# **Northeastern University**

# **ALY6110: Data Management and Big Data**

# **Prof: Daya Rudramoorthi**

# **Module 6 Final Project - Report**

# **Group 7**

**Title:**

Air Quality Analysis and Big Data Insights in the United States

**Group Members:**Harsh Anadkat

Mohammed Abdoul

**Introduction**

Air pollution remains a critical environmental and public health challenge, with carbon monoxide (CO) being a significant pollutant contributing to poor air quality. The Air Quality Index (AQI) is a widely used metric for assessing air pollution levels and potential health risks. This report examines air quality trends by analyzing CO levels and AQI data from various monitoring stations, leveraging predictive modeling techniques to improve air quality forecasting. The study follows the CRISP-DM methodology to systematically analyze and process air quality data, ensuring a structured approach from business understanding to model evaluation.

Understanding air pollution trends is crucial for developing effective mitigation strategies. CO, a byproduct of incomplete combustion processes, can lead to severe health complications, particularly in urban areas with high vehicular traffic and industrial activities. Long-term exposure to elevated CO levels is linked to respiratory diseases, cardiovascular issues, and neurological effects. Monitoring and predicting AQI trends provide actionable insights for policymakers, environmental organizations, and the general public. By employing machine learning techniques, this study aims to enhance the accuracy of AQI forecasting, thereby enabling proactive pollution control measures. Additionally, this research highlights disparities in air quality across different regions and underscores the importance of targeted interventions.

**Business Understanding**

The primary business objective of this study is to improve air quality predictions by analyzing CO concentrations and their impact on AQI. The research seeks to answer key business questions such as how CO concentration affects AQI levels across different geographic locations, what regions exhibit the highest CO concentrations and require urgent policy interventions, whether machine learning models can improve AQI forecasting accuracy compared to traditional statistical methods, and how predictive analytics can be integrated into real-time air quality monitoring systems. By answering these questions, this study provides valuable insights that can support decision-making for public health policies, environmental regulations, and resource allocation for air quality monitoring.

**Data Understanding**

The dataset used in this study was sourced from the U.S. Environmental Protection Agency (EPA) and consists of multiple air quality parameters recorded across various monitoring stations. The dataset includes over 1.8 million observations with ten key attributes related to air quality, covering the period from January 2022 to December 2022. It encompasses nationwide data, including both urban and rural regions. The data was retrieved from the official EPA repository, ensuring reliability and accuracy in the analysis.

**Data Preparation and Cleanup**

Data preprocessing was performed to ensure accuracy and consistency in the analysis. Several steps were undertaken, including handling missing values by removing incomplete records, standardizing CO concentration measurements to a uniform ppm (parts per million) scale, formatting date columns into a standardized datetime format for time-series analysis, detecting and removing extreme outliers using the interquartile range (IQR) method, and generating additional features such as average monthly AQI and seasonal pollution indices to enhance model performance. These steps ensured that the dataset was clean, structured, and suitable for predictive modeling.

**Tools and Technologies Used**

To analyze and model air quality trends effectively, various tools and technologies were utilized. Python served as the primary programming language, with data processing conducted using Pandas and NumPy. Visualization was carried out using Matplotlib, Seaborn, Dash, Plotly, and Tableau to generate insightful charts. Machine learning techniques were implemented using Scikit-learn, while Apache Spark was leveraged for handling large-scale datasets and computational efficiency.

**Dashboards and Visualizations**

A screenshot of a data analysis

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph with a blue bar

Description automatically generated

To provide a comprehensive view of air quality trends and monitoring efforts, multiple visualizations were created. The time-series line chart displays AQI fluctuations over time, highlighting seasonal trends and long-term variations, revealing that AQI levels tend to rise during the summer months due to increased pollution and wildfire activity. The monthly data completeness pie chart showcases the percentage of complete data for each month, revealing inconsistencies in data collection and identifying months with missing data to enhance monitoring station effectiveness. The geospatial heatmap visualizes CO concentration distribution across different regions, emphasizing urban areas with consistently high CO levels, reinforcing the need for targeted air quality management in densely populated regions. The boxplot of CO levels across monitoring stations identifies anomalies and pollution spikes, revealing outliers that indicate periodic surges in CO concentration, possibly due to industrial emissions or traffic congestion. The AQI vs. CO concentration scatter plot examines the correlation between CO levels and AQI values, confirming that increased CO concentrations directly impact air quality deterioration, reinforcing the importance of CO emission control strategies.

