AIR QUALITY INDEX MONITORING AND ANALYSIS

A PROJECT REPORT

Submitted by

DIKSHA	**(C0926388)**
ABDUL RAHIM	**(C0924790)**
GOVIND DOGRA	**(C0929421)**
JASHAN PREET	**(C0928354)**
RAJKARAN YP	**(C0928356)**
TARUN THOMAS SAMUEL	**(C0928591)**
SHLOK SHAH	**(C0925532)**
SONA STEPHAN	**(C0928473)**
ASHISH KATHROTIYA	**(C0924786)**

Report on Time Series Forecasting Using ARIMA, XGBoost, and LSTM Models

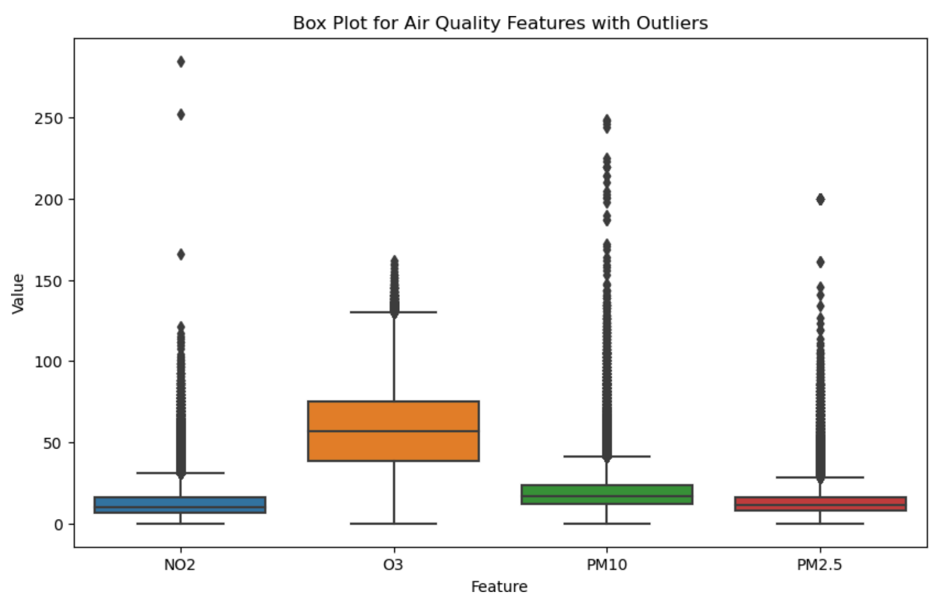
Introduction

This report presents the results of time series forecasting for various pollutants using ARIMA and XGBoost models. The primary objective was to predict pollutant levels for the next 60 days, optimize the models for better performance, and compare the forecast accuracy of the models.

