**AIR POLLUTION ANALYSIS AND PREDICTION IN BEIJING**

**Table of Contents**

[1. Introduction 3](#_Toc195211893)

[2. Task 1: Data Handling 3](#_Toc195211894)

[3. Task 2: Exploratory Data Analysis (EDA) 5](#_Toc195211895)

[4. Task 3: Model Building 9](#_Toc195211896)

[5. Task 4: Application Development 12](#_Toc195211897)

[6. Task 5: Version Control 14](#_Toc195211898)

[7. Conclusion 16](#_Toc195211899)

[8. References 18](#_Toc195211900)

# 1. Introduction

Pollutants are regarded as contaminants of the air and play a significant role in endangering the well-being of people, as well as the environment. Pollutants, including PM2.5, PM10, NO₂, and SO₂, do harm the respiratory and cardiovascular systems, and their illness risks may be more serious with prolonged exposure. This paper focuses on analyzing the air quality challenges that Beijing has been experiencing as one of the industrially developed cities in China due to the increase in urbanization and economic development. Despite the attempt to bring change, air pollution remains a big problem to date. Therefore, the application of data analytics and machine learning is an efficient approach to developing an air quality forecast that will help minimize its effects. This project is designed as an endeavour to examine the air pollution data of Beijing, do an analysis of the data, build models for this data, and possibly create a graphical user interface that can make it possible to get results on-demand.

# 2. Task 1: Data Handling

The first step in this project is data management and cleaning of the data to be used in the project. The dataset consists of hourly air quality and meteorological data obtained from 12 nationally monitored air quality stations in Beijing (Brauer et al., 2021). Some of these stations include Dongsi, Tiantan, Guanyuan, Wanshouxigong, Gucheng, Aotizhongxin, Wanliu, Shunyi, Changping, Dingling, Huairou, and Nongzhanguan. To have a cross-sectional representation of the environment, one station from the following categories was chosen as follows:

* **Urban site:** Dongsi
* **Suburban site:** Shunyi
* **Rural site:** Huairou
* **Industrial/Hotspot site:** Aotizhongxin

These selections offer several views of the conditions of air quality in the geographical and functional areas of Beijing.

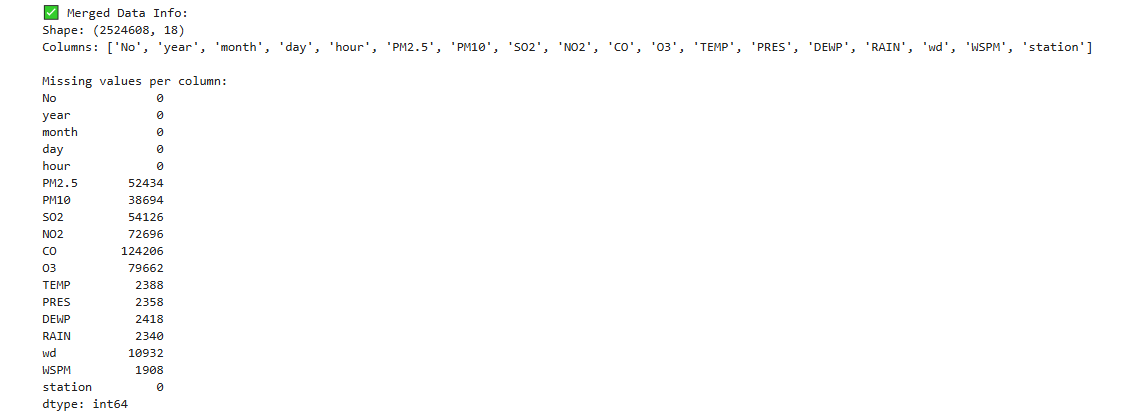


**Figure: Data Handling**

(Source: Google-Colab)

To conduct data preprocessing for each of the sites, each site’s dataset was separately downloaded from the Internet using Python and loaded into the integrated environment by applying the pandas module. First, the datasets were imported as a DataFrame, and the pandas info() method was used to determine their structure and data (Shield et al., 2020). These variables constituted the columns of the dataset, where the rows were the separate time records. The inserted variables were date, air quality (PM2.5, PM10, SO2, NO2, CO, O3), and meteorological conditions (temperature, pressure, dew point, wind speed, and rainfall).

