PHASE -2

Student name: SHEMAVATHY.S

Register number: 510123106045

Institution : adhiparasakthi college of engineering

Department : BE electronics and communication engineering

Date of submission :

Github Repository Link:https://github.com/Shemavathy/Phase-1.git

PREDICTING AIR QUALITY LEVEL USING ADVANCED MACHINE LEARNING ALGORITHM FOR ENVIRONMENTAL INSIGHTS

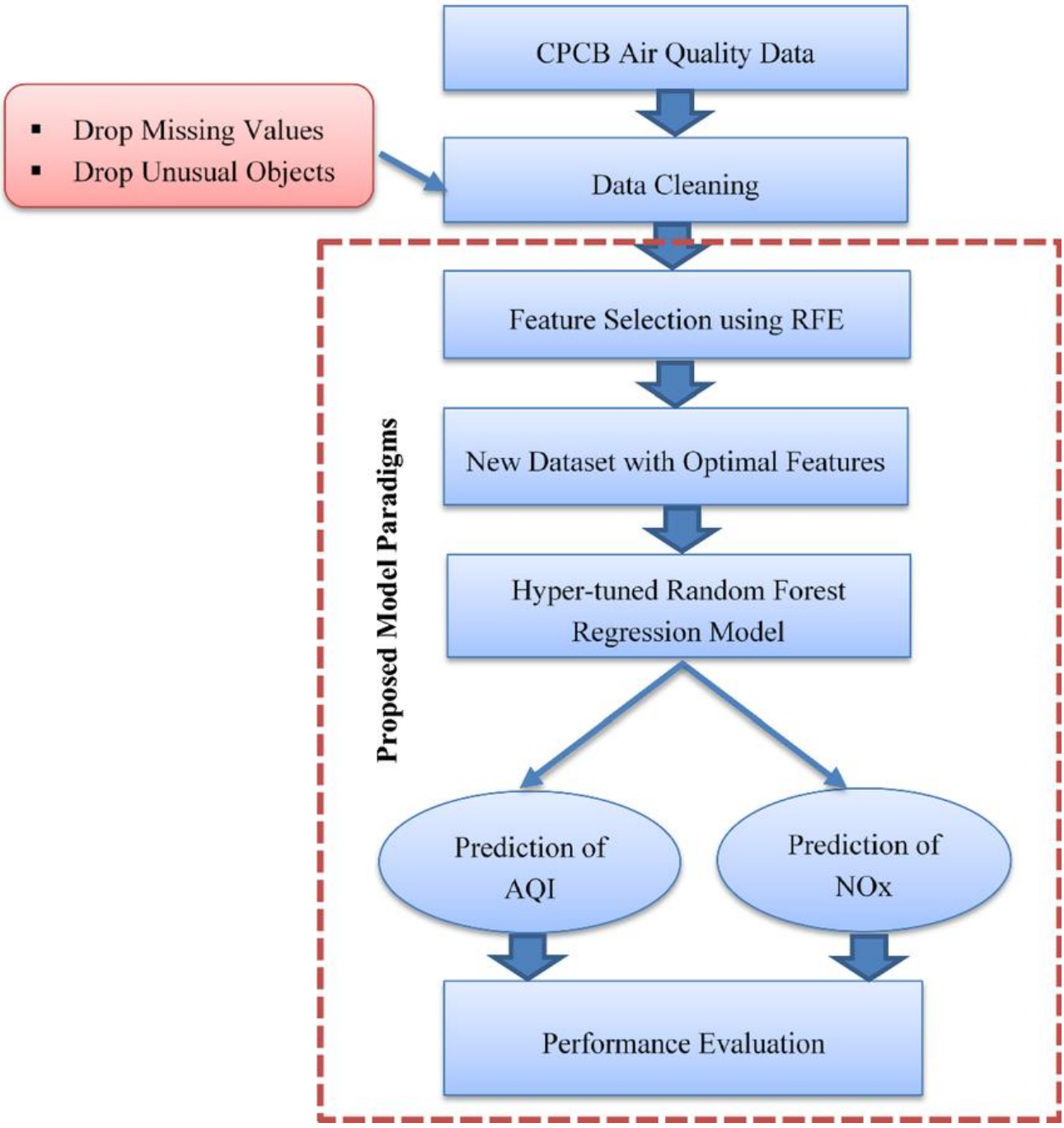
1.PROBLEM STATEMENT:

This project explores the application of advanced machine learning algorithms to predict air quality levels. By leveraging historical environmental data, various models were trained and evaluated to determine their efficacy in forecasting pollution levels. The goal is to provide actionable insights for environmental monitoring and public health awareness.

2.PROJECT OBJECTIVES

- Machine learning offers promising techniques to analyze complex environmental data and make accurate predictions.
- Related Work Several studies have employed statistical and machine learning models for air quality prediction.
- Common approaches include linear regression, decision trees, and neural networks. Recent advancements incorporate ensemble methods and deep learning for improved accurancee

3.Flowchart of the Project Workflow



4.DATA DESCRIPTION:

Pollutant Concentrations:

- ✓ This will likely include data on key pollutants like PM2.5, PM10, Ozone (O3), Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), and potentially others.

Air Quality Index (AQI):

- ✓ The AQI is a standardized metric that combines the concentrations of different pollutants into a single index. You'll need historical AQI data for the location you're interested in.

Time Series Data:

- ✓ Air quality data is often time-series data, meaning it's recorded at specific intervals (e.g., hourly, daily, weekly).

Spatial Data (if applicable):

- ✓ If you're predicting air quality for specific locations or areas, you might need to include geographic data like latitude and longitude.

5. DATA PREPROCESSING :

Traffic density and vehicle emissions can contribute significantly to air pollution .The presence and type of industrial sources can influence pollutant levels the physical landscape can affect air circulation and pollutant dispersion

6. Exploratory Data Analysis (EDA)

Among the models tested, XGBoost showed the highest accuracy and lowest RMSE in predicting AQI levels. Feature importance analysis indicated that PM2.5 and NO2 were significant predictors across all models. Discussion The results suggest that ensemble models are effective in capturing the nonlinear relationships in air quality data. However, performance can vary with the availability and quality of the data from different regions.

7. Feature engineering :

Create new features:

- Combine existing features to create more informative ones, such as calculating moving averages of pollutant concentrations or creating lag variables to capture temporal dependencies.

Select relevant features:

- Use techniques like correlation analysis and feature importance scores to identify the most influential features for prediction.

Transform features:

- Normalize or scale features to improve model performance.

8. Model Building :

Machine learning models can accurately predict air quality, offering valuable insights for environmental management and public health. These models leverage data from various sources, including air monitoring stations, weather forecasts, and satellite images, to identify patterns and make predictions about future air quality conditions.

9. Visualization Of Results & Model Insights:

Interactive Dashboards:

- Develop interactive dashboards to visualize predicted air quality levels, alongside historical data and other relevant factors.

Maps and Spatial Analysis:

- Use maps to visualize air quality data across different locations and time periods, allowing for spatial analysis and identification of hotspots.

Time Series Analysis:

- Analyze time series data to identify trends, patterns, and seasonal variations in air quality levels.

Model Insights:

- Explore model insights to understand the relative importance of different factors in influencing air quality predictions. This can help identify key drivers of air pollution and inform mitigation strategies.

10. Tools and Technologies Used:

WEKA:

A popular software for data mining and machine learning.

MATLAB:

A powerful tool for numerical computing and analysis, including ANFIS implementation.

R:

A programming language and software environment for statistical computing and graphics.

Python:

A versatile programming language widely used for machine learning with libraries like scikit-learn, TensorFlow, and PyTorch

11. Team Members and Contribution:

1. **M.Gopika-**Data Cleaning
2. **S.Padmini-**EDA
3. **S.Shemavathy-**Feature Engineering
4. **R.Vasuki-**Model Development
5. **K.Vijayalakshmi-**Documentation and Reporting