A Project Report on

AIR QUALITY ANALYSIS & PREDICTION

Submitted by

SANDEEP REDDY PANYALA(AP24122060005)

SRI HARSHA VARDHAN GOGISETTY(AP24122060007)

VAMSI KAKUMANU (AP24122060022)

Submitted to

Dr Tapas Kumar Mishra

DSC 501

COMPUTATIONAL ESSENTIALS FOR DATA SCIENCE

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

Dec 2024

TABLE OF CONTENTS

ABSTRACT

Air pollution is a serious worldwide problem that affects ecosystems, human health, and the environment. As pollutant levels rise, precise predicted solutions are needed. The goal of this research is to create a machine learning framework that uses atmospheric and pollutants data to forecast air quality levels. Several feature selection methods, such as Forward Selection, Backward Selection, and Correlation Analysis, were used to assess models including Logistic Regression, Decision Tree, Random Forest, and SVM. When paired with forward selection, logistic regression had the lowest mean squared error (0.0408) and the best accuracy (0.9785), consistently outperforming other models. Appropriate model assessment was ensured by using metrics like confusion matrix, accuracy, and MSE to guide the evaluation procedure. The final Logistic Regression model performed exceptionally well on real-world datasets after being optimized with Forward Selection features.

INTRODUCTION

In the present world, air pollution has become a global issue, having severe impacts on living beings' health and also on the environment. The quality of the air we breathe directly impacts the individuals well-being, because the exposure to pollutants which are harmful leads to facing health issues like respiratory diseases, asthmas, and many other serious health conditions. Along with degradation of human health, the decrease in the air quality also impacts on environmental degradation, climate change and degradation of the ecosystems.

With the rise in industrialization and urbanization, understanding the air pollution metrics has become an important metric. Pollutants like Co,No2, NH3,PM 2.5 contribute significantly in increasing the level of the pollutants in the atmosphere which makes it a complex challenge to address. Stopping these emissions of the pollutants entirely may not be a feasible solution, but we can minimize the extraction of the pollutants which can be an improvement. This can be possible by performing comprehensive analysis on the air quality across different regions.

To understand and assess the impacts of the pollution level of air, the air quality is categorized into different levels, which ranges from good level to hazardous level. These categories help in providing generalization of the pollutant levels within a specific area, allowing for informed decisions with respect to taking health precautions and other outdoor activities. Monitoring air quality and generating predictions are important in preventing the health risks associated with pollution. Especially in densely populated regions like urban areas where the pollution levels tend to be higher.

The main target of this project is to tackle the challenges of forecasting air quality by utilizing the historical data. By assessing the trends and concentration levels of various pollutants, we predict the air quality levels and offer meaningful insights into environmental conditions. These predictions can help individuals, organizations, and policymakers in making well-informed decisions to safeguard public health and promote environmental sustainability.

PROBLEM STATEMENT

Air quality is a critical indicator of environmental and public health. With the increasing levels of pollution driven by urbanization and industrialization, understanding and forecasting air quality have become essential for mitigating its adverse effects. This project focuses on classifying air quality levels by utilizing the historical and pollutant-specific data. The goal is twofold:

Date-Based Prediction: Given a specific date (past, present, or future), the project predicts the air quality level for that day by analyzing historical patterns.

Pollutant-Based Classification: By inputting values of various pollutants (e.g., CO, NO2, PM2.5) determined by the feature selection technique, the project determines the corresponding air quality level, categorizing it into predefined classes such as "Good," "Moderate," or "Hazardous."

By incorporating data analysis, visualization, and machine learning techniques, this project helps in providing actionable insights to individuals and policymakers, which enables us in making informed decisions to protect public health and promote environmental sustainability.

DESCRIPTION OF THE DATASET

This project focuses on analyzing the Air Quality, for better understanding and evaluation we have considered datasets from various cities. These datasets were imported from the Kaggle website. The dataset represents the hourly data of the levels of the pollutants. This high-frequency data helps us in examination of the trends or any potential pollutants which affects the air quality of those regions. The dataset ranges from 25000 to 35000 records and 20 to 25 columns of the data. Depending on the city the pollutants and the records are being altered.

The dataset has features:

FEATURES	DESCRIPTION
From date	The timestamp indicating the start of the hourly data record.
To date	The timestamp indicating the end of the corresponding hourly data record.
PM 2.5(ug/m^3)	Concentration of particulate matter smaller than 2.5 microns.
PM 10 (ug/m^3)	Concentration of particulate matter smaller than 10 microns.
NO (ug/m^3)	Concentration of Nitric Oxide.
NO_2 (ug/m^3)	Concentration of Nitrogen Dioxide.
NO_X (ug/m^3)	Total concentration of Nitrogen Oxides.
NH_3 (ug/m^3)	Concentration of Ammonia.

Benzene (ug/m^3)	Concentration of Benzene.
Toluene (ug/m^3)	Concentration of Toluene.
SO_2 (ug/m^3)	Concentration of Sulfur Dioxide.
CO (ug/m^3)	Concentration of Carbon Monoxide.
Ozone (ug/m^3)	Concentration of Ozone.
Temperature (Degree C)	Temperature in degree Celsius
RH%	Relative humidity percentage.
WS(m/s)	Wind speed in meters per second.
WD(deg)	Wind direction in degrees.
SR(w/m^3)	Solar radiation in watts per square meter.
BP(mm Hg)	Barometric pressure in millimeters of mercury.
VWS (m/s)	Vertical wind speed in meters per second.
AT (degree C)	Apparent temperature in degrees Celsius.
RF(mm)	Rainfall in millimeters.

LITERATURE REVIEW:

Several researches had done regarding air quality analysis in the past

Vamsi et al.[1], the researchers examined changes in the concentration of air pollutants using data from the Central Pollution Control Board (CPCB). They employed descriptive trend analysis to study various air pollutants, including SO2, NO2, PM, O3, CO, and benzene, across multiple cities in India. Based on this analysis, they calculated the Air Quality Index (AQI) for each city. The focus was then narrowed to three major urban areas with severe air quality issues, prioritizing funding efforts to reduce air pollution. Two additional factors were also considered: a) per capita income of the city and b) the population of children under six years old. Cities with low per capita income and high child populations were given the highest priority, as they were deemed to be in urgent need of intervention. The study also assessed the significant reduction in pollution levels during the COVID-19 lockdown.

