y.

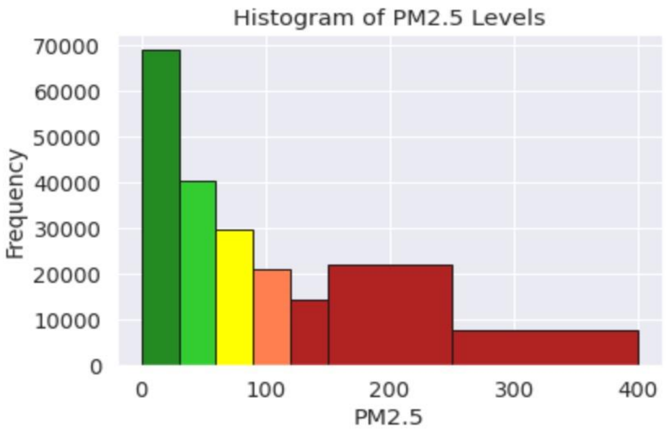
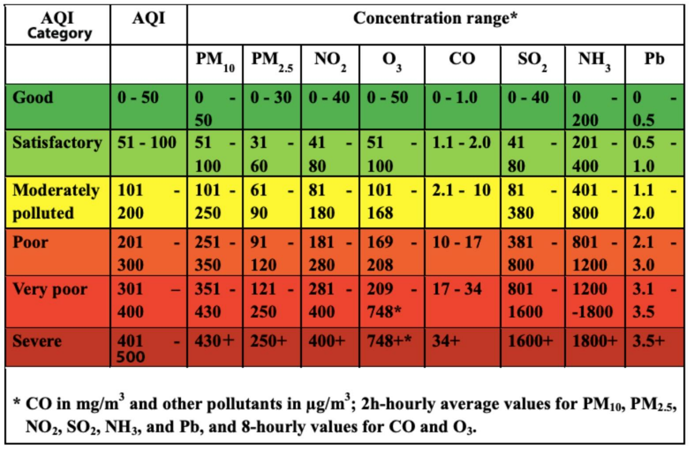


Fig. 4. Histogram of PM2.5 Levels

PM2.5 refers to microscopic particles or droplets that have a diameter of less than 2.5 micrometers. The presence of PM2.5 is a significant air pollutant and has been linked to various health issues like asthma, heart disease, stroke, and lung cancer.To measure the level of PM2.5 in the air, a device called a PM2.5 monitor or particle counter that utilizes a laser beam to count particles of a specific size in the air is used. The concentration of PM2.5 is expressed in micrograms per cubic meter (µg/m³).

PM2.5 is one of several air quality indicators used to measure the quality of the air. Other indicators include ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO) as seen in Fig.5.Governments and organizations around the world monitor these indicators to assess air quality and develop policies and regulations to reduce air pollution and protect public health[7].



| AQI Category | AQI | Concentration range* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $PM_{10}$ | $PM_{2.5}$ | $NO_2$ | $O_3$ | CO | $SO_2$ | $NH_3$ | Pb |
| Good | 0 - 50 | 0 - 50 | 0 - 30 | 0 - 40 | 0 - 50 | 0 - 1.0 | 0 - 40 | 0 - 200 | 0 - 0.5 |
| Satisfactory | 51 - 100 | 51 - 100 | 31 - 60 | 41 - 80 | 51 - 100 | 1.1 - 2.0 | 41 - 80 | 201 - 400 | 0.5 - 1.0 |
| Moderately polluted | 101 - 200 | 101 - 250 | 61 - 90 | 81 - 180 | 101 - 168 | 2.1 - 10 | 81 - 380 | 401 - 800 | 1.1 - 2.0 |
| Poor | 201 - 300 | 251 - 350 | 91 - 120 | 181 - 280 | 169 - 208 | 10 - 17 | 381 - 800 | 801 - 1200 | 2.1 - 3.0 |
| Very poor | 301 – 400 | 351 - 430 | 121 - 250 | 281 - 400 | 209 - 748* | 17 - 34 | 801 - 1600 | 1200 -1800 | 3.1 - 3.5 |
| Severe | 401 - 500 | 430+ | 250+ | 400+ | 748+* | 34+ | 1600+ | 1800+ | 3.5+ |

* CO in $mg/m^3$ and other pollutants in $\mu g/m^3$; 2h-hourly average values for $PM_{10}$, $PM_{2.5}$, $NO_2$, $SO_2$, $NH_3$, and Pb, and 8-hourly values for CO and $O_3$.

