PHASE-3

Predicting Air Quality Levels Using Advanced Machine Learning Algorithm

STUDENT NAME: ROOBINI. S

REGISTER NUMBER: 422223106025

COLLEGE NAME: SURYA GROUP OF INSTITUTIONS

DATE OF SUBMISSION: 14/05/2025

GITHUB LINK:
https://github.com/ROOBINI2707/Air-Quality-.git

1. Problem Statement

- - Air pollution is a significant global concern, adversely affecting public health and the environment. The unpredictability of air quality levels makes it difficult for authorities and individuals to take timely action. This project focuses on developing an intelligent system using advanced machine learning techniques to predict air quality levels, aiding in early warnings and preventive measures. The primary problem statement related to air quality revolves around the detrimental impact of air pollution on human health, the environment, and economic development. Air pollution, caused by various pollutants like particulate matter, gases, and toxic chemicals, leads to respiratory and cardiovascular diseases, and can exacerbate existing health conditions. Furthermore, poor air quality can damage ecosystems, contribute to climate change, and negatively impact economic productivity.

2. Abstract

- - This project utilizes machine learning algorithms to predict air quality levels based on environmental and meteorological data. By incorporating techniques like data preprocessing, exploratory data analysis, and feature engineering, the system is trained to predict the Air Quality Index (AQI). The project evaluates multiple models for accuracy and deploys the best-performing one for real-time use. This solution aims to support urban planning, health advisories, and pollution control strategies. Air quality, a critical aspect of environmental health, refers to the cleanliness of the air we breathe, encompassing various pollutants like particulate matter, gases, and other substances. It's a key factor influencing human health, ecosystems, and the climate. Measuring and monitoring air quality helps in understanding the impact of pollution sources and developing strategies for mitigation.

3. System Requirement

- **Hardware:**

1. Processor: Intel Core i5 or higher
2. RAM: 8 GB minimum
3. Storage: 100 GB
4. GPU (optional for deep learning

- **Software:**

1. OS: Windows/Linux/macOS
2. Programming Language: Python 3.8+
3. Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, TensorFlow/Keras, XGBoost, Flask
4. IDE: Jupyter Notebook or VS Code
5. Tools: Git, Docker (for deployment), Postman (for testing APIs)

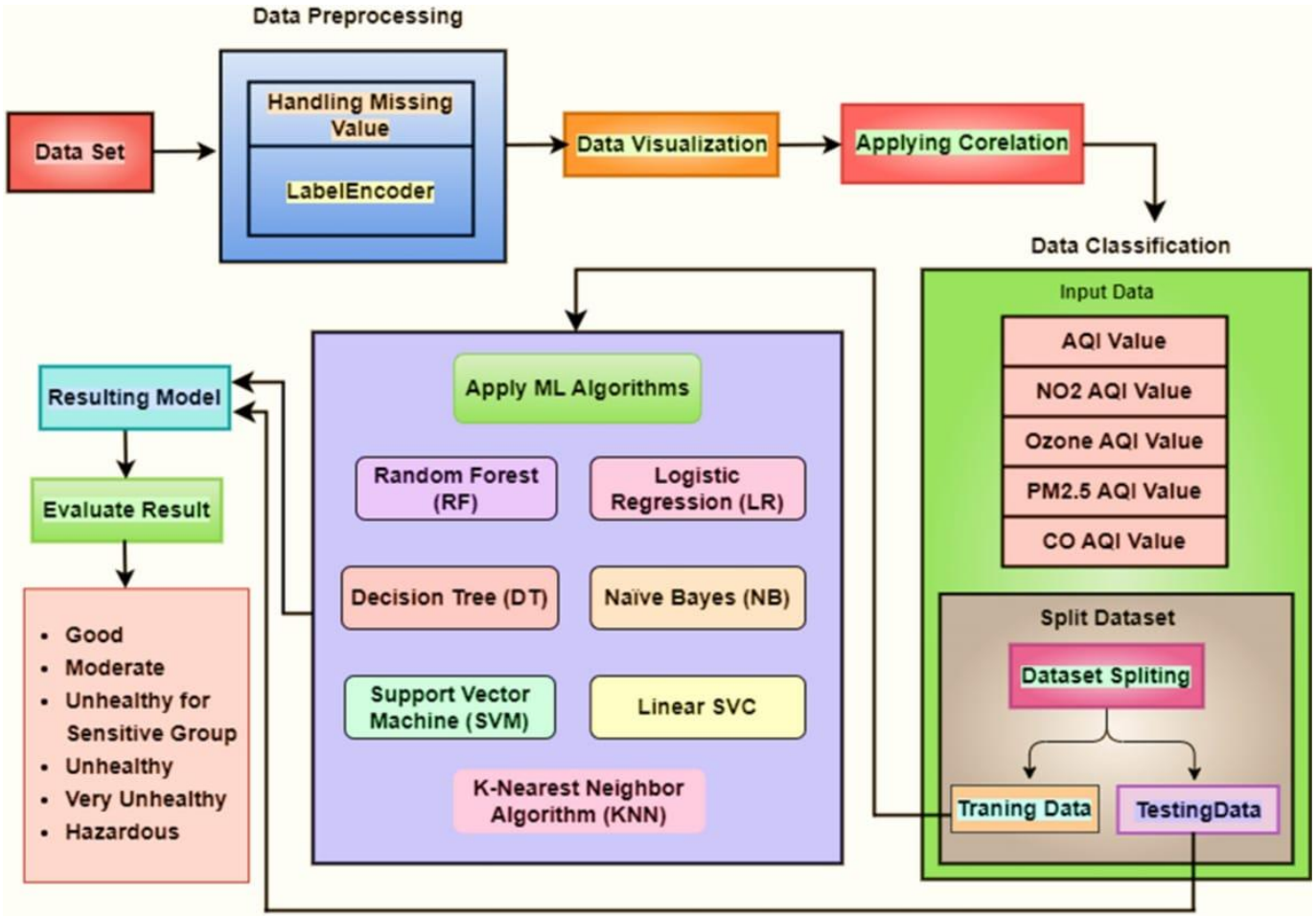
4. Objectives

The primary objectives of air quality focus on preventing, controlling, and reducing air pollution to protect public health and the environment. This involves monitoring air quality, setting standards, and implementing measures to reduce emissions from various sources.

