AIR QUALITY PREDICTION ANALYSIS

-Report BY Y.Harika

1 Problem Statement

Air pollution is a critical environmental and public health issue in India, affecting millions of citizens, particularly in urban regions. Traditional methods of monitoring air quality are reactive, often identifying pollution spikes only after they occur.

There is an urgent need for predictive models that can accurately forecast air quality levels, providing both citizens and authorities with the opportunity to take proactive measures. Machine learning techniques offer the potential to analyse historical data, weather conditions, and pollution sources to predict air quality trends more accurately. This would help decision-makers implement preventive actions in anticipation of pollution spikes, protecting public health and enhancing overall quality of life.

The goal of this project is to develop a machine learning model capable of predicting air quality levels across various Indian cities, using real-time and historical pollution data, along with meteorological factors, to enable timely interventions.

2 Business Needs Assessment

The primary stakeholders affected by air pollution and who stand to benefit from air quality prediction models include:

2.1 Government and Regulatory Bodies

These agencies require accurate, timely data to regulate and enforce pollution controls. Predictive models would allow them to implement measures such as restricting vehicular movement or controlling industrial emissions before air quality reaches dangerous levels.

Solution: Machine learning-driven predictions can inform better air quality management policies, early warnings, and action plans.

2.2 Industries and Corporations

Businesses, especially those contributing to emissions, need predictions to comply with regulatory requirements and reduce their carbon footprint. Accurate forecasting would help industries plan operations during low-pollution periods.

Solution: Predictive models can guide operational adjustments, ensuring compliance with environmental laws and reducing their environmental impact.

2.3 Public Health Organizations

Need: Health organizations need air quality forecasts to mitigate health risks, especially for vulnerable populations such as children, the elderly, and those with pre-existing respiratory conditions.

Solution: Predictive data would enable these organizations to issue health advisories, recommend preventive care, and focus on public awareness campaigns about pollution hazards.

3 Target specifications

The success of this project depends on meeting key performance metrics related to the accuracy, timeliness, and applicability of the air quality prediction model. Below are the target specifications for this project:

Metric	Target Specification	Rationale
Prediction Accuracy (RMSE)	RMSE ≤ 10 $\mu g/m^3$ for PM2.5 predictions	Ensures that the model provides accurate predictions, minimizing errors and improving trust in the system.
Prediction Accuracy (R^2 Score)	$R^2 \geq 0.8$ for overall prediction model	Indicates that the model captures 80% or more of the variance in air quality data, reflecting strong predictive power.
Timeliness of Prediction	Predictions to be available 24 hours in advance	Provides stakeholders with adequate time to prepare for air quality deterioration and take necessary actions.
Geographic Coverage	Predictions for all Tier 1 and Tier 2 cities in India, with potential for expansion	Ensures that the model covers major population centres and is scalable for rural or under-monitored areas.
Update Frequency	Hourly updates using real-time data from air quality and weather monitoring stations	Provides up-to-date and relevant data to enhance the accuracy of short-term air quality predictions.
Data Sources	Integrate data from CPCB, OpenAQ, and meteorological APIs	Ensures comprehensive data collection, combining historical pollution data with real-time weather variables.
Model Interpretability	Provide insights into the contributing factors affecting air quality predictions	Allows decision-makers to understand which factors (e.g., weather, traffic) are driving the pollution predictions.
User Interface	User-friendly dashboard for government agencies, citizens, and industries	Ensures that the model's results are accessible, easy to interpret, and actionable for non-technical users.

Table 1.Target Specification list obtained from observations

4. External Search

To develop a robust air quality prediction model for India, an extensive external search was conducted to identify relevant datasets, best practices in machine learning models for air quality forecasting, and previous studies in this field. Key areas of focus included data sources, model selection, and methodologies applied in similar air quality projects globally. Below are the major findings from the external search:

4.1 Air Quality Data Sources

- **Central Pollution Control Board (CPCB)**: The official national source for air quality data in India. The CPCB monitors key pollutants such as PM2.5, PM10, NO2, and ozone in various cities across the country. Historical data from CPCB is publicly available and widely used for research.
- **OpenAQ**: A global open-source platform that aggregates real-time air quality data from sensors worldwide, including India. OpenAQ provides access to pollution data from multiple cities and allows real-time monitoring.
- **Kaggle Datasets**: Kaggle hosts several datasets related to air quality in India, including detailed historical data on PM2.5, PM10, and other pollutants, as well as meteorological data. These datasets are commonly used for training machine learning models.