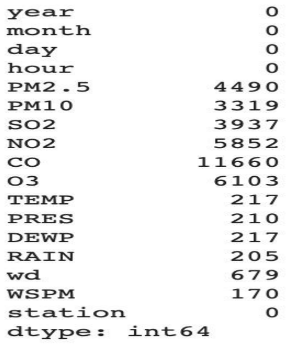
Fig. 5. AQI information

## 3.3 Data Cleaning and Pre- Processing

Data encoding: As some algorithms can not handle categorical variables, and the dataset contains features such as “wind direction”, feature encoding needs to be done. One Hot Encoder was used to encode the features. Each category is converted into a binary feature that indicates whether or not the observation belongs to that category. This approach creates new columns for each possible category and assigns a value of 1 or 0 to each column depending on whether the observation belongs to that category or not.

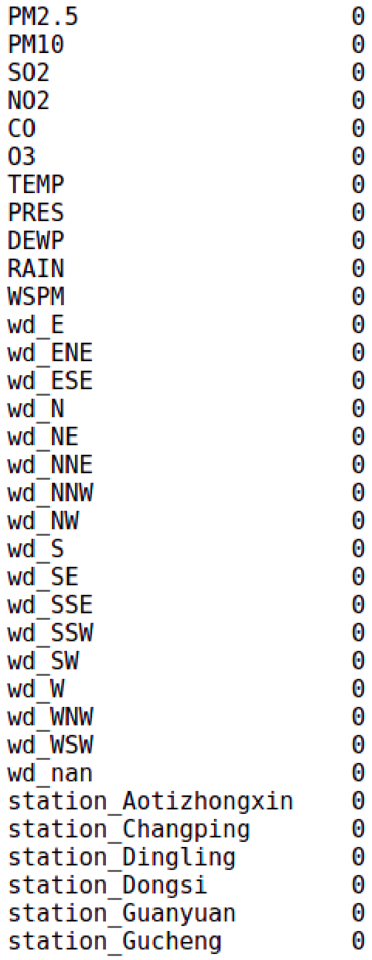
Standardization: The data are scaled with a mean of 0 and a standard deviation of 1 using standardization. This is done to make sure that all the characteristics have the same range and size, which has helped some machine learning algorithms perform better.

Handling missing values: In order to clean the data, the columns were checked for missing values which can be viewed in Fig. 6. Once identified, the null values were handled using the K-Nearest Neighbors (KNN) imputation technique. The algorithm identifies the k nearest neighbors of each missing value based on the similarity of their other features, and then takes the average or of their corresponding values to fill in the missing value. This technique was used after standardization, to ensure that the distances between data points are based on the actual differences between them and not on differences in scale.This method helps to reduce the bias in the dataset caused by the missing values and preserves the relationships between the variables in the dataset.



```
year             0
month            0
day              0
hour             0
PM2.5         4490
PM10          3319
SO2           3937
NO2           5852
CO           11660
O3            6103
TEMP           217
PRES           210
DEWP           217
RAIN           205
wd             679
WSPM           170
station          0
dtype: int64
```

