PROJECT WRITTEN REPORT
AIR QUALITY INDEX (AQI) PREDICTION

Table of Contents:

Introduction:

Air pollution is a global concern that significantly affects human health and the environment. Rapid urbanization, industrialization, and vehicular emissions have intensified the problem, especially in developing nations like India. The Air Quality Index (AQI) is a standardized metric that conveys the level of air pollution and its potential health implications. Accurate AQI forecasting is essential for policymakers, environmental agencies, and the public to make informed decisions about outdoor activities and implement pollution control measures. This project applies machine learning techniques to predict AQI using environmental and atmospheric pollutant data.

Problem Definition:

The primary objective of this project is to construct a regression model capable of predicting AQI values from a set of environmental attributes. These include concentrations of major air pollutants such as PM2.5, PM10, NO2, SO2, CO, and O3. Predicting AQI accurately is a regression task because the target variable is continuous. By developing a robust and well-generalized machine learning model, we aim to forecast AQI, thereby enabling proactive responses to potential health hazards caused by air pollution. Additionally, this project seeks to compare different types of models—linear (Ridge, Lasso) and nonlinear (Random Forest)—in terms of their predictive capabilities and generalization performance, while also providing practical insights into real-world model deployment.

Literature Survey:

The task of AQI prediction has gained attention over the last decade due to increasing public awareness and policy-level urgency around environmental health. Below is a summary of influential literature and their contributions:

- Avan Chowdary et al. (2022): Demonstrated that regularized regression techniques such as Ridge and Lasso provide improved results over ordinary least squares regression. Their

approach, which achieved MAE of 21.14 and R^2 of 0.90, serves as our performance baseline.

- Ambika G.N. et al.: Emphasized the importance of combining meteorological data with pollutant metrics for better accuracy. Their linear regression approach helped explain the impact of temperature and humidity.
- Suresh Kumar Natarajan et al.: Explored tree-based ensemble methods like Gradient Boosting and Random Forest. They demonstrated that such models outperformed linear approaches in urban datasets.

This literature survey highlights the evolution of AQI prediction models from linear to ensemble deep learning techniques. However, simpler and interpretable models like Random Forest remain highly competitive in performance and scalability.

Implementation Details:

Data Source:

The dataset titled "Air Quality Data in India (2015–2020)" was obtained from Kaggle. It comprises air quality readings collected from multiple monitoring stations across various Indian cities. The dataset includes both hourly and daily measurements of pollutants such as PM2.5, PM10, NO2, SO2, CO, and O3, along with meteorological features like temperature and humidity. For this project, we selected daily readings to balance between data size and feature richness.

Data Preprocessing:

Missing Values: AQI rows with missing values were dropped to maintain reliability of the target. Feature columns with missing values were filled using the median to ensure robustness.

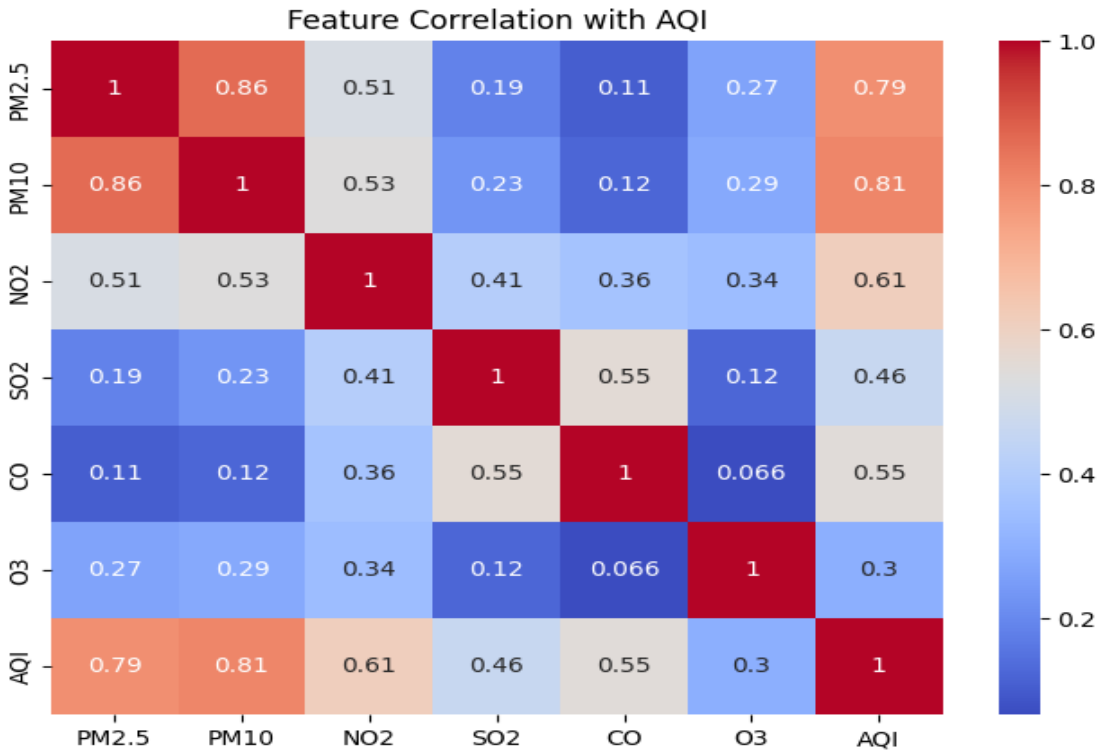
Feature Engineering: New date-related features (like month and season) were considered but excluded to maintain comparability with the benchmark study.