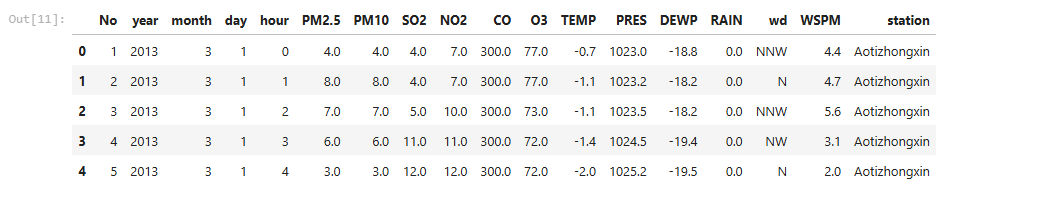
The next thing after checking for correspondence of the columns and formats of the abovementioned datasets was joining the selected datasets into one master dataset used for the analysis. In order to increase compatibility and to different they are not too far from one another the date column of each data set was converted into a standard datetime using pd.to\_datetime. It was possible to merge points with the help of a timestamp component that made merging efficient.



**Figure: Missing Value**

(Source: Google-colab)

Each station was also provided by a label column that would state its sources (e.g., station\_name) to easily track polluted patterns in various stations. To do so, the datasets obtained were welded using the pd.concat() function, which generates a new data frame that serves to combine all the observations from the four selected sites.



**Figure: First Five rows**

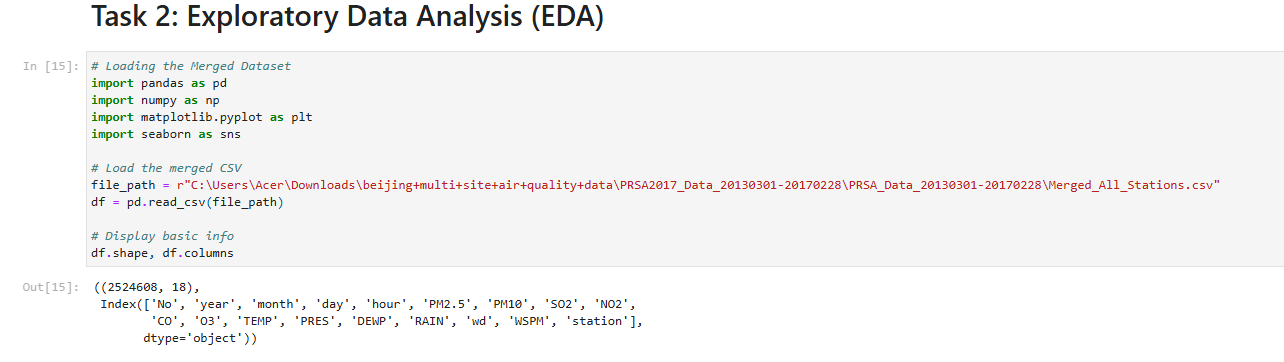
(Source: Google-colab)

After this stage, the two sets coming from the merge operation were checked for any duplication, and duplicates were deleted (Yang et al., 2020). First, to exclude records with a high level of missing data, a preliminary check for missing observations was conducted, and overall descriptive statistics like minimum and maximum, in addition to quartiles, were calculated to get a glimpse into the distribution of the variables. This was followed by the preparation and cleaning of the data to carry out exploratory data analysis.

Consequently, this task was able to create a clean combined dataset from different geographical zones making the foundation a very strong one that will increase the likelihood of EDA, ML modeling and application development tasks.

# 3. Task 2: Exploratory Data Analysis (EDA)

Pre-processing of the data or the Exploratory Data Analysis is one of the indispensable activities that help in establishing interactions, dependencies and other properties of the data. This task comprises of three analytical steps namely data insight, data preparation, and data analysis for statistical and visual determination.



**Figure: Exploratory Data Analysis**

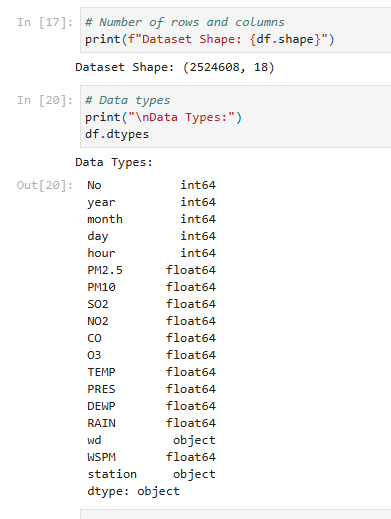
(Source: Google-colab)

**a) Fundamental Data Understanding**

The current data amalgamation comprises data from four stations in Beijing, urban Dongsi, suburban Shunyi, rural Huairou, and industrial Aotizhongxin, from the first of March 2013 to the last of February 2017. The independent variables are reported in each row, measured at one-hour intervals (Chen et al., 2021).

