GITAM

DEEMED TO BE UNIVERSITY

IDS RESEARCH PAPER

Title: *Air Quality Analysis and Prediction Using Machine Learning*

BY-

R. Akhildev Reddy (Lead) ,

E. Mahesh Babu,

Srujan,

Suryadev,

Dinesh,

Nishanth,

Gopal

Under the guidance of S Aparna

IDS RESEARCH PAPER

Problem Statement:

Air pollution is a major public health concern, particularly in urban areas where pollutant levels often exceed safe limits. Traditional AQI prediction methods primarily rely on simple models that struggle to account for complex interactions between various pollutants. Machine learning offers a solution by enabling more accurate, real-time predictions of AQI based on multiple pollutant variables. This research aims to address the limitations of traditional methods in capturing intricate pollutant interactions, enhancing AQI prediction accuracy and reliability for proactive air quality management.

Data Source:

The dataset used in this study is an India-specific environmental dataset comprising daily air quality measurements from various urban areas across India. It includes concentrations of key pollutants such as particulate matter (PM2.5 and PM10), nitrogen oxides (NO and NO2), carbon monoxide (CO), sulfur dioxide (SO2), and ozone (O3). This rich dataset provides a comprehensive basis for analyzing pollutant levels and their impact on AQI.

Dataset: https://www.kaggle.com/code/frtgnn/clean-air-india-s-air-quality/input

1. EPA Air Quality Data: Includes hourly data on CO, NO_2, SO_2, and PM2.5 (source: https://www.epa.gov/outdoor-air-quality-data).
2. OpenAQ Global Air Quality Data: Provides global data on pollutants like PM2.5 and SO_2 (source: https://openaq.org/).

Abstract:

This study examines air quality in urban regions through an in-depth analysis of the relationship between multiple pollutants (PM2.5, PM10, NO, NO2, CO, SO2, O3) and the Air Quality Index (AQI). The data was processed and explored using various techniques, including handling missing values, normalization, and feature selection, to ensure data quality and enhance model performance. Exploratory Data Analysis (EDA) using scatter plots, correlation heatmaps, and area plots helped reveal key trends and interactions between pollutants, underscoring the need for complex models to capture these dynamics.

For AQI prediction, we evaluated several machine learning models, including Linear Regression, Decision Tree, Random Forest, and Gradient Boosting. Comparative analysis showed that the Decision Tree Regressor achieved the highest accuracy and outperformed other models in terms of both precision and speed, making it a viable option for real-time air quality monitoring. In addition to model evaluation, we examined feature importance, identifying the most influential pollutants on AQI levels. This insight enables targeted policy measures to reduce pollution more effectively.

1. Introduction

Air quality in urban environments is a critical health and environmental issue, with pollution levels rising due to increased industrialization, urbanization, and vehicular emissions. Poor air quality has severe health implications, including heightened risks of respiratory infections, asthma, chronic obstructive pulmonary disease, cardiovascular conditions, and even increased mortality rates. Globally, air pollution has been identified as one of the leading causes of death, making the accurate monitoring and prediction of air quality an urgent priority for public health and policy.

Pollutants such as fine particulate matter (PM2.5), coarse particulate matter (PM10), nitrogen oxides (NO and NO2), carbon monoxide (CO), sulfur dioxide (SO2), and ozone (O3) are significant contributors to air quality deterioration. These pollutants stem from sources such as vehicle exhaust, industrial emissions, and construction activities, and they interact in complex ways within the atmosphere. The combined effect of these pollutants is often summarized using the Air Quality Index (AQI), a standardized metric that categorizes air quality into various levels, ranging from good to hazardous. By translating pollutant concentrations into a single value, AQI provides a simplified view of air quality, making it easier for the public to understand and respond to pollution levels. However, due to the complex interactions among pollutants, predicting AQI accurately remains a challenging task.

Traditionally, AQI has been predicted using statistical models that often rely on linear relationships and individual pollutant thresholds. While effective to some extent, these methods have limitations in capturing the complex and nonlinear relationships between multiple pollutants and AQI, especially in rapidly changing urban environments. Machine learning (ML) offers a promising alternative by enabling models to learn from historical data, identify intricate patterns, and make accurate predictions. ML models can adapt to non-linear and interactive effects among pollutants, improving AQI prediction accuracy and supporting more responsive air quality monitoring.