Kang et al.[2], their focus was on the application of machine learning (ML) algorithms in air quality modeling, specifically for the following tasks: a) Predicting atmospheric pollutant concentrations, b) Forecasting air quality by predicting future pollutant levels and AQI values, c) Enhancing spatial resolution of pollutant data, particularly through the use of ensemble learning techniques, d) Applying dimensionality reduction and feature selection using Principal Component Analysis (PCA). The study also noted that the percentage of training data varied with each ML model. Among the models evaluated, Random Forest demonstrated superior accuracy in predictions compared to traditional methods such as decision trees and multiple linear regression.

Ameer et al.[3], an architecture for air quality prediction using machine learning (ML) techniques is presented. The workflow includes data collection from IoT-based sensors, followed by communication, data management, and application layers to enable real-time air pollution monitoring and informed decision-making. The study evaluates four ML regression techniques: Decision Tree, Random Forest, Gradient Boosting Regression (GBR), and Multi-Layer Perceptron. Among these, Random Forest delivered the best performance, achieving the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) across various cities, with an MAE of approximately 10.5% in Chengdu and 13.1% in Guangzhou. While the Decision Tree model had the advantage of faster processing times, it exhibited higher error rates. GBR, on the other hand, showed the weakest performance in both accuracy and processing time.

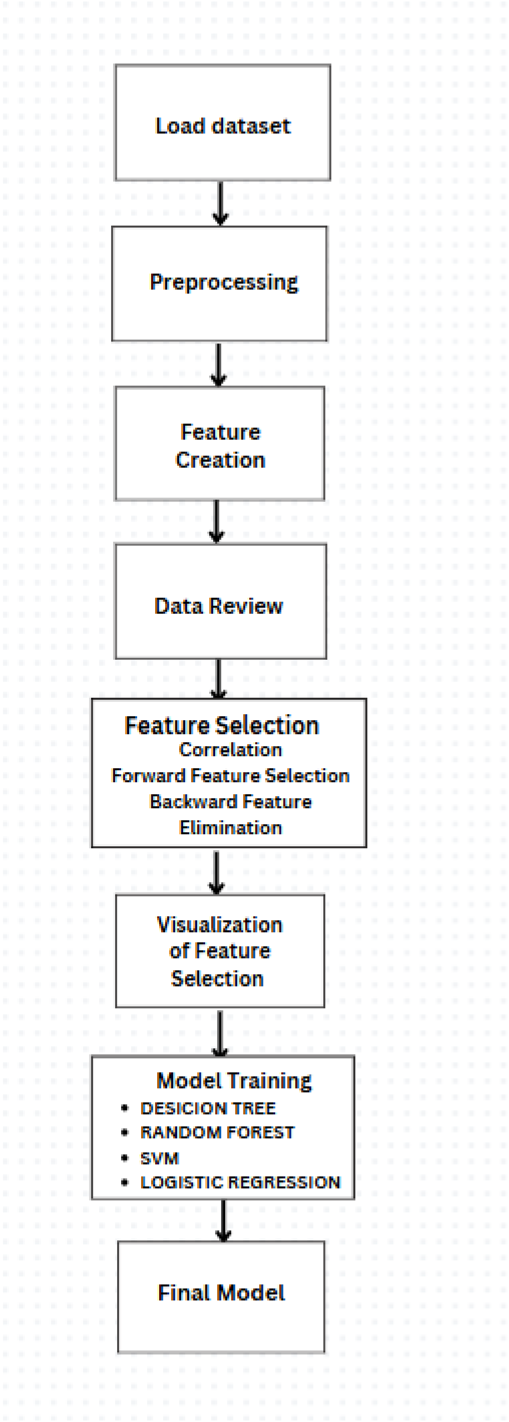
Madan et al.[4], explores how machine learning (ML) techniques address challenges in predicting air quality levels more accurately than traditional methods. It compares various ML approaches, including linear regression, decision trees, random forests, support vector machines, neural networks, and hybrid models like light gradient boosting, applied to pollutant and meteorological datasets. The integration of meteorological factors like wind speed, temperature, and humidity enhanced prediction accuracy, with neural networks and ensemble methods such as boosting identified as the most effective. Despite advancements, challenges like sensor data quality and computational costs remain. The study concludes that while ML has significantly improved air quality prediction, further research is needed to address data variability and enhance robustness across diverse environmental conditions.

Kök et al.[5], presented an LSTM-based deep learning model for air quality prediction in smart cities using IoT data. It leverages time-series pollution data, focusing on ozone and nitrogen dioxide levels, to forecast air quality. The LSTM model, optimized with key hyperparameters, outperforms the Support Vector Regressor (SVR) in predicting air quality categories such as Good, Moderate, and Unhealthy-Hazardous based on AQI thresholds. Experimental results highlight LSTM's superior accuracy, precision, and recall, showcasing its capability for processing large-scale data and aiding urban planning. The study calls for further exploration of advanced deep learning techniques for improved air quality forecasting.

METHODOLOGY

Workflow

Forecasting air quality levels involves a systematic approach that encompasses data



