Predicting Air Quality Index (AQI) in India Using Machine Learning Techniques

Abhishek Aggarwal : 740077752

1 Introduction

This paper uses machine learning techniques to analyze India's Air Quality Index (AQI). The main objective is to use regression and classification models, specifically RandomForestRegressor and Support Vector Machine (SVC), to predict AQI values across different locations. Policy-making, environmental preservation, and public health all depend on an understanding of AQI. Proactive steps to enhance air quality and lower health hazards are aided by AQI prediction.

In recent years, the rapid industrialization and urbanization in India have led to a significant increase in air pollution levels. Monitoring and predicting the AQI is essential to safeguard the health of millions of people and to develop effective strategies to mitigate pollution. By employing machine learning models, we can analyze historical data, identify patterns, and make accurate predictions about future AQI levels, which can inform decision-making processes and public health advisories.

2 Dataset Description

The file, "IndiaAQI.csv," includes data on AQI levels and several pollutants (PM2.5, PM10, SO2, NO2, OZONE, CO, NH3) that were gathered from monitoring stations all over India. The last update date, pollution levels, AQI, major pollutant, and geographic locations are among the important features. To deal with missing values and create an appropriate format analysis, the data was preprocessed.

The dataset is large, spanning several areas and offering a thorough understanding of the air quality in various Indian locations. The concentration levels of different contaminants, which are known to have detrimental impacts on human health, are included in every entry in the collection. Cleaning the data, resolving missing values with methods like imputation, and converting the data into a format that was appropriate for model training and assessment were all part of the preprocessing stages. This guarantees that the machine learning models can be trained efficiently, producing predictions that are more accurate and trustworthy. [1]

3 Machine Learning Techniques

Two machine learning techniques were used to analyze the data but only Random Forest Regressor was used as it was more accurate:

- **Random Forest Regressor (RFR)**: A regression model used to predict continuous AQI values. The model combines multiple decision trees to improve predictive accuracy and control over-fitting. Since it has shown the most accurate findings when it comes to AQI value prediction, I prefer the Random Forest Regressor.
- **Support Vector Machine (SVC)**:A classification model that classifies AQI levels using a radial basis function (RBF) kernel. Because of its efficacy in high-dimensional spaces and capacity to represent intricate feature interactions, SVC was selected. And i also selected few other model for this then mostly all of them were giving me a low accurate result that's why i gone with Random Forest Regressor

4 Data Preparation

Data preparation involved several critical steps to ensure the dataset was suitable for model training:

- **Handling Missing Values**: I have Used SimpleImputer to fill in missing pollutant values by replacing them with the median value of each pollutant column.
- **Data Transformation**: Pivoted the data to have pollutants as columns, making it easier to analyze. Removed rows with missing AQI values to ensure data quality.
- **Feature and Target Preparation**: Prepared the features (pollutants) and the target variable (AQI) for model training.
- **Data Splitting**: Split the dataset into training and testing sets using train_test_split, with 80% of the data for training and 20% for testing.

5 Results and Insights

The RandomForestRegressor model provided reliable predictions with a Mean Squared Error (MSE) of 54.81% and an R^2 score of 0.99 and achieved the accuracy of 99.43%, indicating strong performance. The Support Vector Machine (SVC) achieved an accuracy of 0.00% and an R^2 score of 0.85% demonstrating its effectiveness in classifying AQI levels. These results drastically underscore the potential of machine learning techniques in accurately predicting and classifying AQI.

Strong relationships between particular pollutants and AQI values were also found in the investigation, suggesting that some pollutants have a greater influence on air quality than others. The AQI values' geographic distribution indicated locations with continuously high pollution levels, indicating the need for focused measures in these areas. Policymakers were able to create specialized policies for various locations because to the visualization of the most common contaminants among states, which offered insightful information about the causes of pollution.

6 Conclusion

I have learned a lot about the challenges of utilizing machine learning techniques to predict air quality thanks to this work. I discovered the RandomForestRegressor's advantages in managing actual environmental data and producing precise AQI forecasts by experimenting with it. As managing missing values and converting data into an appropriate format were essential steps in guaranteeing the model's efficacy, I also realized the significance of data preparation and preprocessing. I was better able to comprehend the connections between various pollutants and AQI values as well as the geographic spread of pollution in India once I visualized the data. The need of ongoing data monitoring and the requirement for reliable validation methods were both emphasized by this study. Future models can attain greater accuracy and dependability by utilizing sophisticated validation techniques and integrating more extensive datasets.

7 Limitations

- **Data Limitations**: The dataset contained missing values and potential data quality issues. The geographic scope was was insufficient, and the geographic breadth was restricted to particular areas of India not all areas were taken into consideration.
- **Method Limitations**: In order to produce more accurate results, the models first make assumptions about the relationships between variables, which may not accurately reflect the complexity of the real world. Second, without much fine-tuning, the Random Forest Regressor may have trouble with extremely complicated patterns, even though it can handle non-linear data.

