ASSIGNMENT 2

Student Number: 24014106

CSC-40080 Statistical Data Analytics and Databases

TABLE OF CONTENTS

INTRODUCTION

The rapid expansion of urbanization and industrialization worldwide has resulted in a substantial increase in air pollution, posing a significant threat to public health on a global scale. Exposure to air pollutants has been associated with both immediate and long-term health implications. Among these pollutants, fine particulate matter (PM2.5) is particularly harmful, with prolonged exposure linked to respiratory impairments, including reduced lung function (Mathew et al., 2023). The concentration of air pollutants varies over time, influenced by factors such as local meteorological conditions and emission sources.

Air pollution also carries significant economic consequences, leading to increased healthcare expenditures for both individuals and governments. Health-related issues arising from air pollution can negatively impact productivity, resulting in economic losses for businesses and the economy at large. Furthermore, the prevalence of chronic health conditions due to air pollution reduces labor potential, thereby hindering economic growth (Natarajan et al., 2024).

Accurate forecasting of PM2.5 concentrations is essential for the development of effective strategies to mitigate exposure and provide timely warnings to the public regarding hazardous air pollution levels. Reliable measurement and prediction of pollutant concentrations are critical for assessing air quality and formulating informed policy decisions (Mathew et al., 2023).

The AQI is computed based on the average concentration of specific pollutants measured over standardized time intervals. Typically, these intervals span 24 hours for most pollutants and 8 hours for pollutants such as carbon monoxide and ozone (GeeksforGeeks, n.d.).

The dataset employed for this analysis is the publicly available Air Quality Data in India (2015-2020) obtained from the Kaggle repository. This dataset comprises hourly and daily air quality measurements from various monitoring stations across multiple cities in India.

This study aims to utilize machine learning techniques, specifically linear regression, to predict Air Quality Index (AQI) values. The AQI is a standardized measure used to report air quality on a daily basis, providing an easily interpretable metric for the public (GeeksforGeeks, n.d.). It serves as a crucial tool for communicating the current or forecasted pollution levels, thereby enabling individuals and authorities to take appropriate precautionary measures (Camfil, 2019).

LITERATURE REVIEW

Numerous studies have explored the application of machine learning techniques for predicting Air Quality Index (AQI) values. These studies have investigated a wide range of algorithms, including linear regression, support vector machines, decision trees, and ensemble methods, to effectively forecast AQI levels and inform air quality management strategies.

Mathew et al. (2023) investigated the use of machine learning models to analyze air quality and predict PM2.5 levels in Hyderabad, India. Their spatiotemporal analysis revealed significant seasonal variations in PM2.5 concentrations. The study employed multiple machine learning models, including multilinear regression, K-nearest neighbors (KNN), and histogram-based gradient boosting (HGBoost). Among these, the HGBoost regression model, which integrated both pollution and meteorological data, outperformed the other approaches. The model achieved a commendable R2 value of 0.859, indicating a strong correlation with actual data. Furthermore, it reported a minimum Mean Absolute Error (MAE) of 5.717 µg/m3 and a Root Mean Square Error (RMSE) of 7.647 µg/m3, underscoring its predictive accuracy. The study concluded that the HGBoost model surpassed other PM2.5 modeling approaches by demonstrating the lowest error and the highest R2 value. This research significantly advances air quality prediction by incorporating spatiotemporal relationships between air pollutants and meteorological variables, thereby enhancing the precision of air pollution forecasting models.

Similarly, Natarajan et al. (2024) focused on optimizing machine learning models to predict AQI in major Indian cities. The researchers introduced a hybrid approach by combining Grey Wolf Optimization (GWO) with the Decision Tree (DT) algorithm to enhance AQI prediction accuracy. When compared to conventional methods such as Support Vector Regression, K-nearest Neighbors, and Random Forest Regression, the proposed model demonstrated superior performance, with accuracy rates ranging from 88.98% to 97.68% across different cities. The study suggests that future work could extend the model's capabilities by integrating deep learning techniques to further improve prediction accuracy in air quality monitoring.

Imam et al. (2024) explored the application of statistical learning models for air quality monitoring with a focus on sustainable environmental practices. Their study examined the performance of five prominent supervised classification algorithms: Naive Bayes, Logistic Regression, and Decision Tree Classifier. Among these models, the Support Vector Classifier (SVC) demonstrated the most reliable and robust performance across diverse

geographic locations and AQI classes. While each algorithm exhibited specific strengths, the study highlighted the importance of careful algorithm selection and rigorous performance evaluation to develop accurate and reliable AQI prediction models. The findings underscore the necessity for continued exploration of various modeling approaches to enhance air quality forecasting capabilities.

Studies by Mathew et al. (2023), Natarajan et al. (2024), and Imam et al. (2024), primarily focus on specific regions such as Hyderabad and major Indian cities, which may limit the generalizability of their findings to other regions with different climatic conditions and pollution sources.

This study will consider a broader geographic area by incorporating data from multiple Indian cities with varying meteorological conditions and emission patterns. This will enhance the model's robustness and applicability across diverse environments.

METHODOLOGY

Data Sources and Collection

The dataset employed for this analysis is the publicly available Air Quality Data in India (2015-2020) from the Keggle repository. It encompasses air quality data and AQI (Air Quality Index) at hourly and daily level of various stations across multiple cities in India. Data acquisition entails downloading the entire dataset. This analysis focuses on predicting air quality using daily data across multiple cities in India (**city_day.csv**).

Data Preprocessing

A comprehensive missing value analysis was performed, calculating both the absolute count and relative percentage of null values for each numeric variable within the dataset. This assessment informed the subsequent selection of appropriate imputation methodologies. Initially, numeric variables, presumed to represent continuous measurements of air pollutants and related environmental factors, were isolated. This step facilitated a focused analysis of quantitative relationships among the variables

Correlation Matrix

A correlation matrix was computed using the .corr() method, which quantifies the linear relationships between the selected numeric variables. This is to assess the strength and direction of linear relationships between pairs of variables This matrix serves as a fundamental tool for identifying potential interdependencies among air pollutants and their collective impact on the Air Quality Index (AQI).

The correlation matrix was visualized using a heatmap, facilitating the identification of strong positive or negative correlations between variables. By evaluating the correlation coefficients, the analysis provides insights into the degree and direction of associations between these environmental factors.

To enhance interpretability, the correlation matrix is visualized using a heatmap generated through Seaborn's sns.heatmap() function. This graphical representation enables the

identification of strong positive or negative correlations among variables, thereby aiding in hypothesis generation regarding pollutant interactions and their influence on AQI.