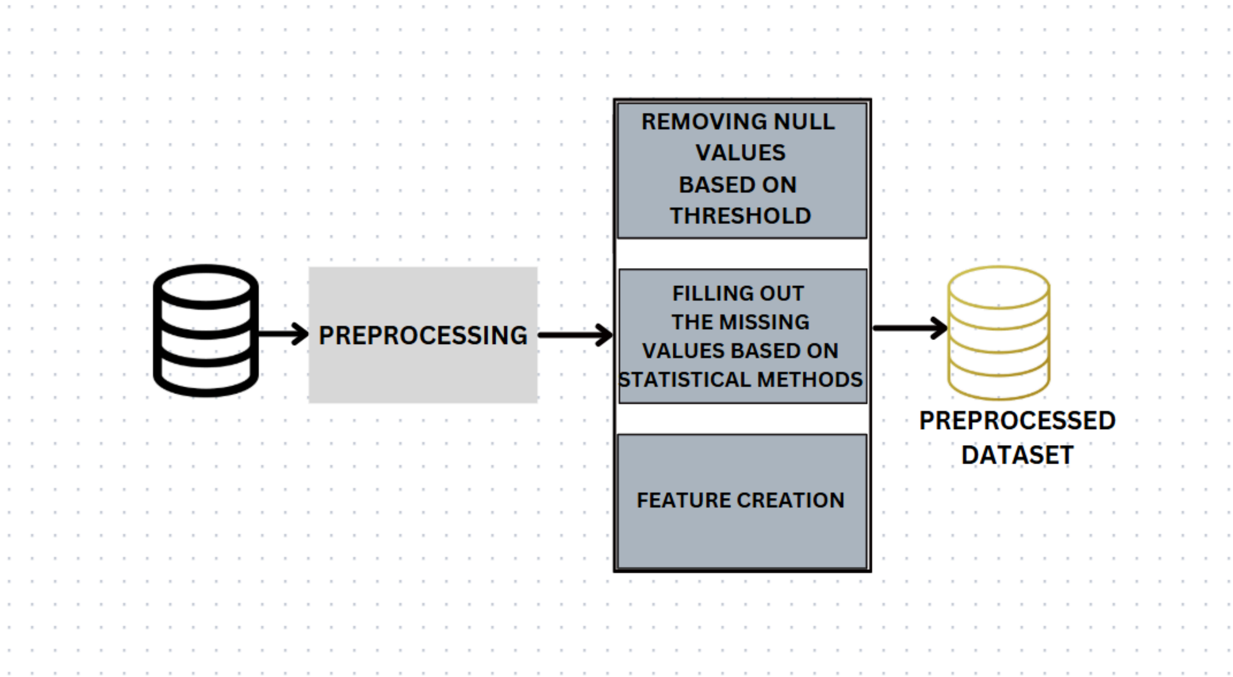
collection, preprocessing, model selection, training, evaluation, and deployment. For any project we would be following the Data Science life cycle process. The initial process is Discovery of the data, which involves gathering of the data from the identified internal and external sources.

Data Collection

We collected the air quality datasets of various cities, which contain the release of pollutants in an hourly manner, for analyzing the pollution trends and predicting the air quality levels. These datasets were obtained from the platform Kaggle which is publicly available.

The dataset is downloaded from kaggle, and loads the dataset which is in the CSV format, into the Python environment by using the Pandas library. Once loaded, the dataset can once be checked for understanding the structure and the quantities present in the dataset. This includes checking for the summarization of the data, identifying some important features, checking for the presence of the null or missing values. These steps help in getting an idea of the dataset before implementing the analysis and modelling phases.

Data Preprocessing


PREPROCESSING
REMOVING NULL VALUES BASED ON THRESHOLD
FILLING OUT THE MISSING VALUES BASED ON STATISTICAL METHODS
FEATURE CREATION
PREPROCESSED DATASET

Data preprocessing is the main step in data analysis or machine learning projects. Ths step ensures the data is clean, consistent and can be ready for applying any methods or performing analysis. For preprocessing of our dataset, timestamp columns (From Date and To Date) were converted to datetime objects for accurate handling. Columns were classified into date-related and pollutant-related data for further analysis. We have repaired the missing values present in our dataset. If there is a column with more than 55% of the missing values we dropped that column from the dataset. This helps in not affecting the remaining data present in the dataset. For the columns with less than 20% of missing values, linear interpolation method method was used to fill in the missing values. This method estimates the missing values based on the nearby observations, maintaining the data continuity. Remaining data was cleaned to ensure consistency and readiness for analysis. In case if there are any records which are having residual missing values even after performing above steps, then they were dropped to create a completely clean dataset. We have converted the hourly data samples into a day format. We have also created a feature Weekday with respect to the date/month/year format present in the dataset record the day of that date is created at that point. After performing all the preprocessing methods, the dataset is stored into a new csv file. This preprocess ensures free of missing values, aggregated to a meaningful daily frequency which helps for easier analysis and additional features for contextual insights.

Data Review

It is an activity through which the correctness conditions of the data are verified. It validates the data types, ensures each column in the dataset has the expected data type. Incorrect data leads to errors in computation and inaccurate results during analysis. It also includes the specification of the type of the error or condition not met, the qualification of the data, and its division into the "error-free" and "erroneous" data. Validating each column is with the expected data type and also checks if the associated values are within their specified range or not(range check).

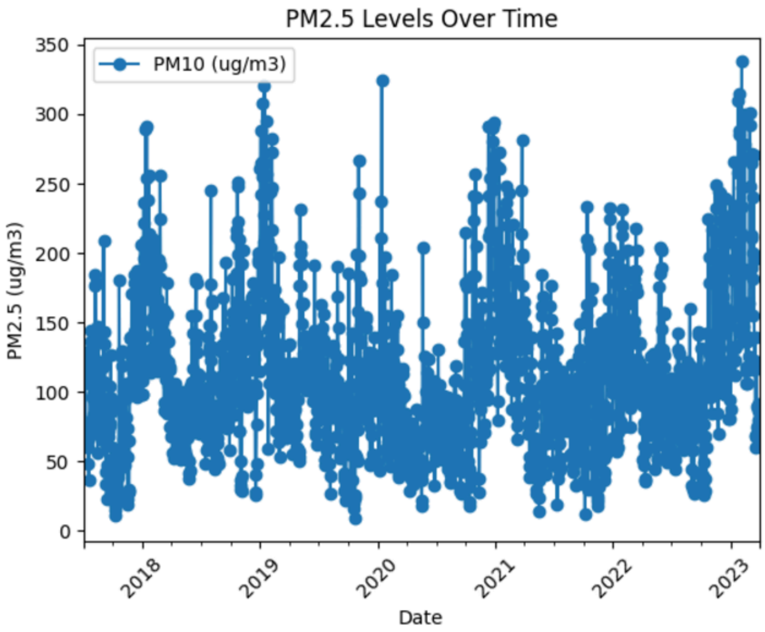
Exploratory Data Analysis (EDA)

The primary objective of EDA is to gain an in-depth understanding of the dataset by visually and statistically summarizing its key characteristics. This process is essential for uncovering hidden patterns, relationships, trends, and potential anomalies in the data that may not be immediately obvious. The insights derived from EDA help shape the direction for further model development.