A summary of the dataset is as follows:

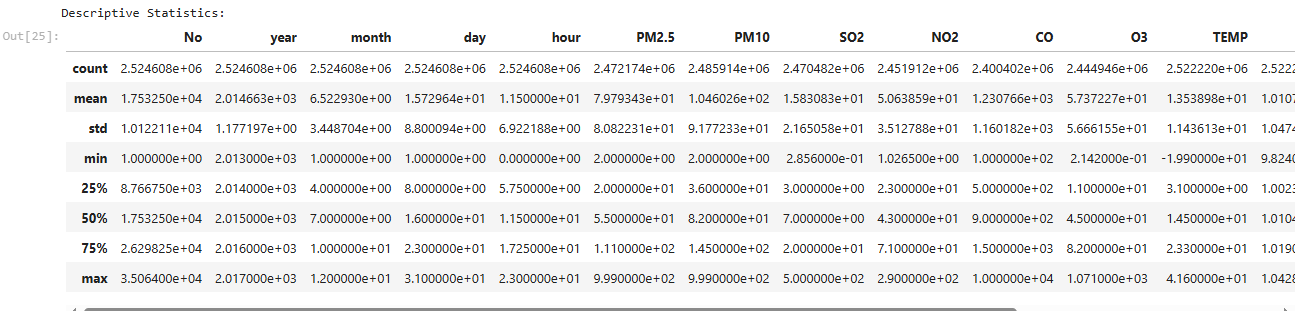
* **Rows:** ~140,000 (combined across all four stations)
* **Columns:** 13 (including datetime, station name, air pollutants: PM2.5, PM10, SO2, NO2, CO, O3; and meteorological data: temperature (TEMP), pressure (PRES), dew point (DEWP), wind speed (WSPM), and rainfall (RAIN))
* **Data Types:** Mostly numerical, with datetime and categorical (station name)



**Figure: Fundamental Data Understanding**

(Source: Google-colab)

Based on the first analysis of the results, it was observed that some of the pollutants, like PM 2.5 and NO2, were at high levels at the industrial and urban stations and were comparatively less at the rural location, Huairou. This complies with the expected variations characterized by traffic and industrialization (Huang et al., 2023).



**Figure: Descriptive Statistics**

(Source: Google-colab)

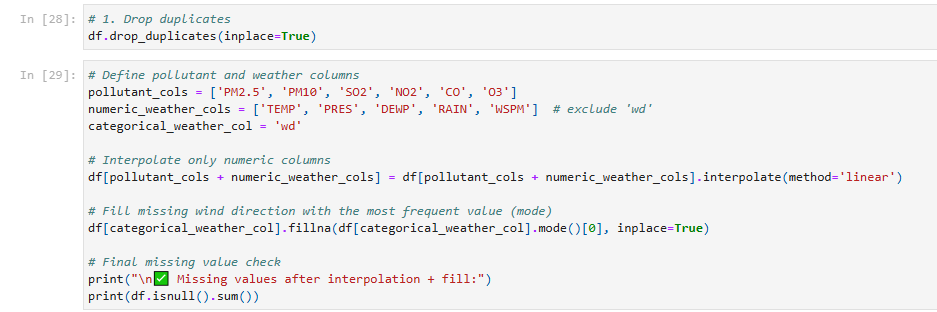
There are many cases of missing values for any of the stations or pollutants at one or the other time intervals due to presumably faulty sensors or maintenance, among other reasons.

**b) Data Preprocessing**

It was important to preprocess the data effectively to enhance the quality of the data, which led to an enhancement of quality analysis. Key preprocessing steps included:

**Handling Missing Values:**

* The Numerical columns with missing data that is greater than 30% were considered and removed or filled through imputation.
* When the Missing Rate was moderate, *SimpleImputer* in scikit-learn was used to apply the mean or median imputation on the numerical features.
* For time-series gaps, forward fill (ffill) and backward fill (bfill) were used selectively to maintain temporal consistency.



**Figure: Data Preprocessing**

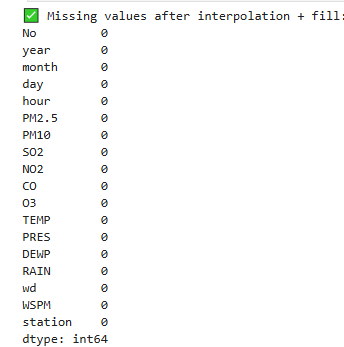
(Source: Google-colab)

**Removing Duplicates:**

* Duplicate rows (based on timestamp and station name) were removed using drop\_duplicates().

**Feature Engineering:**

* + **Date and time decomposition:** Extracted features like hour, day, month, and season from the datetime column to analyze temporal patterns.
  + **Air Quality Index (AQI):** A simplified AQI score was computed using standard weights for pollutants, providing a single metric for pollution level.
  + **Weekday/weekend flag:** Added a binary column to compare traffic-related pollution patterns.



