A SYNOPSIS OF MINOR PROJECT I

ON

Air Quality Index Prediction System for Madhya Pradesh Cities using Machine Learning

Submitted In Partial Fulfillment of the Requirements for The Award of The Degree Of

BACHELOR OF TECHNOLOGY

Computer Science and Engineering Specialization

Artificial Intelligence and Machine Learning

By

Jeet Hemnani (0002AL231036)
Varun Chandanala (0002AL231073)
Rohan Kamble (0002AL231053)
Ansh Shrinag (0002AL231016)

GUIDE

Mr. Vipin Verma

Co-GUIDE

Dr Sagar Choudhary

School of Information Technology
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal

January- June 2025

Introduction

Overview

Environmental monitoring has become a cornerstone of modern urban planning, with the Air Quality Index (AQI) serving as the primary metric for assessing atmospheric health. As urbanization accelerates across Madhya Pradesh, particularly in cities like Bhopal, Indore, and Gwalior, the need for advanced technological interventions to monitor and forecast pollution levels has never been more critical. This project, "Air Quality Index Prediction for Madhya Pradesh Cities," utilizes the power of Machine Learning to transition from simple observation to proactive forecasting.

Air Quality Index (AQI)

The Air Quality Index (AQI) is a standardized tool used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become. The higher the AQI value, the greater the level of air pollution and the greater the health concern. The AQI is calculated based on the concentration of major air pollutants regulated by the Clean Air Act, specifically Ground-level Ozone, Particle Pollution (also known as Particulate Matter, including PM 2.5 and PM 10, Carbon Monoxide, Sulfur Dioxide, and Nitrogen Dioxide. In the context of this project, AQI serves as the dependent variable—the target value we aim to predict to help citizens understand the potential health risks in their environment.

Machine Learning (ML)

To achieve accurate forecasting, this project employs Machine Learning (ML), a dynamic subset of Artificial Intelligence (AI). Machine Learning involves training algorithms to learn patterns directly from data. By feeding historical data into the system, the machine "learns" about the complex, non-linear relationships between various environmental factors—such as temperature, humidity, and wind speed—and the resulting pollution levels. This project specifically utilizes Supervised Learning, where the model is trained on a labeled dataset containing past pollution records.

Prediction Models

A Prediction Model is a mathematical engine built using Machine Learning algorithms to forecast future outcomes. In this project, the prediction model functions as a regression tool. It analyzes historical trends to estimate a continuous numerical value—the future of AQI. We are utilizing the Random Forest Regressor, a robust ensemble technique that operates by constructing a multitude of "decision trees" during training. Each tree offers a prediction, and the model aggregates these to produce a final, highly accurate forecast. This approach minimizes errors and ensures that the model remains effective even when handling fluctuating weather patterns typical of Madhya Pradesh.

Rationale

While cities like Indore have consistently been recognized for cleanliness under the *Swachh Survekshan*, air quality remains a volatile parameter. For instance, recent data shows that despite good waste management, AQI in cities like Bhopal and Gwalior often dips into the "Poor" or

"Very Poor" categories during winter months and festival seasons. Existing monitoring systems provide real-time data, but there is a lack of accessible, localized *forecasting* tools that tell a common citizen what the air quality will be like tomorrow or next week.

This project is needed to bridge the gap between raw data and actionable insight. By using Machine Learning, we can identify hidden patterns in pollution data that are not immediately obvious to human observers. This tool will be particularly useful for sensitive groups (like children and the elderly) in Madhya Pradesh to plan their outdoor activities, thereby reducing their exposure to harmful pollutants.

Objectives

- To collect and preprocess historical air quality data for major Madhya Pradesh cities (Bhopal, Indore, Gwalior) from reliable sources like the Central Pollution Control Board (CPCB).
- To analyze the correlation between different pollutants (PM2.5, PM10, NO2) and the overall AQI.
- To train and test a Machine Learning model (such as Random Forest or LSTM) to predict future AQI values with high accuracy.
- To develop a simple interface where users can view the predicted AQI for their specific city.

