**Air Quality Analysis Report**

**Abstract**

Air pollution is a major environmental and public health concern worldwide. This study analyzes air quality data from various global cities with the goal of predicting the Air Quality Index (AQI) using different machine learning algorithms. The comparison of models demonstrated that Random Forest provides the most accurate predictions of AQI.

**Background**

The Air Quality Index (AQI) is a standardized indicator that summarizes the concentration of pollutants such as CO, CO2, NO2, SO2, O3, PM2.5, and PM10 into a single value reflecting air quality. Accurate AQI prediction enables timely interventions to protect public health. This study uses real-world air quality measurements to examine the relationship between key pollutants and AQI and applies statistical and machine learning methods to predict AQI levels.

**Methods**

**Dataset**

The dataset contains air quality measurements from six major cities: Brasilia, Cairo, Dubai, London, New York, and Sydney. Key variables include CO, CO2, NO2, SO2, O3, PM2.5, PM10, AQI, and timestamps.

**Preprocessing**

* Missing values for CO2 were imputed using the median.
* Timestamps were converted to datetime format, with additional features extracted (Year, Month, Day, Hour).
* Categorical city names were encoded numerically.
* Duplicate records were removed.
* Outliers were identified using the Interquartile Range (IQR) method.

**Exploratory Data Analysis (EDA)**

* Boxplots, histograms, and heatmaps were used to explore data distribution and correlations.
* Scatterplots visualized relationships between pollutants and AQI.
* Time series plots tracked pollutant evolution over time.

**Machine Learning Models**

The following regression models were trained and evaluated to predict AQI:

* Linear Regression
* Decision Tree Regressor
* Random Forest Regressor
* XGBoost Regressor

Model performance was assessed using:

* Root Mean Squared Error (RMSE)
* Coefficient of Determination (R²)

**Results**

| **Model** | **RMSE** | **R²** |
| --- | --- | --- |
| Linear Regression | 11.22 | 0.8142 |
| Decision Tree | 10.86 | 0.8258 |
| Random Forest | 7.65 | 0.9137 |
| XGBoost | 8.43 | 0.9000 |

**Feature Importance Analysis** revealed the following as the most influential variables for AQI prediction:

* **PM2.5** and **PM10** were the most critical factors.
* Followed by **O3**, **CO2**, and **NO2**.

Scatterplots of actual vs. predicted values confirmed that complex models, particularly Random Forest and XGBoost, provide accurate AQI predictions.

**Conclusions**

This study demonstrates the potential of machine learning models in accurately predicting the Air Quality Index based on pollutant concentrations. The Random Forest model achieved the best performance, making it a suitable tool for air pollution monitoring and forecasting applications. Feature importance analysis highlighted **PM2.5** and **PM10** as the most significant contributors to AQI levels. Future work could focus on hyperparameter tuning, incorporating additional data sources, and exploring deep learning techniques to further enhance prediction accuracy.