Fig. 6.. Sum of missing values in dataset



```
PM2.5                 0
PM10                  0
SO2                   0
NO2                   0
CO                    0
O3                    0
TEMP                  0
PRES                  0
DEWP                  0
RAIN                  0
WSPM                  0
wd_E                  0
wd_ENE                0
wd_ESE                0
wd_N                  0
wd_NE                 0
wd_NNE                0
wd_NNW                0
wd_NW                 0
wd_S                  0
wd_SE                 0
wd_SSE                0
wd_SSW                0
wd_SW                 0
wd_W                  0
wd_WNW                0
wd_WSW                0
wd_nan                0
station_Aotizhongxin  0
station_Changping     0
station_Dingling      0
station_Dongsi        0
station_Guanyuan      0
station_Gucheng       0
```

Fig. 7. Missing values after encoding and imputation

In Fig6 and Fig7 the results before and after imputation with KNN and encoding are shown. Dropping columns: The columns 'year', 'month', 'day', and 'hour' were dropped during the modeling process as they were found to be less relevant.

Feature Engineering: The following top 9 features were selected for analysis based on their relevance to air quality and their availability in the dataset:

1. PM10: Particulate matter with a diameter of 10 micrometers or less
2. CO: Carbon monoxide
3. DEWP: Dew point temperature
4. SO2: Sulfur dioxide
5. TEMP: Temperature
6. NO2: Nitrogen dioxide
7. PRES: Pressure
8. O3: Ozone
9. WSPM: Wind speed

These features were selected after conducting exploratory data analysis and building decision trees. The features that are chosen for the splits in Decision Trees are considered to be the most important ones, as they have the greatest impact on the final decision made by the tree. The goal was to include features that significantly impact air quality. Additionally, it was made sure that the selected features did not have high multiple correlations to avoid issues of overfitting.

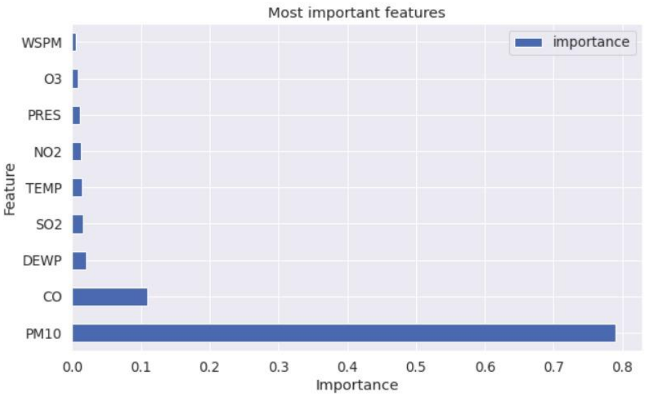


Fig. 8. Top 9 Features

Handling Outliers: As a result of performing handling outliers, it was observed that many important values were getting eliminated which would hamper the environmental sustainability cause. Hence, this approach was discarded.

# 4 Problem solution

## 4.1 Modeling

In the modeling phase of this project, a number of regression models were built for the prediction of the target feature "PM2.5".

### 4.1.1 LGB Regressor:

LGB Regressor is a machine learning algorithm based on gradient boosting. It is an efficient and effective algorithm that uses a combination of decision trees and boosting to improve the accuracy of predictions. LGB Regressor can handle a large number of observations and it gave the best result with 93.5 R2.

### 4.1.2 Gradient Boosting Regressor:

Gradient Boosting Regressor is an ensemble machine learning algorithm that builds decision trees one at a time, where each new tree helps to correct errors made by previously trained trees. The Gradient Boosting Regressor is a powerful model that is often used for regression tasks because it can handle complex datasets and is less prone to overfitting than other models.

### 4.1.3 XGB Regressor:

The XGB Regressor is a type of gradient boosting algorithm that is optimized for speed and performance. It uses a similar approach to the Gradient Boosting Regressor, but with some modifications to the algorithm to make it more efficient. The XGB Regressor is highly accurate and can handle large datasets with many features.

### 4.1.4 KNN Regressor:

The KNN Regressor is a non-parametric machine learning algorithm that is used for regression tasks. It works by identifying the k-nearest neighbors to a given data point in the training set and then predicting the target variable based on the average of their values. KNN Regressor is useful when the dataset is small and has few features.

