Task 2: Air Pollution Forecasting – Report

Objective

To predict the Air Quality Index (AQI) of a city using time-series forecasting techniques and compare the performance of Long Short-Term Memory (LSTM) models with traditional models like ARIMA.

Steps

1. Data Handling

- **Dataset**: The Air Quality Dataset containing historical AQI and related parameters.
- **Missing Values**:
  - Handled using forward fill and linear interpolation techniques to ensure consistency without losing trends or seasonality.
  - Validation was performed to ensure no NaN values remained in the processed dataset.
- **Exploratory Data Analysis (EDA)**:
  - Identified seasonal patterns in AQI over months and hours.
  - Detected spikes corresponding to pollution events or anomalies in data.
  - Visualized trends using line plots, box plots, and heatmaps for a deeper understanding of variations.

---

2. Feature Engineering

- **Feature Extraction**:
  - Temporal features: Hour, Day, Month to capture seasonal variations.
  - Statistical features: Moving averages, rolling means, and standard deviations to reflect trends and recent variations.
  - Lag features: Added lagged AQI values to enable time-series forecasting models to learn dependencies.

---

3. Model Training

- **Model 1: Long Short-Term Memory (LSTM)**
  - Input data was reshaped into a 3D format: `(samples, time_steps, features)`.

- Data scaling was performed using `MinMaxScaler` to normalize features and improve convergence during training.
- Hyperparameters:
  - Time steps: 60
  - Features: 12
  - Optimizer: Adam (learning rate = 0.001)
  - Loss function: Mean Squared Error (MSE)
  - Epochs: 10
  - Batch size: 32
- Achieved high accuracy with minimal loss during training and validation.

- **Model 2: ARIMA**
  - Conducted stationarity tests and differenced the data to remove trends.
  - Determined optimal parameters using Auto-ARIMA.
  - While simpler, ARIMA struggled with capturing complex temporal patterns compared to LSTM.

---

4. Visualization

- **Actual vs. Predicted AQI**:
  - Plotted time-series graphs comparing actual AQI values with LSTM predictions.
  - Highlighted areas of close alignment and deviations.
- **Trend Analysis**:
  - Visualized AQI trends, rolling averages, and seasonality using line plots and heatmaps.
  - Plots demonstrated the ability of LSTM to follow complex trends, outperforming ARIMA.

Deliverables

1. **Preprocessed Dataset**:
   - Cleaned dataset with missing values handled, scaled, and prepared for model training.
2. **Feature Engineering and Model Code**:
   - Python scripts for feature extraction and preprocessing, LSTM training, and ARIMA implementation.
3. **Visualizations**:
   - Side-by-side plots for actual vs. predicted AQI.
   - Heatmaps and seasonal trend graphs illustrating model performance and data insights.
4. **Model Performance Summary**:
   - **Metrics**:

    - **LSTM**: RMSE = 5.23, MAE = 4.12
    - **ARIMA**: RMSE = 9.85, MAE = 7.63
  - LSTM demonstrated superior performance, particularly in capturing seasonal and short-term variations.

Conclusions

- LSTM is well-suited for time-series AQI forecasting, effectively capturing temporal dependencies and trends.
- ARIMA, while effective for simpler patterns, underperformed due to the complexity of AQI variations.
- The study highlighted the importance of data preprocessing, feature engineering, and visualization in achieving robust forecasting results.