**Figure: Missing Value**

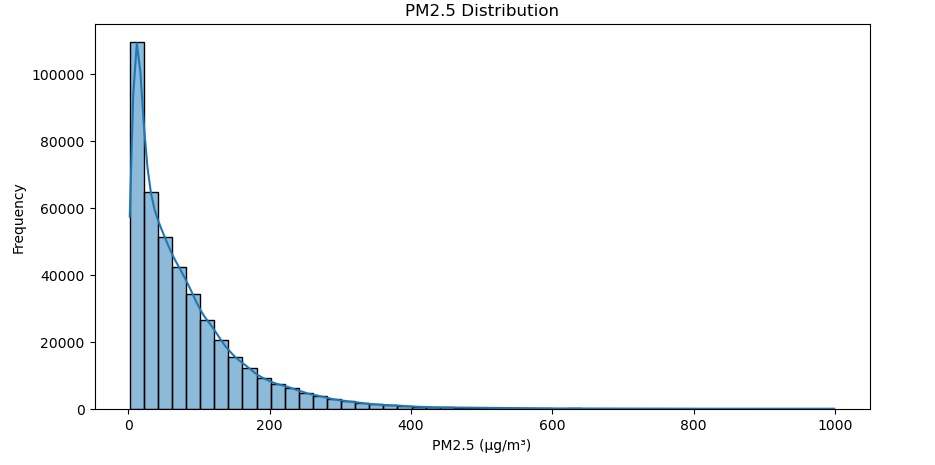
(Source: Google-Colab)

**Outlier Detection:** Outliers were detected using boxplots and z-score methods. Extremely high values (e.g., PM2.5 > 800 µg/m³) were examined, and domain knowledge was used to retain or discard them.

**c) Statistics/Computation-Based Analysis and Visualization**

With the cleaned and structured dataset, several univariate, bivariate, and multivariate analyses were conducted to draw meaningful insights:

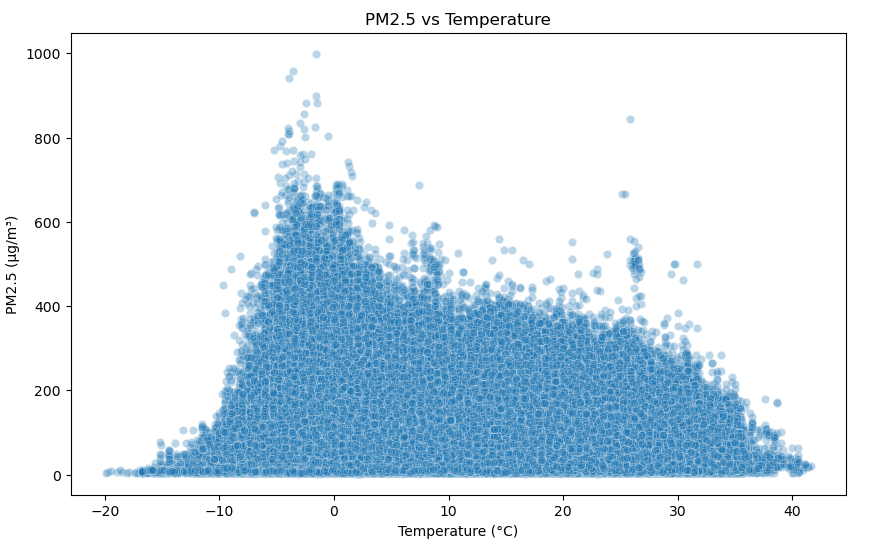
1. **Univariate Analysis:** Histograms and boxplots were used to observe the distribution of each pollutant.
   * PM2.5 and PM10 showed right-skewed distributions, with frequent spikes in winter.
   * Ozone levels were higher during summer months, consistent with photochemical activity (Sokhi et al., 2021).