### 4.1.5 Random Forest Regressor:

Random Forest Regressor is a type of ensemble learning algorithm that creates a large number of decision trees and aggregates their predictions to produce a final output. It is similar to the Extra Trees Regressor but uses a different approach to split the nodes of the decision trees. The Random Forest Regressor is powerful for large datasets with many features and can handle missing values.

### 4.1.6 Extra Trees Regressor:

The Extra Trees Regressor is a machine-learning model that is similar to the Random Forest Regressor. However, instead of constructing individual decision trees, the Extra Trees Regressor generates a large number of randomized decision trees and then aggregates their predictions. This approach can help to reduce overfitting and improve model performance.

Overall, each of these models has its strengths and weaknesses depending on the dataset and the specific task at hand. In this project, we applied each model to the air quality dataset to predict PM2.5 levels. The performance of each model was evaluated based on its accuracy, and the best-performing model was selected for further analysis.

## 4.2 Evaluation metrics

The regression models can be evaluated using several metrics such as RMSE, MAE, Adjusted R2 score and R2 score.

### 4.2.1 Root Mean Square Error (RMSE):

RMSE is a common metric used to evaluate the performance of regression models. It measures the difference between the predicted and actual values and provides a measure of how well the model fits the data. A lower RMSE value indicates that the model is better at predicting PM2.5 levels.

### 4.2.2 Mean Absolute Error (MAE):

MAE is another metric used to evaluate the performance of regression models. It measures the average absolute difference between the predicted and actual values.

### 4.2.3 R-squared (R2) score:

R2 score is a statistical measure that represents the proportion of variance in the dependent variable (PM2.5 levels) that is predictable from the independent variables (other air quality factors). In other words, R2 score measures how well the model fits the data. A higher R2 score indicates that the model is better at predicting PM2.5 levels. However, R2 score alone does not indicate the goodness of fit of the model, as it increases with the addition of more variables.

### 4.2.4 Adjusted R-squared (Adjusted R2) score:

Adjusted R2 score is a reformed version of the R2 score that takes into account the number of independent variables in the model. Adjusted R2 score penalizes the model for adding unnecessary independent variables that do not improve the model's performance. Therefore, Adjusted R2 score is a more reliable measure of the model's goodness of fit than the R2 score.

In Table 1, evaluation metrics are shown for all the models built.

| **Regression Models** | **RMSE** | **MAE** | **R2 score** | **Adjusted R2 score** |
|---|---|---|---|---|
| KNN | 23.1 | 13.9 | 91.7 | 91.7 |
| Extra Trees | 21.7 | 13.9 | 92.7 | 92.7 |
| LGB | 20.3 | 12.8 | 93.6 | 93.6 |

| | | | | |
|---|---|---|---|---|
| Random Forest | 24.4 | 15.7 | 90.7 | 90.7 |
| Gradient Boosting | 24.2 | 15.4 | 90.9 | 90.8 |
| XGB | 25.3 | 16.2 | 90.0 | 90.0 |

Table 1: Evaluation results of the models

### 4.3 Model Comparison

In the table above, it can be observed that LGB Regressor outperforms all the models predicting PM2.5 with a R2 score of nearly 94% which is a significant increase compared to other models.

### 4.4 Languages Used

Python was used to complete this project in its entirety. Different libraries based in python have been utilized to code effectively. HTML was used to create the user interface and web pages for the model demo and python flask was used to create the web application that provided an interface for users to interact with the model demo.

### 4.5 Tools Used

Jupyter Notebook was utilized for implementing the models. Sklearn is used for preprocessing of data, matplotlib and Seaborn are used for visualizations, sklearn models for comparison of performance. GitHub and Jira were used as communication tools to help manage team collaboration, code development, and task assignments. GitHub provided a platform for version control and code sharing, allowing team members to work together on code development and track changes over time. Jira was used for project management, enabling team members to create and assign tasks, track progress, and communicate updates. By utilizing these tools, the team was able to work efficiently and effectively towards achieving project goals.