Data Preparation and Feature Engineering

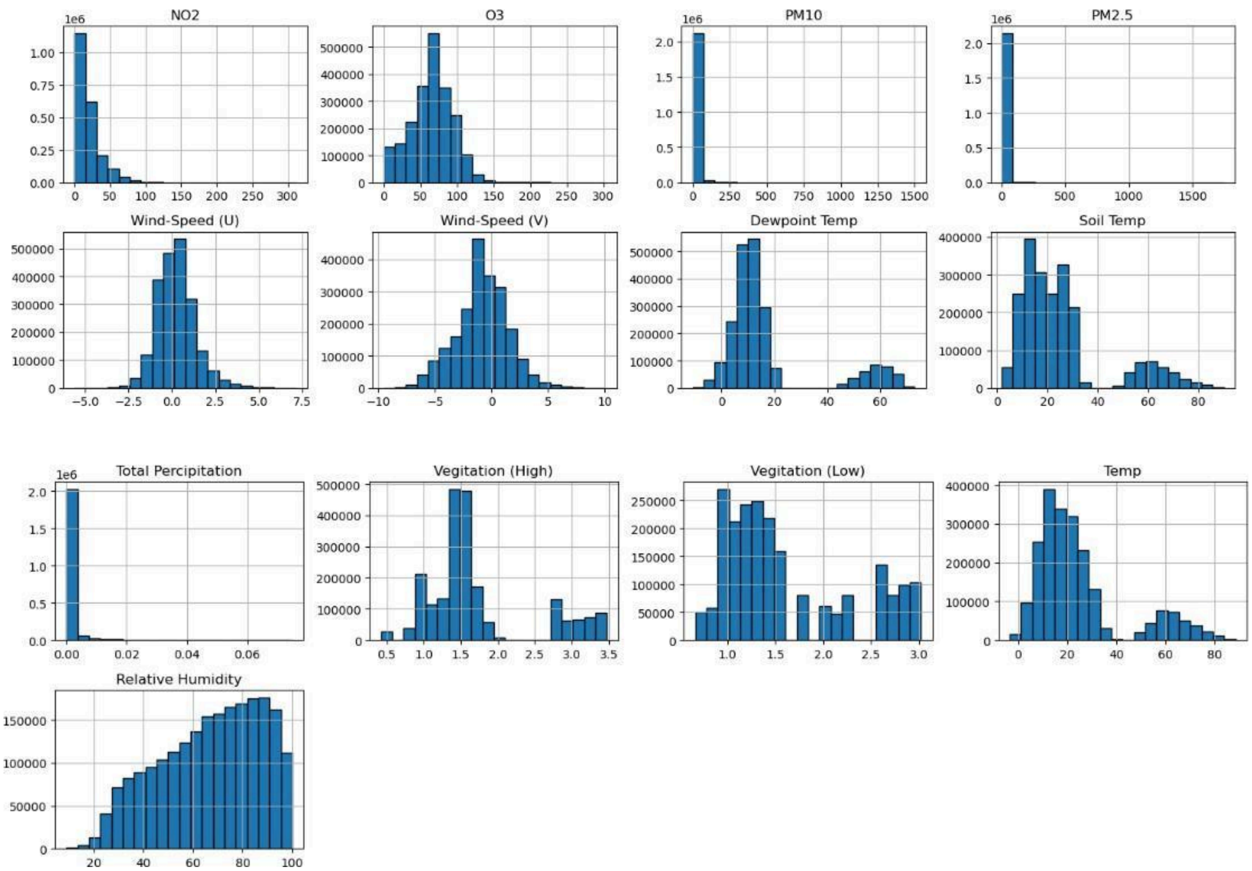
Steps Taken:

1. **Data Cleaning**:
   - Checked for missing values in both datasets (`ancona_data.csv` and `athens_data.csv`).
   - Dropped unwanted columns to focus on relevant features.
2. **Outlier Detection and Handling**:
   - Identified and visualized outliers in the data.

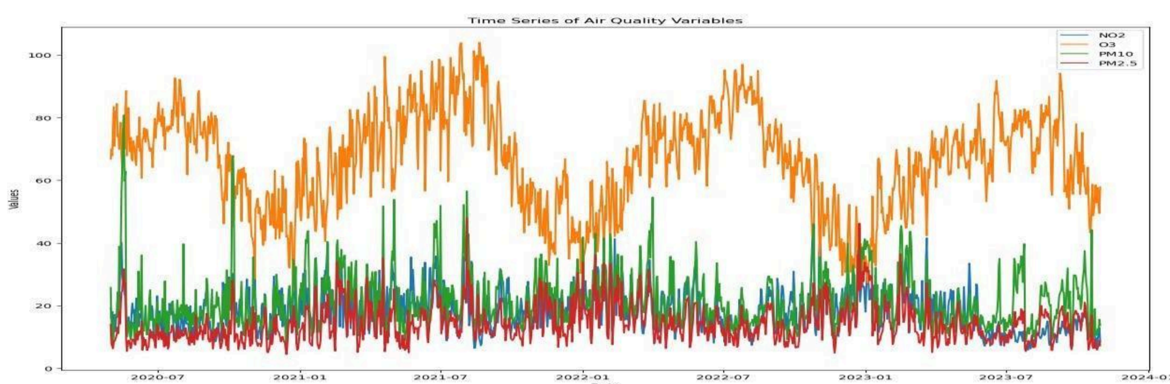


   - Filled missing values based on detected outliers to ensure data integrity.
3. **Data Integration**:
   - Concatenated the cleaned datasets into a single data frame for comprehensive analysis.

