**PHASE 1**

**Student Name :** Balaji.K

**Register Number :** 621523121009

**Institution :** Mahendra college of engineering

**Department :** BE.BME

**Date of Submission :** 24.04.2025

**1.Problem Statement:**

Prediction of Air Quality using Supervised Machine Learning Algorithm for Environmental Insight

**2.Objectives of the Project**

* To develop a machine learning model that accurately predicts air pollution levels.
* To explore how pollutant concentrations and meteorological factors contribute to air quality.
* To analyze trends and patterns in air quality data using EDA.
* To compare various supervised algorithms and evaluate their effectiveness.
* To provide actionable insights for environmental policy makers and public health agencies.

**3.Scope of the Project**

**Features to Analyze:**

* Pollutant levels: PM2.5, PM10, NO2, CO, O3, SO2
* Meteorological variables: Temperature, Pressure, Wind Speed, Humidity
* Temporal features: Date, Time

**Limitations:**

* Only supervised learning algorithms are considered.
* The dataset is static and does not support real-time prediction.
* Model performance depends on the completeness and quality of the dataset.
* External environmental factors not captured in the dataset may affect accuracy.

**4.Data Sources**

* **Source:** Kaggle [Predicting Air Pollution using ML and NN](https://www.kaggle.com/code/faridtaghiyev/predicting-air-pollution-using-ml-and-nn" \t "_new)
* **Type:** Public and Static
* **Description:** The dataset contains hourly air quality measurements from various monitoring stations, including pollutants and weather conditions.

**5.High-Level Methodology**

* **Data Collection:** The data is downloaded from Kaggle as a CSV file and used in its static form for training and evaluation.
* **Data Cleaning:**
* Handling missing values using imputation techniques.
* Dropping duplicates.
* Converting datetime features.
* Encoding categorical variables if any.
* **Exploratory Data Analysis (EDA):**
* Histograms, pair plots, and correlation heatmaps to understand variable relationships.
* Time series plots to visualize pollutant trends.
* Seasonal decomposition to identify variations in pollution levels.
* **Feature Engineering:**
* Extraction of date/time features like hour, month, day.
* Standardization or normalization of numerical variables.
* Creation of lag features or moving averages for temporal patterns.
* **Model Building:**
* Supervised ML algorithms such as:
* Linear Regression
* Random Forest Regressor
* XGBoost Regressor
* Support Vector Regressor (SVR)
* These models are suitable due to their ability to model complex, non-linear relationships and work well with tabular data.
* **Model Evaluation:**
* Mean Absolute Error (MAE)
* Root Mean Squared Error (RMSE)
* R² Score
* K-Fold Cross Validation
* **Visualization & Interpretation:**
* Predicted vs Actual AQI plots
* Feature importance charts
* Residual plots to assess error distribution
* **Deployment (Optional):** A Streamlit app could be created for user interaction with the prediction model (future scope).

**6.Tools and Technologies**

**Programming Language:**

* Python

**Notebook/IDE:**

* Google Colab
* Jupyter Notebook

**Libraries:**

* pandas, numpy – Data handling
* matplotlib, seaborn, plotly – Visualization
* scikit-learn – Modeling and evaluation
* xgboost, lightgbm – Gradient boosting models
* tensorflow (for optional neural network comparison)

**Optional Deployment Tools:**

* Streamlit or Flask (if deployment is pursued)

**7.Team Members and Roles**

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Responsibility** |
| Balaji.K | Project Lead &  Responsibilities | Data collection NLP model training intent classification |
| Jeevanandham**.**A | NLP Developer &Data Analyst | EDA data visualization feature extraction |
| Dhanush.B | Backend Developer&  Data collection NLP | API integration chatbot logic and response handling |
| Vignesh**.**M | Deployment Engineer | Web app development deployment testing |