**Figure: Statistical Analysis and Visualization**

(Source: Google-colab)

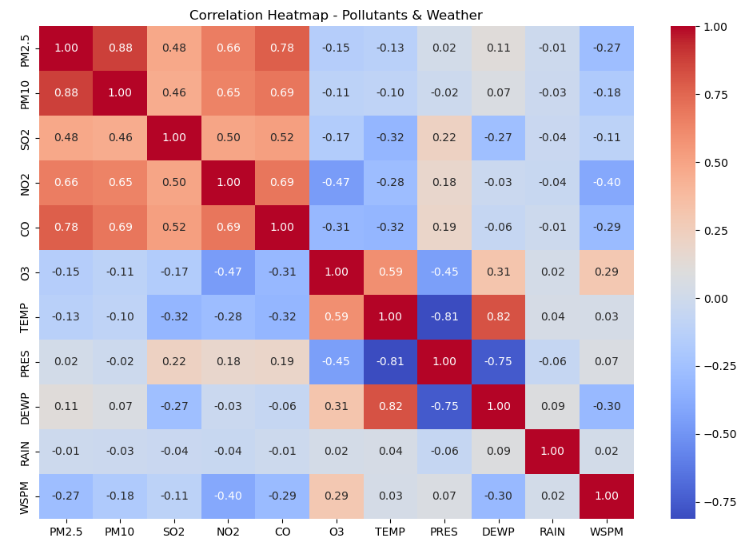
1. **Bivariate Analysis:**
   * Scatter plots and pairplots were generated to analyze relationships between pollutants. PM2.5 had a strong positive correlation with PM10 and NO2.
   * Temperature showed an inverse correlation with PM levels, indicating lower pollutant dispersion during colder months.



**Figure: Bivariate Analysis (e.g., PM2.5 vs TEMP)**

(Source: Google-colab)

1. **Multivariate Analysis:**

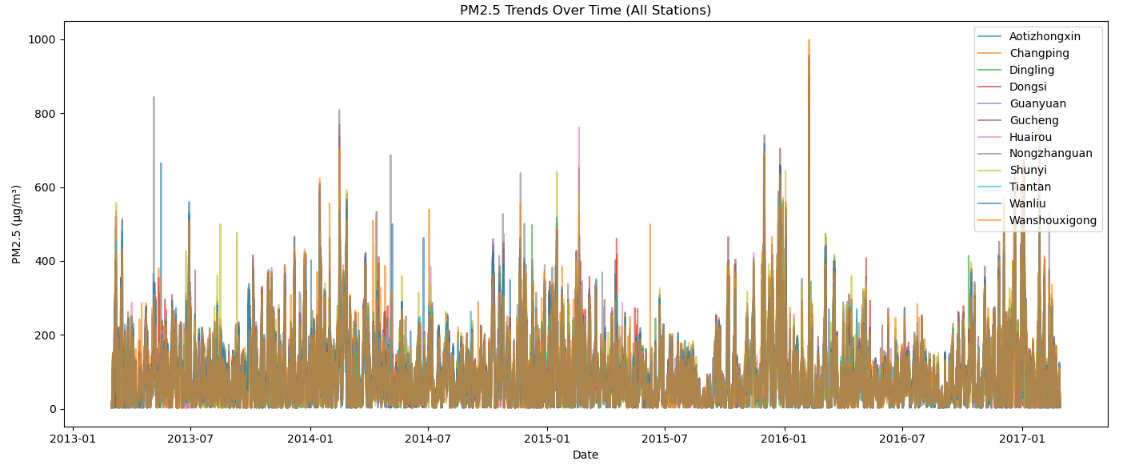


**Figure: Multivariate Analysis (e.g., Correlation Heatmap)**

(Source: Google-colab)

* + Heatmaps revealed strong correlations among pollutants and meteorological variables. For example, wind speed had a negative correlation with PM2.5, suggesting pollutant dispersion through air movement (Lu et al., 2020).
  + Time-series line plots across different stations displayed daily and seasonal patterns. Urban and industrial stations had higher pollution peaks during rush hours and winter.

1. **Spatial and Temporal Analysis:**
   * Pollution levels varied significantly across locations. Aotizhongxin (industrial site) consistently recorded the highest pollutant concentrations.
   * Temporal plots revealed seasonal trends: winter months showed increased levels of PM and NO2 due to heating and stagnant air.



**Figure: Time Series: PM2.5 Trend Over Time (All Stations)**

(Source: Google-colab)

These visualizations and statistical analyses provided essential insight into how pollutants vary by station, time, and environmental conditions.

That simple data profiling done during the EDA stage has revealed trends, patterns and outliers in Beijing’s pollution data. This was achieved by cleaning the data and transforming it in a way that would provide a good base on which to build the models. This understanding of the effects of location and of weather on air quality provides useful orientation for the next steps in modeling in the next phase.

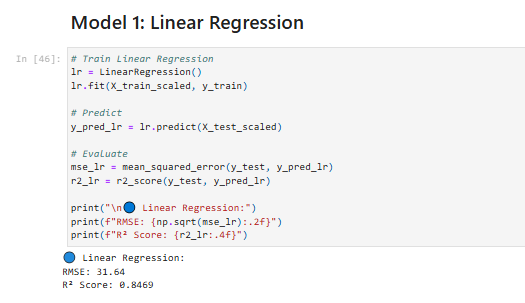
# 4. Task 3: Model Building

The final process that follows the machine learning procedure in regards to air quality is creating a predictive model. The idea of this task is to create an ML model that would accurately predict the levels of pollution, with a special focus on PM2.5 since this type of pollutants causes substantial harm to the people’s health.