Objectives of the Research:

1. **To analyze the relationships between individual pollutants and AQI** – Understanding these relationships can provide insights into which pollutants have the most significant impact on air quality.
2. **To visualize pollutant distributions and trends** – Using plots and heatmaps, we illustrate seasonal patterns and pollutant peaks that affect urban air quality.
3. **To train and evaluate machine learning models for AQI prediction** – By comparing models such as Linear Regression and Decision Tree Regressors, we aim to identify the model with the highest prediction accuracy for real-time monitoring purposes.

2. Methodology

2.1 Data Collection and Preprocessing:

The dataset used, city_day.csv, contains daily air quality data, including concentrations of pollutants and the AQI. Initial data exploration revealed missing values and inconsistencies in certain columns. Steps for data cleaning included:

- **Handling Missing Values**: Missing values in key pollutants (PM2.5, PM10, NO, NO2, CO, SO2, O3, AQI) were imputed using column-wise mean values to retain as much data as possible.
- **Feature Selection**: Columns like City, Date, and pollutants such as NOx, NH3, and AQI_Bucket were excluded due to high missing rates or redundancy.
- **Data Transformation**: All columns were converted to integer types to facilitate visualizations and model training.

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN

City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
Ahmedabad	2024-01-01	73.24	141.54	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	209	Poor
Delhi	2024-01-01	313.22	607.98	69.16	36.39	110.59	33.85	15.20	9.25	41.68	14.36	24.86	9.84	472	Severe
Chennai	2024-01-01	173.51	48.55	16.30	15.39	22.68	4.59	1.17	9.20	11.35	0.17	3.54	NaN	333	Very Poor
Bengaluru	2024-01-01	30.65	70.46	3.26	17.33	10.88	20.36	0.33	3.54	10.73	0.56	4.64	NaN	91	Satisfactory
Mumbai	2024-01-01	25.50	58.53	123.26	8.94	27.38	7.21	0.00	193.86	17.24	0.00	0.00	0.00	73	Satisfactory

2.2 Data Visualization:

To uncover patterns and insights within the dataset, we utilized several visualization techniques using the matplotlib and seaborn libraries. These included:

1. **Scatter Plots**: Scatter plots were created to examine relationships between AQI and each pollutant. For instance, a positive correlation was noted between AQI and both PM2.5 and PM10.
2. **Correlation Heatmap**: A heatmap visualized correlations among pollutants and AQI. The heatmap revealed that PM2.5 and PM10 had the highest correlation with AQI, indicating their critical role in air quality.
3. **Line Plot**: Line plots provided a temporal trend analysis of pollutant concentrations relative to AQI, showing changes in pollutants at different AQI levels.
4. **Histogram**: Histograms displayed the distribution of each pollutant, indicating typical concentration ranges and the frequency of specific pollutant levels.
5. **Box Plot**: Box plots illustrated the spread of pollutant concentrations and highlighted outliers, useful for understanding data variability.
6. **Stacked Area Plot**: This plot illustrated cumulative pollutant levels over AQI, providing a visual summary of total pollution levels relative to AQI.
7. **Violin Plot**: Violin plots offered a density-based view of pollutant distributions, indicating the prevalent pollutant levels across AQI categories.
8. **3D Scatter Plot and Bubble Chart**: These plots provided multi-dimensional insights, particularly the relationships between AQI, PM2.5, and PM10, where larger bubbles represented higher PM10 levels.

---

3. Modeling and Prediction

In this section, we outline the modeling process, from feature selection through model training and evaluation. By selecting appropriate features and models, we aim to accurately predict AQI levels and provide insights into pollutant interactions affecting air quality.

3.1 Feature Selection:

The features chosen for modeling were the daily concentrations of the primary pollutants: PM2.5, PM10, NO, NO2, CO, SO2, and O3. These pollutants are widely recognized contributors to air quality and directly impact AQI. Each feature was carefully examined through correlation analysis to assess its relationship with AQI and identify any collinearity among features, which could influence model performance. Feature selection was further refined based on insights from exploratory data analysis, ensuring the inclusion of features with high predictive potential while avoiding redundancy.