Key technical objectives :

This involves establishing a network of air quality monitoring stations, using advanced sensors and technologies to detect and quantify pollutants like particulate matter (PM), nitrogen dioxide (NO2), ozone (O3), sulfur dioxide (SO2), and carbon monoxide (CO). Data collected is then used to assess air quality, track trends, and identify pollution hotspots.

5. Flowchart of the Project Workflow



6. Data Description

Parameters	Station	Average	Median	S.D.	Minimum	Maximum	RMAQG
PM_{10} (μgm^{-3})	Pasir Gudang	49.86	49.98	10.56	30.32	83.76	150
	Johor Bahru	42.12	39.01	12.03	23.05	109.14	
	Muar	51.61	49.78	13.81	29.77	105.56	
CO (ppm)	Pasir Gudang	0.689	0.682	0.151	0.374	1.173	30
	Johor Bahru	0.626	0.631	0.175	0.282	1.079	
	Muar	0.541	0.517	0.141	0.305	1.186	
NO_2 (ppm)	Pasir Gudang	0.013	0.014	0.004	0.005	0.031	0.17
	Johor Bahru	0.015	0.016	0.004	0.008	0.023	
	Muar	0.009	0.008	0.002	0.004	0.013	
SO_2 (ppm)	Pasir Gudang	0.007	0.006	0.005	0.001	0.022	0.13
	Johor Bahru	0.007	0.007	0.004	0.001	0.020	
	Muar	0.002	0.002	0.001	0.000	0.004	
O_3 (ppm)	Pasir Gudang	0.013	0.013	0.003	0.008	0.023	0.10
	Johor Bahru	0.014	0.014	0.002	0.009	0.021	
	Muar	0.019	0.018	0.003	0.007	0.029	
API	Pasir Gudang	47.51	48.00	7.98	31	66	50
	Johor Bahru	42.18	41.50	8.77	26	79	
	Muar	47.65	47.00	8.78	30	77	

7. Data Preprocessing

Handling missing values in air quality data is a critical step, as it directly affects model accuracy and reliability. Here's breakdown of best practices and techniques**.**

Duplicate records: The dataset contains duplicates data. It is irrelevant to the data. Since, the duplicate data is an dependent data therefore it can also be removed for the purpose of the project.

duplicate = df.duplicate()

8. Exploratory Data Analysis (EDA)

- **1. Data Collection and Preparation:**
- Gather data from sources like air quality monitoring stations, meteorological data, and other relevant datasets. Clean and prepare the data by handling missing values, outliers, and inconsistencies.
- **2. Univariate Analysis:**
- Examine individual variables (e.g., pollutant concentrations, temperature, humidity) to understand their distributions, ranges, and statistical properties.
- **3. Bivariate Analysis:**
- Explore relationships between pairs of variables using scatter plots, correlation matrices, and other techniques to identify potential correlations and associations.
- **4. Multivariate Analysis:**
- Investigate relationships among multiple variables, such as using clustering or dimensionality reduction techniques to identify groups of pollutants or regions with similar air quality characteristics.
- **5. Visualizations:**
- Use charts, maps, and other visualizations to effectively communicate findings and insights.

9. Feature Engineering

- - Time-based features: hour, day, month
- - Lag features for pollutants (for time-series modeling)
- - Composite pollution index
- - Interaction terms (e.g., pollutant × temperature)
- - Dimensionality reduction (e.g., PCA if needed)

10. Model Building

- Air quality model building involves using mathematical and numerical
- Techniques to simulate the processes that affect air pollutants, including their
- dispersion and chemical reactions in the atmosphere.

OUTPUT OF MODEL BUILDING

11. Model Evaluation

- - Evaluation metrics:
- - Mean Absolute Error (MAE)
- - Root Mean Squared Error (RMSE)
- - R^2 Score
- - Visual comparison of predicted vs actual AQI
- - Model benchmarking to choose the best performer

12. Deployment

Deployment public link : https://1a3f13dfc6b83ae4ca.gradio.live

Deploying air quality monitoring systems involves various approaches, from using low-cost sensors in urban areas to integrating air quality monitoring into smart city initiatives. The deployment can be for different purposes, like assessing the impact of microgrids on air quality, or for general monitoring of air pollution in cities. It's also crucial to consider the best practices for deploying low-cost sensors, including network planning, validation, and partnership with stakeholders.

13. Future Scope

- - Integration with live IoT sensor data
- - Mobile application for real-time AQI monitoring
- - Predictive models for pollution spikes and health advisory alerts
- - Geo-mapping of AQI predictions using GIS tools
- - Expansion to multiple cities or regions

14. Team Members and Roles

- - Project Manager: Coordinates timelines, deliverables, and resources – **ROOBINI.S**
- - Data Scientist: Data cleaning, EDA, and model building – **MANISHA. A**
- - ML Engineer: Model tuning, evaluation, and deployment – **TAMIZHVANI.R**
- - Documentation Lead: Reports, presentation, and final documentation – **KAVITHA.S**