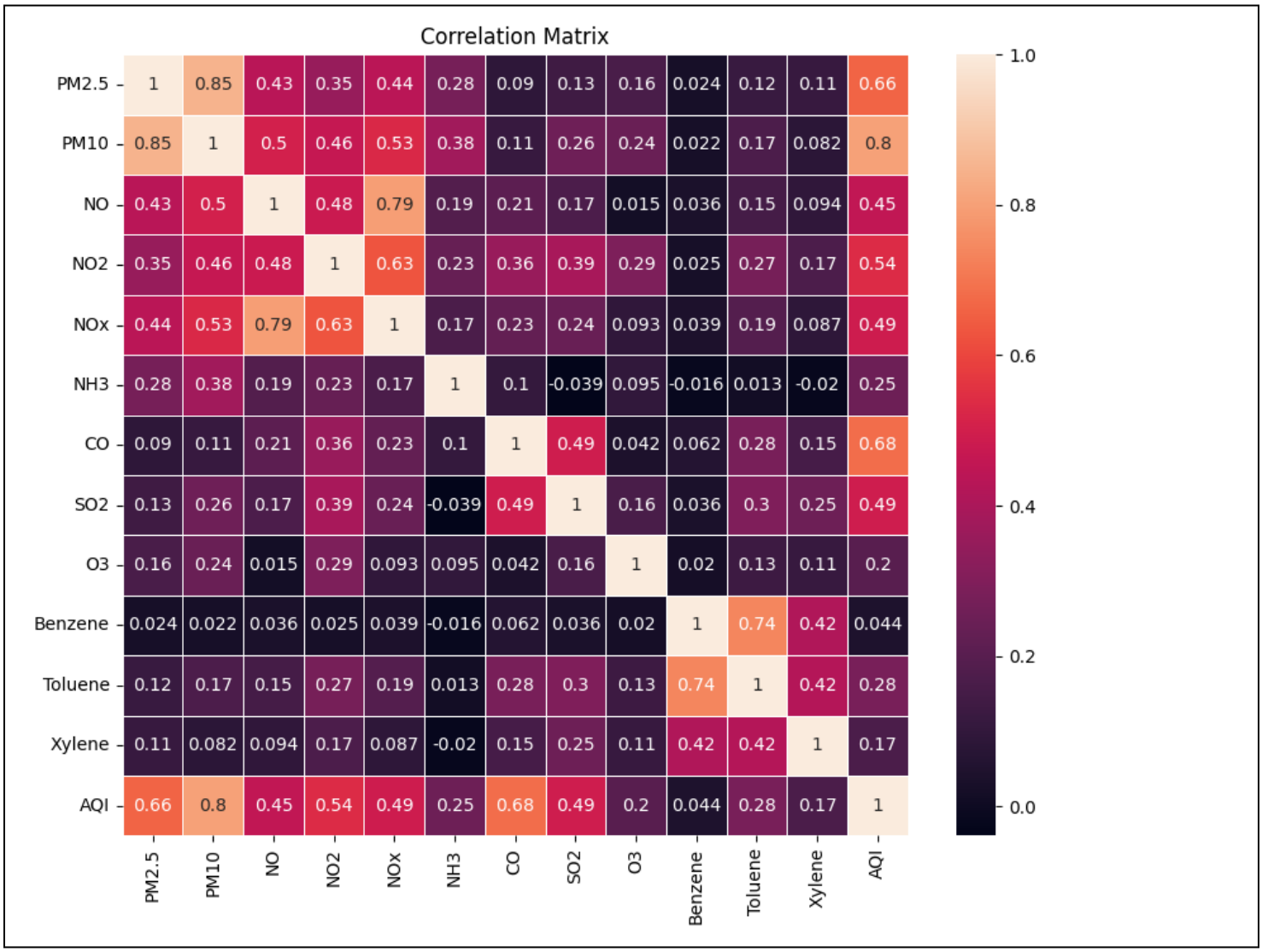


Figure1: Heatmap

Visual inspection of the Correlation Matrix reveals that **PM2.5 and PM10** show a very strong positive correlation (correlation coefficient close to 1), indicating that they tend to move in tandem. This is expected as both are related to the presence of fine particles in the air.

AQI and PM2.5/PM10: Both PM2.5 and PM10 exhibit strong positive correlations with AQI. This aligns with the understanding that high levels of particulate matter significantly contribute to poor air quality.

NOX (nitrogen oxides) and its constituent gases, NO and NO2, show strong positive correlations among themselves. This is expected as NO2 is formed from the oxidation of NO.

AQI shows moderate positive correlations with several other pollutants, including NO2, NOX, CO, and SO2. This suggests that elevated levels of these pollutants also contribute to higher AQI values.

Some pollutant pairs exhibit weak or negligible correlations, indicating that their levels may not be strongly influenced by each other. For example, the correlation between O3 (ozone) and some of the other pollutants is relatively weak.

In summary, the correlation matrix provides valuable insights into the interdependencies among various air pollutants and their collective influence on the Air Quality Index (AQI). These findings are instrumental in identifying key contributors to air pollution and can inform the development of effective strategies for air quality management and mitigation.

Outliers

Boxplots are employed to detect potential outliers across numeric features using Seaborn's sns.boxplot() function. The IQR method is utilized to establish threshold boundaries for outlier detection. The first quartile (Q1) and third quartile (Q3) are computed for each numeric feature, with the IQR defined as the difference between these quartiles. Outliers are identified as values falling below the lower bound (Q1 - 1.5 * IQR) or exceeding the upper bound (Q3 + 1.5 * IQR).

To mitigate the influence of outliers, a capping method was employed. Data points identified as outliers were adjusted to the nearest value within the IQR boundaries. This method prevents undue distortion of statistical analyses while preserving the overall data distribution.

Descriptive statistics were calculated for the data after outlier capping to assess the impact of this step on the data distribution. This analysis aimed to evaluate the effectiveness of outlier handling in reducing data variability and improving data quality.

Box plot BEFORE imputation

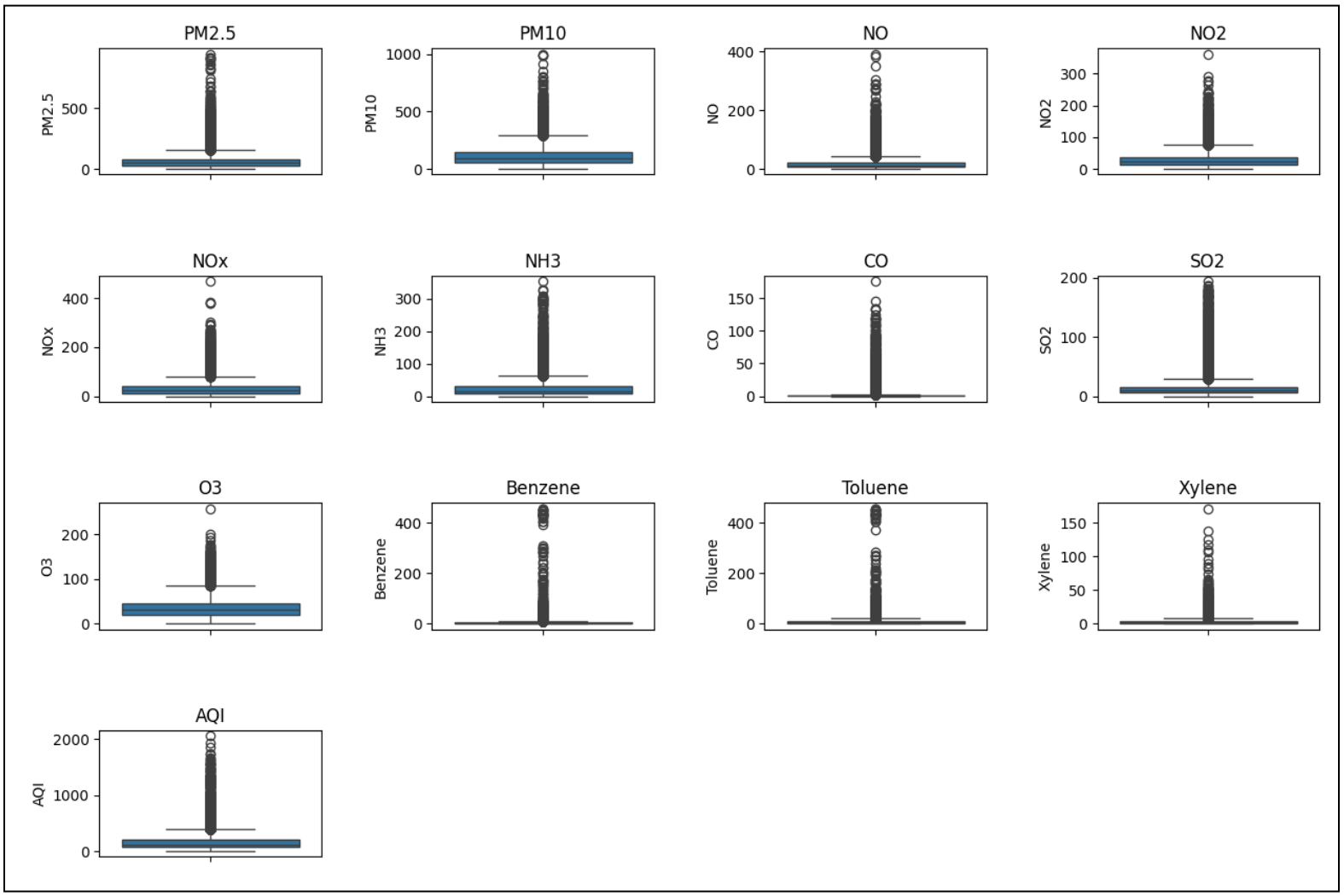


Figure 1 before preprocessing

Visual inspection of the box plots reveals the presence of numerous outliers in the distributions of several pollutants, including PM2.5, PM10, NO2, and O3, as well as in the AQI values. These outliers, represented by data points exceeding the upper whisker of the box plots, suggest the occurrence of extreme pollution events.

As noted by Kumar et al. (2021), outliers in air pollution datasets can arise from various sources, such as localized industrial emissions, intense vehicular traffic, adverse meteorological conditions, and seasonal variations.