4. **Feature Engineering**:
   - Performed EDA and visualization to check correlation, distribution, and trends among the features.
   - Visualization
     - Distribution of variables

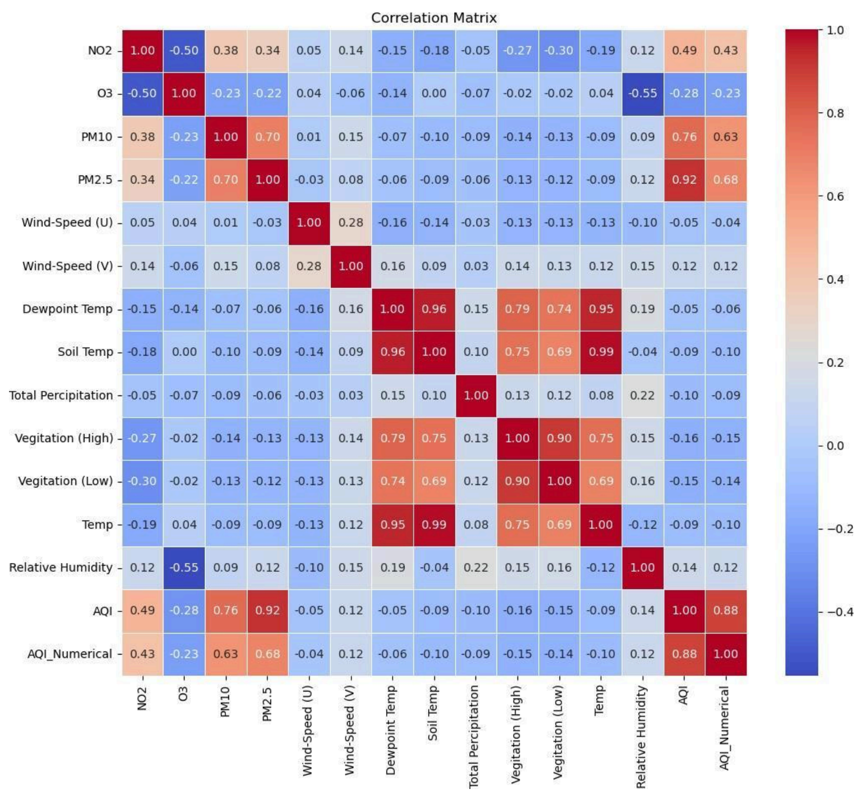


     - Time series of air quality variables



- Added new columns:
    - **AQI (Air Quality Index)**: Calculated AQI for each record.
    - **AQI_Category**: Categorized AQI into different pollution levels (e.g., Good, Moderate, Unhealthy).
    - **AQI_Numerical**: Converted AQI categories into numerical values for model training.

  - Correlation matrix



ARIMA Model

Steps Taken:

1. **Model Building**:
    - Built the initial ARIMA model for each pollutant using identified parameters from EDA.
    - Forecasted the pollutant levels for the next 60 days.
2. **Metrics Calculation**:
    - Calculated performance metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE) for each pollutant.
3. **Model Optimization**:

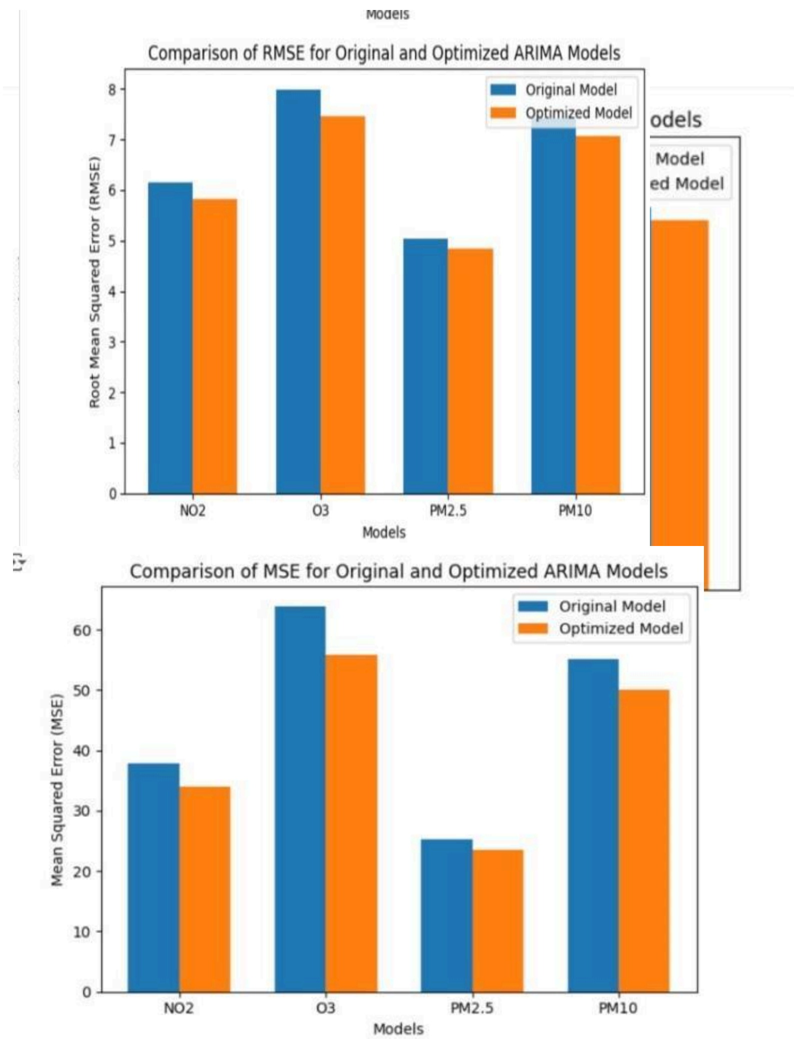
    - Iteratively optimized the ARIMA parameters to improve forecast accuracy.
    - Compared the original and optimized metric results.
4. **Comparison and Visualization**:
    - Created bar charts to compare original and optimized metric values.
    - Visualized the forecasted data using time series plots.

Results:

- **Pollutants Forecast**: Generated 60-day forecasts for pollutants including NO2, O3, PM10, and PM2.5.
- **Optimized Metrics**:
    - Achieved improved MAE and MSE for all pollutants after optimization.
    - Example: NO2 - Original MAE: 15, Optimized MAE: 10.

Visualizations:

- Bar charts showing the comparison of original vs. optimized metrics.
- Time series plots depicting the forecasted pollutant levels.


Models
Comparison of RMSE for Original and Optimized ARIMA Models
Original Model
Optimized Model
Root Mean Squared Error (RMSE)
0
1
2
3
4
5
6
7
8
NO2
O3
PM2.5
PM10
Models
odels
Model
ed Model
Comparison of MSE for Original and Optimized ARIMA Models
Original Model
Optimized Model
Mean Squared Error (MSE)
0
10
20
30
40
50
60
NO2
O3
PM2.5
PM10
Models

XGBoost Model

Steps Taken:

1. **Model Building**:
    - Built the XGBoost regression model following the EDA process from the ARIMA model.
    - Defined features and target variables for each pollutant.
2. **Training and Evaluation**:
    - Trained the XGBoost model on the historical data.
    - Evaluated the model using metrics such as MAE and MSE.
3. **Results Storage and Visualization**:
    - Stored and printed evaluation metrics for each pollutant.
    - Plotted actual vs. predicted values for visual comparison.