4.2 Machine Learning Models for Air Quality Prediction

- **Time-Series Forecasting Models**: Studies on air quality prediction often utilize time-series models such as ARIMA (Autoregressive Integrated Moving Average) and Prophet, which are designed for forecasting based on historical data.
- **Random Forest and Gradient Boosting**: Research indicates that ensemble models like Random Forest and Gradient Boosting are highly effective in predicting air quality by handling complex, non-linear relationships between input variables (e.g., weather data and pollution levels).
- **Deep Learning Models**: Neural networks, particularly Long Short-Term Memory (LSTM) networks, are increasingly used for time-series forecasting. These models can capture long-term dependencies between variables and are particularly useful for handling large datasets.

4.3 Previous Research and Case Studies

- **Beijing Air Quality Forecasting**: A prominent case study conducted in Beijing, China, used Random Forest models combined with weather data to predict air pollution levels. The study emphasized the importance of combining meteorological factors such as wind speed and temperature with pollution data for better prediction accuracy.
- **Pollution Mapping in European Cities**: European studies have focused on real-time pollution mapping and air quality forecasting using IoT-based sensors and deep learning techniques. These models have helped provide real-time alerts and warnings to citizens.

4.4 Tools and Technologies

- **Python Libraries for Machine Learning**: Libraries such as Scikit-learn, TensorFlow, and XGBoost are widely used for building machine learning models for air quality prediction. Scikit-learn is ideal for implementing Random Forest and Gradient Boosting models, while TensorFlow supports deep learning techniques like LSTM.
- **Meteorological APIs**: Weather data is essential for air quality forecasting. APIs like OpenWeatherMap provide real-time weather updates (temperature, humidity, wind speed) which can be integrated with pollution data for more accurate predictions.

5 Benchmarking

Benchmarking was conducted to compare the proposed machine learning-based air quality prediction model with existing solutions both in India and internationally. This evaluation helped identify the strengths, weaknesses, and potential improvements that the new model can offer over current approaches. The following benchmarks were assessed based on accuracy, timeliness, geographic coverage, user interface, and scalability.

5.1 Existing Air Quality Prediction Systems

- **SAFAR (System of Air Quality and Weather Forecasting and Research) – India**
  - Accuracy: SAFAR provides air quality forecasts using a mix of satellite data, ground monitoring, and meteorological models. However, accuracy is often limited by the complexity of pollution dynamics, with moderate precision for short-term forecasting.
  - Benchmark Comparison: The proposed model offers a more specialized focus on forecasting using predictive analytics, aiming to combine the best of real-time data and forecast accuracy.

5.2 Key Performance Benchmarks

System	Accuracy	Timeliness	Geographic Coverage	User Interface	Scalability
Proposed Model	High (RMSE ≤ 10 μg/m³)	Hourly and daily forecasts	All Tier 1 and Tier 2 cities	User-friendly dashboard	Easily scalable through open data
SAFAR (India)	Moderate	Daily forecasts	Limited to metro cities	Basic mobile/web	Limited regional scalability
OpenAQ (Global)	Moderate	Real-time data	Wide global coverage	API-based	Global data aggregator
Beijing AQ System	High	Hourly and daily forecasts	Focus on Beijing	User-friendly	Scalable to other cities in China

System	Accuracy	Timeliness	Geographic Coverage	User Interface	Scalability
CAMS (Europe)	High	Daily and real-time	Europe and select global regions	API-based	Highly scalable via satellite data

Table 2. Details of Key Performance Benchmarks

The benchmarking analysis highlights that existing solutions, such as SAFAR and OpenAQ, provide either real-time monitoring or limited forecasting, but lack high-resolution, city-level predictions tailored to India's diverse regions. International systems, such as Beijing's Air Quality Forecast System and CAMS, offer advanced prediction models and real-time data but are designed for different geographic contexts. The proposed model seeks to improve upon these benchmarks by delivering accurate, hourly predictions for multiple Indian cities, combining real-time data with advanced machine learning techniques, and providing a user-friendly interface for both technical and non-technical stakeholders.

6 Applicable patents

- **Air Quality Monitoring and Prediction Systems**
  - Patent Title: Air Quality Monitoring and Predictive Analytics System
  - Patent Number: US9674179B2
  - This patent covers a system that integrates data from multiple air quality monitoring devices and applies predictive analytics to forecast future pollution levels.
- **Real-Time Air Quality Monitoring Using Sensors**
  - Patent Title: Real-Time Air Quality Monitoring and Forecasting System Using IoT Sensors
  - Patent Number: US10401587B1
  - This patent describes a system for real-time air quality monitoring using Internet of Things (IoT) sensors and data aggregation.
- **Distributed Air Quality Monitoring System**
  - Patent Title: Distributed Air Quality Monitoring System Using Machine Learning
  - Patent Number: US11270684B2
  - This patent covers a distributed system for monitoring air quality using machine learning to analyze data from multiple sensors deployed in various locations. It includes methods for predictive analysis and issuing real-time alerts.