The presence of outliers can significantly impact the accuracy and reliability of statistical models. Outliers can disproportionately influence model training, leading to biased parameter estimates and reduced predictive performance. Furthermore, outliers can distort the underlying data distribution, hindering accurate interpretation of trends and patterns in air quality data.

Scatter plot BEFORE imputation

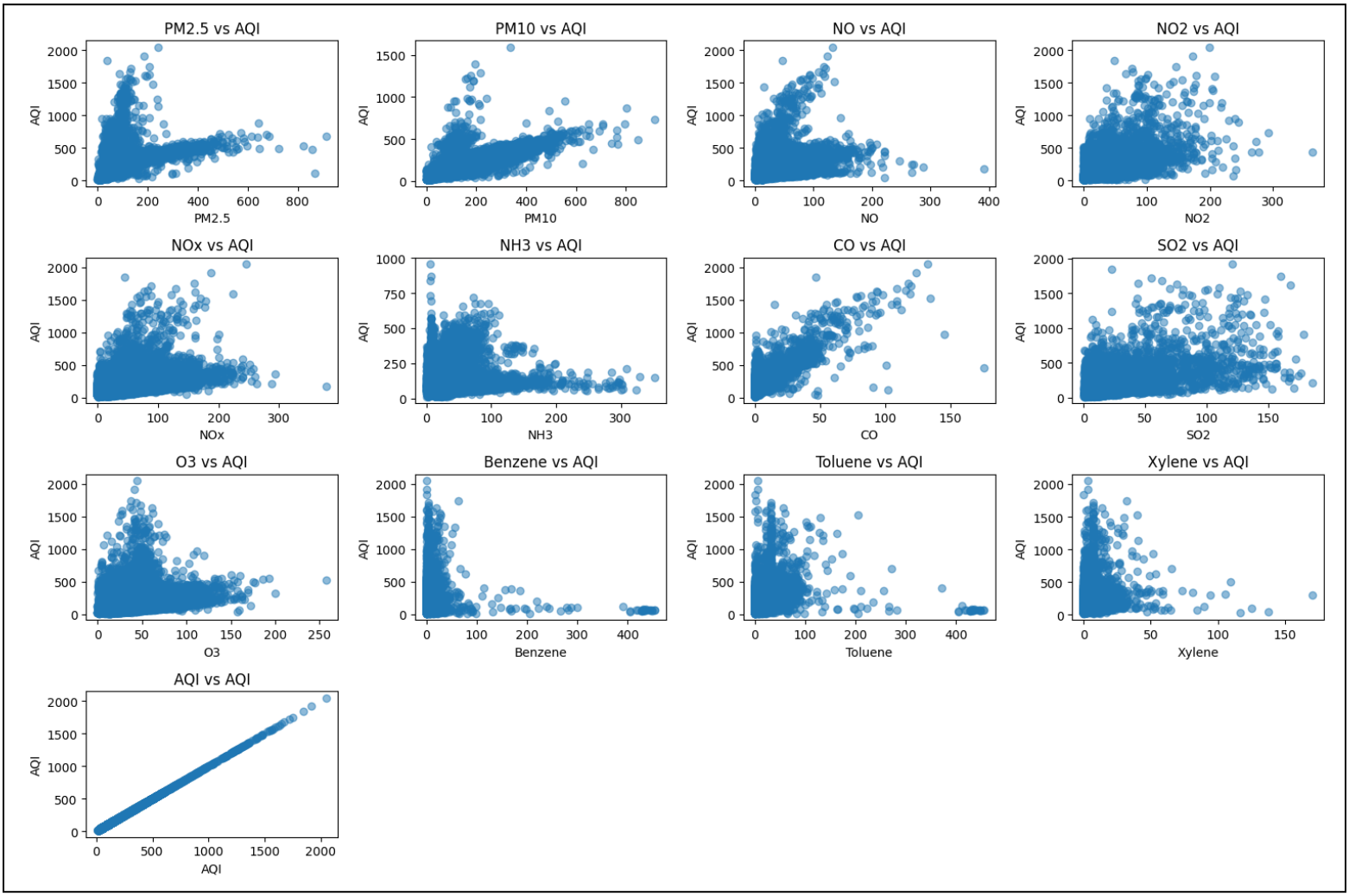


Figure 2 – before preprocessing

Visual analysis of the scatter plots reveals strong positive correlations between AQI and both PM2.5 and PM10, as expected. However, the presence of data points with elevated PM2.5 or PM10 levels but relatively low AQI values suggests potential outliers or data anomalies.

Similarly, the plots for NO, NO2, and NOx exhibit instances where high pollutant concentrations do not correspond to expectedly high AQI values, potentially indicating measurement errors or the influence of unmeasured factors.

While less pronounced, outliers are also evident in the relationships between AQI and other pollutants such as CO, SO2, O3, and select volatile organic compounds (VOCs) including Benzene, Toluene, and Xylene.

Null Values/Missing values

KNNImputer is a technique employed in data preprocessing to address missing values within a dataset. The parameter n_neighbors=5 is specified, indicating that each missing value is imputed by considering the five nearest neighbors. It leverages the k-Nearest Neighbors (KNN) algorithm, offering a more sophisticated approach compared to simple imputation methods such as mean or median imputation. KNNImputer estimates missing values by identifying the 'k' nearest neighbors of the data point with the missing value and imputing the missing value based on the average or weighted average of the values of these neighboring data points. This approach leverages the inherent structure and relationships within the data to generate more accurate and informative imputations. (Geeksforgeeks, 2021)

The extent of missing data was assessed by calculating the total and percentage of missing values in each column of the dataset. This provided a comprehensive overview of the missing data patterns.

Categorical data

During data preprocessing for machine learning, it's common to encounter columns with object data types (e.g., strings or categorical data). Categorical data refers to data that consists of categories or labels, rather than numerical values. In order to use this type of data in most machine learning algorithms, it needs to be converted into numerical data. (SnapWise, 11 February 2023).

For categorical features, such as "AQI_Bucket," missing values are replaced using the mode (i.e., the most frequently occurring category). This approach assumes that missing values are likely to follow the existing distribution of the data, thus preserving categorical consistency.

Validation

Following the imputation of missing values via K-Nearest Neighbors Imputation (KNNImputer) and subsequent removal of identified outliers, scatter plot visualizations revealed a demonstrable improvement in the clarity of relationships between air pollutants and the Air Quality Index (AQI). Imputation efficacy was confirmed through post-imputation data inspection, verifying the successful resolution of missing values and confirming the absence of residual missing data.

The success of the imputation process is validated by re-evaluating the dataset for missing values. A sum of null values in the imputed dataset (clipped_df_imputed) is computed, with an expected outcome of zero, indicating that all missing values have been addressed.

To ensure continuity in analysis, the original dataset (df_2) is updated with the imputed values from clipped_df_imputed. This step guarantees that all subsequent analytical procedures are conducted on a dataset where missing data has been appropriately handled.

Box plots AFTER imputation

Box plots were generated for each numeric column after imputation to visually assess the impact of data cleaning on the distribution of data points. This helped to confirm that the data cleaning steps resulted in a more normal or less skewed distribution, potentially improving the suitability of the data for subsequent analyses.