Other visualizations include the daily observation count bar chart, which presents the frequency of observations across different monitoring stations, highlighting discrepancies in data collection and inconsistencies in station coverage. The regional comparison table provides a tabular view of AQI, CO levels, and monitoring frequency across different regions, offering an overview of air quality variations and helping identify areas that require improved air quality management interventions. The CO concentration distribution histogram displays the frequency distribution of CO levels across the dataset, identifying pollution spikes. The CO level clusters over time categorize different time periods based on CO levels into low, moderate, and high pollution days using clustering techniques, enabling authorities to predict high-risk periods. Finally, the choropleth map of monitoring station coverage illustrates the distribution of air quality monitoring stations across different regions, confirming that urban areas have a higher concentration of stations while revealing monitoring gaps in rural locations.

**Key Insights from the Visualizations**

The dashboards provided several critical insights. Seasonal AQI trends indicate higher AQI levels in the summer months, suggesting a need for seasonal intervention strategies such as stricter emission regulations and public awareness campaigns. The monthly data completeness pie chart reveals gaps in data collection, emphasizing the necessity of improved monitoring station operations. Geographic disparities in air quality were evident, with certain urban areas consistently experiencing high CO concentrations, necessitating targeted regulatory measures. Monitoring station coverage gaps were identified through the choropleth map, indicating that some regions lack sufficient air quality tracking infrastructure, necessitating the expansion of monitoring efforts. Data anomalies detected through boxplots and histograms suggest temporary pollution spikes that require further investigation to determine underlying causes. Lastly, discrepancies in data collection efforts, as indicated by the daily observation count bar chart, highlight the need for standardized air quality monitoring practices.

By integrating these visualizations into an interactive dashboard, stakeholders can enhance data-driven decision-making and take proactive measures to manage air quality effectively. These insights allow policymakers and environmental agencies to prioritize interventions, allocate resources strategically, and implement measures that improve air quality monitoring systems.

**Predictive Modeling**

To predict AQI values based on CO levels and other monitoring factors, two machine learning models were developed: the Random Forest Regressor and the Decision Tree Regressor. The Random Forest Regressor is an ensemble learning method that utilizes multiple decision trees to improve accuracy and robustness. It reduces variance by averaging predictions from multiple trees, making it highly effective for air quality prediction. The Decision Tree Regressor serves as a simpler baseline model, offering high interpretability but being prone to overfitting, making it less reliable for generalization.

**Train-Test Splitting and Feature Selection**

The dataset was split into 80% training and 20% testing sets to effectively evaluate model performance. The features selected for model training included Daily Maximum 8-hour CO Concentration, Daily Observation Count, Percent Data Completeness, and Site Latitude and Longitude. Feature scaling was applied to standardize numerical variables, ensuring uniform influence across different features. This step helped enhance model performance and prevent certain features from disproportionately impacting predictions.

**Model Evaluation**

To assess the performance of the models, several evaluation metrics were used, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R²), and Cross-Validation Scores. RMSE measures the average magnitude of prediction errors, while MAE calculates the absolute differences between predicted and actual AQI values. R² assesses how well the independent variables explain the variance in AQI values, and cross-validation ensures consistency in model performance across multiple training subsets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | RMSE | MAE | R² | Cross-Validation R² |
| Random Forest | 0.0305 | 0.0006 | 0.99992 | 0.99949 |
| Decision Tree | 0.0164 | 0.0003 | 0.99998 | 0.99967 |

**Residual Analysis**

A screen shot of a graph

Description automatically generated

Residual analysis was performed to assess the distribution of prediction errors. The Random Forest model's residuals followed a near-normal distribution, indicating a strong predictive performance. However, the Decision Tree model's residuals exhibited higher variance, suggesting occasional overfitting. This confirmed that while the Decision Tree model could capture the trends in AQI, it was more susceptible to noise in the dataset.