7 Applicable Regulations

The development and deployment of an air quality prediction model in India must adhere to several regulations and legal frameworks, ensuring that the system complies with environmental standards, data protection laws, and technological guidelines.

7.1 Environmental Regulations

7.1.1 Air (Prevention and Control of Pollution) Act, 1981

This is the primary legislation in India governing air pollution. It outlines the duties and powers of regulatory bodies like the Central Pollution Control Board (CPCB) and State Pollution Control Boards (SPCBs) to monitor and control air pollution. The Act establishes air quality standards and mandates regular reporting and action plans for polluted regions.

7.2 Data Protection and Privacy Regulations

7.2.1 Information Technology Act, 2000

This Act outlines the rules and regulations for the collection, storage, and processing of electronic data in India. It includes provisions for data protection, cybersecurity, and penalties for data breaches.

7.3 International Standards and Guidelines

7.3.1 General Data Protection Regulation (GDPR) (Applicable if Serving Global Users)

GDPR is an EU regulation governing data protection and privacy for all individuals within the European Union. If the air quality prediction system collects data from or serves EU users, it must comply with GDPR requirements regarding data processing, storage, and consent.

7.4 Public Health Regulations

7.4.1 World Health Organization (WHO) Air Quality Guidelines

The WHO provides internationally recognized guidelines for acceptable levels of air pollutants (PM2.5, PM10, NO_2, ozone). These standards inform regulatory limits for air pollution that protect public health.

Complying with these regulations ensures that the air quality prediction model operates within legal frameworks and meets environmental, data security, and privacy standards.

8 Applicable Constraints:

Several constraints need to be considered during the development, deployment, and maintenance of the air quality prediction model. These constraints include technological, environmental, data-related, regulatory, and economic factors. Understanding these constraints ensures that the model remains realistic, feasible, and functional under given conditions.

8.1 Data Availability and Quality

8.1.1 Constraint: Inconsistent or incomplete air quality data

Reliable air quality data is essential for accurate predictions, but data availability across different regions in India is inconsistent. Some areas may have limited monitoring stations, leading to gaps in historical and real-time data.

8.2 Technological Constraints

8.2.1 Computational resources for model training

Machine learning models, especially complex algorithms like Random Forest or Long Short-Term Memory (LSTM) networks, require substantial computational power for training and real-time forecasting.

9 Business model

9.1 Actionable Alerts and Recommendations

Delivers actionable recommendations (e.g., outdoor activity advisories, operational adjustments for industries) based on predictive analytics, allowing users to take proactive measures.

9.2 Data-Driven Public Health Insights

Helps public health organizations identify at-risk populations and issue timely warnings to mitigate the effects of poor air quality on public health.

9.3 Compliance with Environmental Regulations

Enables industries and businesses to meet compliance standards set by regulatory bodies, thus avoiding fines and improving corporate responsibility.

9.4 Subscription-Based Services:

Offer tiered subscriptions for access to different levels of data, including daily forecasts, real-time alerts, and advanced analytics. Different pricing models could cater to small businesses, large industries, and government agencies.

9.5 Data Licensing:

Sell or license real-time and historical air quality data to third-party applications, researchers, or corporations that require data for their internal models or products.

9.6 Consultation and Integration Services:

Provide consultation to cities, industries, and urban planners to integrate the model into their operations or smart city projects.

9.7 Sensor Manufacturers and IoT Providers:

Partner with manufacturers of air quality sensors and Internet of Things (IoT) devices to create a robust, real-time data collection network. This could help improve data granularity and expand geographic coverage.

9.8 Cloud Computing Providers (AWS, Google Cloud, Microsoft Azure):

Collaborate with cloud service providers to host the model and manage large-scale data processing, ensuring scalability and seamless data delivery.

9.9 Government and Municipal Bodies:

Establish partnerships with CPCB, SPCBs, and local municipalities to access data from existing air quality monitoring stations and integrate predictive models into public health initiatives.

10 Concept Generation

10.1 Real-Time Air Quality Forecasting Using IoT Sensors

This concept involves deploying a network of IoT-based air quality sensors across major cities and regions. The sensors will gather real-time data on pollutants like PM2.5, PM10, NO2, and O3, as well as meteorological data such as wind speed, temperature, and humidity. Machine learning algorithms, such as Random Forest or Gradient Boosting, will process the data and provide real-time air quality forecasts.