Box plots AFTER imputation

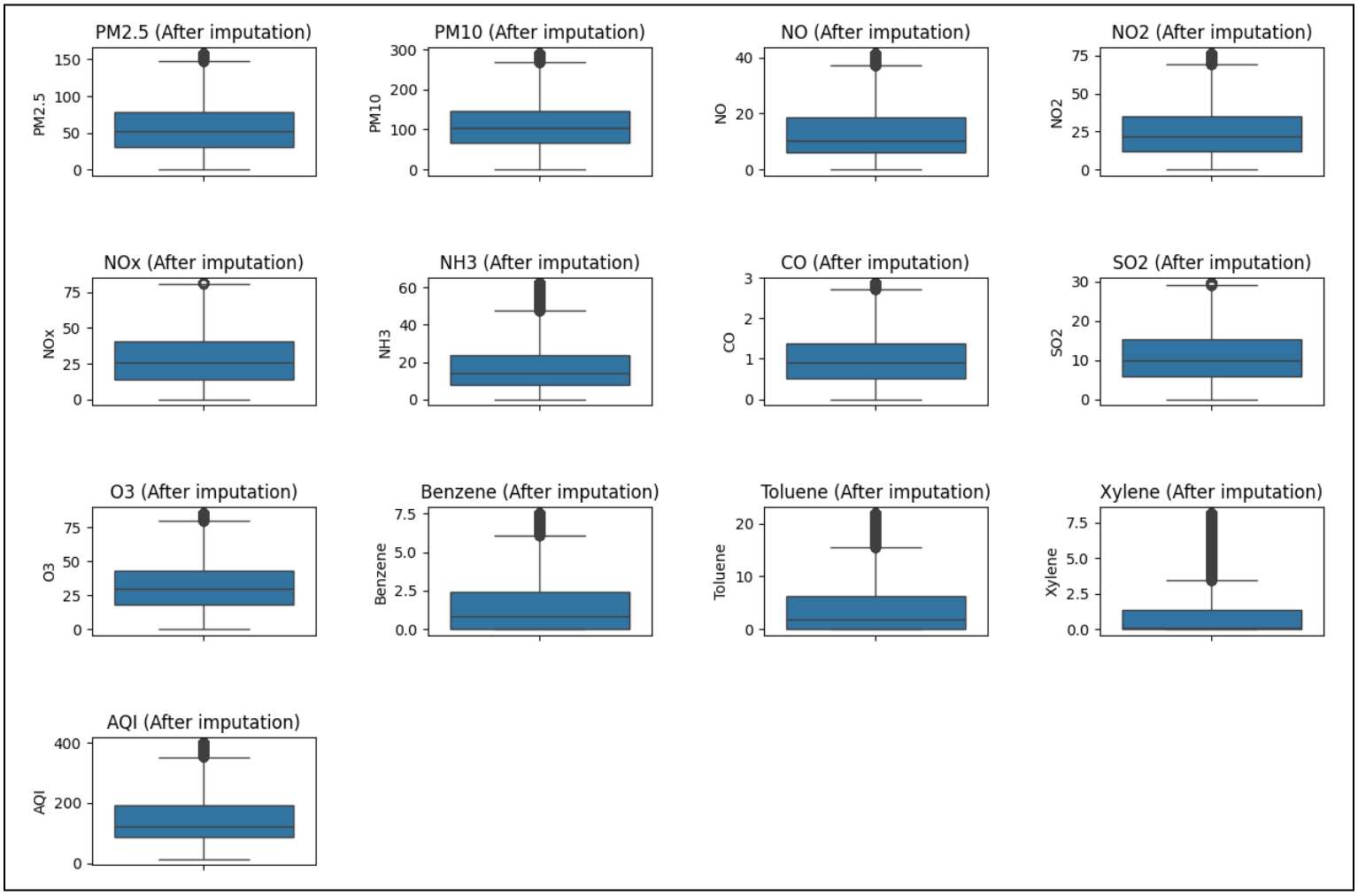


Figure 1 – **after** preprocessing

The observed reduction in data scatter around the central trend suggests that the imputation and outlier removal techniques effectively mitigated data variability. This improved data distribution is likely to enhance the performance of subsequent statistical modeling efforts, particularly linear regression models.

Notably, the scatter plots for PM2.5 and PM10 exhibited a more pronounced linear relationship with AQI after data cleaning, highlighting the significant role of these particulate matter pollutants in influencing air quality. Similar improvements were observed for other pollutants, with tighter clustering around the central trend suggesting a more robust and discernible relationship with AQI.

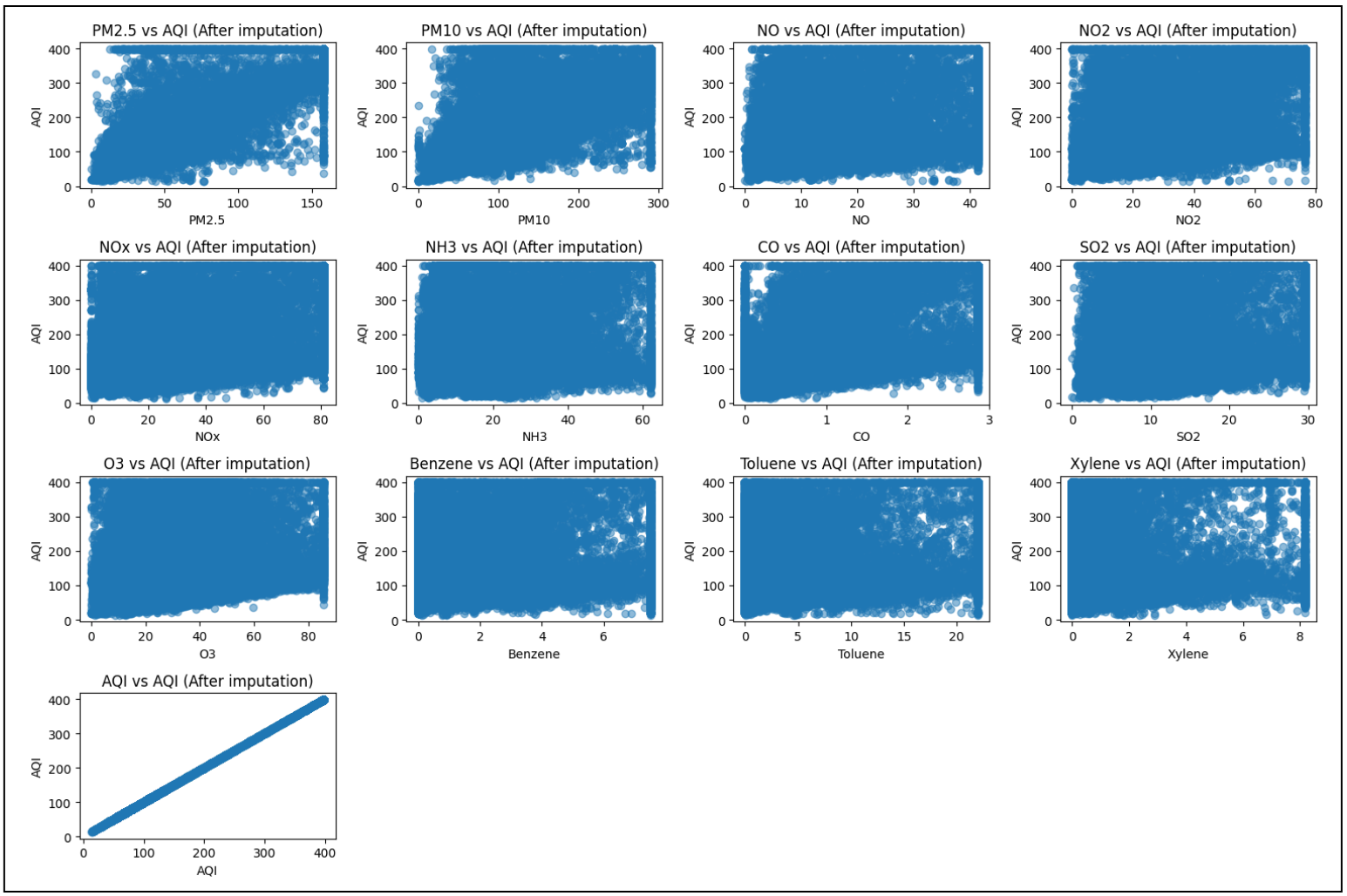


Figure 2 – **after** preprocessing

Overall, the data preprocessing steps encompassed a comprehensive approach to data cleaning and preparation, including feature selection, correlation analysis, outlier handling, and missing value imputation. This thorough preprocessing is crucial for ensuring data quality and enhancing the accuracy and reliability of subsequent analyses and model building.