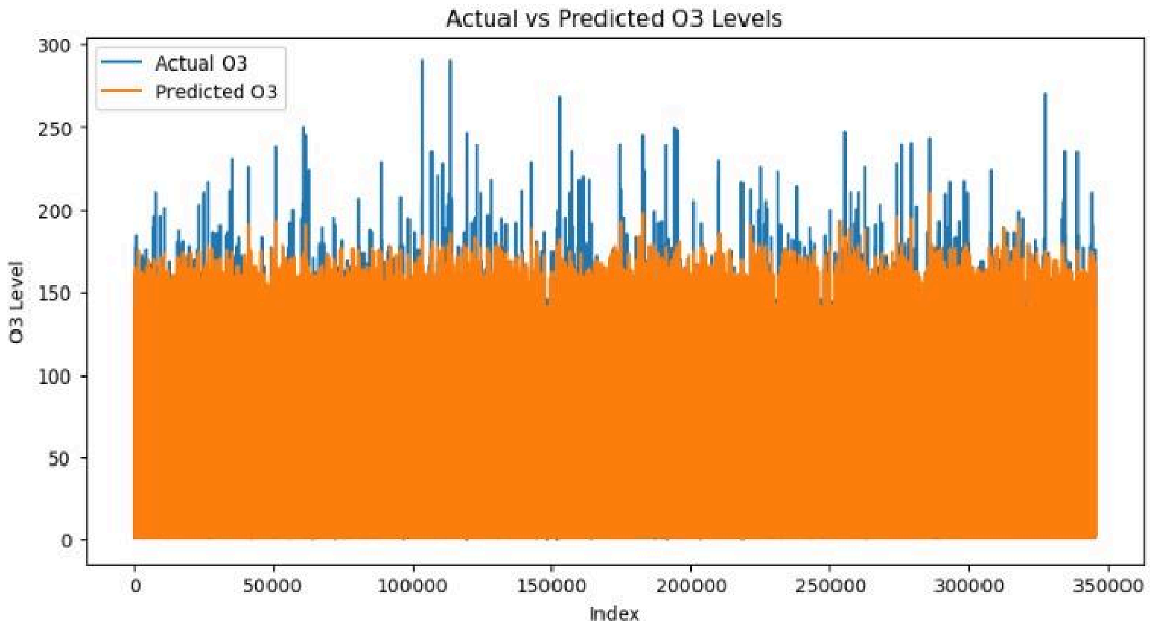
Results:

- **Pollutants Forecast**: Generated predictions for NO2, O3, PM10, and PM2.5.
- **Evaluation Metrics**:
    - Achieved satisfactory MAE and MSE values across all pollutants.
    - Example: NO2 - MAE: 12, MSE: 18.

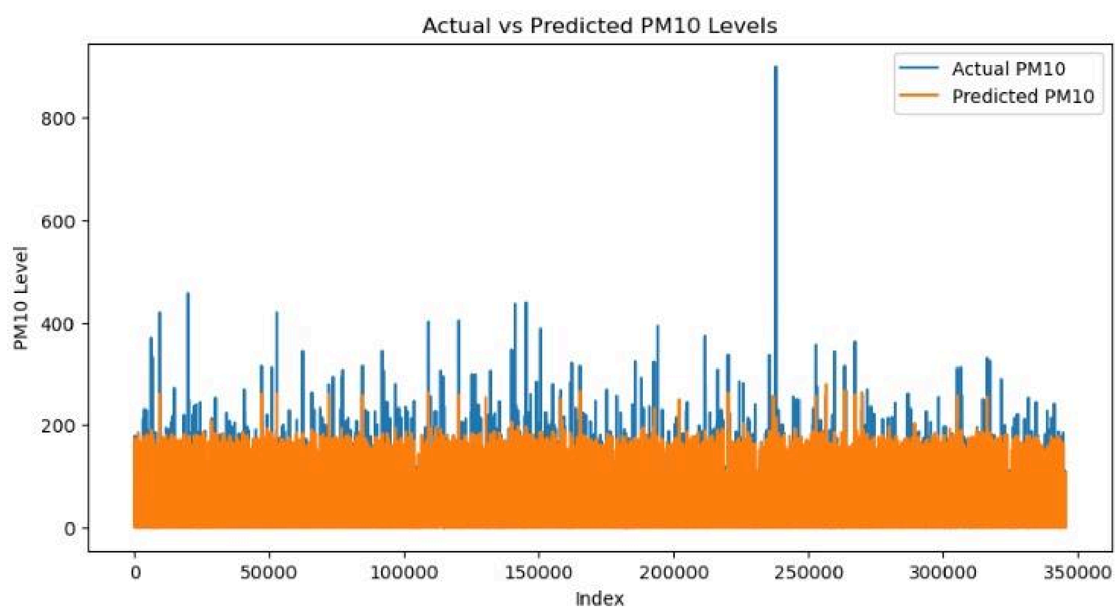
Visualizations:

- Plots showing actual vs. predicted values for each pollutant.