## 5 Conclusion

These results can be taken into consideration for impactful decision making related to air quality. The result obtained proves that historical data can be utilized to create models that can predict PM 2.5 levels of the city in the future. Rapid developments in the field of AI and Machine Learning has proven to be a boon for a variety of fields. This model can aid lawmakers to preemptively take actions to prevent any situation getting out of hand, keeping in mind the health of the citizens.

## 6 Acknowledgement

We would like to express our gratitude to Professor Dr. Vishnu Pendayala for his valuable assistance and support during the course of this project. His guidance has been instrumental in helping us achieve our goals.

## References

[1] M.Yang, "A Machine Learning Approach to Evaluate Beijing Air Quality," Senior Thesis, University of California, Davis, (2018)

[2] P. Bhalgat, S. Pitale, and S. Bhoite, "Air quality prediction using machine learning algorithms," Int. J. Comput. Appl. Technol. Res., vol. 8, no. 9, pp. 367-370, 2019, ISSN: 2319-8656

[3] T. V. Vu, Z. Shi, J. Cheng, Q. Zhang, K. He, S. Wang, and R. M. Harrison, "Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique," Environmental Science & Technology, vol. 53, no. 13, pp. 7593-7602, (2019)

[4] S. Mahanta, T. Ramakrishnudu, R. Raj Jha, and N. Tailor, "Urban Air Quality Prediction Using Regression Analysis," Dept. of CSE, NIT Warangal, Warangal, India

[5] T. M. T. Lei, S. W. I. Siu, J. Monjardino, L. Mendes, and F. Ferreira, "Using Machine Learning Methods to Forecast Air Quality: A Case Study in Macao," Atmosphere, vol. 13, no. 9, p. 1412, Sep. 2022.'

[6] Z. Liu *et al.*, “Characteristics of PM mass concentrations and chemical species in urban and background areas of China: emerging results from the CARE-China network,” *Atmospheric Chemistry and Physics*, vol. 18, no. 12, pp. 8849–8871,Jun.2018,doi:https://doi.org/10.5194/acp-18-8849-2018.

[7] “AQI calculation update,” *airveda* https://www.airveda.com/blog/AQI-calculation-update

## Appendix A

### Criteria met in Rubrics

1. Visualization – In the exploratory data analysis phase, various visualizations such as heatmaps, bar charts, line chart, etc are employed to gain insights into the data and demonstrate associations.
2. Presentation Skills - Prepared explanatory slides, used clear language, and added interesting pictures to explain model performance. Used a structured story with evidence from data, and reflection to improve time delivery and practiced presentation skills.
3. Relates to sustainability - The project contributes to sustainability by offering information on the amount of air pollutants in various regions. This information helps policymakers make informed decisions, which can reduce pollution and improve air quality, thus creating a healthier and more sustainable environment. The work is towards the cause Climate action among 17 SDGs (Department of Economic and Social Affairs Sustainable Development).By minimizing the air pollutants in the air, the health and well-being of people will be improved which directly contributes in making a sustainable environment.
4. Saving the model for quick demo - Models are saved for testing and demonstration purposes using a pickling approach. This involves saving the trained model as a file that can later be loaded and tested with new data.
5. Code Walkthrough - The Jupyter Notebook contains the entire code with clear explanations in comments, making it easy to comprehend the operations carried out.
6. Report - IEEE format was followed with self-written clear language for the report. The report contains all the necessary information to comprehend the problem requirement and approach to solving it.
7. Version Control - All the data and code is stored in a public accessible GitHub repository, along with a readme file for instructions and guidance. JIRA board was used to monitor the status of every author’s story where the Project Name of the Project created is – DATA245 - Group5, URL is https://datacampers.atlassian.net/jira/software/projects/DG/boards/2/backlog

8. Discussion / Q&A - During the presentation, an open discussion will be encouraged, answering any raised questions. Additionally, there will be a designated time for a Q&A session at the end of the demo.