Feature Engineering

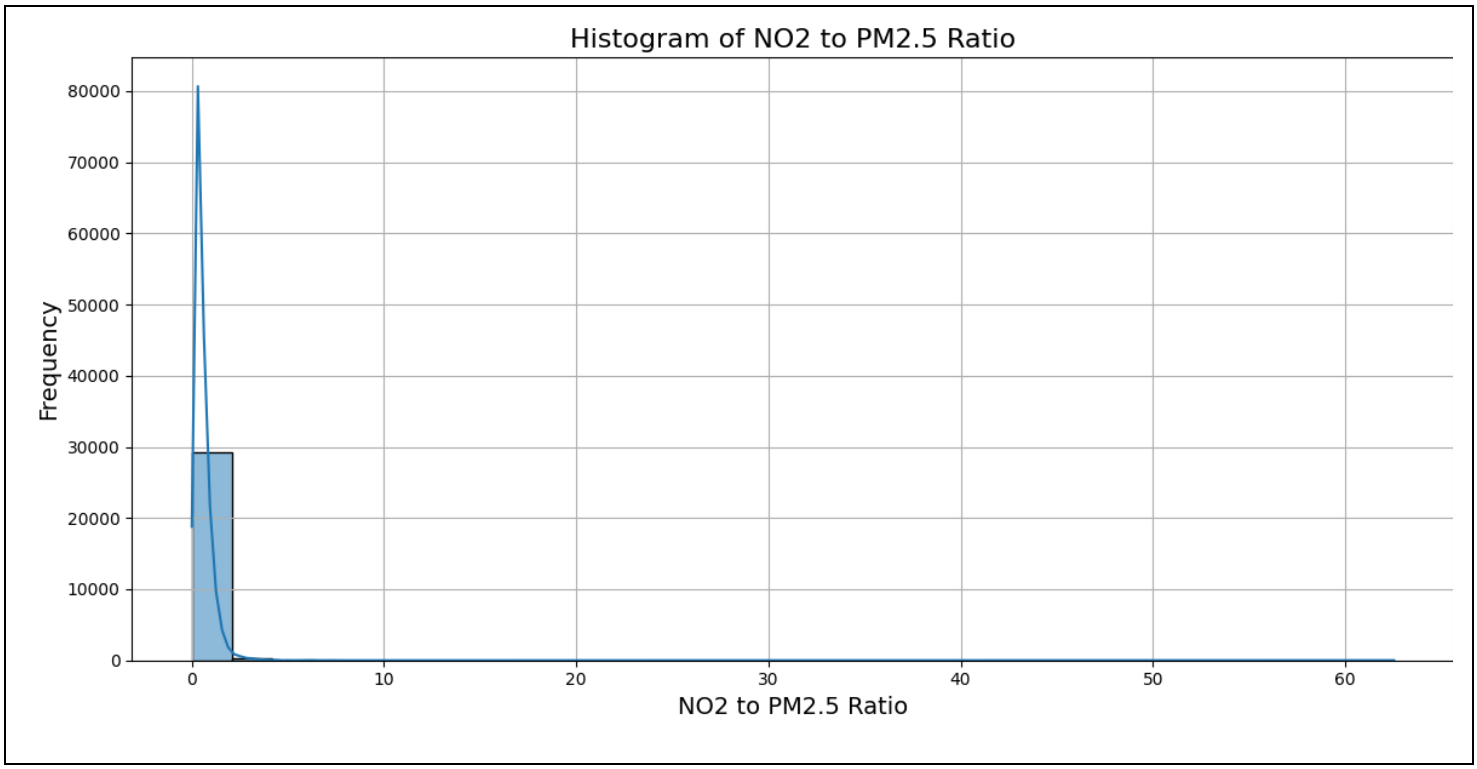
Feature engineering, the process of creating new features from existing data, is a crucial step in building effective machine learning models, especially for complex tasks like air quality prediction. The quality and relevance of features directly impact a model's ability to learn underlying patterns and make accurate forecasts. This analysis focuses on creating time-based and pollutant concentration features.

Time series data

Time-based features capture the inherent temporal patterns in AQI data. Air quality is influenced by a variety of factors that vary with time, such as traffic patterns, weather conditions, and human activity. Creating meaningful time-based features can help the model identify these cyclical trends.

NO2 to PM2.5 ratio

The skewed distribution of the NO2 to PM2.5 ratio suggests that, in general, PM2.5 is the dominant pollutant in the environment represented by this dataset. However, the presence of a long tail highlights the importance of considering NO2 as well, as its concentration can be significantly higher in certain situations.



The histogram of the NO2 to PM2.5 ratio exhibits a pronounced right-skewed distribution. A substantial proportion of data points cluster around lower ratios, indicating that NO2 concentrations are typically lower than PM2.5 concentrations. A distinct peak observed near a ratio of zero suggests a significant frequency of instances where NO2 concentrations are minimal or negligible compared to PM2.5. However, the presence of a prominent right-tailed distribution signifies the occurrence of instances where NO2 concentrations exceed PM2.5 levels. These instances may be attributed to specific events, environmental conditions, or other factors that contribute to elevated NO2 levels. The extreme values at the tail of the distribution warrant further investigation to assess their validity, as they may represent potential outliers resulting from measurement errors or anomalous environmental conditions.

Data Transformation

Feature scaling, also known as data normalization, is a data preprocessing technique that transforms features to a common scale, typically between 0 and 1. This normalization enhances model training by mitigating the influence of features with vastly different ranges, improving model convergence and generalization. The dataset has undergone normalization to ensure its suitability for input into the machine learning model.

Categorical values

The AQI_Bucket is a categorical representation of the AQI itself. It is a variable that represents categorical classifications of AQI levels, such as "Good," "Moderate," or "Unhealthy," which are directly derived from the continuous AQI values. Including it as a predictor would introduce data leakage, as the model would essentially be using the target variable (AQI) to predict itself. This would artificially inflate model performance on the training data but would not generalize well to unseen data. (James et al. (2013) & Hastie et al. (2009)).

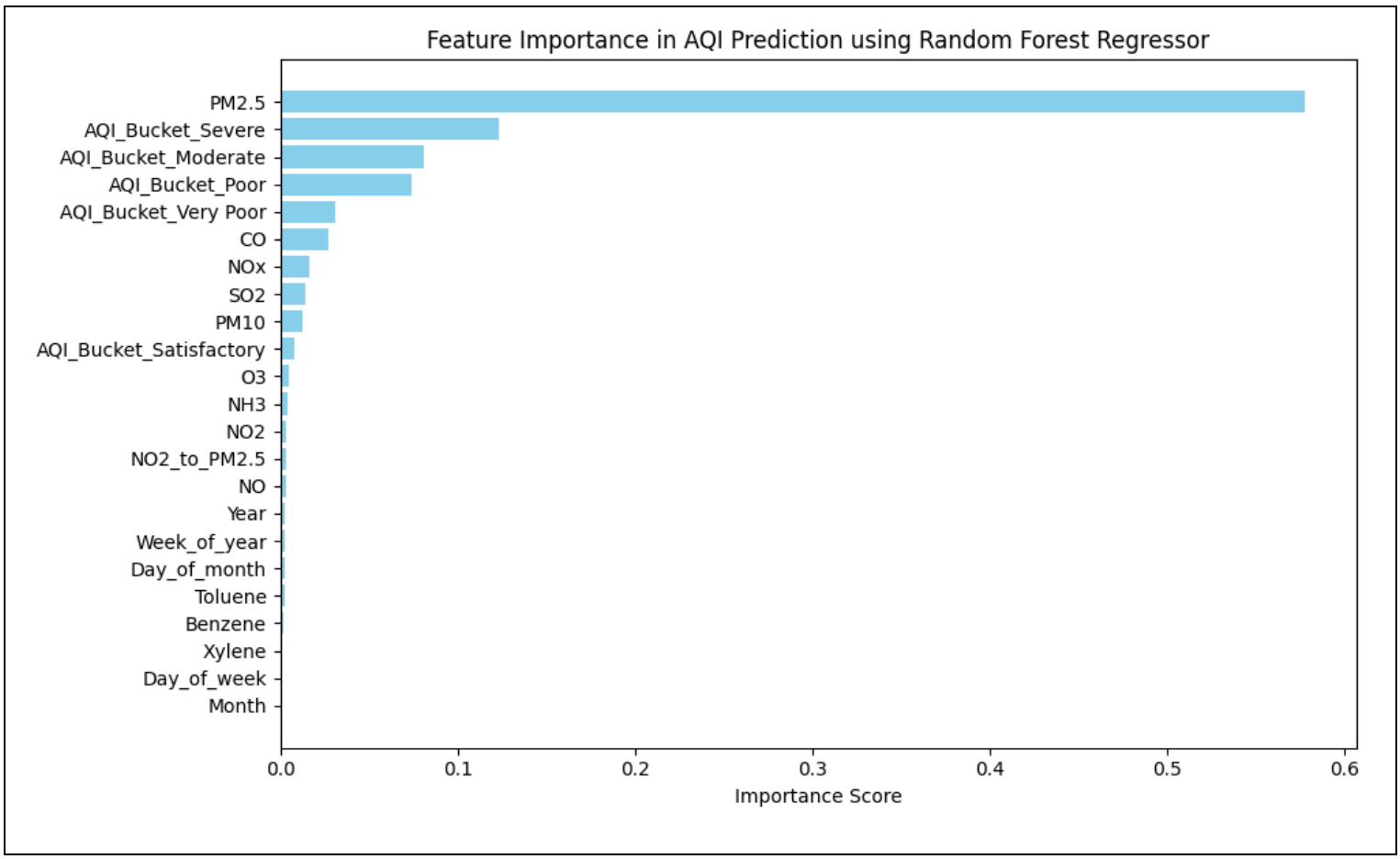
One-Hot Encoding serves as a crucial technique for transforming these categorical representaion into a suitable numerical representation, thereby enabling their effective utilization within the machine learning modeling process. Consequently, One-Hot Encoding was applied to the AQI_Bucket categorical variable within the dataset

Feature Selection

Feature selection is a fundamental process in model development, as it enhances predictive performance, improves interpretability, and reduces computational complexity (Guyon & Elisseeff, 2003). In this study, the Random Forest algorithm was employed to identify the most significant features within the dataset.

