Haldia Institute of Technology

Department of Information Technology

SESSION: 2025-2026

A PROJECT REPORT ON

Smart AQI Prediction

(AirLytics)

Submitted To:

Mr. Debasish Sahoo

Submitted By:

Sayantan Mondal - 10300223144

Soumya Byabarta - 10300223165

Shuvam Mondal - 10300223156

Affiliated to MAKAUT, West Bengal

Abstract

AirLytics is an end-to-end machine learning-based system designed to predict and analyze the Air Quality Index (AQI) using historical air pollution data combined with weather and temporal parameters. Air pollution stands as one of the most critical environmental challenges affecting public health and climate today. This project addresses this challenge by transforming raw environmental data into actionable insights through predictive analytics, visualization, and explainable machine learning techniques. The system utilizes a three-tier architecture comprising a Machine Learning layer trained on Google Colab, a backend hosted on Render, and a frontend deployed on Netlify. Key highlights include the use of Random Forest algorithms for regression and classification, and the implementation of SHAP (SHapley Additive exPlanations) for model interpretability.

Chapter 1: Introduction

1.1 Overview

The rapid industrialization and urbanization of the modern world have led to a significant deterioration in air quality. AirLytics—Smart Air Quality Prediction System—is a comprehensive solution designed to monitor and forecast air quality. It is an end-to-end system that leverages historical data to predict AQI values and classify them into categories such as 'Good,' 'Moderate,' or 'Poor'.

1.2 Objectives

The primary focus of this project is to bridge the gap between raw data and understandable environmental insights. The specific objectives are:

- To predict precise AQI values using regression models.
- To classify air quality into standard safety buckets.
- To integrate pollutant data with weather and temporal factors for higher accuracy.
- To ensure the system is fully deployed and accessible via a web interface.

Chapter 2: Problem Statement

2.1 The Challenge

Air quality has a direct and profound impact on human health, agricultural productivity, and urban sustainability. However, accurately predicting AQI is inherently complex due to several factors:

- **Interactions:** The presence of multiple interacting pollutants.
- **Variability:** Seasonal changes and fluctuating weather conditions.

• **Non-linearity:** The complex, non-linear relationships between various environmental features.
• **Data Volume:** The challenge of processing large-scale, time-dependent datasets.

2.2 Limitations of Existing Systems

Traditional environmental monitoring systems often suffer from significant limitations:
• They are primarily reactive rather than predictive.
• They lack interpretability, making it difficult to understand why AQI is high.
• There is a lack of real-time analytics support and poor handling of extreme pollution events.

2.3 Proposed Goal

To overcome these limitations, this project aims to build a data-driven, scalable AQI forecasting system using machine learning. The system is designed to support environmental monitoring, enhance public health awareness, and facilitate smart decision-making.

Chapter 3: System Architecture

3.1 Architectural Design

AirLytics follows a robust Client–Server Architecture designed for scalability, modularity, and production readiness. The system is divided into three distinct layers:

1. Machine Learning Layer: Developed within Google Colab. Responsible for data preprocessing, feature engineering, and model training/evaluation.

2. Backend Layer: Hosted on the Render cloud platform. Utilizes FastAPI/Flask to create a REST API for model inference and data exchange.

3. Frontend Layer: Deployed on Netlify. Provides interactive dashboards for AQI visualization and analytics.

Chapter 4: Methodology and Technologies

4.1 Technology Stack

The project utilizes a modern tech stack to ensure efficiency and performance:
• **Machine Learning:** Python, Pandas, NumPy, Scikit-learn, and SHAP for Explainable AI.
• **Models:** Random Forest Regressor and Random Forest Classifier.
• **Visualization:** Matplotlib, Seaborn, and advanced charts like Streamgraphs and Network Graphs.
• **Web Technologies:** HTML, CSS, JavaScript/React for the frontend, and FastAPI for the backend.

4.2 Dataset Description

The system is trained on a state-wise AQI dataset from India. The dataset includes:
• Pollutants: PM2.5, PM10, NO, NO2, NOx, SO2, CO, O3, Benzene, Toluene, and Xylene.
• Weather Parameters: Temperature, Humidity, and Wind Speed.
• Temporal Features: Year, Month, and Season.

4.3 Data Preprocessing

To ensure model accuracy, rigorous preprocessing steps were undertaken:
• **Imputation:** Handling missing values using KNN and statistical imputation methods.
• **Formatting:** Parsing and sorting date-time information.
• **Scaling:** Applying StandardScaler to normalize feature values.
• **Encoding:** One-hot encoding for categorical features like 'State'.
• **Splitting:** Implementing a time-based train–test split to strictly avoid data leakage.

Chapter 5: Machine Learning Models

5.1 Model Selection

The project employs Random Forest algorithms for both regression and classification tasks.
• **Regression:** The Random Forest Regressor is used to predict specific AQI numerical values.
• **Classification:** The Random Forest Classifier categorizes air quality into buckets (e.g., Good, Severe).

5.2 Justification

Random Forest was selected because it performs exceptionally well on tabular environmental data. It is capable of handling non-linear relationships, is robust against noise, and offers high predictive accuracy without requiring heavy hyperparameter tuning. Furthermore, it is less prone to overfitting compared to single decision tree models.

5.3 Explainable AI (XAI)

A key feature of AirLytics is its implementation of SHAP (SHapley Additive exPlanations). This ensures the model is not a 'black box.'
• **Global Explanation:** Identifies the overall impact of features. The top contributors to AQI prediction were identified as PM2.5, CO, PM10, NO2, and SO2.
• **Local Explanation:** Provides a breakdown for single AQI predictions, improving transparency and trust.

Chapter 6: Implementation and Analytics

6.1 Advanced Analysis

Beyond basic prediction, the system includes optimization for Extreme AQI Handling. By using sample weighting for high AQI values, the model improves predictions for severe pollution events and reduces bias toward normal ranges.

6.2 Visualization

The frontend provides advanced visual analytics to derive deep environmental insights:
• State-wise AQI pie charts.
• AQI gauge charts for real-time status.
• Streamgraphs showing AQI trends over time.
• Feature relationship network graphs.

Chapter 7: Conclusion and Future Scope

7.1 Conclusion

AirLytics successfully delivers a full-stack, machine learning-powered air quality prediction system. The project demonstrates strong predictive performance, interpretable machine learning through SHAP, and end-to-end deployment capability. It effectively showcases the practical application of data science in environmental monitoring.

7.2 Future Enhancements

To further improve the system, the following enhancements are proposed:
• Integration of IoT sensors for real-time local data collection.
• GIS-based pollution mapping for geographic analysis.
• Implementation of Deep Learning models like LSTM and Transformers for better time-series forecasting.
• Development of mobile applications for alerts and dashboards.