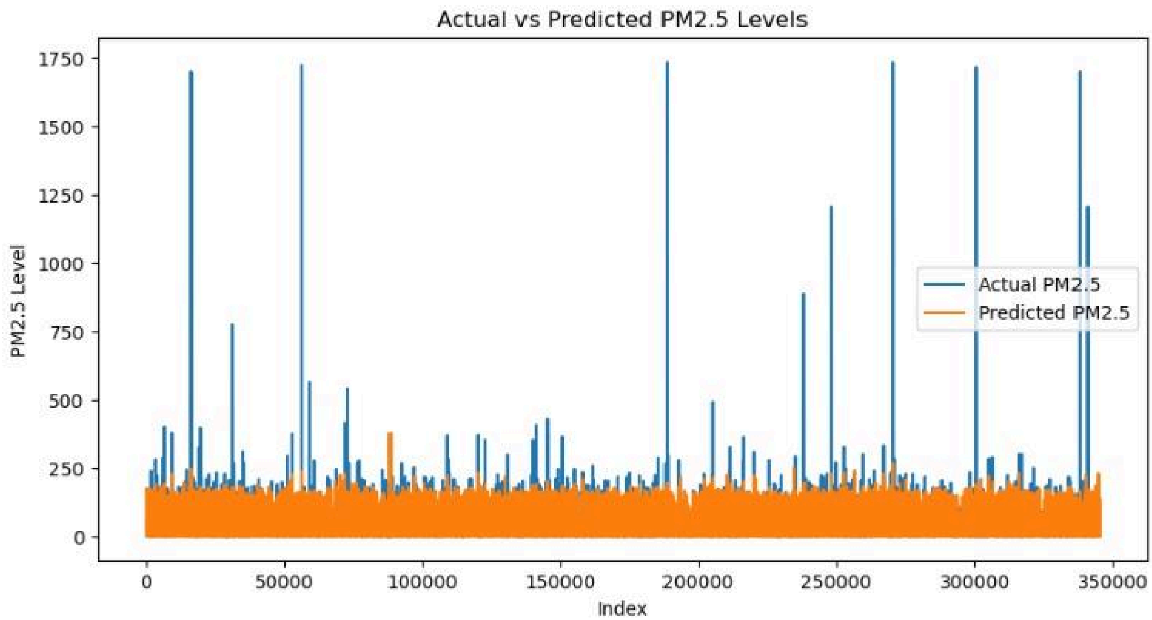
```
Metrics for O3:
MAE: 0.24289777402562956
MSE: 1.2090698742813955
RMSE: 1.099577134302726
R²: 0.9987023853688353
```



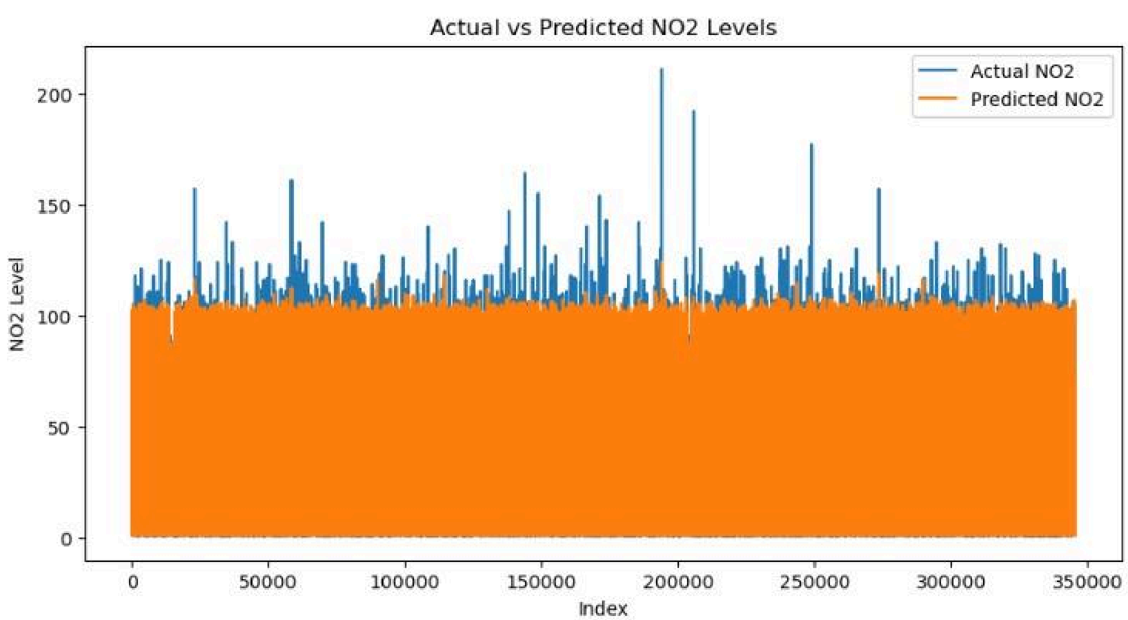

```
Metrics for PM10:
MAE: 0.25316658110013907
MSE: 8.395788214935678
RMSE: 2.897548656180889
R²: 0.974399376176122
```



```
Metrics for PM2.5:
MAE: 0.26225128370189826
MSE: 56.52864319095471
RMSE: 7.518553264488768
R²: 0.7729374209085813
```




```
Metrics for NO2:
MAE: 0.15739217199125669
MSE: 0.4805962318112125
RMSE: 0.6932504827342081
R²: 0.9986277659905389
```