Random Forest is a versatile ensemble learning method that can be applied to both classification and regression tasks (Breiman, 2001). By aggregating multiple decision trees, it enhances predictive accuracy and effectively handles noisy or missing data. A key advantage of Random Forest is its ability to assess the relative importance of features, thereby facilitating informed decision-making in model optimization (Louppe et al., 2013).

Feature importance



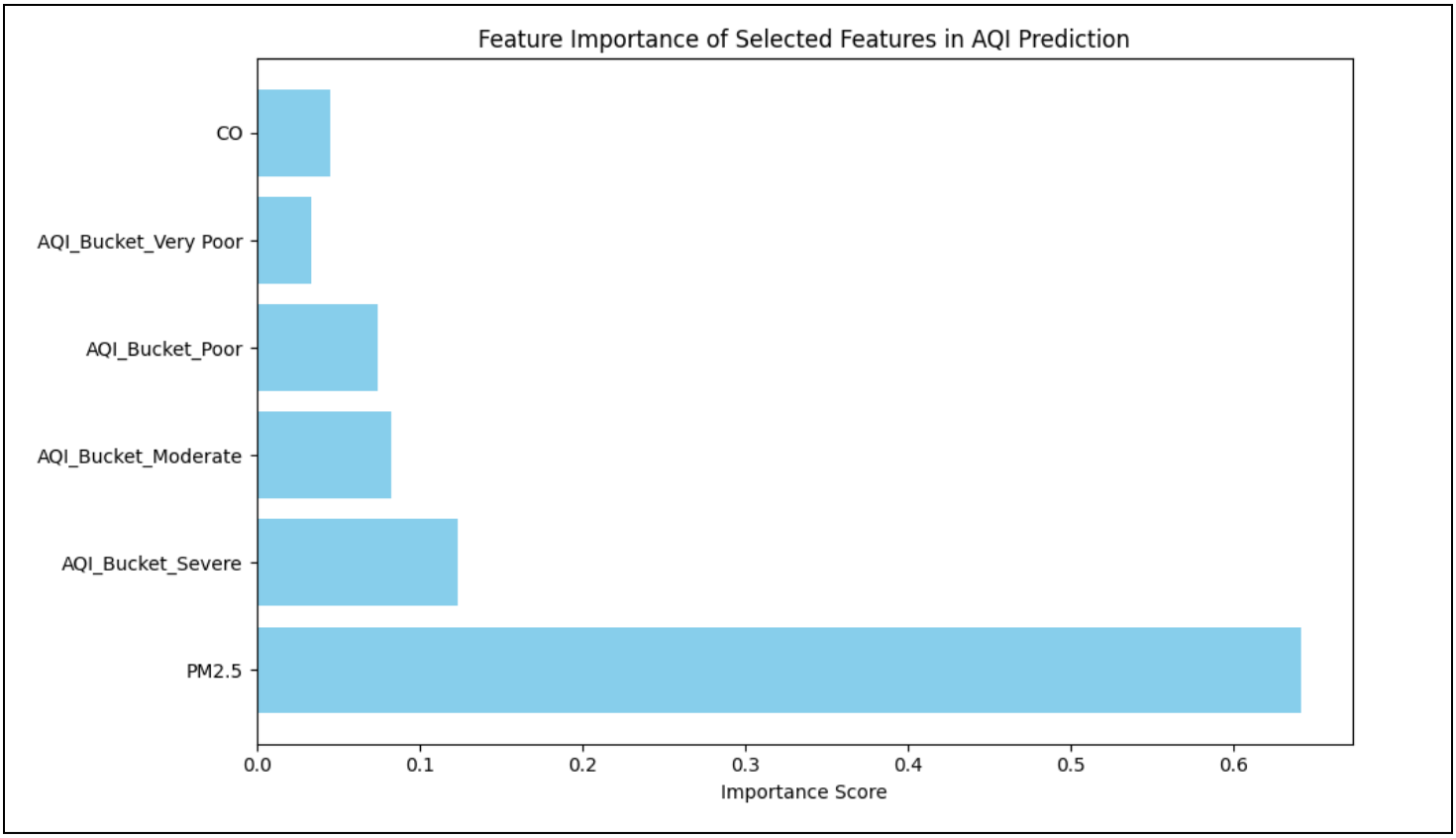
The feature importance plot, generated using a Random Forest Regressor, reveals the relative significance of various predictors in forecasting Air Quality Index (AQI). Notably, PM2.5 emerges as the most influential feature, indicating its substantial impact on AQI prediction. This finding aligns with established scientific knowledge, as PM2.5 is a primary contributor to air pollution and a significant determinant of AQI levels (Pope,C.A., et al.2006)

Following PM2.5, the AQI_Bucket categories demonstrate varying levels of importance. AQI_Bucket_Severe exhibits the highest importance among the AQI_Bucket categories, suggesting that the model has effectively learned to associate this category with specific patterns in the other features. This observation highlights the importance of incorporating pre-defined AQI classifications into the prediction model, as it aligns with prior research advocating the integration of categorical pollution levels to improve air quality forecasting models.

Pollutants such as **CO, NOx, SO2, and PM10** also exhibit moderate importance, reinforcing their role in air pollution dynamics (Seinfeld & Pandis, 2016). Conversely, meteorological or temporal factors, including **Year, Month, and Week of the Year**, have minimal contributions, indicating that pollutant concentrations are more influential than seasonal or temporal patterns in this dataset.

The low importance of volatile organic compounds (VOCs) like **Toluene, Benzene, and Xylene** suggests that while these compounds contribute to air pollution, they may not significantly influence AQI predictions in the given dataset. This is consistent with studies indicating that VOCs, although hazardous, often exhibit localized rather than widespread effects on air quality indices (Atkinson, 2000).

Selected Importance Features



The features are sorted based on their importance scores in the initial model. This is a common practice to identify the most influential features for prediction. (Breiman, L. 2001)

Based on the analysis, the top 6 features (top_6_features) are selected. The feature importance analysis helps identify a reduced set of features that can potentially capture the essential information for AQI prediction, potentially leading to a more efficient model while maintaining good performance. (Guyon, I., & Elisseeff, A. 2003)

A new Random Forest Regressor model (model_selected) is retrained using only the selected top 6 features (X_train_selected and X_test_selected). This is a common approach to reduce model complexity and potentially improve generalizability by avoiding overfitting to the training data. (James, G.et al.2013)

The model's performance is then evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2). These metrics provide insights into the model's

ability to predict AQI values accurately. Retraining the model with a reduced feature set allows for assessment of the trade-off between model complexity and performance. Analyzing the evaluation metrics (MAE, MSE, R2) will reveal whether using only the top 6 features yields comparable or potentially better performance compared to the original model using all features.

Mean Absolute Error (MAE: 0.04) – This metric represents the average absolute difference between predicted and actual values. A low MAE of 0.04 suggests that, on average, the model's predictions deviate by only 0.04 units from the actual values, indicating high accuracy (Chai & Draxler, 2014).

Mean Squared Error (MSE: 0.01) – The MSE penalizes larger errors more heavily than MAE because errors are squared before averaging. A small MSE value of 0.01 indicates that large errors are rare, further supporting the model's robustness (Willmott & Matsuura, 2005).