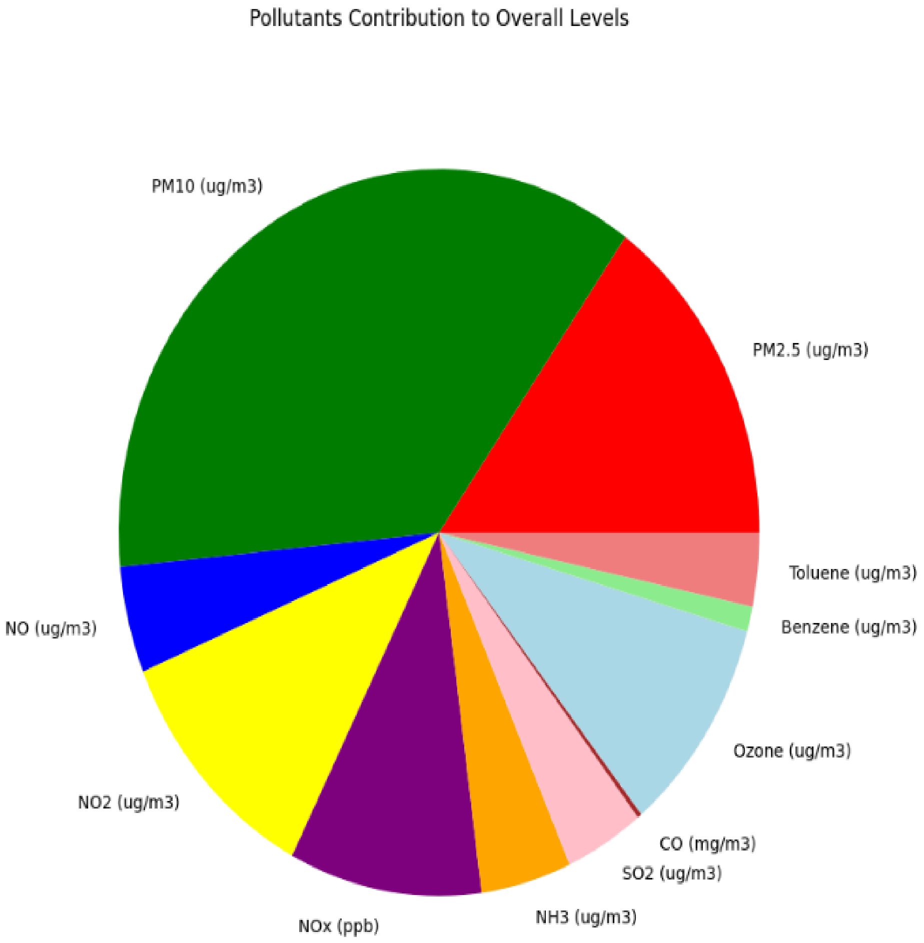
To understand the variance associated with each feature we have analyzed the variance for each numerical feature. This helps in identifying the features with wide spread of their values, which could provide insights for patterns of behaviours with respect to pollutants. A bar graph is plotted for the visualization of the variance of these features.

The distribution of numeric features was visualized using histograms. This helped in understanding the spread, skewness, and overall shape of the data for each variable. By plotting these distributions, we could identify normal or non-normal trends and detect potential outliers that could influence further analyses or predictions.

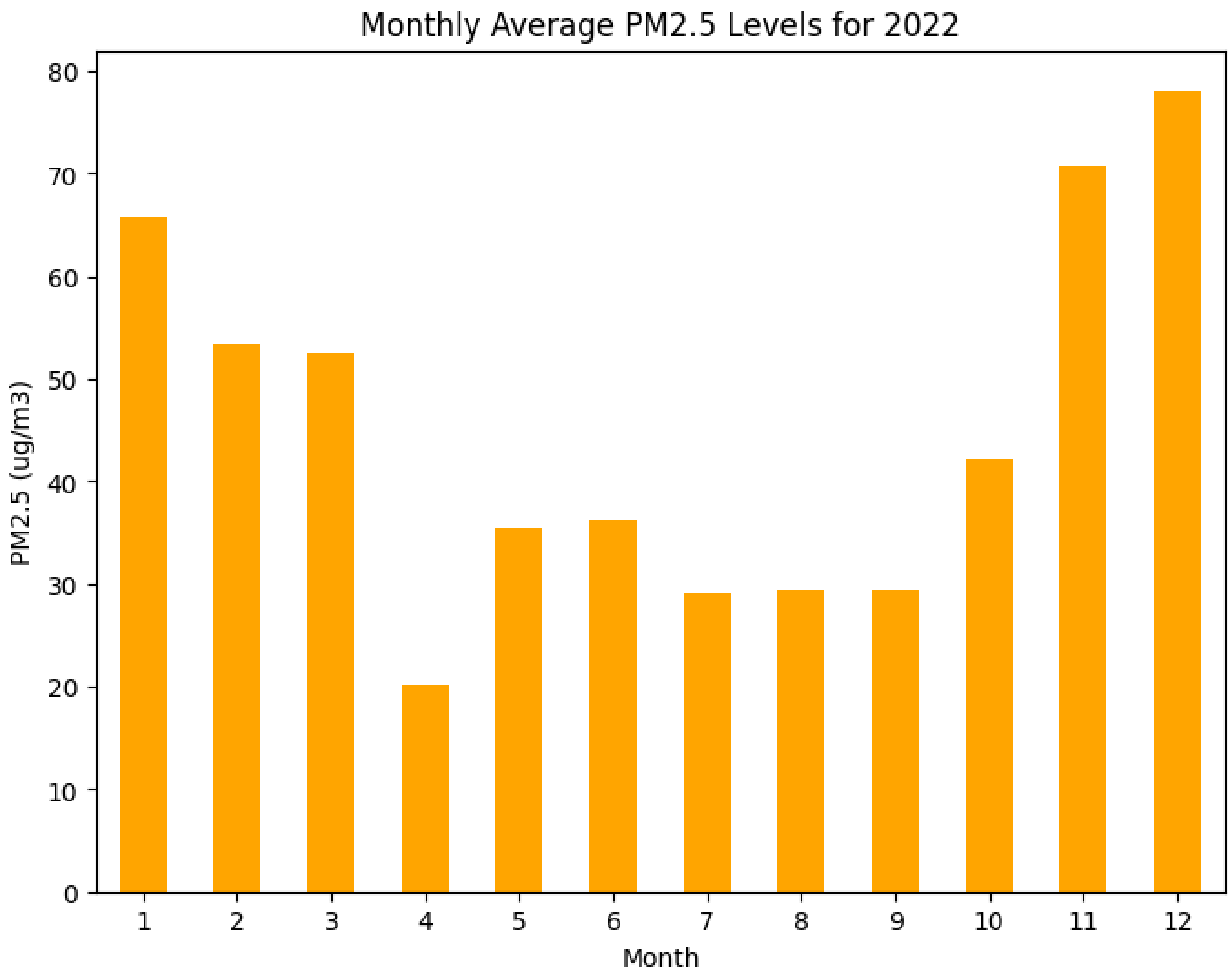
We plotted time series graphs for pollutants like PM2.5 to observe their behavior over time. These plots helped identify seasonal or periodic trends, spikes during specific times, or long-term increases or decreases. For example, PM2.5 levels showed notable variations across different periods, emphasizing the need for further investigation into contributing factors.