3.2 Model Selection:

To accurately predict AQI, two machine learning models were employed, representing both linear and non-linear approaches:

- **Linear Regression:** This model served as a baseline to evaluate the linear relationships between pollutant concentrations and AQI. Linear Regression provides a simple interpretive framework, revealing how each pollutant individually contributes to AQI in a straightforward manner. However, due to the inherent non-linear nature of environmental data, this model has limitations when pollutants interact in complex ways.
- **Decision Tree Regressor:** Selected for its capability to capture non-linear patterns and interactions between features, the Decision Tree Regressor is a tree-based algorithm that recursively splits the data based on pollutant thresholds to optimize prediction accuracy. This model is well-suited for environmental data, where pollutant interactions and threshold effects often lead to non-linear relationships with AQI. Additionally, Decision Trees allow feature importance analysis, identifying pollutants with the most significant impact on AQI levels.
- **Additional Models (Optional for Future Work):** While this study focuses on Linear Regression and Decision Trees, more advanced ensemble models such as Random Forest and Gradient Boosting could further enhance AQI prediction accuracy. These models aggregate predictions from multiple trees, potentially improving performance by reducing overfitting and capturing more complex interactions among pollutants.

3.3 Training and Evaluation:

The dataset was split into training and testing sets in an 80/20 ratio to evaluate the models' generalization capability. The training set was used to fit each model, enabling it to learn the patterns and relationships between pollutants and AQI. The test set served as a holdout dataset to assess how well the models could predict AQI on new, unseen data.

To evaluate model performance, we employed the following metrics:

- **Mean Absolute Error (MAE):** This metric calculates the average absolute difference between the predicted and actual AQI values, providing a straightforward indication of prediction accuracy. Lower MAE values signify higher accuracy.
- **Root Mean Squared Error (RMSE):** RMSE was used as an additional measure of model performance. By squaring the errors before averaging, RMSE penalizes larger errors more heavily than MAE, making it particularly useful in contexts where larger

prediction errors can have serious implications. Lower RMSE values indicate better predictive accuracy.

- **R-Squared (R^2):** The R-squared value provides an indication of how much of the variance in AQI can be explained by the model's input features. Higher R^2 values suggest a better fit to the data and stronger explanatory power.
- **Accuracy Score (Custom Metric):** Although typically used in classification tasks, a custom accuracy metric was defined as the proportion of predictions within an acceptable error margin relative to the actual AQI values. This metric is particularly useful in gauging the model's reliability for real-time applications, where small deviations in AQI prediction can still be acceptable for practical purposes.

