ABSTRACT

This project focuses on analyzing and predicting the Air Quality Index (AQI) using machine learning algorithms. The dataset includes multiple environmental parameters such as pollutant concentrations, temperature, and humidity, which are used to determine the air quality category. Through data preprocessing, visualization, and feature correlation analysis, significant factors affecting AQI were identified. Machine learning models, including CatBoost Classifier, were implemented and fine-tuned using RandomizedSearchCV to achieve optimal prediction performance. The model achieved high accuracy in predicting AQI categories such as Good, Moderate, and Poor, providing a reliable tool for environmental monitoring and decision-making.

INTRODUCTION

Air pollution has become a major concern in urban areas due to rapid industrialization and increasing vehicle emissions. The Air Quality Index (AQI) serves as an essential metric to communicate the level of air pollution and its potential health effects. This project aims to develop a data-driven approach for analyzing and predicting AQI levels using advanced data science techniques.

The workflow involves data collection, exploratory data analysis (EDA), correlation mapping, and predictive modeling. Various visualizations were used to understand pollutant trends and category distributions. The CatBoost Classifier was selected for its efficiency and high performance on categorical data. Model optimization was performed using RandomizedSearchCV to fine-tune hyperparameters and improve accuracy. This project demonstrates how machine learning can effectively contribute to environmental data analysis and air quality prediction.

CONCLUSION

The developed AQI prediction model successfully classified air quality levels with high accuracy. The CatBoost Classifier, after hyperparameter tuning, provided the most reliable results compared to baseline models. The findings highlight the strong influence of key pollutants and meteorological factors on AQI. Visualization and correlation analysis provided valuable insights into the patterns of air pollution. Overall, the project underscores the potential of data science in environmental monitoring, enabling policymakers and citizens to make informed decisions to reduce pollution and improve air quality.