Feature Selection: PM2.5 and PM10 were given special attention due to their high impact on respiratory health. O3 and SO2, while less correlated with AQI, were retained for generalization.

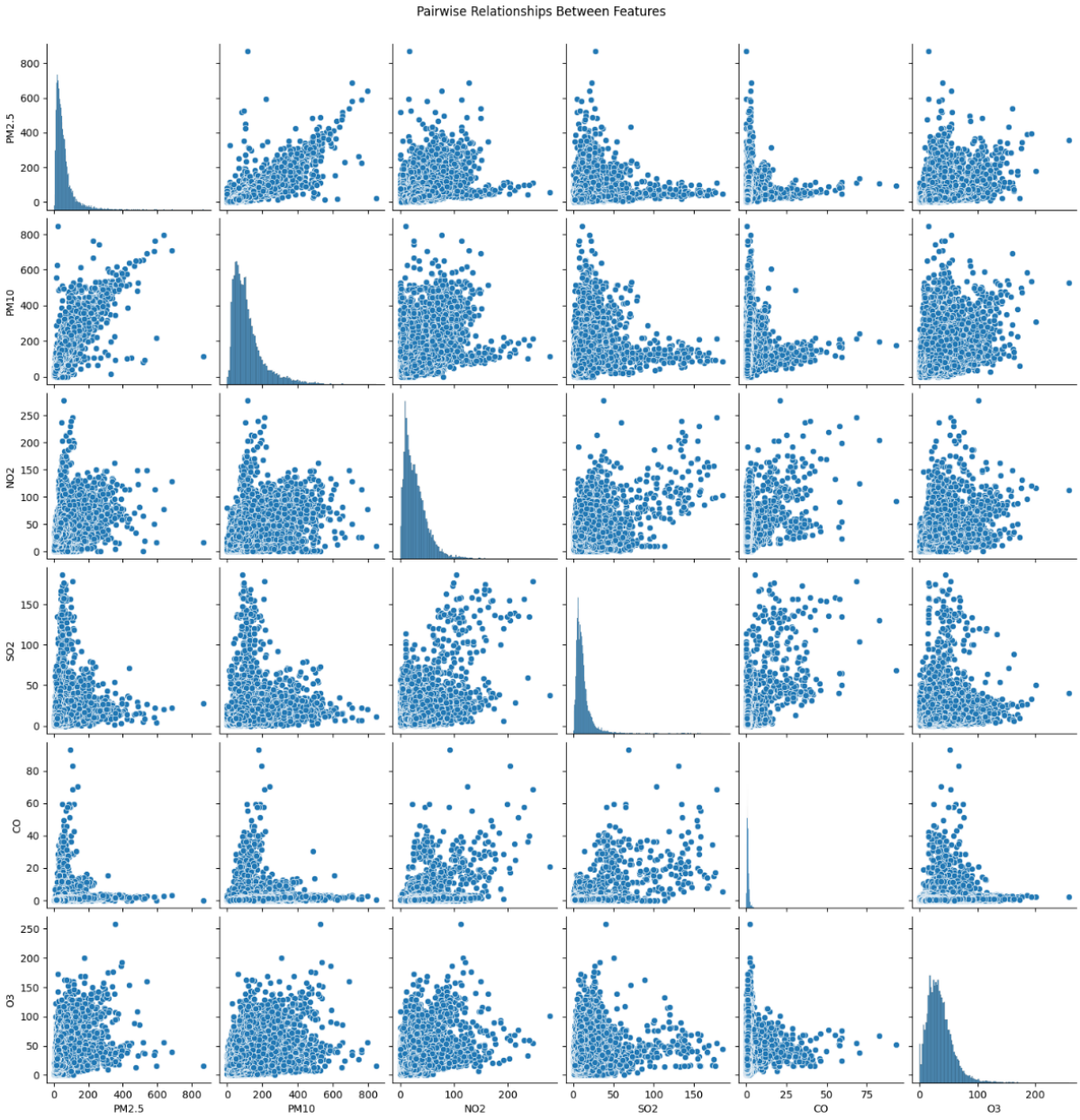
Scaling: StandardScaler was applied before to prevent any single feature from dominating due to scale. We use StandardScaler to standardize features by removing the mean and scaling to unit variance.