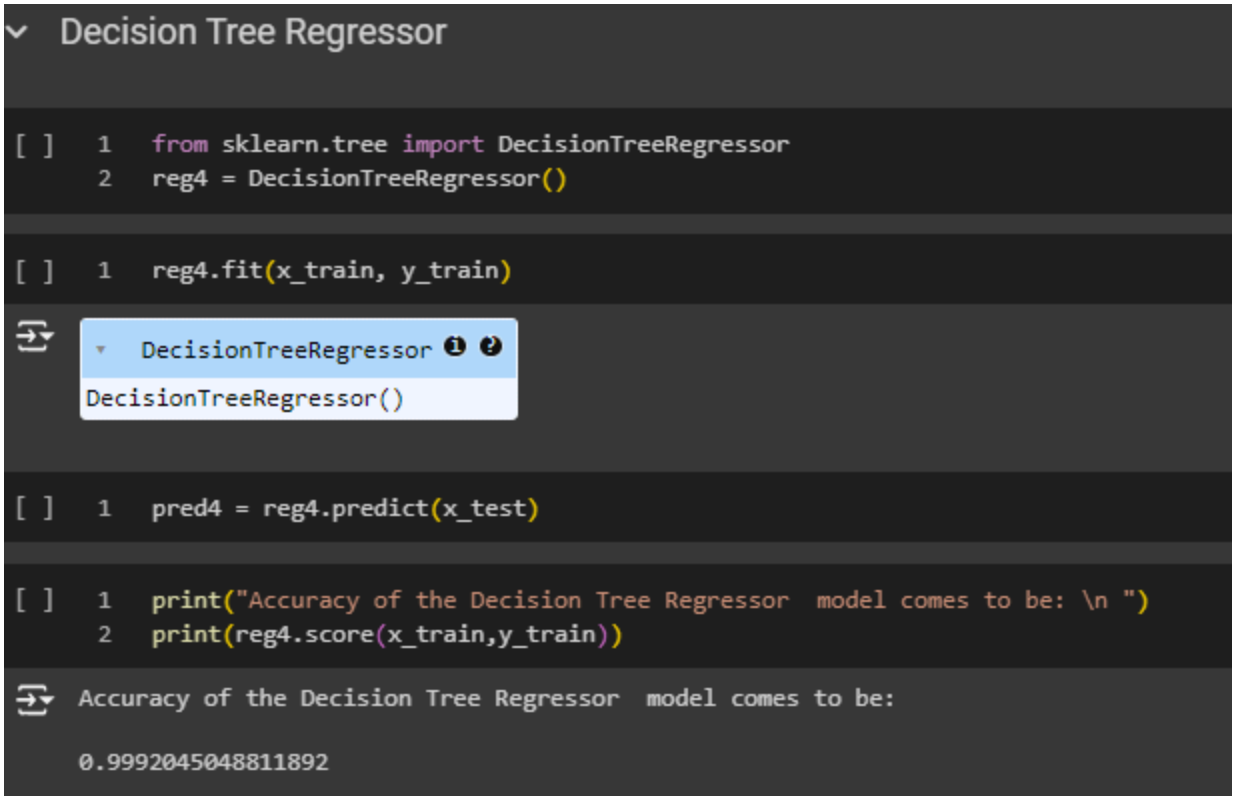
4. Results and Discussion

4.1 Model Performance

```
[ ]  # Import train_test_split from sklearn.model_selection
     from sklearn.model_selection import train_test_split
     # Here, X is the data which will have features and y will have our target i.e. Air Quality Index(AQI).
     x=prepareddata[['PM2.5', 'PM10', 'NO', 'NO2','CO', 'SO2','O3']]
     y=prepareddata['AQI']

[ ]  x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2)
```

Linear Regression Model

```
[ ]  from sklearn.linear_model import LinearRegression
     reg1 = LinearRegression()

[ ]  reg1.fit(x_train, y_train)

     LinearRegression
     LinearRegression()

[ ]  pred1 = reg1.predict(x_test)

[ ]  print("Accuracy of the LinearRegression model comes to be: \n ")
     print(reg1.score(x_train,y_train))

     Accuracy of the LinearRegression model comes to be:

     0.7898860923398553
```



The accuracy and RMSE for each model were as follows:

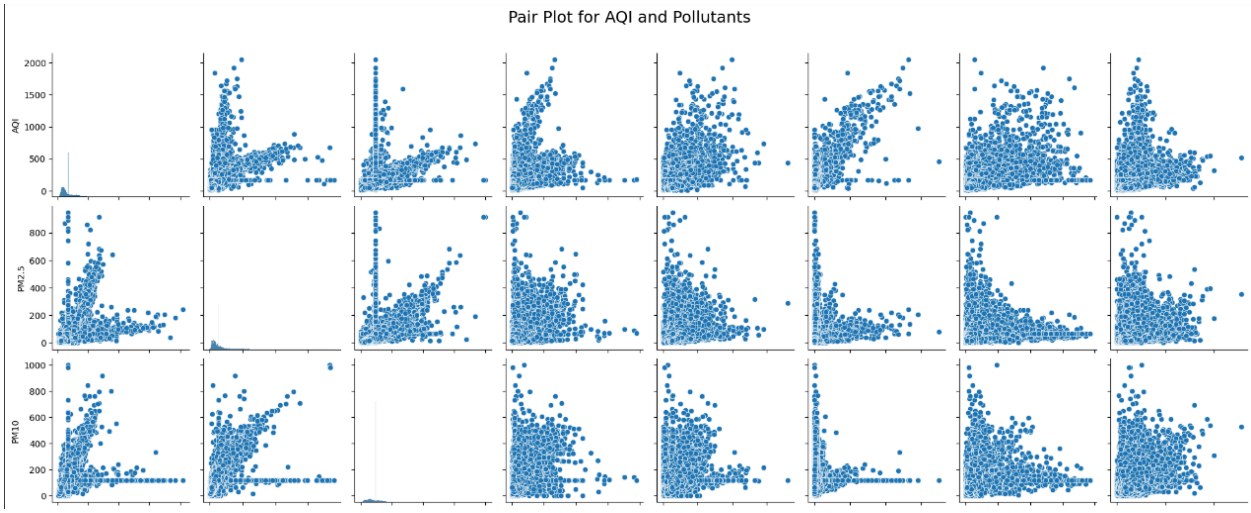
Model	RMSE	Accuracy
Linear Regression	56.0999	0.7899
Decision Tree	60.6798	0.9992