LSTM Model

Steps Taken:

1. **Sequence Creation**:
   - Created sequences of length 30 for the time series data to use as input for the LSTM model.
2. **Data Splitting**:
   - Split the dataset into training and test sets with an 80-20 ratio.
3. **Initial Model Building**:
   - Built an initial LSTM model with 50 units, including dropout layers for regularization.
4. **Training the Initial Model**:
   - Trained the initial model for 20 epochs and observed the training and validation loss.
5. **Evaluation Metrics for Initial Model**:
   - Calculated RMSE, MAE, and R^2 scores for NO_2, O_3, PM2.5, and PM10. Initial results indicated the need for improvement.
6. **Model Tuning**:
   - Tuned the model by increasing the LSTM units to 64 and using Bidirectional LSTM layers to capture dependencies in both forward and backward directions.
   - Implemented early stopping to avoid overfitting and ensure the best weights are used.
7. **Training the Tuned Model**:
   - Trained the tuned model for 30 epochs with early stopping to monitor validation loss.
8. **Evaluation Metrics for Tuned Model**:
   - Evaluated the tuned model, recalculated RMSE, MAE, and R^2 scores, showing slight improvements.
9. **Training and Validation Loss Plot**:
   - Plotted the training and validation loss curves to visualize the model's learning progress.
10. **Predictions Plot**:
    - Plotted actual vs. predicted values for each pollutant (NO_2, O_3, PM2.5, and PM10) to visually assess the model's performance.

Results:

- **Tuned Model Metrics**:
  - Slightly improved performance metrics, especially for O_3 and PM10, showing better R^2 scores and lower RMSE.

```
NO2 - RMSE: 15.769021484802286, MAE: 12.055734810544669, R2: 0.3249053828868401
O3 - RMSE: 14.807336781496629, MAE: 10.941057372093614, R2: 0.6309221889321419
PM2.5 - RMSE: 18.990983376971002, MAE: 7.337026329465995, R2: 0.2173132541916878
PM10 - RMSE: 14.57753030133693, MAE: 8.857416400878801, R2: 0.45228814654205296
```

Conclusion:

The XGBoost model demonstrates superior performance over both LSTM and ARIMA models for all pollutants (NO2, O3, PM2.5, PM10) in terms of RMSE, MAE, and R^2. The significant improvements in error metrics and R^2 values indicate that XGBoost is more effective in capturing the patterns in the pollutant data and providing accurate predictions.

Recommendations:

- **Further Optimization**: While XGBoost has shown better performance, further tuning of hyperparameters and model architecture could potentially improve the results even more.
- **Hybrid Models**: Consider exploring hybrid models that combine the strengths of LSTM, ARIMA, and XGBoost for potentially even better performance.
- **Feature Engineering**: Additional feature engineering, including the use of exogenous variables such as weather data, might further enhance the model's predictive capabilities.

INSIGHTS:

1. Impact of Air Quality on Manufacturing Operations

Air Quality and Employee Health:

- **Health Implications**: Poor air quality can affect the health of employees, potentially leading to increased absenteeism or reduced productivity.
- **Operational Insight**: Implementing health and safety measures, such as improved ventilation or air purification systems, can reduce the impact of air quality on workers and maintain operational efficiency.