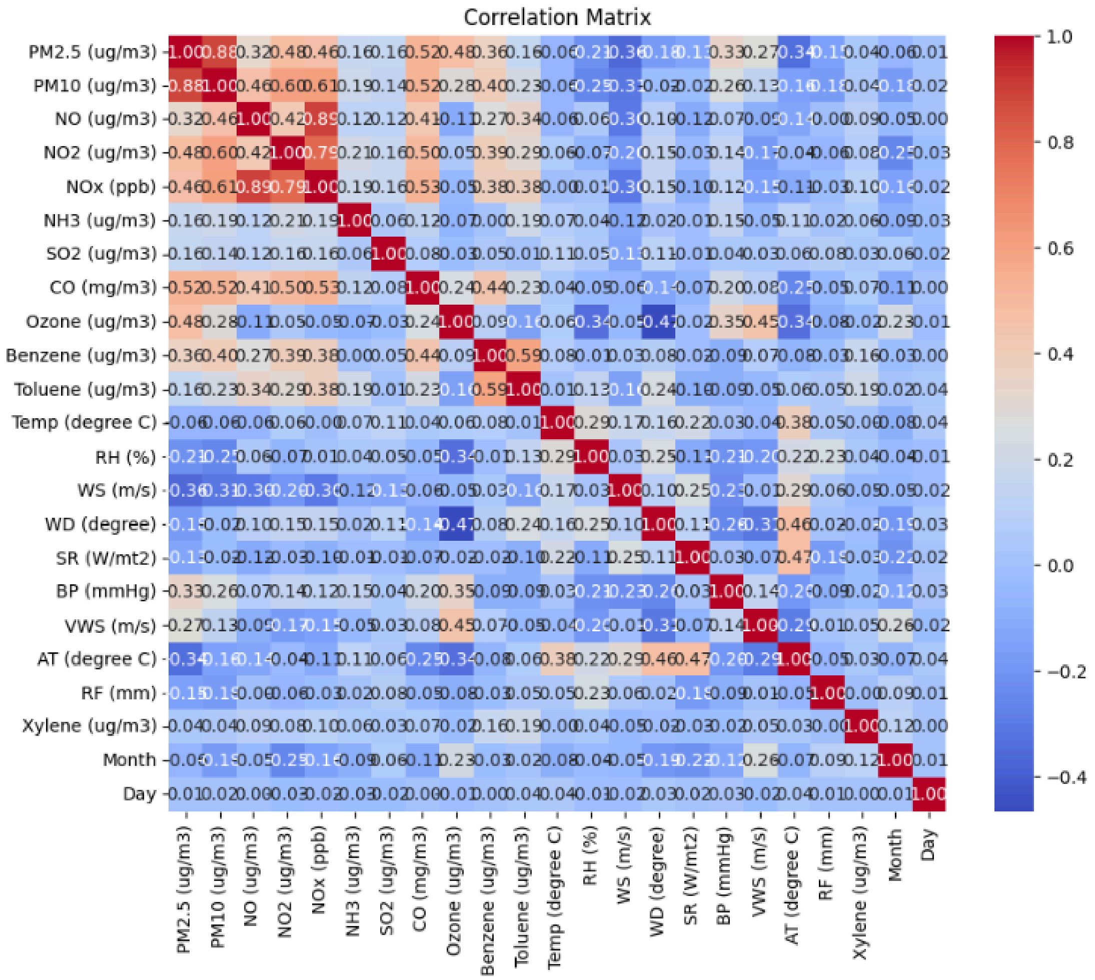
To analyze the contribution of different pollutants to overall air quality, we calculated the total pollutant levels for each pollutant and visualized them using a pie chart. This visualization provided insights into the dominant pollutants in the dataset, allowing us to focus on pollutants contributing significantly to poor air quality.