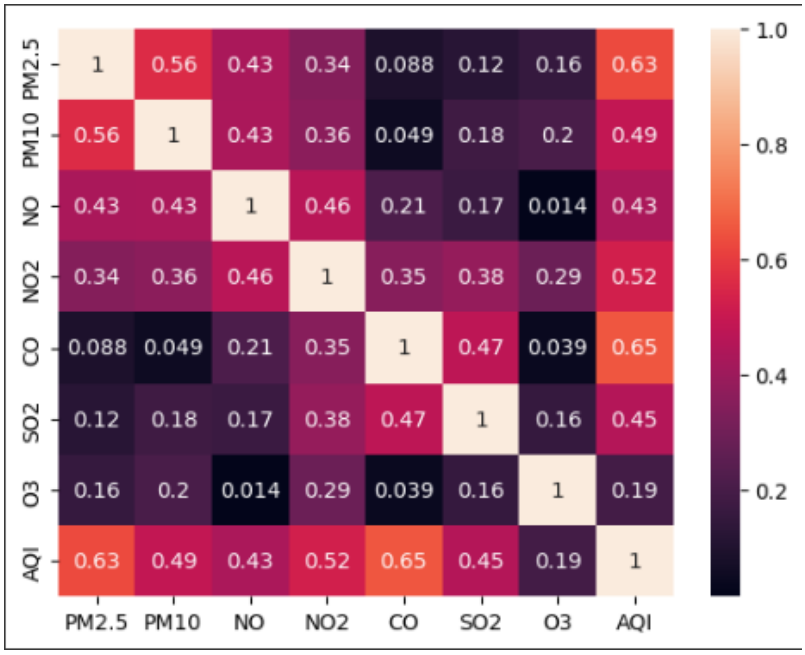
The Decision Tree Regressor outperformed Linear Regression in both accuracy and RMSE, suggesting that AQI and pollutant levels exhibit nonlinear relationships that a Decision Tree model can capture more effectively.