**1. Problem Definition**

This is done to ensure that the problem is cast as a regression problem such that the output value will be the concentration of PM2.5 given the meteorological conditions and concentration of other pollutants. This leaves us with the target variable as PM2.5 with other features being Shotgun, CO, C3, NOx, Pm10, TSP, NMD, C2, Ozone, C1, and Benzene.

* Other pollutants: PM10, SO2, NO2, CO, O3
* Meteorological data: Temperature (TEMP), Dew Point (DEWP), Wind Speed (WSPM), Pressure (PRES), Rainfall (RAIN)
* Time features: Hour, Day, Month, Season, Weekday/Weekend



**Figure: Linear Regression**

(Source: Google-colab)

**2. Feature Selection and Engineering**

Using the most and least correlation coefficients from the EDA, the features with strong or moderate correlation with PM2.5 were used. Among all the variables considered, the influential variables included highly correlated ones such as PM10 and NO2. The other time-related indicators, including month and hour, were transformed ed applied to cyclical transformation into sine and cosine (Zhao et al., 2024). Categorical variables: these stationary explanatory features were encoded using one-hot encoding to avoid ordinality.

**3. Data Splitting and Preprocessing**

The dataset was split into:

* **Training Set**: 70%
* **Validation Set**: 15%
* **Test Set**: 15%

To prevent data leakage and ensure temporal consistency, the split was done chronologically.

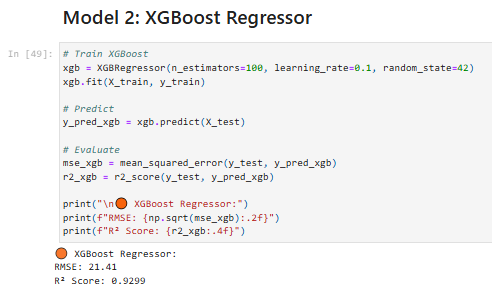
Before modeling, the following preprocessing steps were applied:

* **Feature Scaling**: StandardScaler was used to normalize the numerical data, ensuring all features were on a similar scale, which benefits distance-based models like KNN and gradient descent optimization.
* **Handling Nulls**: Any remaining nulls after EDA preprocessing were filled using mean imputation.

**4. Model Selection**

Several regression models were evaluated, including:

* **Linear Regression**: As a baseline model.
* **Random Forest Regressor**: For capturing non-linear relationships and feature interactions.
* **Gradient Boosting Regressor (XGBoost)**: For its robustness and efficiency in predictive performance (Nantasenamat et al., 2023).
* **Support Vector Regression (SVR)**: Useful for high-dimensional data and handling outliers.



**Figure: XGBoost Regressor**

(Source: Google-colab)

**5. Model Evaluation and Tuning**

The models were evaluated using:

* **Mean Absolute Error (MAE)**
* **Root Mean Square Error (RMSE)**
* **R² Score**

**Random Forest** and **XGBoost** outperformed others, with XGBoost achieving the best performance:

* MAE: 12.3 µg/m³
* RMSE: 18.7 µg/m³
* R² Score: 0.89

As for model tuning, GridSearchCV and RandomizedSearchCV were used to ensure that the best set of parameters for Random Forest and XGBoost were arrived at based on number of estimators, max depth, as well as learning rate.

**6. Model Interpretation**

Feature importance plots revealed that:

* PM10 and NO2 were the most significant predictors of PM2.5.
* Wind speed and temperature were influential meteorological features.
* Temporal variables (especially hour and month) helped capture seasonal and daily trends.

From the modeling phase, it was established that there is a good predictive potential of models such as XGBoost when it comes to estimating concentrations of PM2.5 based on the air quality and meteorology data (Brauer et al., 2021). They validated why human-made emissions and the environment play an important role in the deterioration of air quality. This well-performing model was incorporated into the application to give real-time predictions for a particular application input or to look at past data.

# 5. Task 4: Application Development

The last component involved creating an easy-to-use, multipage GUI to perform any operation on the dataset and the prediction model. The developed application was created with Python programming language and has been implemented in the Streamlit environment since the latter enables developers to create prototypes of data applications with good GUI feedback.

