**Coding Implementation**

The coding contributions to the air quality prediction project encompass critical aspects of data preprocessing, feature engineering, and machine learning model development. These contributions include handling missing values, scaling datasets, creating new features, and implementing machine learning algorithms like Random Forest, SVM, and LSTM. Each component is thoroughly documented, with a clear focus on enhancing the model's predictive accuracy. All scripts have been organized and committed to the GitHub repository, ensuring transparency and ease of review.

**Data Preprocessing**

***Handling Missing Values***

To address missing values in the dataset, a combination of imputation strategies was employed. Numerical columns with missing values were handled using median imputation, ensuring robustness against outliers, while categorical variables were processed using mode imputation to maintain data consistency. This approach preserved the dataset’s integrity and minimized data loss, ensuring optimal conditions for model training.

***Normalization/Scaling***

To prepare the dataset for machine learning algorithms, all numerical features were scaled using the Min-Max normalization technique. This method transformed each feature’s range to [0, 1], ensuring uniform input scales and improving algorithm convergence rates. The consistent input ranges facilitated fair contributions of features to the models, particularly in algorithms sensitive to input magnitude, such as Support Vector Machines and Neural Networks.

***Outlier Detection and Removal***

Outliers in numerical features were identified using the Interquartile Range (IQR) method, which flags values lying 1.5 times above the upper quartile or below the lower quartile. These outliers were carefully removed to reduce noise and enhance model accuracy without compromising the dataset’s core structure.

**Feature Engineering**

***Feature Transformations***

To enhance the model's ability to capture temporal patterns, time-based features were extracted from the dataset. Variables such as ‘hour of the day’, ‘day of the week’, and ‘season’ were derived from the timestamp data to identify trends in air quality fluctuations over time. These transformations allowed the model to incorporate periodic patterns, improving its accuracy in predicting pollution levels at specific times.

***New Feature Creation***

To address complex interactions between environmental factors, new composite features were engineered. For example, a ‘heat index’ variable was created by combining temperature and humidity values, which are critical in understanding pollution levels during extreme weather. Additionally, the ‘pollutant ratio’ (e.g., PM2.5/PM10) was introduced to highlight the proportional relationship between fine and coarse particulate matter, further refining the dataset for improved model predictions.

**Model Training and Evaluation**

***Model Implementation***

The project employed a combination of machine learning and deep learning algorithms to maximize predictive performance. Random Forest was selected for its robustness to overfitting, ease of interpretability, and strong performance with structured data. Support Vector Machine (SVM) was chosen due to its capability to model complex, non-linear relationships in the dataset. Additionally, Long Short-Term Memory (LSTM) networks were incorporated to capture temporal dependencies and trends in air quality over time, leveraging the sequential nature of the data. These models were carefully chosen to balance accuracy, computational efficiency, and adaptability to different data characteristics.

Each algorithm was implemented using Python with libraries such as scikit-learn for Random Forest and SVM, and TensorFlow/Keras for LSTM. Hyperparameter tuning was conducted for all models to optimize performance, employing techniques like grid search for Random Forest and kernel optimization for SVM.

***Evaluation Metrics***

To assess the models’ predictive accuracy and reliability, multiple evaluation metrics were employed. The Root Mean Squared Error (RMSE) was used to quantify the average magnitude of prediction errors, focusing on penalizing larger deviations. Mean Absolute Error (MAE) was chosen to provide an intuitive measure of the average error magnitude, irrespective of direction. Additionally, the R² score (coefficient of determination) was utilized to evaluate the proportion of variance in air quality levels explained by the model.

These metrics were aligned with the project’s goal of delivering an accurate and interpretable prediction system. RMSE and MAE offered insight into the model’s practical applicability, while R² provided a measure of overall model reliability. Collectively, they ensured a thorough and objective evaluation of model performance.