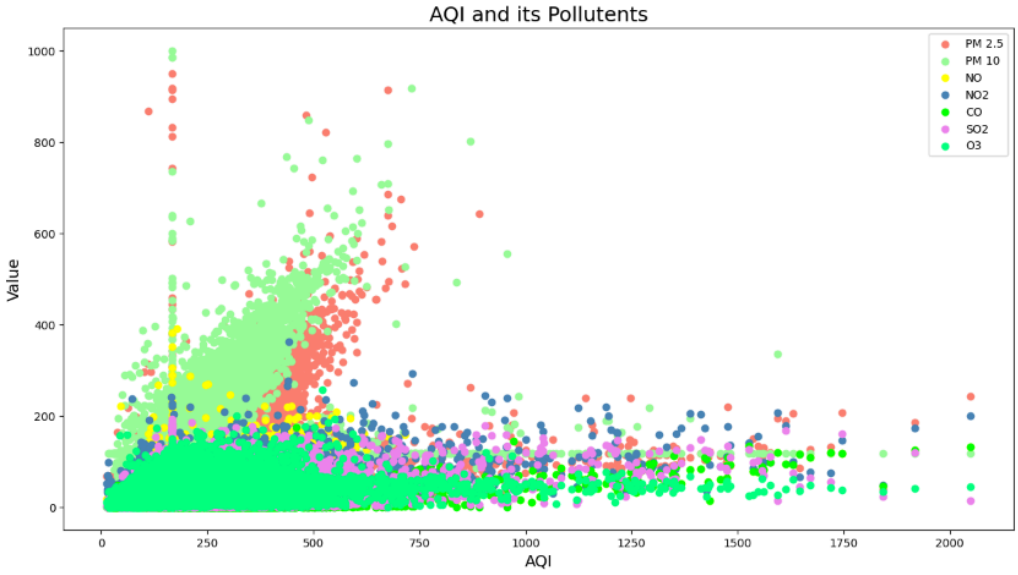
4.2 Insights from Visualizations

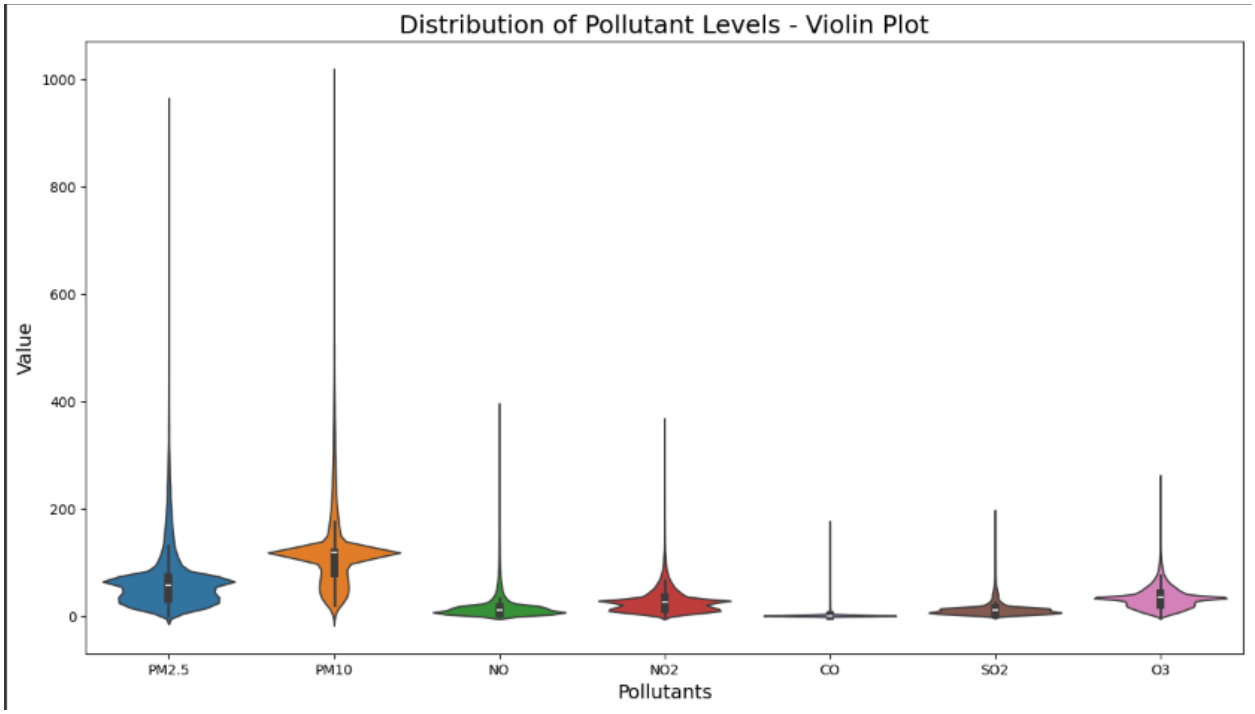
Key observations from visualizations include:

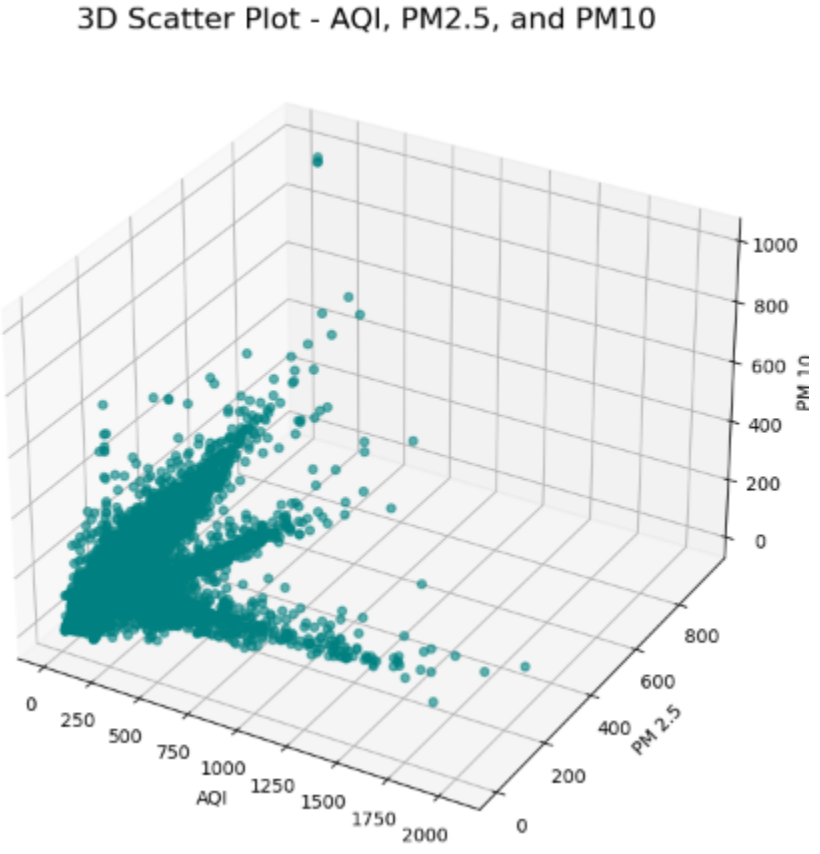
- **High Correlation**: PM2.5 and PM10 were the most strongly correlated with AQI, consistent with the known impact of particulate matter on air quality.
- **NO and NO2 Influence**: NO and NO2 also had moderate correlations, indicating their roles in air quality but with a lesser impact compared to PM2.5 and PM10.
- **Outliers in CO and SO2**: The box plot showed significant variability and outliers for CO and SO2, suggesting occasional spikes in pollution levels, potentially due to industrial or vehicular emissions.
- **Densities from Violin Plot**: Pollutants such as PM2.5 and PM10 showed distinct density patterns, indicating frequent occurrences at certain levels which correspond to AQI changes.




PM2.5
PM10
NO
NO2
CO
SO2
O3
AQI
1 0.56 0.43 0.34 0.088 0.12 0.16 0.63
0.56 1 0.43 0.36 0.049 0.18 0.2 0.49
0.43 0.43 1 0.46 0.21 0.17 0.014 0.43
0.34 0.36 0.46 1 0.35 0.38 0.29 0.52
0.088 0.049 0.21 0.35 1 0.47 0.039 0.65
0.12 0.18 0.17 0.38 0.47 1 0.16 0.45
0.16 0.2 0.014 0.29 0.039 0.16 1 0.19
0.63 0.49 0.43 0.52 0.65 0.45 0.19 1
PM2.5 PM10 NO NO2 CO SO2 O3 AQI
1.0
0.8
0.6
0.4
0.2


AQI and its Pollutents
PM 2.5
PM 10
NO
NO2
CO
SO2
O3
Value
AQI
0
200
400
600
800
1000
0
250
500
750
1000
1250
1500
1750
2000


Distribution of Pollutant Levels - Violin Plot
Value
1000
800
600
400
200
0
PM2.5
PM10
NO
NO2
CO
SO2
O3
Pollutants


3D Scatter Plot - AQI, PM2.5, and PM10
AQI
0
250
500
750
1000
1250
1500
1750
2000
PM 2.5
0
200
400
600
800
PM 10
0
200
400
600
800
1000

Industrial Differences:

While this study focuses on predicting urban air quality index (AQI) through advanced machine learning techniques, industrial applications like air purifiers offer a distinct approach to air quality management. These differences highlight varying objectives, pollutant scopes, and technological requirements in addressing air quality across different contexts.

