Air Quality Index (AQI) Prediction

1. Introduction

This project aims to predict the Air Quality Index (AQI) for Lahore using historical weather and pollution data from 2019 to 2024. We explored multiple machine learning models, performed extensive feature engineering, and optimized hyperparameters to improve predictive performance.

---

2. Challenges Faced

During the project, we encountered several challenges:

- **High Multicollinearity**: Many features had a high Variance Inflation Factor (VIF), leading to multicollinearity.
- **Feature Selection**: Identifying the most relevant features without losing critical information.
- **Hyperparameter Tuning**: Finding the best parameters for models like Random Forest and XGBoost.
- **Overfitting**: Ensuring that models generalized well on unseen data rather than memorizing training patterns.
- **Performance Variation**: Different models performed variably, requiring multiple iterations to identify the best.

---

3. Feature Engineering

Feature engineering played a crucial role in improving model performance. The key steps included:

3.1 Feature Creation

- **Lag Features**: Created aqi_lag_1, aqi_lag_5, and aqi_lag_10 to capture temporal dependencies.
- **Rolling Averages**: Used aqi_3day_avg and aqi_5day_avg to smooth short-term fluctuations.

- **Weather Interactions**: Generated wind_humidity (wind speed × humidity) and wind_temp (wind speed × temperature) to capture interdependencies.
- **Squared Features**: Added avg_wind_speed_sq and humidity_sq to capture non-linear relationships.

3.2 Feature Selection

To combat multicollinearity, we:

1. Computed **VIF scores** and dropped features with extreme values.
2. Used **correlation analysis** to remove highly correlated variables.
3. Retained only **essential features** with independent predictive power.

---

4. Models Used and Why

We experimented with multiple regression models to compare their effectiveness:

Model	Reason for Use
Linear Regression	Baseline model to understand linear relationships.
Lasso Regression	Helps with feature selection by applying L1 regularization.
Random Forest	Handles non-linearity well, reduces overfitting.
XGBoost	Boosting-based ensemble method for high accuracy.

We used **Random Forest and XGBoost** for their superior handling of **non-linearity and feature interactions**.

---

5. Model Evaluation & Results

We evaluated models using:

- **R^2 Score**: Measures the variance explained by the model.
- **Mean Absolute Error (MAE)**: Measures average prediction error.
- **Root Mean Squared Error (RMSE)**: Penalizes larger errors more heavily.

Final Model Results

Model	R^2 Score	MAE	RMSE
Linear Regression (Before Feature Engineering)	0.4127	0.2347	0.3335
Linear Regression (After Feature Engineering)	0.8299	0.1258	0.1814
Lasso Regression	0.8328	-	-
Random Forest (Before Hyperparameter Tuning)	0.8835	0.1109	0.1501
Random Forest (After Hyperparameter Tuning)	0.8879	0.1100	0.1473
XGBoost (Cross-Validation)	0.8316	-	-

---

6. Conclusion & Next Steps

We successfully built an AQI prediction model with high accuracy using Random Forest (Optimized), which achieved the best performance with an R^2 of 0.8879 and the lowest RMSE.

Next Steps:

1. **Expanding Data Sources**: Incorporating real-time AQI and satellite data.
2. **Further Hyperparameter Optimization**: Fine-tuning model parameters for even better accuracy.
3. **Deploying the Model**: Creating a real-time AQI forecasting system.

This project highlights the importance of feature engineering, model selection, and hyperparameter tuning in improving AQI prediction accuracy.