10.2 Mobile Application for Personalized Air Quality Alerts

This concept focuses on developing a mobile application that provides citizens with personalized air quality updates and health advice based on their location. The app will use GPS data and access real-time air quality information from central monitoring stations, providing alerts when pollution levels reach harmful thresholds. Machine learning algorithms will tailor health recommendations based on user profiles, including factors like age, respiratory conditions, and activity levels.

11 Concept Devolopment

11.1 Real-Time Air Quality Forecasting Using IoT Sensors

11.1.1 Objective

To develop a real-time air quality forecasting system using a network of IoT sensors deployed across urban and semi-urban areas in India, combined with machine learning algorithms for predictive analytics. The system will provide accurate, real-time pollution forecasts and health advisories for government bodies, industries, and citizens.

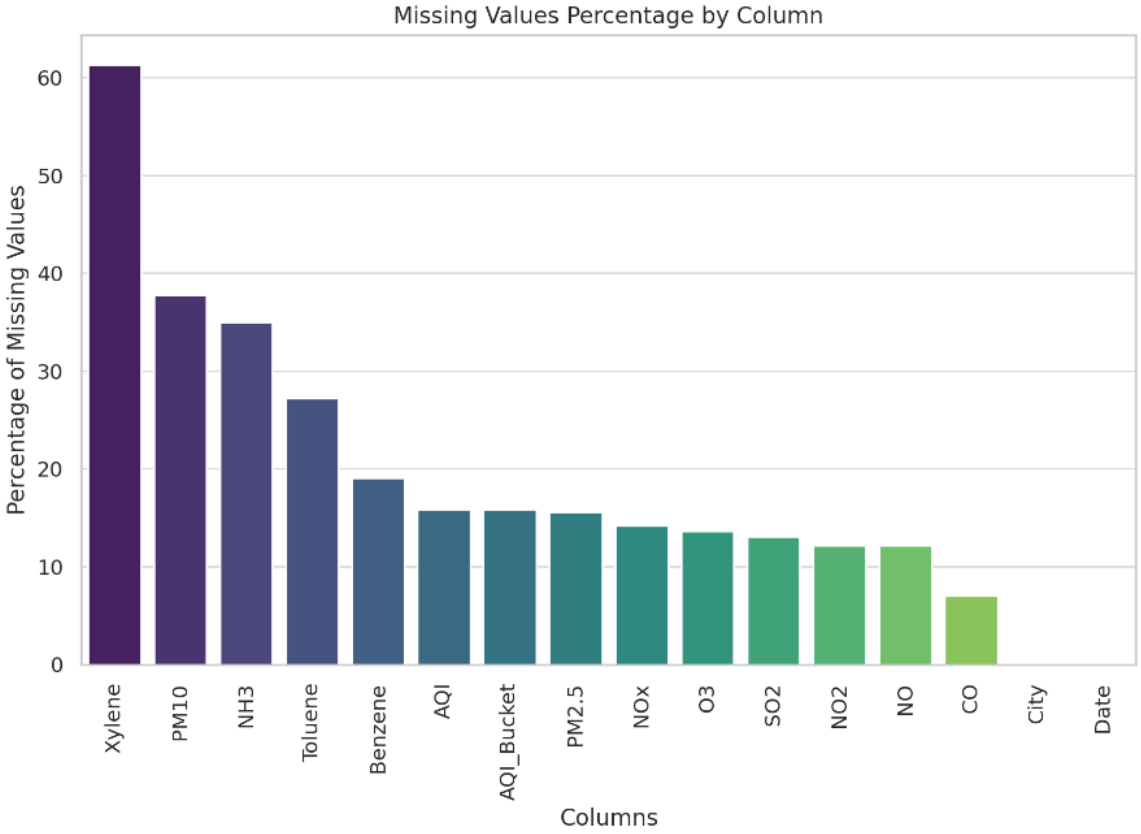
11.1.2 Key Components

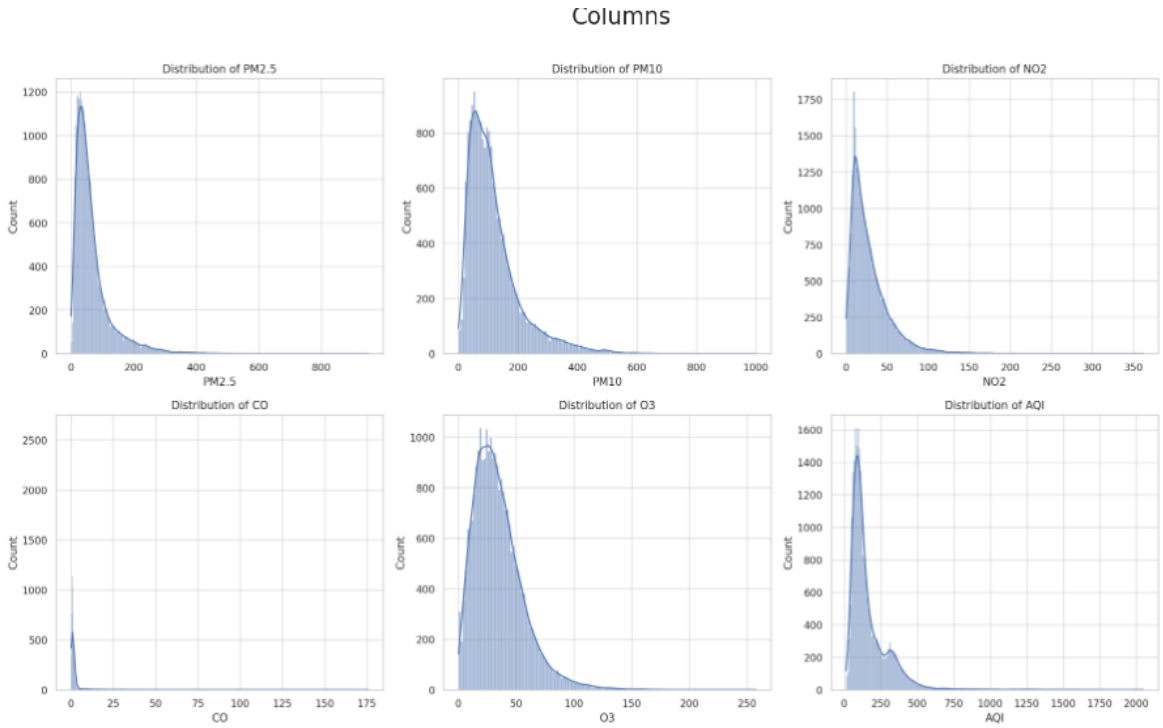
- IoT Sensors: Low-cost, real-time air quality sensors deployed at strategic locations to monitor pollutants such as PM2.5, PM10, NO2, and ozone, as well as meteorological data like temperature, humidity, and wind speed.
- Machine Learning Models: Algorithms such as Random Forest, Gradient Boosting, and Long Short-Term Memory (LSTM) networks will be used to process the collected data and generate real-time and short-term forecasts (up to 24-48 hours).
- Data Processing Infrastructure: Cloud-based data storage and processing infrastructure to handle the large volume of data from IoT sensors. Scalable cloud platforms (e.g., AWS, Google Cloud) will enable seamless data analysis and forecasting.
- User Interface (UI): A user-friendly dashboard and mobile app for real-time air quality data access. Alerts, recommendations, and visualizations will be delivered to government officials, industries, and citizens.

12 Code Implementation

GitHub link: *https://github.com/HarikaMKHS/Air-Quality-Prediction-Model*

I'll perform a deeper analysis, including missing data handling, visualizations for key features, and an investigation into potential correlations between variables. This will help us understand the data structure and decide on an appropriate machine learning approach.





The exploratory data analysis that I had done with the datasets obtained from Kaggle

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.stattools import adfuller
from sklearn.metrics import mean_absolute_error, mean_squared_error
import numpy as np
```

In [2]:
```
# Load and prepare data
data = pd.read_csv("C:/Users/Harika/Documents/feynn labs/city_day.csv")
data['Date'] = pd.to_datetime(data['Date'])
data.set_index('Date', inplace=True)
```

In [4]:
```
# Interpolate missing AQI values
data['AQI'] = data['AQI'].interpolate(method='time')