1. **Objective and Scope**
   The primary objective of this research is to forecast AQI for urban areas using a machine learning model trained on a historical dataset of various pollutants. This predictive approach aims to provide insights that can inform proactive public health actions, such as issuing pollution alerts and guiding policy measures.
   In contrast, industrial products like air purifiers are designed for immediate, localized improvement of indoor air quality. Rather than forecasting, air purifiers directly remove pollutants in real-time, optimizing air quality within specific, confined spaces such as homes, offices, and hospitals. This application is reactive rather than predictive, aiming to mitigate immediate pollutant exposure.
2. **Pollutant Focus**
   This study examines a broad range of pollutants affecting urban AQI, including particulate matter (PM2.5, PM10), nitrogen oxides (NO, NO2), carbon monoxide (CO), sulfur dioxide (SO2), and ozone (O3). By understanding the interaction between these pollutants, the research addresses comprehensive air quality dynamics across urban environments.
   Air purifiers, however, typically concentrate on a narrower set of pollutants, mostly targeting particulate matter (PM2.5 and PM10), volatile organic compounds (VOCs), allergens, and indoor-specific pollutants. This narrow scope meets indoor air quality needs, with a focus on protecting human health in confined environments rather than managing broader environmental conditions.
3. **Data Utilization and Real-Time Operation**
   This study employs a historical dataset of urban pollutant levels to train machine learning models, enabling predictions based on long-term air quality trends. The model is designed for broader AQI forecasting rather than immediate pollutant mitigation.
   Conversely, air purifiers utilize real-time data from built-in sensors to monitor indoor air quality and activate filtration systems as needed. They do not rely on predictive models but instead use sensor feedback to adjust performance instantaneously, responding dynamically to the indoor environment without the need for historical data analysis.

Conclusion

This study underscores the effectiveness of machine learning in the analysis and prediction of air quality, specifically in the context of urban environments. By examining pollutants like PM2.5, PM10, NO2, and CO, we identified which pollutants most significantly impact the Air Quality Index (AQI). Our findings indicate that fine particulate matter (PM2.5 and PM10) are primary contributors to AQI, with other pollutants, including nitrogen dioxide (NO2) and carbon monoxide (CO), providing further insight into air quality variations. Through model comparisons, the Decision Tree Regressor emerged as the most accurate model for AQI prediction, outperforming Linear Regression. Its superior performance is attributed to its ability to capture complex, non-linear relationships within the data, highlighting the importance of using non-linear models to handle the multifaceted and dynamic nature of pollution data.

The implications of this research are notable for urban air quality management, as accurate AQI predictions can inform timely actions to protect public health and improve community awareness. This study also showcases the potential of machine learning models to provide actionable insights for policymakers, who can leverage data-driven insights to design effective pollution reduction strategies. By identifying the most impactful pollutants, targeted interventions can be implemented to improve air quality more effectively.

Future Work:

To further enhance the accuracy and applicability of AQI prediction models, several future research directions are suggested:

1. **Advanced Modeling Techniques:** While this study focused on Decision Trees and Linear Regression, exploring ensemble methods like Random Forest and Gradient Boosting could provide even higher accuracy by capturing a broader range of patterns within the data. Ensemble models can reduce overfitting while maintaining predictive power, making them particularly useful in complex environmental datasets.
2. **Feature Expansion:** The inclusion of additional contextual factors, such as weather variables (e.g., temperature, humidity, wind speed) and traffic or industrial activity data, could significantly improve the model's ability to predict AQI by accounting for factors that influence pollutant dispersion and concentration. These added features would allow the model to better capture the multifactorial aspects of air quality variations.
3. **Real-Time Prediction and Monitoring System:** To support proactive air quality management, integrating the developed model into a real-time monitoring system with alert and notification capabilities could enable timely responses to sudden changes in air quality. Such a system could inform the public and local authorities, helping to mitigate exposure during periods of poor air quality. This would be particularly beneficial in high-risk areas, where pollution levels fluctuate rapidly due to traffic or industrial activities.

References:

Guttikunda, S. K., & Gurjar, B. R. (2012). Role of atmospheric modeling in air pollution exposure estimates and management in mega cities. *Environmental Pollution*, 162, 36-48.

- Discusses the significance of air pollution modeling in urban areas.

Li, X., et al. (2017). Deep air quality forecasting using hybrid deep learning frameworks. *IEEE Transactions on Knowledge and Data Engineering*, 30(12), 2203-2214.

- Examines machine learning frameworks for air quality forecasting.

Castell, N., et al. (2015). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 78, 186-195.

- Explores low-cost sensors for air quality monitoring and their applications in data-driven modeling.

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *The Lancet*, 360(9341), 1233-1242.

- Reviews the impact of various pollutants on human health, relevant to studies analyzing AQI.

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

- Provides foundational information on ensemble methods like Random Forest, often used in AQI prediction models.