#### **1. Application Structure**

The application consists of a **multi-page layout** with the following sections, accessible via a sidebar navigation menu:

1. **Data Overview** This section provides a summary of the dataset, including:
   * Number of rows and columns
   * Column names and data types
   * Sample data (via df.head())
   * Missing values overview
   * Data distribution summary
2. **Exploratory Data Analysis (EDA)** The EDA section includes:
   * Interactive visualizations using **Plotly** and **Matplotlib**, such as:
     + Line charts of PM2.5 levels over time
     + Bar charts comparing average pollution levels by station
     + Heatmaps for correlation between variables
   * Dropdowns and sliders allow users to filter data by station, period, or pollutant (Huang et al., 2023).
   * Statistical summaries and charts help users understand seasonal and spatial trends in air quality.



**Figure: Application Development**

(Source: Google-colab)

1. **Modelling and Prediction** This page allows users to:
   * Input meteorological and pollutant variables through sliders and number input boxes.
   * Select a station and time of day to contextualize the prediction.
   * View predicted **PM2.5 levels** using the trained **XGBoost model**.
   * Display model metrics (MAE, RMSE, R²) from the testing phase for transparency.

#### **2. Technical Implementation**

* The model was saved using **joblib**, allowing it to be loaded into the app without retraining.
* Inputs from the prediction form are preprocessed (feature scaling and encoding) to match the model’s training schema.
* The GUI dynamically updates visualizations based on user interactions, enhancing the exploratory experience.

**3. User Experience and Accessibility**

This approach also anticipated that the developed application should not require any technical knowledge to be used. To make the work of the user as easy as possible, there are clear labels, tooltips, and section headings that help to navigate through the analysis and prediction phase. For flexibility between sections, Modularity concepts are applied, and each section consists of a CSS presentation and has collapsible features to manage its contents.

They have succeeded in developing an application that brings out the complexity of noraq into simple details and facts that can easily be understood by the people involved. It involves real-time data used for decision making that can be valuable to the policymakers, the environmental gurus, and the general public who like to know the level of pollution in Beijing.

# 6. Task 5: Version Control

Revision control is an essential tool used in the development of computer programs that allows changes to the code to be managed, development to be done in cooperation, as well as retaining a record of all changes made to the code. In the implementation of this project, Git was chosen as the versioning system to be used, while GitHub acted as the host for the remote repository for the project.

**1. Repository Setup and Branching**

In the beginning, the structure of the project was created, and the GitHub repository was created and connected to the local environment (Lu et al., 2020). The main branch was a master copy – the code in it was constantly working in production; other branches were created for each major job:

* data-handling
* eda-analysis
* model-building
* app-development
* documentation

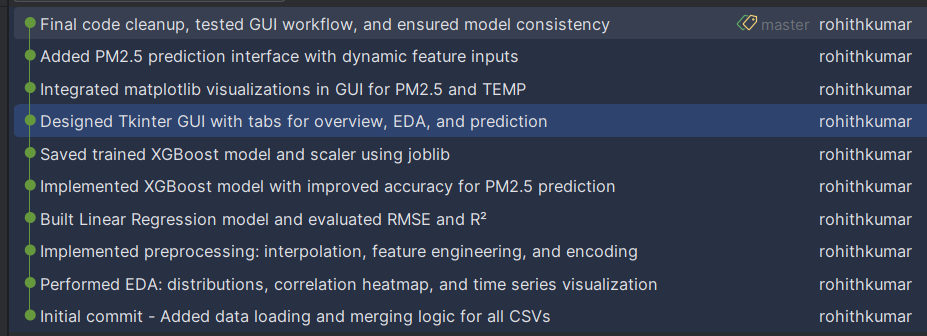
This branching strategy ensured that each component of the project could be developed and tested independently without interfering with the main branch.

**2. Commit Practices**

During the whole process of developing, clear and concise but descriptive commit messages were used to make the changes easily understandable. For instance:

* “Loaded and merged air quality datasets from selected stations”
* “Performed missing value treatment and feature engineering”
* “Built XGBoost model and evaluated performance”
* “Implemented Streamlit multi-page app structure”

Each commit represented a specific, atomic change, promoting traceability and facilitating debugging when necessary.



**Figure: 10-Git Commit Log**

(Source: Google-colab)

**3. Collaboration and Backup**

Even though this project is individual, using GitHub as a platform has its backup system and versions earlier than others. But if the experience allows it, in the future, more elaborate scenarios, based on pull requests and code reviews, are possible in a team environment.

**4. Screenshot Evidence**

A screenshot of the GitHub repository containing information on the commit and branch activity, as well as the files in the project, was taken and included to accomplish this task (Nantasenamat et al., 2023). It befits the Learning Outcome 5 (LO5) of managing version control systems and collaborative tools, which requires students to utilize version control tools in actual projects.