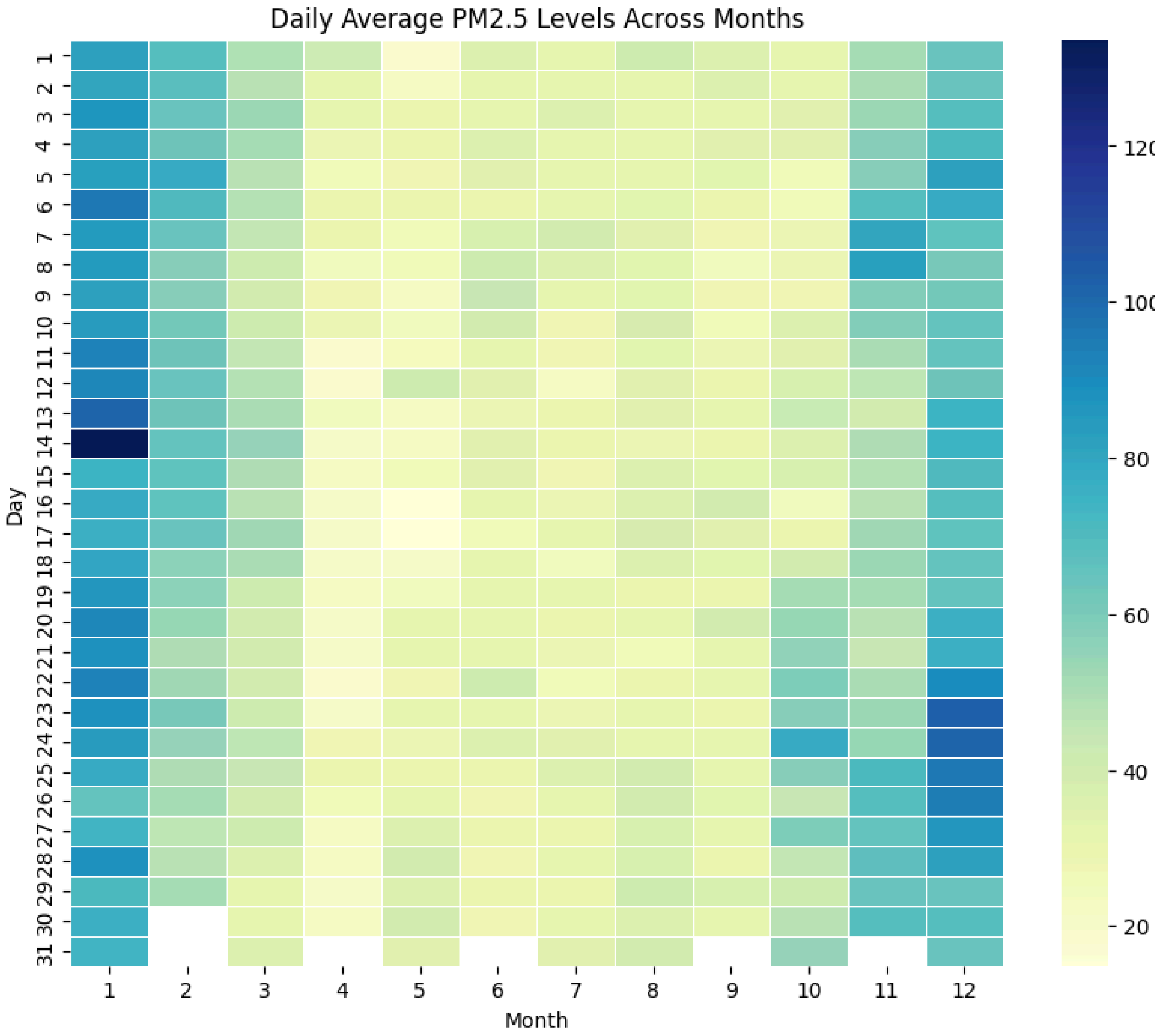
The data was categorized by month, and the average PM2.5 concentrations for each month were computed in order to analyze the monthly patterns in PM2.5 levels. The findings made it easier to spot trends and seasonal changes in the quality of the air over time. For instance, seasonal weather patterns or other environmental factors may cause PM2.5 concentrations to be greater in some months. Below a bar plot of the monthly averages helped draw attention to these variations and gave information about the times of year when air quality might be more troublesome. Planning interventions or educating the public about possible pollution hazards require such knowledge.



A correlation matrix was visualized using a heatmap to examine relationships between pollutants and meteorological variables. Strong positive correlations, such as between PM2.5 and PM10, indicate shared sources or similar conditions, while negative correlations, like those with wind speed, suggest pollutant dispersion effects. This analysis highlights key factors influencing air pollution, aiding in targeted interventions, feature selection for predictive models, and informed policymaking to mitigate pollution effectively.



The heatmap generated below offers a comprehensive visualization of daily average PM2.5 levels across months, revealing significant insights. Seasonal trends can be identified, highlighting months or seasons with consistently high or low PM2.5 levels and detecting periodic patterns aligned with climatic or weather changes. Critical pollutant days with unusually high concentrations may correlate with specific events like industrial activities, festivals, or adverse weather conditions such as fog or stagnant air. Monthly averages allow comparison of PM2.5 levels across months, pinpointing periods with the most or least pollution. Daily variations within a month can help identify recurring causes of elevated or reduced pollutant levels. Understanding air quality dynamics through these trends aids in public health advisories and policymaking. Furthermore, the analysis supports data-driven planning, recommending optimal timings for outdoor activities, urban projects, or awareness campaigns. This heat map serves as a valuable tool for air quality management and strategic interventions.



Using daily averages of PM2.5 levels, a heatmap was created to visualize trends across days and months. The heatmap revealed patterns of pollution intensity over the year, highlighting days or months with consistently high or low pollution levels. This seasonal analysis provided insights into periods of heightened pollution and potential causes.

Feature Selection

To identify the most influential features for air quality classification and prediction, we have used three distinct feature selection techniques were employed:

1. Correlation-Based Feature Selection
2. Forward Feature Selection
3. Backward Feature Elimination

Correlation-Based Feature Selection

Calculated the correlation matrix of all features and their correlation with the target variable (Air Quality Level). Ranked features based on their absolute correlation values with the target. Iteratively selected the top features while ensuring that the pairwise correlation among the selected features was below a predefined threshold (0.8).

Forward Feature Selection

We used the Sequential Feature Selector from sklearn in forward selection mode. Trained a Random forest classifier iteratively, adding one feature at a time that improved the model's cross-validated accuracy. Continued until 10 features were selected.

Backward Feature Elimination

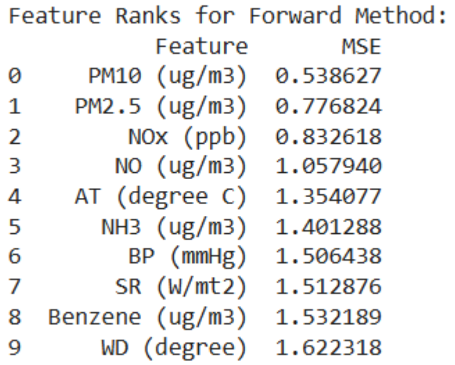
Used the SequentialFeatureSelector in backward elimination mode. Started with all features and removed features one by one based on their negative impact on model accuracy. stopped when 10 features remained.