**Feature Importance Analysis**

**A screen shot of a computer

Description automatically generated**

Feature importance analysis revealed that Daily Max 8-hour CO Concentration was the most critical predictor, contributing 99.99% of the predictive power. Other variables, such as site latitude and longitude, daily observation count, and data completeness, had minimal influence on AQI predictions. These findings reinforced the direct relationship between CO concentration and air quality, emphasizing the importance of controlling CO emissions to improve air quality.

**Model Deployment and Integration**

Deploying the predictive model in a real-world setting is essential for maximizing its impact and ensuring effective air quality monitoring. One of the key strategies for deployment is the integration of the trained Random Forest model into an interactive Air Quality Dashboard. By embedding the model within a user-friendly dashboard, real-time AQI predictions can be visualized and analyzed, allowing stakeholders to monitor air quality trends more efficiently. This dashboard can serve as a vital tool for policymakers, environmental agencies, and the general public to make informed decisions regarding pollution control measures.

Another crucial aspect of deployment is establishing an Automated Data Pipeline to ensure continuous updates of the model with newly collected air quality data. This system would allow for real-time adjustments and improvements in model accuracy by incorporating the latest observations. By automating the data flow, the model remains up to date, ensuring reliable predictions that reflect current pollution trends. Such an approach minimizes manual intervention, improving efficiency and scalability for large-scale air quality monitoring.

To further enhance the practical applicability of the model, implementing Alert Mechanisms for Pollution Spikes is recommended. These alert systems would generate notifications whenever AQI levels exceed safe thresholds, allowing authorities to take immediate action. For example, if the model predicts an imminent increase in pollution levels beyond regulatory limits, environmental agencies can deploy mitigation strategies such as issuing public health advisories or enforcing temporary restrictions on high-emission activities. The ability to proactively address pollution spikes can significantly reduce the adverse effects of poor air quality on public health.

**Conclusion**

This study demonstrates that AQI predictions can be effectively modeled using machine learning techniques, particularly the Random Forest Regressor, which outperformed the Decision Tree model in terms of accuracy and generalization. By leveraging machine learning, we have enhanced air quality forecasting, providing a more robust and data-driven approach to environmental monitoring. These findings hold significant implications for policymakers, urban planners, and environmental agencies striving to mitigate air pollution and improve public health outcomes.

The research confirms that CO levels serve as the primary determinant of AQI variations, emphasizing the importance of stringent CO emission controls. The superior performance of the Random Forest model highlights its potential as a reliable tool for real-time AQI forecasting, making it a preferable choice for deployment in air quality monitoring systems. Additionally, the study underscores seasonal and regional variations in pollution levels, reinforcing the need for targeted policy interventions that account for periodic fluctuations in air quality. Moreover, gaps in monitoring infrastructure have been identified, indicating that certain regions lack sufficient sensor coverage, which may compromise the accuracy of air quality assessments. Addressing these gaps through expanded monitoring efforts will be essential for enhancing the effectiveness of predictive models.

**Future Work**

Future research efforts should focus on several key areas to further improve the predictive accuracy and applicability of AQI forecasting models. One critical direction is the Integration of Additional Pollutants such as NO2, PM2.5, and SO2, which also significantly impact air quality. By incorporating a broader range of pollutants, the model can provide a more comprehensive assessment of air quality trends, leading to better-informed policy decisions.

Another essential aspect for improvement is the Incorporation of Meteorological Variables such as temperature, humidity, and wind speed. These meteorological factors play a crucial role in air pollution dispersion and accumulation, making their inclusion in the model highly valuable. By integrating weather data, the predictive power of the model can be significantly enhanced, allowing for more accurate forecasts.

Lastly, a major area of future implementation involves Real-Time Deployment of the model into a live monitoring system. A fully operational real-time system would enable continuous tracking of air quality and provide instant feedback on pollution levels. Such deployment would allow for more effective intervention strategies, ensuring that environmental agencies and policymakers can respond proactively to emerging pollution threats. By developing and refining real-time monitoring capabilities, air quality prediction models can be transformed into indispensable tools for safeguarding public health and promoting sustainable environmental policies.

**References**

* U.S. Environmental Protection Agency. (2022). Air Quality Data. Retrieved from <https://www.epa.gov/airdata>
* Cohen, A. J., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution. The Lancet, 389(10082), 1907–1918.
* Brook, R. D., et al. (2010). Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement from the American Heart Association. Circulation, 121(21), 2331–2378.