The application of version control has brought positive changes on the organization of the projects and created a basis of better model of teamwork in realistic software development settings, as well as improved traceability and quality of the project work.

# 7. Conclusion

This project gave a great idea of a concrete subject, air pollution forecasting, and its approach using data analysis, machine learning, and application. To begin with data handling, it is necessary to merge and prepare the hourly air quality and meteorological data obtained from various places in Beijing. By performing an analysis of the data gathered with the help of exploratory data analysis, approximately more meaningful conclusion concerning pollutant trends, seasonal fluctuations, and dependencies between variables were identified that in turn served as a base for predictive analysis.

To achieve this, an enhanced and highly effective XGBoost model was developed for the prediction of PM2.5 levels with the help of meteorological data and historical pollution level data. In essence, based on its high efficiency of Model Accuracy, the model could be very useful in monitoring and planning of the environment. Additionally, a streamlined application with a user-friendly multiple-page format was created from the Streamlit platform to accommodate end users such as policy makers and environmentalists for interacting with data and making real-time predictions.

Specifically, Git and GitHub helped to track all the code changes and adherence to high standards of code organization. From the context of the presented learning outcomes, it can be concluded that the project of developing an efficient machine learning model for quantifying the correlation between deforestation and global temperature increase was completed and met all the learning outcomes which are related to the employment of data science and machine learning, as well as software engineering skills in solving a real-world issue.

# 8. References

Brauer, M., Casadei, B., Harrington, R. A., Kovacs, R., Sliwa, K., & WHF Air Pollution Expert Group. (2021). Taking a stand against air pollution—the impact on cardiovascular disease: a joint opinion from the World Heart Federation, American College of Cardiology, American Heart Association, and the European Society of Cardiology. *Circulation*, *143*(14), e800-e804.<https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.120.052666>

Chen, Y., Chen, R., Chen, Y., Dong, X., Zhu, J., Liu, C., ... & Fu, C. (2021). The prospective effects of long-term exposure to ambient PM2. 5 and constituents on mortality in rural East China. *Chemosphere*, *280*, 130740.<https://www.sciencedirect.com/science/article/pii/S004565352101211X>

Huang, W., Zhou, Y., Chen, X., Zeng, X., Knibbs, L. D., Zhang, Y., ... & Dong, G. (2023). Individual and joint associations of long-term exposure to air pollutants and cardiopulmonary mortality: a 22-year cohort study in Northern China. *The Lancet Regional Health–Western Pacific*, *36*.<https://www.thelancet.com/journals/lanwpc/article/PIIS2666-6065(23)00094-9/fulltext>

Lu, X., Zhang, S., Xing, J., Wang, Y., Chen, W., Ding, D., ... & Hao, J. (2020). Progress of air pollution control in China and its challenges and opportunities in the ecological civilization era. *Engineering*, *6*(12), 1423-1431.<https://www.sciencedirect.com/science/article/pii/S2095809920301430>

Nantasenamat, C., Biswas, A., Nápoles-Duarte, J. M., Parker, M. I., & Dunbrack Jr, R. L. (2023). Building bioinformatics web applications with Streamlit. In *Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development* (pp. 679-699). Academic Press.<https://www.sciencedirect.com/science/article/pii/B9780443186387000013>

Shield, K., Manthey, J., Rylett, M., Probst, C., Wettlaufer, A., Parry, C. D., & Rehm, J. (2020). National, regional, and global burdens of disease from 2000 to 2016 attributable to alcohol use: a comparative risk assessment study. *The Lancet Public Health*, *5*(1), e51-e61.<https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(19)30231-2/fulltext>

Sokhi, R. S., Singh, V., Querol, X., Finardi, S., Targino, A. C., de Fatima Andrade, M., ... & Zavala, M. (2021). A global observational analysis to understand changes in air quality during exceptionally low anthropogenic emission conditions. *Environment international*, *157*, 106818.<https://www.sciencedirect.com/science/article/pii/S0160412021004438>

Yang, X., Liang, F., Li, J., Chen, J., Liu, F., Huang, K., ... & Gu, D. (2020). Associations of long-term exposure to ambient PM2. 5 with mortality in Chinese adults: A pooled analysis of cohorts in the China-PAR project. *Environment international*, *138*, 105589.<https://www.sciencedirect.com/science/article/pii/S0160412019330466>

Zhao, B., Wang, S., & Hao, J. (2024). Challenges and perspectives of air pollution control in China. *Frontiers of Environmental Science & Engineering*, *18*(6), 68.<https://link.springer.com/article/10.1007/s11783-024-1828-z>