EXPERIMENTAL RESULTS AND ANALYSIS

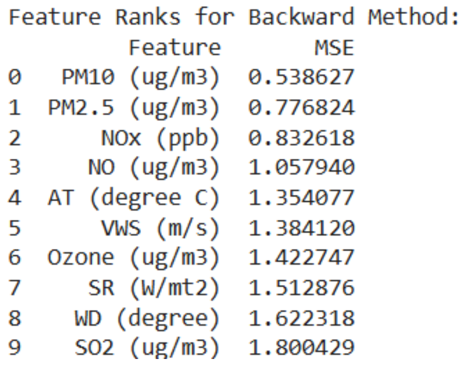
Feature Ranking and Performance Comparison

After selecting features with each method:

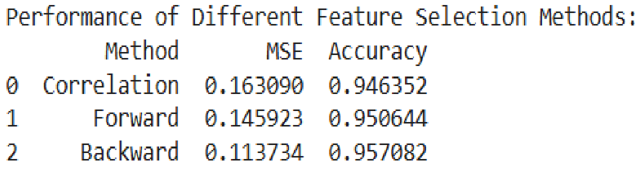
1. Rank Features:
   - Calculated the MSE of the model trained on each individual feature.
   - Ranked features based on their contribution to reducing prediction error.
2. Comparison:
   - Plotted the accuracy and MSE for all three methods.
   - Visualized the trade-offs between accuracy and error reduction for each technique.



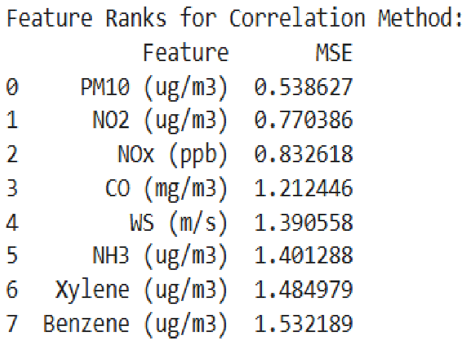
```
Feature Ranks for Forward Method:
           Feature       MSE
0      PM10 (ug/m3)  0.538627
1     PM2.5 (ug/m3)  0.776824
2        NOx (ppb)  0.832618
3       NO (ug/m3)  1.057940
4     AT (degree C)  1.354077
5      NH3 (ug/m3)  1.401288
6        BP (mmHg)  1.506438
7       SR (W/mt2)  1.512876
8   Benzene (ug/m3)  1.532189
9      WD (degree)  1.622318
```



```
Feature Ranks for Backward Method:
           Feature       MSE
0    PM10 (ug/m3)  0.538627
1   PM2.5 (ug/m3)  0.776824
2       NOx (ppb)  0.832618
3      NO (ug/m3)  1.057940
4   AT (degree C)  1.354077
5       VWS (m/s)  1.384120
6   Ozone (ug/m3)  1.422747
7      SR (W/mt2)  1.512876
8     WD (degree)  1.622318
9     SO2 (ug/m3)  1.800429
```



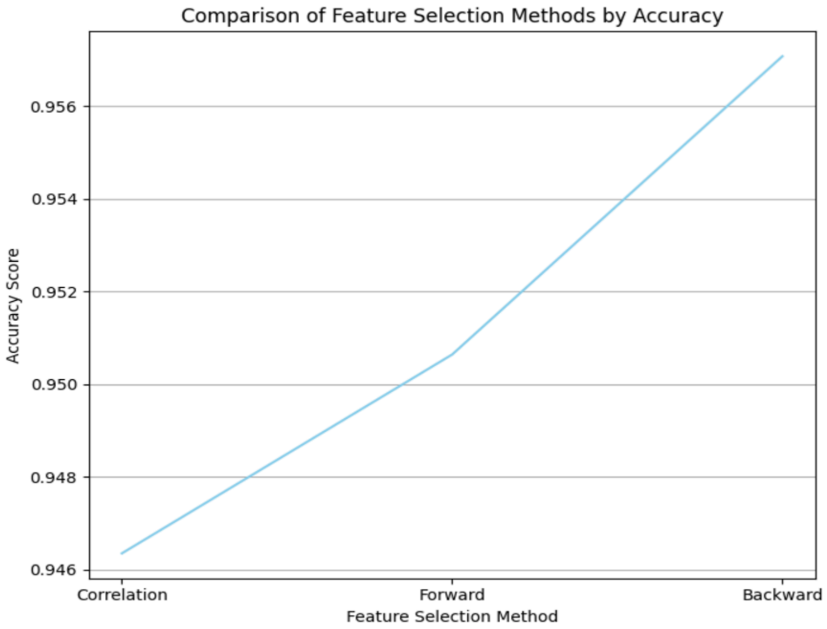
```
Performance of Different Feature Selection Methods:
        Method       MSE  Accuracy
0  Correlation  0.163090  0.946352
1      Forward  0.145923  0.950644
2     Backward  0.113734  0.957082
```



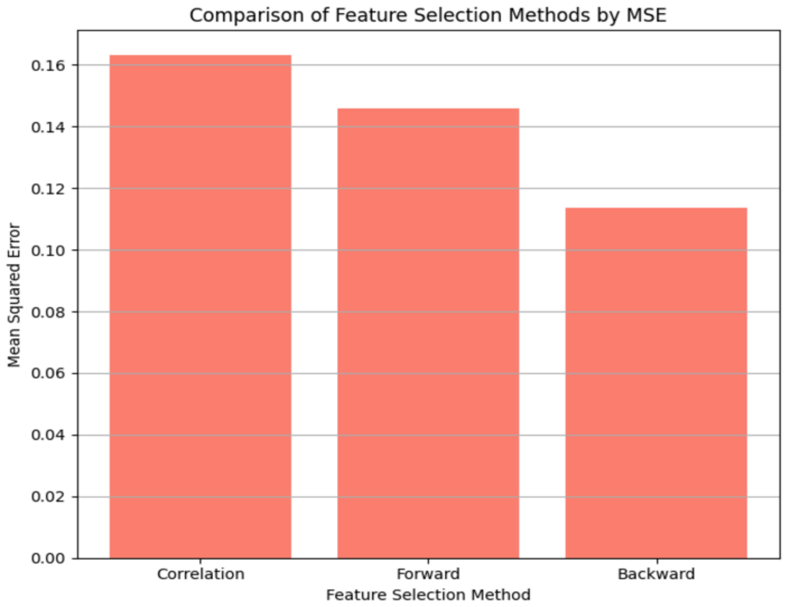
```
Feature Ranks for Correlation Method:
            Feature       MSE
0      PM10 (ug/m3)  0.538627
1       NO2 (ug/m3)  0.770386
2         NOx (ppb)  0.832618
3        CO (mg/m3)  1.212446
4          WS (m/s)  1.390558
5       NH3 (ug/m3)  1.401288
6    Xylene (ug/m3)  1.484979
7   Benzene (ug/m3)  1.532189
```

Created visualizations to compare the performance of the feature selection methods:

Accuracy Plot: Line graph showing accuracy scores for each method.