# ADF Test to check stationarity
adf_test = adfuller(data['AQI'].dropna())
```

In [5]:
```
# Differencing if needed
if adf_test[1] > 0.05:
    data_diff = data['AQI'].diff().dropna()
else:
    data_diff = data['AQI']
```

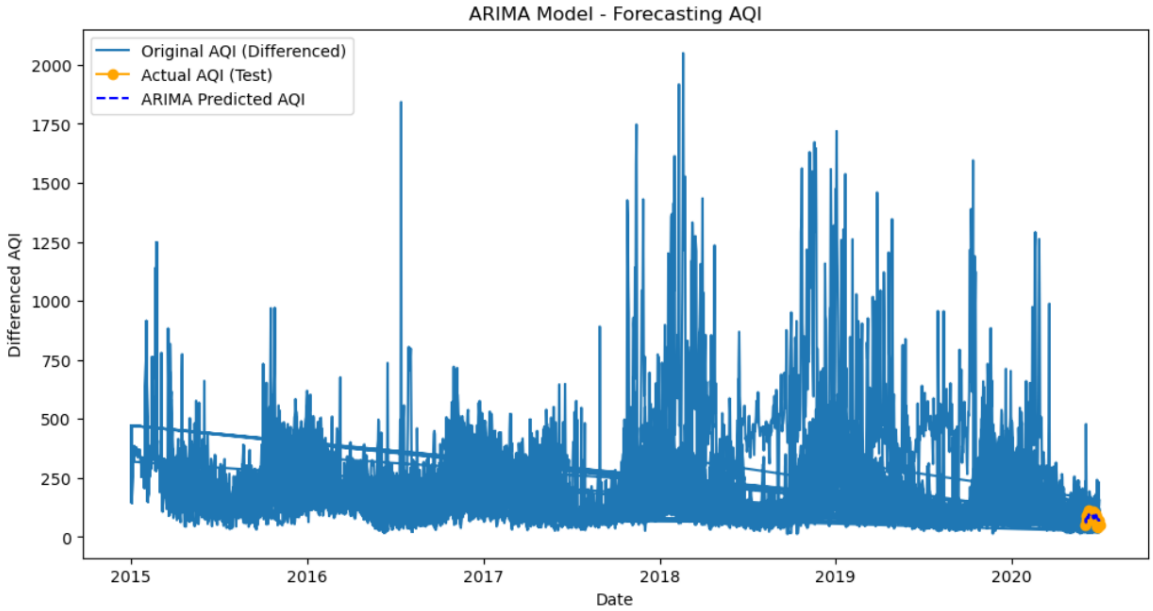
In [6]:
```
# Train the ARIMA model (p=1, d=1, q=1 for simplicity)
model = ARIMA(data_diff, order=(1, 1, 1))
arima_result = model.fit()
```

In [8]:
```
# Evaluate model
y_train = data_diff[:-forecast_steps]
y_test = data_diff[-forecast_steps:]
y_pred = arima_result.predict(start=len(y_train), end=len(data_diff) - 1)

mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
```

In [9]:
```
# Plot results
plt.figure(figsize=(12, 6))
plt.plot(data_diff, label="Original AQI (Differenced)")
plt.plot(y_test.index, y_test, color='orange', marker='o', label="Actual AQI (Test)")
plt.plot(y_test.index, y_pred, color='blue', linestyle='--', label="ARIMA Predicted AQI")
plt.legend(loc="upper left")
plt.title("ARIMA Model - Forecasting AQI")
plt.xlabel("Date")
plt.ylabel("Differenced AQI")
plt.show()