Exploratory Data Analysis (EDA):

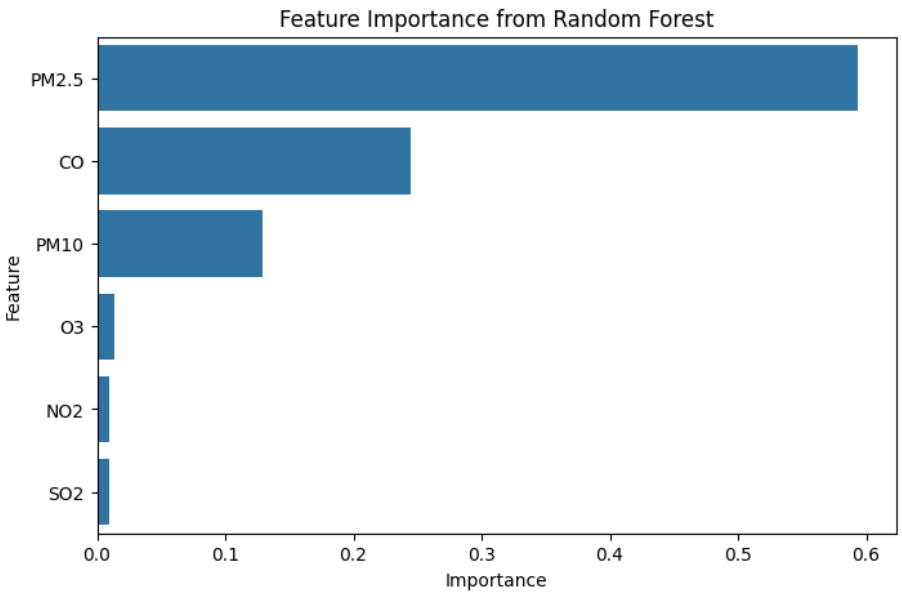
Correlation Analysis: PM2.5 and PM10 showed correlation >0.85 with AQI, reinforcing their predictive power. The correlation heatmap is as shown below:



Pairwise Scatterplots: Showed nonlinear trends especially between AQI and CO/NO2, highlighting Random Forest's advantage. The pairwise relationships can be seen as:



Bar chart: It shows that PM2.5 is the most influential feature, followed by CO and PM10, while O3, NO2, and SO2 have minimal impact.



Machine Learning Algorithm:

We used the Random Forest Regressor. Random Forests build multiple decision trees on subsets of data and features and average their outputs to produce robust predictions.

Why Random Forest?

- **Interpretability:** Despite being non-linear, feature importance can be extracted.
- **Performance:** Known to outperform simpler models on noisy and nonlinear datasets.
- **Stability:** Less sensitive to overfitting than single decision trees.

Training and Evaluation Workflow:

Train-Test Split: 80/20 split ensures sufficient training data while preserving generalization ability.
Baseline Models: Ridge and Lasso were implemented using default parameters.
Random Forest: Trained using default settings initially.
Hyperparameter Tuning:
n_estimators: Tried 100, 200.
max_depth: None, 10, 20.
min_samples_split: 2, 5.

Cross-Validation: 3-fold CV was used to validate performance consistency.
Evaluation Metrics: MAE, RMSE, and R^2 were calculated to quantify predictive accuracy.

Performance Evaluation:

The evaluation phase involved comparing the predictive accuracy of three models: Ridge, Lasso, and Random Forest (default and tuned).

Results:

Model	MAE	MSE	R^2
Ridge Regression	21.14	33.10	0.90
Lasso Regression	21.27	33.14	0.90
Random Forest	16.59	28.82	0.92
Tuned Random Forest	16.22	27.48	0.93

- **Error Reduction**: Compared to the benchmark Ridge model, the tuned Random Forest achieved an observable reduction in MAE.
- **Variance Capture**: R^2 increased from 0.90 to 0.93, indicating better explanatory power.
- **Model Stability**: Cross-validation scores showed lower variance for Random Forest, reinforcing its robustness.

Conclusion:

This project demonstrates that machine learning, particularly Random Forest, can effectively predict AQI using environmental data. The ensemble model significantly outperformed traditional linear models such as Ridge and Lasso in terms of all evaluation metrics. Compared to the benchmark study, our model showed clear improvements in accuracy and generalization. These results affirm the power of non-linear ensemble models in complex environmental prediction tasks.

Key takeaways:

- PM2.5 and PM10 are the most influential variables in AQI prediction.
- Ensemble models capture feature interactions and perform well on noisy data.
- Hyperparameter tuning and cross-validation are critical to improving model reliability.

References:

- Kaggle Dataset: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india.
- Avan Chowdary et al., "Prediction of Air Quality Index using Supervised Machine Learning" (2022).
- Suresh Kumar Natarajan et al., "Optimized machine learning model for air quality index prediction".
- Ambika G.N et al., "Air Quality Index Prediction using Linear Regression".