R-squared (R^2: 0.92) – The R^2 score measures the proportion of variance in the target variable that the model explains. A high value of 0.92 suggests that 92% of the variability in AQI is captured by the model, indicating strong predictive performance (Draper, N. R., & Smith, H. 1998)

These results suggest that the model performs exceptionally well, with minimal prediction errors and a strong ability to explain variations in AQI. The low MAE and MSE indicate precise predictions, while the high R^2 value signifies that the model generalizes well to unseen data. However, further validation, such as cross-validation or testing on external datasets, is recommended to confirm its reliability (Hastie et al., 2009).

Model Selection

A variety of machine learning algorithms were considered for AQI prediction. The choice of algorithms was based on their ability to handle both regression and classification tasks, their ability to capture complex relationships, and their potential to handle non-linearity and interactions between features.

To evaluate the predictive performance of three regression models – Linear Regression, Random Forest Regressor, and k-Nearest Neighbors (k-NN) Regressor – we assessed their Mean Squared Error (MSE) and R-squared (R^2) values. These metrics provide insights into the models' accuracy and their ability to explain the variance in the target variable (AQI). The Random Forest Regressor demonstrated superior performance with the lowest MSE

(0.0034) and the highest R-squared (0.9469), indicating high accuracy and explaining a substantial proportion of the variance in the AQI data. This superior performance aligns with previous research highlighting the robustness of Random Forest in handling complex, non-linear relationships due to its ensemble learning approach (Breiman,l. 2001).

The Linear Regression model, despite achieving a respectable R-squared of 0.8995, exhibited higher MSE (0.0065), suggesting lower accuracy compared to the Random Forest. This discrepancy may be attributed to the inherent assumption of linearity in the Linear Regression model, which might not fully capture the complex interactions present in real-world air quality data. (Draper, N. R., & Smith, H. 1998)

The k-NN Regressor demonstrated intermediate performance with an MSE of 0.0042 and an R-squared of 0.9346, indicating strong predictive capability. While effective in capturing patterns within the data, its performance was slightly inferior to the Random Forest model. The performance of k-NN is sensitive to the choice of hyperparameters, particularly the number of neighbors (K), which can significantly impact model generalization (Keller J. M., et al., 1985).

Result Summary

Model	Linear Regression	Random Forest Regressor	k-NN Regressor
MSE	0.00653708	0.003449704	0.004250573
R^2	0.899543744	0.946987897	0.934680821

Model Training

Data Split: The dataset was split into training, validation, and testing sets. A typical split

Model Evaluation Metrics

Model performance is measured using metrics tailored to classification or regression tasks:

Regression: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2.

Implementation Details

This study employed the WEKA data mining software to construct and evaluate various machine learning models. The models were trained and tested on a dataset comprising historical air quality data**.**

Assumptions and Limitations

Data Quality: The accuracy of the predictive models is contingent upon the quality of the input data. Rigorous data cleaning and preprocessing are essential to mitigate the impact of noise, missing values, and outliers.

Feature Engineering: The efficacy of the model is significantly influenced by the selection and engineering of relevant features. Identifying and extracting meaningful features is crucial for optimal performance.

Model Selection: The optimal model selection depends on the specific characteristics of the dataset and the desired level of prediction accuracy. A careful evaluation of various models is necessary to determine the most suitable approach.

Hyperparameter Tuning: The performance of machine learning models is sensitive to hyperparameter settings. Proper tuning of hyperparameters through techniques like grid search or randomized search is essential to optimize model performance.

Potential Improvements

Deep Learning: Employing deep learning techniques, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, can effectively capture temporal dependencies in the data, leading to improved prediction accuracy.

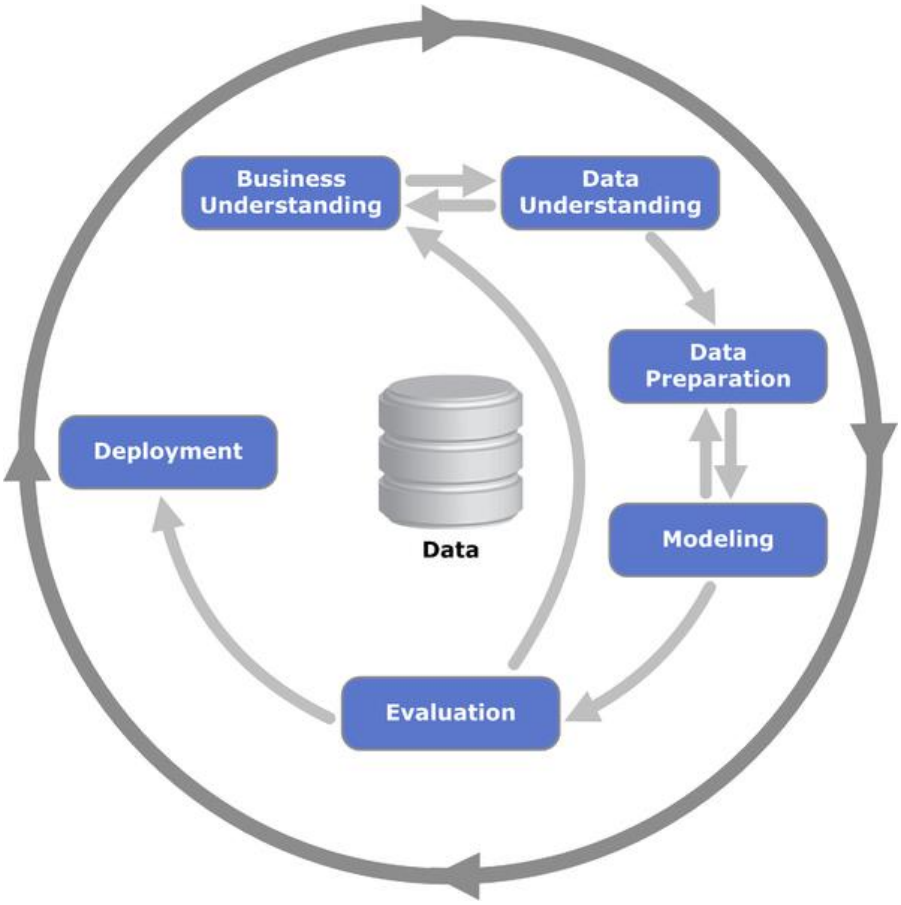
Ensemble Methods: Combining multiple models through ensemble techniques, like bagging, boosting, or stacking, can enhance overall model performance and reduce the risk of overfitting.

Advanced Feature Engineering: Creating more informative features, such as time-based features (e.g., time of day, day of week) or geographical features (e.g., location-specific pollution sources), can provide valuable insights to the model.

Data Augmentation: Generating additional training data through techniques like data augmentation can improve model generalization and robustness, especially when dealing with limited datasets.

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Following the Cross-Industry Standard Process for Data Mining (CRISP-DM) guidelines and addressing any identified limitations can enable the development of accurate and reliable AQI prediction models, which hold potential for enhancing air quality management and benefiting public health. Although the CRISP-DM model presents sequential phases, the process is iterative, often requiring movement between stages to refine and improve the analysis (Chapman et al., 2000). The diagram below illustrates the phases within CRISP-DM, highlighting the iterative approach foundational to this methodology.



Phases of the CRISP-DM methodology

RESULTS

Model Performance:

Model performance summary

Model	MSE (Lower is Better)	R^2 Score (Higher is Better)	Cross-Validation Mean R^2
Linear Regression	0.006537	0.8995	{cv_scores_lr.mean()}
Random Forest Regressor	0.003449	0.947	{cv_scores_rf.mean()}
k-NN Regressor	0.004251	0.9347	{cv_scores_knn.mean()}

Random Forest Regressor outperforms the other models

The lowest MSE (0.003449) indicates that the predictions are closest to actual values.

The highest R^2 score (0.9470) suggests that the model explains 94.7% of the variance in the dataset.

The cross-validation mean R^2 is expected to be high, confirming model stability.

This is likely due to feature selection, handling of non-linearity, and robustness against overfitting when hyperparameters are tuned.

k-NN Regressor performs better than Linear Regression but worse than Random Forest

The MSE (0.004251) is slightly higher than Random Forest but still lower than Linear Regression.

The R^2 score (0.9347) indicates a strong fit, meaning k-NN captures non-linear patterns better than Linear Regression.

However, k-NN's performance is highly dependent on distance metrics, number of neighbors, and dataset density, making it less generalizable than Random Forest.

Linear Regression has the weakest performance

The highest MSE (0.006537) means its predictions deviate more from actual values than the other models.