Feasibility Study

Technical Feasibility-:

- The project is technically feasible as it utilizes well-established Machine Learning libraries in Python, such as Scikit-Learn and Pandas. The team has the necessary programming skills to implement these algorithms.

Operational Feasibility-:

- The data required for this project is publicly available through the CPCB and open-source repositories like Kaggle. No specialized sensors are required as we are working with secondary data.

Economic Feasibility-:

- This project is zero-cost. All software tools (Python, VS Code) and datasets are open-source and free to use. It requires no expensive hardware beyond a standard laptop.

Methodology/ Planning of work

Methodology / Planning of Work

Technical Framework and Tools The core development of this project is built upon a supervised Machine Learning approach, executed using the **Python** programming language. Python is selected for its extensive ecosystem of data science libraries. We will leverage **Pandas** for high-performance data manipulation and preprocessing, **NumPy** for numerical computations, and **Scikit-Learn** for implementing the machine learning algorithms. The primary algorithm chosen for this project is the **Random Forest Regressor**. This is a robust "Ensemble Learning" method that operates by constructing a multitude of decision trees during training time. For prediction, the model outputs the mean prediction of the individual trees. We selected this specific algorithm because of its high accuracy and its ability to handle non-linear relationships between meteorological data and pollutant concentrations. Furthermore, Random Forest effectively corrects the common habit of decision trees to "overfit" to their training set, ensuring our predictions for cities like Bhopal and Gwalior remain accurate even when analyzing new, unseen data.

Execution Pipeline The project will follow a structured data science pipeline divided into the following phases:

1. **Data Collection:** The first phase involves acquiring historical air quality data. We will extract daily AQI records for major Madhya Pradesh cities from reliable repositories such as **Kaggle** and the **Central Pollution Control Board (CPCB)** archives. This dataset will cover the past 3-5 years to ensure the model learns from long-term seasonal trends.

2. **Data Preprocessing:** Raw environmental data often contains noise. We will perform rigorous cleaning to handle missing values caused by sensor downtime and remove outliers that could skew the results. We will also apply **Normalization** techniques to ensure that all pollutant values (which vary in magnitude) are scaled to a similar range for efficient processing.

3. **Feature Engineering:** To improve model efficiency, we will select the most impactful features—such as **PM2.5, PM10, NO2,** and **Humidity**—that contribute most significantly to the AQI calculation. This step ensures the model focuses on the signals that matter most.

4. **Data Splitting:** To validate our results scientifically, the dataset will be partitioned into training and testing sets using a standard **80:20 ratio**. The training set (80%) will be used to teach the model the relationship between pollutant levels (independent variables) and the AQI (dependent variable). The remaining 20% will be kept strictly "unseen" by the

model and used only for final validation.

5. **Model Training & Evaluation:** The Random Forest Regressor will be trained on the processed data. Once trained, the model will be evaluated using standard performance metrics such as **Mean Absolute Error (MAE)**. This will quantify exactly how close our predicted AQI values are to the actual recorded values, providing a data-driven foundation for smarter urban planning.

Software/Hardware required for the project's development

Software:

- **Language:** Python 3.x
- **IDE:** Visual Studio Code or Jupyter Notebook
- **Libraries:** Pandas (for data handling), NumPy (for calculations), Scikit-learn (for ML algorithms), Matplotlib (for graphs).

Hardware:

- **Processor:** Intel Core i5 or equivalent
- **GPU**: RTX 3050 6GB (min requirement)
- **RAM:** 8GB or higher.
- **Storage:** 256GB SSD.

Expected outcomes

- A functional Machine Learning model capable of predicting AQI for Bhopal and Indore with at least 85% accuracy.
- A comparative study showing which pollutants are the primary contributors to poor air quality in MP cities.
- A graphical report visualizing the trend of air pollution in MP over the last 5 years.

Signature(s) of student

Signature(s) of student

Signature(s) of student

Signature(s) of student