print(f"MAE: {mae}, MSE: {mse}, RMSE: {rmse}")
```



MAE: 13.539576870610661, MSE: 291.04917337222025, RMSE: 17.060163345414377

13 Final Product Prototype

13.1 Prototype Development

13.1.1 Sensor Network Setup

- Initial deployment of IoT air quality sensors in selected cities (e.g., Delhi, Mumbai, Bangalore) with a focus on pollution hotspots.
- Sensors will monitor pollutants and collect real-time data, which will be transmitted to a central cloud server.

13.1.2 Data Collection and Integration

- Integration of pollution data from existing sources such as CPCB and OpenAQ, combined with the real-time sensor data.
- Collection of historical data to train machine learning models.

13.1.3 Model Training and Testing

- Train machine learning models (Random Forest, Gradient Boosting) on historical and real-time data.
- Validate model accuracy using evaluation metrics such as Root Mean Squared Error (RMSE) and R^2 score.
- Pilot tests in selected cities to validate the prediction accuracy of the model

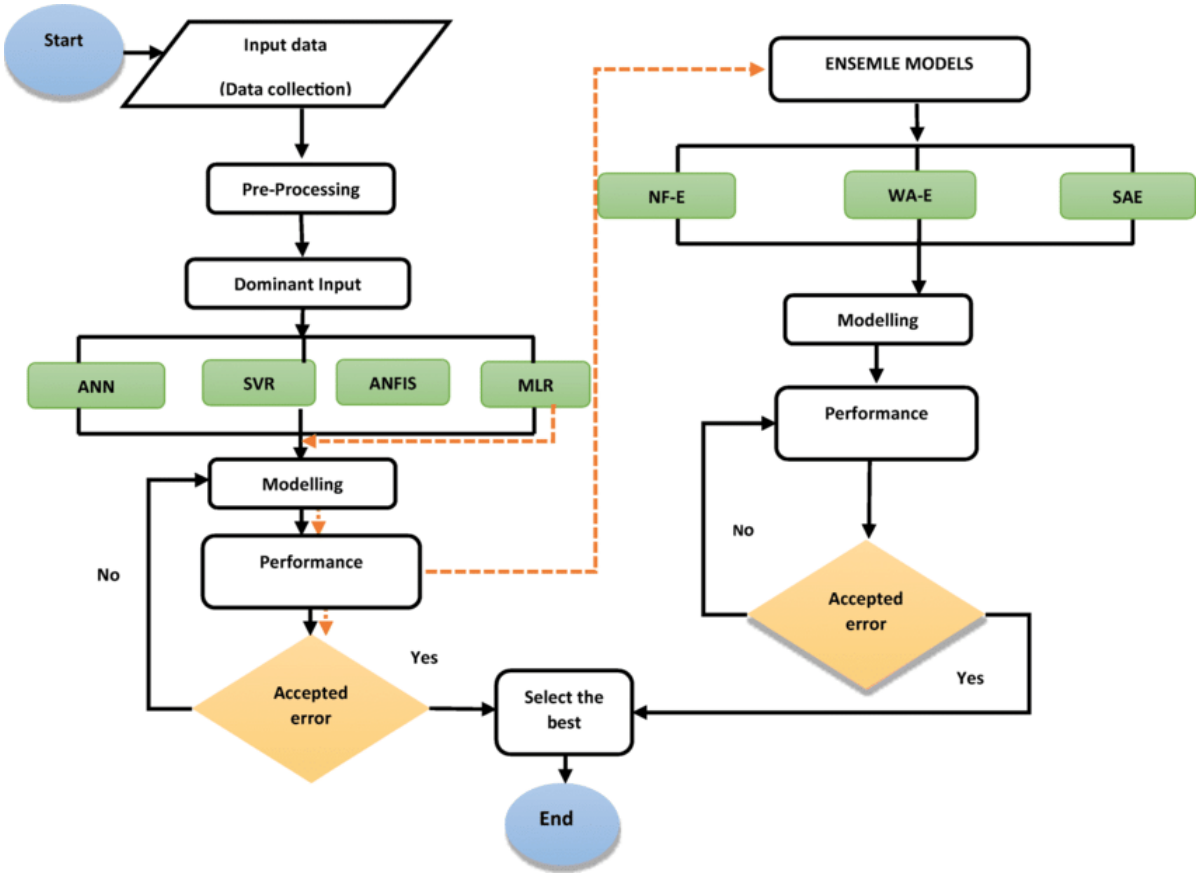


Fig 1: Schematic Diagram for " Air Quality Prediction" model

14 Product Details

14.1 How does it work

14.1.1 Data Collection

- **Sensors**: Deploy a network of air quality sensors to collect real-time data on pollutants (e.g., PM2.5, NO2, SO2, O3) and environmental factors (e.g., temperature, humidity).
- **Meteorological Data**: Integrate weather data (e.g., wind speed, direction, temperature) from local meteorological stations or satellites.

14.1.2 Data Integration

- **Data Aggregation**: Combine data from multiple sources (sensors, weather forecasts, historical records) into a centralized database.
- **Preprocessing**: Clean and preprocess the data to ensure accuracy and consistency. This may involve filtering out noise and handling missing values.

14.1.3 Analysis and Modeling

- **Machine Learning Algorithms**: Use machine learning models to analyze historical data and identify patterns in air quality. Common algorithms include regression models, decision trees, or neural networks.
- **Predictive Modeling**: Develop predictive models that forecast future air quality levels based on current conditions, historical data, and identified trends.

14.1.4 Real-Time Monitoring

- **Continuous Updates**: Continuously monitor incoming data from sensors and update predictions in real time.
- **Alerts and Notifications**: Implement an alert system to notify users when air quality levels are predicted to exceed safe thresholds.

14.1.5 User Interface

- **Dashboard:** Create an interactive dashboard that visualizes real-time air quality data, predictions, and trends for users, whether through a mobile app or web interface.
- **Personalization**: Allow users to customize notifications and reports based on their preferences or health conditions.