MSE Plot: Bar chart illustrating the mean squared error for each method.



Model Building

Model building is the foundational phase where data preparation, feature selection, and model training are carried out. The aim is to construct machine learning models that can effectively predict air quality levels based on pollutant data.

Training Multiple Models

The code trains several machine learning algorithms, which includes:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)

Each model is trained on the selected features, and its performance is evaluated on the test data.

Best Model Selection

After training, the best model for each feature selection method is identified based on its accuracy. Subsequently, the overall best model across all methods is selected. The corresponding features used by this model are noted for later predictions.

Retraining Best Model: The overall best model is retrained on the entire training dataset using its associated features to prepare it for deployment.

Model Evaluation

Model evaluation is essential for assessing the generalization ability of the trained models and identifying the most reliable one for prediction.

Performance Metrics

The code evaluates each model using the following metrics:

- **Accuracy**: Measures the proportion of correct predictions over total predictions.
- **Mean Squared Error (MSE)**: Quantifies the average squared difference between predicted and true values.
- **Confusion Matrix**: Provides a detailed breakdown of prediction results, showing true positives, true negatives, false positives, and false negatives.

- **Classification Report**: Offers detailed performance metrics for each class, including precision, recall, F1-score, and support.

Comparison of Models

The results are summarized in a table, highlighting the best-performing model for each feature selection method. This table includes:

- Model name
- Feature selection method
- Accuracy
- Mean Squared Error (MSE)

The comparison enables a systematic selection of the best overall model, which balances accuracy and robustness.

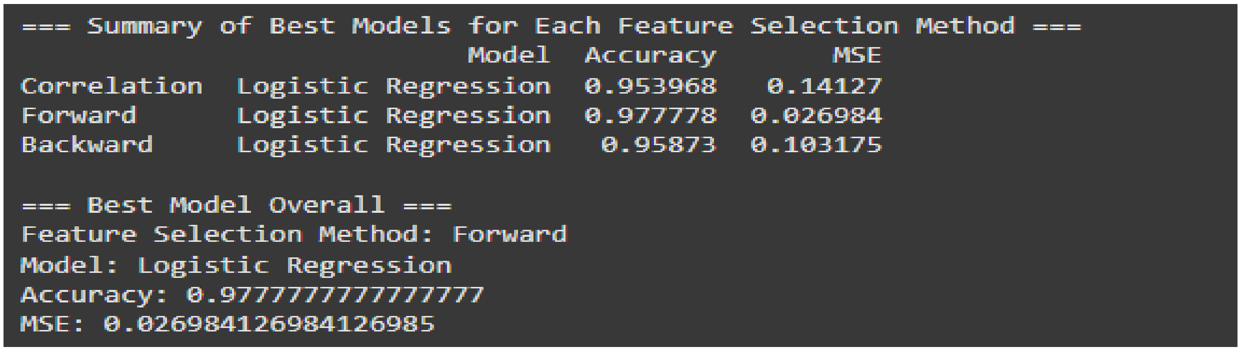
Selection of Best Overall Model

The best-performing model across all feature selection methods is chosen as the final model. This selection is based on its ability to achieve the highest accuracy, ensuring reliable performance in real-world applications.

MODEL - DECISION TREE		
Feature Selection Method	Accuracy	MSE
Correlation	0.9142	0.2476
Forward	0.9268	0.2063
Backward	0.9142	0.2428

MODEL - RANDOM FOREST		
Feature Selection Method	Accuracy	MSE
Correlation	0.9380	0.1904
Forward	0.9539	0.1222
Backward	0.9539	0.1222

MODEL - SVM		
Feature Selection Method	Accuracy	MSE
Correlation	0.9380	0.2
Forward	0.9539	0.1317
Backward	0.9571	0.1238



```
=== Summary of Best Models for Each Feature Selection Method ===
                           Model  Accuracy       MSE
Correlation  Logistic Regression  0.953968   0.14127
Forward      Logistic Regression  0.977778  0.026984
Backward     Logistic Regression   0.95873  0.103175

=== Best Model Overall ===
Feature Selection Method: Forward
Model: Logistic Regression
Accuracy: 0.9777777777777777
MSE: 0.026984126984126985
```

CONCLUSION

Logistic Regression consistently performed well across all feature selection methods, achieving high accuracy and low MSE in each case. Forward Selection emerged as the most effective feature selection method, yielding the highest accuracy (0.9785) and the lowest MSE (0.0408).

Backward Selection and Correlation Selection also showed strong performance, with Logistic Regression performing best in both methods. Other models such as Random Forest and SVM showed varied performance depending on the feature selection method, but Logistic Regression generally outperformed them in terms of accuracy and MSE.

The final model was trained using Logistic Regression with the selected features from Forward Selection, and it showed excellent performance with high accuracy and low error.

REFERENCES

[1]. Vamshi, B Nitesh. (2021). Air Quality analysis. Journal of Data Analysis.

[2]. Kang, G.K., Gao, J.Z., Chiao, S., Lu, S. and Xie, G., 2018. Air quality prediction: Big data and machine learning approaches. *Int. J. Environ. Sci. Dev*, *9*(1), pp.8-16.

[3]. Ameer, S., Shah, M.A., Khan, A., Song, H., Maple, C., Islam, S.U. and Asghar, M.N., 2019. Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE access*, *7*, pp.128325-128338.

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

[5]. Kök, İ., Şimşek, M.U. and Özdemir, S., 2017, December. A deep learning model for air quality prediction in smart cities. In *2017 IEEE international conference on big data (big data)* (pp. 1983-1990). IEEE.