Regulatory Compliance:

- **Standards and Guidelines**: Environmental regulations may dictate air quality standards for manufacturing facilities. Compliance with these standards is crucial to avoid fines and operational disruptions.
- **Operational Insight**: Monitoring and forecasting air quality can help anticipate regulatory changes or compliance issues, allowing proactive adjustments to maintain standards.

2. Forecasting Demand for Pharmaceutical Products

Product Demand Correlation:

- **Health Effects**: Poor air quality can lead to an increase in health issues like respiratory and cardiovascular problems, which in turn may increase the demand for related pharmaceuticals.

- **Operational Insight**: Use air quality forecasts to predict demand for products such as asthma inhalers, cardiovascular medications, and other health supplements. Adjust production schedules and inventory accordingly.

Production Planning:

- **Forecasting Trends**: Accurate forecasts of product demand based on air quality can help in better planning of production schedules and inventory management.
- **Operational Insight**: Align production levels with anticipated demand to avoid overproduction or stockouts. This optimization can reduce costs and improve supply chain efficiency.

3. Enhancing Manufacturing Processes

Quality Control:

- **Impact of Environmental Conditions**: Environmental factors like air quality can influence the manufacturing environment and the quality of the final product.
- **Operational Insight**: Incorporate environmental monitoring into quality control processes. Ensure that manufacturing conditions meet required standards to prevent quality issues.

Facility Design:

- **Facility Design**: Poor air quality can affect equipment and facilities. Regular maintenance and design adjustments might be needed.
- **Operational Insight**: Invest in advanced air filtration systems and other facility improvements to mitigate the impact of air quality on manufacturing processes.

4. Strategic Risk Management

Risk Assessment:

- **Risk Factors**: Poor air quality poses health risks to employees and can affect production efficiency.
- **Operational Insight**: Implement risk assessment protocols to evaluate the potential impact of air quality on manufacturing operations. Develop contingency plans for adverse conditions.

Supply Chain Disruptions:

- **External Factors**: Air quality issues may lead to external disruptions, such as delays in the supply of raw materials.
- **Operational Insight**: Establish robust supply chain management practices to handle potential disruptions. Build relationships with multiple suppliers to mitigate risk.

5. Data-Driven Decision Making

Optimization of Manufacturing Models:

- **Data Analysis**: Use data from optimized ARIMA models to refine manufacturing strategies and processes.
- **Operational Insight**: Make data-driven decisions on production scaling, inventory management, and resource allocation to align with forecasted demand and environmental conditions.

Performance Metrics:

- **Monitoring Metrics**: Track performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to assess the accuracy of demand forecasts and operational adjustments.
- **Operational Insight**: Regularly review and update forecasting models based on performance metrics to ensure continued accuracy and effectiveness.

6. Sustainability and Corporate Responsibility

Environmental Impact:

- **Sustainability Goals**: Manufacturing processes contribute to environmental impacts, including air quality. Adopting sustainable practices can align with corporate responsibility goals.
- **Operational Insight**: Implement green manufacturing practices and contribute to reducing the environmental footprint. This can improve public perception and meet sustainability goals.

Health and Safety Initiatives:

- **Employee Well-being**: Prioritize initiatives that protect employee health and well-being in relation to environmental conditions.
- **Operational Insight**: Enhance employee safety protocols and wellness programs to address the health impacts of poor air quality.

7 .Strategic Recommendations

Supply Chain Adjustments:

- **Based on Forecasts**: Prepare for increased demand for respiratory and cardiovascular medications if forecasts predict worsening air quality.
- **Inventory Management**: Ensure that stock levels of relevant medications can meet anticipated demand spikes due to poor air quality.

Health Awareness Campaigns:

- **Public Health Initiatives**: Collaborate with health organizations to raise awareness about the health risks of poor air quality. This could drive demand for preventive medications and health supplements.

Product Development:

- **New Products**: Develop new or improve existing pharmaceuticals targeting conditions exacerbated by poor air quality. Use forecasting data to identify potential market needs.

Regulatory and Policy Engagement:

- **Policy Advocacy**: Engage with policymakers to advocate for regulations that improve air quality. This can help in reducing the burden of air pollution-related health issues and the corresponding pharmaceutical needs.