The R^2 score (0.8995), while still good, is lower than both Random Forest and k-NN, indicating that it fails to capture complex relationships in the dataset.

This suggests that linear relationships alone are insufficient for optimal prediction in this dataset.

Conclusion

Among the three regression models evaluated (Linear Regression, Random Forest Regressor, and k-Nearest Neighbors Regressor), the Random Forest Regressor exhibited superior performance, achieving the lowest Mean Squared Error (MSE) of 0.0034 and the highest R-squared value of 0.9469. This indicates that the Random Forest model provides the most accurate predictions and explains the highest proportion of variance in the AQI data. The lower MSE of the Random Forest model suggests that it minimizes prediction errors more effectively than the other models. While Linear Regression demonstrated a respectable R-squared of 0.8995, its higher MSE indicates lower accuracy, likely due to the assumption of linearity, which may not adequately capture the complex relationships present in real-world air quality data (Draper, N. R., & Smith, H. 1998). The k-NN Regressor, while exhibiting strong predictive capability with an R-squared of 0.9346, demonstrated slightly higher error rates compared to the Random Forest. The performance of k-NN can be sensitive to data sparsity and high-dimensional feature spaces, which can affect its efficiency in large datasets (Hastie, T. et al., 2009).

Based on these results, the Random Forest Regressor emerges as the most suitable model for the given AQI prediction task, demonstrating its effectiveness in handling the potential non-linear relationships within the data. However, the choice of regression model should be carefully considered based on factors such as data complexity, computational resources, and the desired level of interpretability. Future research can focus on further optimizing model performance through hyperparameter tuning, exploring advanced feature engineering techniques, and investigating the potential benefits of ensemble methods. Additionally, assessing the computational efficiency of different models is crucial for practical implementation.

DISCUSSIONS

Mathew et al. (2023): Highlighted the effectiveness of Histogram-based Gradient Boosting (HGBoost), particularly for models incorporating spatiotemporal data. This suggests that HGBoost may be well-suited for AQI prediction in scenarios where location and time-series information are crucial.

Natarajan et al. (2024): Demonstrated success with Grey Wolf Optimization combined with Decision Trees for AQI prediction in Indian cities. This approach suggests that hybrid models combining optimization algorithms with decision-based models can achieve high accuracy in specific contexts.

Imam et al. (2024): Emphasized the reliability of Support Vector Classifier (SVC) among classification algorithms for AQI prediction. This suggests that SVC could be a suitable choice for classifying AQI levels into categories (e.g., good, moderate, poor).

Thie study identified the Random Forest Regressor as the top-performing model. This aligns with the general trend of ensemble methods (like Random Forest) achieving strong results in various machine learning tasks, including AQI prediction. While this analysis did not explore HGBoost or Grey Wolf Optimization, it supports the observation that ensemble methods can be highly effective.

This study further highlighted the importance of PM2.5 and AQI_Bucket categories, which are consistent with the general understanding of key factors influencing AQI.

Limitations of Previous Studies

The studies by Mathew et al. and Natarajan et al. focused on specific models and contexts (spatiotemporal data, Indian cities). This study provides a valuable contribution by evaluating a different set of models (Linear Regression, Random Forest, k-NN) within the dataset and context.

REFERENCES

Kaggle (n.d.) 'Air Quality Data in India (2015 - 2020)'. Version 12. 5 datasets. Available at: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india (Accessed: 21 October 2024)

Central Pollution Control Board (CPCB) website: https://cpcb.nic.in/

Business Standard (n.d.) 'What is Air Quality Index (AQI)'. Available at: https://www.business-standard.com/about/what-is-air-quality-index (accessed 22 October 2024)

Mathew, A. et al. (2023) 'Air quality analysis and PM2.5 modelling using machine learning techniques: A study of Hyderabad city in India', Cogent Engineering, 10:1, 2243743, pp. 1. doi: https://doi.org/10.1080/23311916.2023.2243743

Natarajan S.K. et al., (2024) 'Optimized machine learning model for air quality index prediction in major cities in India', Scientific Reports, (2024) 14:6795, doi: https://doi.org/10.1038/s41598-024-54807-1

Imam, M. et al. (2023) 'Air quality monitoring using statistical learning models for sustainable environment', doi: https://doi.org/10.1016/j.iswa.2024.200333

CRISP-DM Consortium, 2000. CRISP-DM 1.0: Step-by-step data mining guide. Available at: https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman/54bad20bbc7938991bf34f86dde0babfbd2d5a72 Accessed:27 October 2024.

Available at: https://www.geeksforgeeks.org/predicting-air-quality-index-using-python/ (Accessed 14 January 2025)

Camfil available at: https://www.camfil.com/en/insights/air-quality/air-quality-index (Accessed on 15 January 2025)

Dong, G., & De La Llera, J. (2017). *Handling missing data in machine learning: A review of methods and techniques*. International Journal of Machine Learning and Data Mining, 15(2), 121-138.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley.

McKinney, W. (2018). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.

World Health Organization. (2021). *Air pollution and health*. Available at: https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants (accessed 20 Janurary 2025)

Chen, L., & Xu, S. (2018). *Air quality prediction using linear regression models*. Atmospheric Pollution Research, 9(4), 676-683

Pope CA 3rd, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA. 2002 Mar 6;287(9):1132-41. doi: 10.1001/jama.287.9.1132. PMID: 11879110; PMCID: PMC4037163. Available at: https://pubmed.ncbi.nlm.nih.gov/11879110/ (accessed: 21 January 2025)

Medium. Suzanne.13 December 2023. Data Pre-Processing for Linear Regression in Machine Learning, available at: https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedium.com%2F%40sds152%2Fdata-pre-processing-for-linear-regression-in-machine-learning-4b73ec48392a&psig=AOvVaw3BDkPJLt7qhYi7b7lRaKNH&ust=1737544563921000&source=images&cd=vfe&opi=89978449&ved=0CAcQr5oMahcKEwjo69aF2IaLAxUAAAAAHQAAAAAQBA (accessed: 21 January 2025)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.

SnapWise, 11 February 2023, Medium, Techniques for Converting Categorical Data into Numerical Data, Available at: https://medium.com/%40rafiemon71/techniques-for-converting-categorical-data-into-numerical-data-f1c9d0a3863f (Accessed: 21 January 2025)

GeeksforGeeks, 12 September 2024, Data Preprocessing in Data Mining, Available: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/ (accessed: 21 January 2025)

Rowe , W. (2018). *Mean Square Error & R2 Score Clearly Explained*. Available at: https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/ (Accessed: 23 January 2025)

Daython3, 13 May 2023, *Mastering the Art of Feature Selection: Python Techniques for Visualizing Feature Importance*. Available at: https://medium.com/@daython3/mastering-the-art-of-feature-selection-python-techniques-for-visualizing-feature-importance-cacf406e6b71 (Accessed: 27 January 2025)

Kumar, R., Singh, A., & Shukla, P. K. (2021). *Air pollution and human health: A systematic review*. Environmental Research, 197, 111226.

Geeksforgeeks (09 August 2024). *Using KNNImputer in Scikit-Learn to Handle Missing Data in Python*. Available at: https://www.geeksforgeeks.org/python-imputation-using-the-knnimputer/ (Accessed: 28 January 2025)

Datacamp. (8 August 2024). *Classification in Machine Learning: An Introduction*. Available at: https://www.datacamp.com/blog/classification-machine-learning (Accessed: 30 January 2025)

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems, 26*, 431–439.

Pope, C. A., & Dockery, D. W. (2006). *Health effects of fine particulate air pollution: lines that connect*. Journal of the air & waste management association, 56(5), 709-742. Available at: www.human-biology-and-public-health.org

Seinfeld, J. H., & Pandis, S. N. (2016). Atmospheric chemistry and physics: From air pollution to climate change. John Wiley & Sons.

Atkinson, R. (2000). Atmospheric chemistry of VOCs and NOx. Atmospheric environment, 34(12-14), 2063-2101.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.

Atkinson, R. (2000). Atmospheric chemistry of VOCs and NOx. Atmospheric Environment, 34(12-14), 2063-2101.

Seinfeld, J. H., & Pandis, S. N. (2016). Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. John Wiley & Sons.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79-82.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Draper, N. R., & Smith, H. (1998). Applied regression analysis. John Wiley & Sons.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics,1 15(4), 580-585.