9. Lessons learned - The team learned how to use complex machine-learning models to identify significant features related to the target feature through correlation analysis. Also, gained skills in effectively visualizing data and creating a logical narrative around it. In addition, the team learned how to write a report in the IEEE format and how to use GitHub and Git for version control. Additionally, also learned to develop proficiency in searching relevant research papers using websites such as Google Scholar.

10. Prospects of winning competition/publication – The project has the potential to win competitions and be published due to its data-driven approach, accurate and high model accuracy, and relevance to important sustainability issues.

11. Innovation - The features used in the project were found to be more relevant than those used in the literature. Additionally, KNN Imputer was utilized, which was not commonly used. The models implemented were also unique compared to the commonly used models. Hyperparameter tuning was conducted for a few models during the training process.

12. Evaluation of performance – Based on the evaluation metrics, Extra Tree Regression was finalized as the final model, as it achieved an R2 score of approximately 94%. Please refer to Table 1 for details.

13. Teamwork - The entire team participated in all project phases and had weekly meetings to discuss progress toward the goal. Appendix B contains a clear credit statement.

14. Technical difficulty – The modeling process was time-consuming due to the need for hyperparameter tuning. Regressors were particularly challenging and required significant amounts of time. Additionally, traditional laptops were not powerful enough to handle the workload.

15. Practiced pair programming? - The team worked together on a programming task using Google Colab and Jupyter Notebook. The tasks were divided based on issues created on the Jira board. The team was responsible for completing the task and if there were any impediments, they could mention it in their Jira stories and everyone could look together into the issue. The pair programming helped the team to excel as everyone was giving regular updates on their stories and increased efficiency. The pair programming helped in improving the code quality and better collaboration skills.

16. Practiced agile / scrum (1-week sprints) - The team followed a 5-week sprint agile framework using JIRA. Stories for each week were distributed among team members, and weekly meetings were held to track progress on deliverables.

17. Used Grammarly / other tools for language? – Grammarly was used to check the language and grammar rules of the documents.

18. Slides – A slide presentation was prepared that covered the important aspects of the project.

19. Demo - A demo structure was prepared by the team to ensure a well-organized and effective presentation, highlighting the functionality of the working model.

20. In order to format the paper according to IEEE standards, the official IEEE LaTeX template provided on the IEEE website was used. This ensured that the paper adhered to the guidelines for font size, margins, and other formatting requirements. The LaTeX editor "Overleaf" facilitates collaboration among the group members and streamlines the writing and editing process. Overall, using LaTeX and the IEEE template helped produce a professional paper.

21. Used creative presentation techniques - The presentation was created using the Prezi tool, which includes engaging animations to make it more interesting.

22. Literature Survey - The team referred to relevant papers and research on air quality predictions and classifications. The literature was organized into meaningful subsections and concepts were introduced appropriately. All cited works were properly referenced.

## Appendix B

### Author Contributions

Hruthik Vinnakota: Data collection and preprocessing he was responsible for collecting and cleaning the data to ensure it is suitable for machine learning tasks. The tasks involved were removing missing values, outliers, and duplicates.

Maria Hovhannisyan: Model selection and development she was responsible for selecting the appropriate machine learning algorithms and developing models that can accurately predict the target variable.

Sahanti Samarth Zade: Model evaluation she was responsible for evaluating the performance of the machine learning models developed by the team. She used techniques such as cross-validation and grid search to determine the best performing model.

Shubham Singh: Deployment and integration he was responsible for deploying the machine learning models in a production environment and exploring new techniques and algorithms that could be applied to the project.

Utsav Rastogi: Research and innovation he was responsible for keeping up with the latest developments in machine learning and integrating them with other systems as necessary.