14.1.6 Community Engagement

- **Reporting and Feedback**: Enable users to report local air quality issues and contribute data, fostering community involvement and awareness.
- **Educational Resources**: Provide users with information on air quality, health impacts, and best practices for reducing exposure to pollutants.

14.1.7 Continuous Improvement

- **Model Refinement**: Regularly update and refine predictive models based on new data, user feedback, and emerging research to improve accuracy. Collaborate with local governments, NGOs, and research institutions to enhance data sources

15 Conclusion

The air quality prediction model leverages IoT sensors and machine learning to provide accurate, real-time air quality forecasts. It empowers government agencies, industries, and citizens to make informed decisions to mitigate pollution and protect public health. The scalable system offers personalized alerts, actionable insights, and compliance support, contributing to proactive pollution management and improved air quality across India. This solution marks a vital step toward creating healthier environments and addressing the country's air quality challenges.

16 Financial Equation

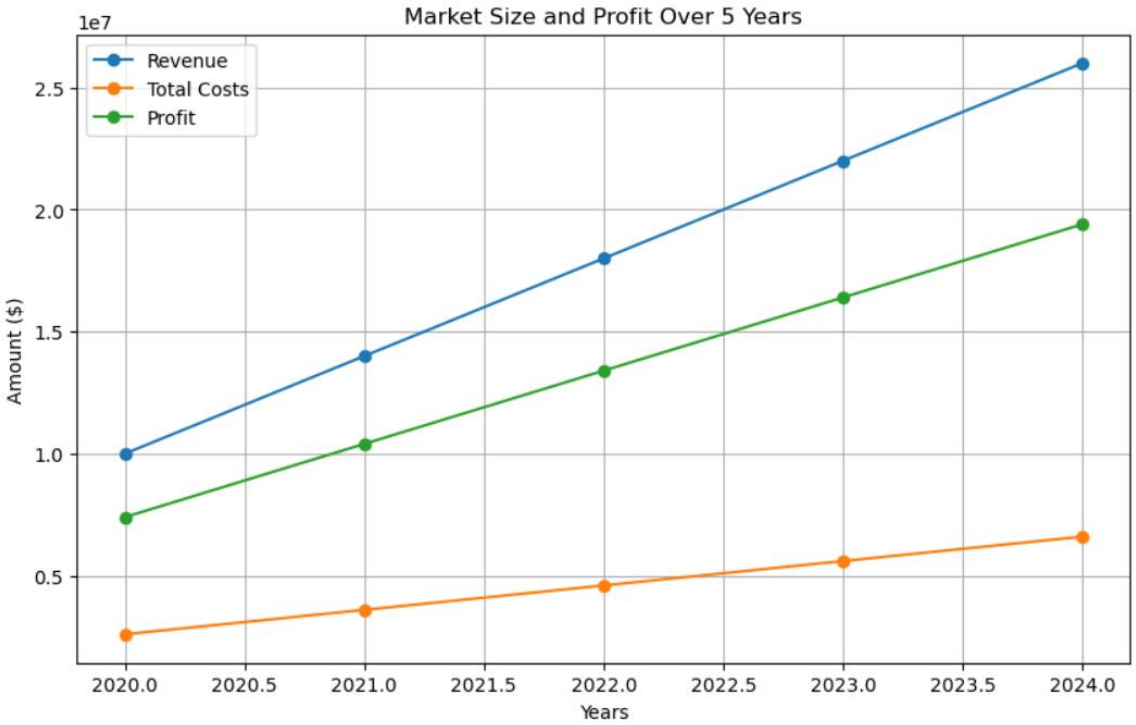


Fig : Air Quality Analysis chart for 5 Years

Estimating the market size and profit for an air quality-related project would depend on several assumptions, such as the target market, pricing, expected sales, and operational costs. Let's go through a hypothetical scenario:

Example Scenario: Market Size and Profit Estimation for an Air Purification Business

Assumptions:

Target Market: Urban areas with high AQI (e.g., cities with AQI > 100)

Population: 1,000,000 people in the target area

Market Penetration Rate: Expected 5% of the population may purchase an air purifier

Price per Unit: $200 per air purifier

Fixed Costs: $100,000 per year (rent, salaries, etc.)

Variable Costs per Unit: $50 (manufacturing, shipping, etc.)

Calculations:

- **Market Size**: Population × Penetration Rate × Price per Unit
- **Revenue**: Price per Unit × Expected Units Sold
- **Total Costs**: Fixed Costs + (Variable Costs per Unit × Units Sold)
- **Profit**: Revenue - Total Costs

Example calculation:

1 Year 1

- Penetration Rate = 5%
- Units Sold = 1,000,000×0.05=50,0001,000,000 \times 0.05 = 50,0001,000,000×0.05=50,000
- Revenue = 50,000×200=10,000,00050,000 \times 200 = 10,000,00050,000×200=10,000,000
- Variable Costs = 50,000×50=2,500,00050,000 \times 50 = 2,500,00050,000×50=2,500,000
- Total Costs = 2,500,000+100,000=2,600,0002,500,000 + 100,000 = 2,600,0002,500,000+100,000=2,600,000
- Profit = 10,000,000−2,600,000=7,400,00010,000,000 - 2,600,000 = 7,400,00010,000,000−2,600,000=7,400,000