Figures and References

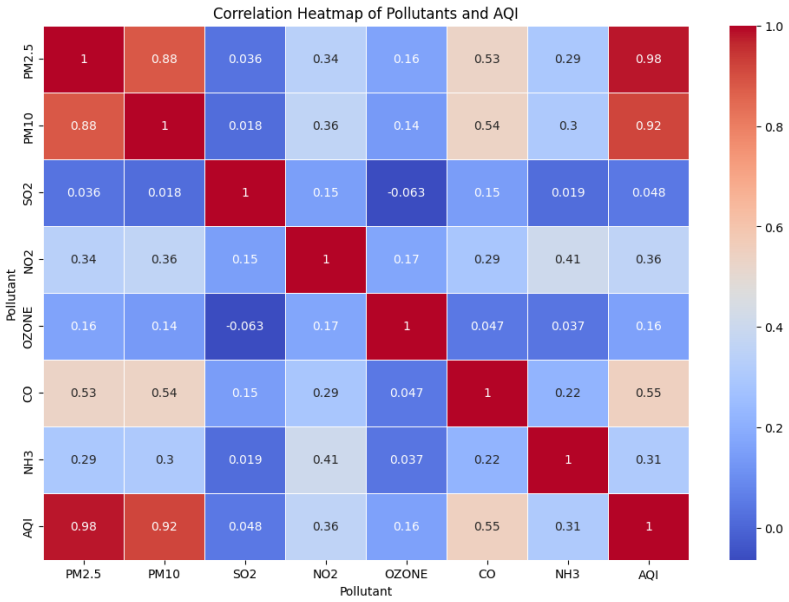

Figure 1: Correlation Heatmap of Pollutants and AQI

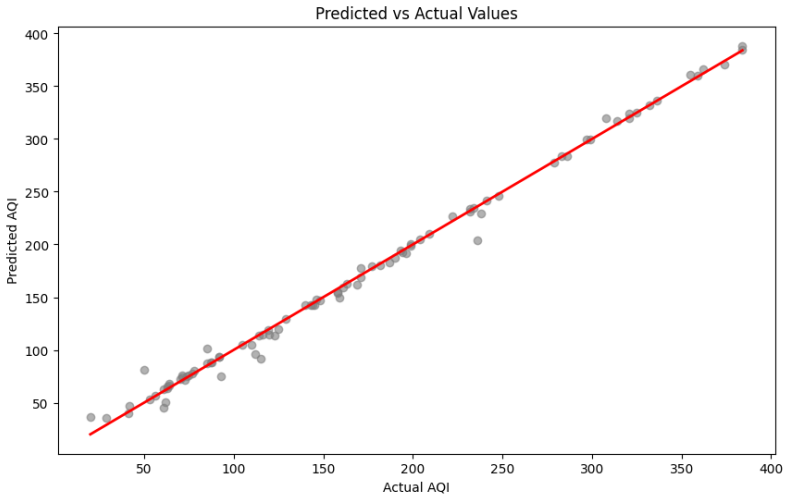

Figure 2: Predicted vs Actual AQI Values

References

[1] Bhadra Mohit. *India Air Quality Index(2024) Dataset*. URL: `https://www.kaggle.com/datasets/bhadramohit/india-air-quality-index2024-dataset/data`.

[2] Varun Paidipelli. *AIR QUALITY PREDICTION MODEL*. URL: `https://www.kaggle.com/code/varunpaidipelli/air-quality-prediction-model#Prediction-Model`.

Appendix

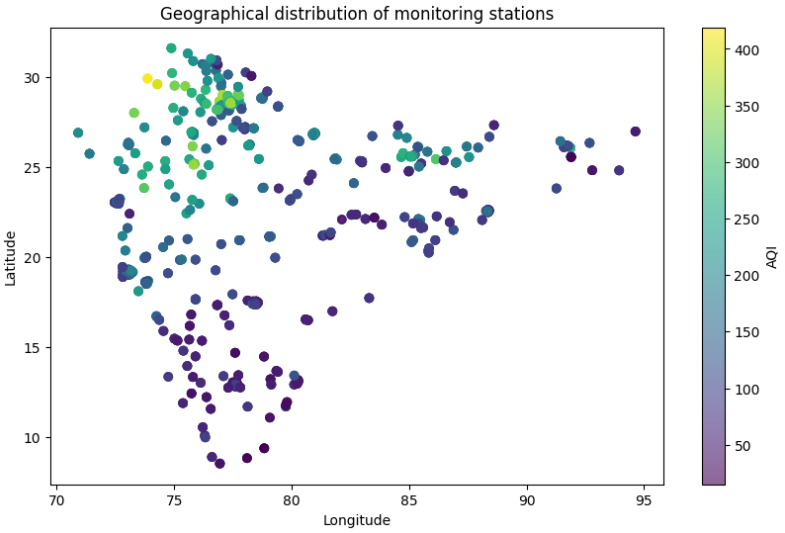

Figure 3: AQI on basis of Geographical Distribution in India

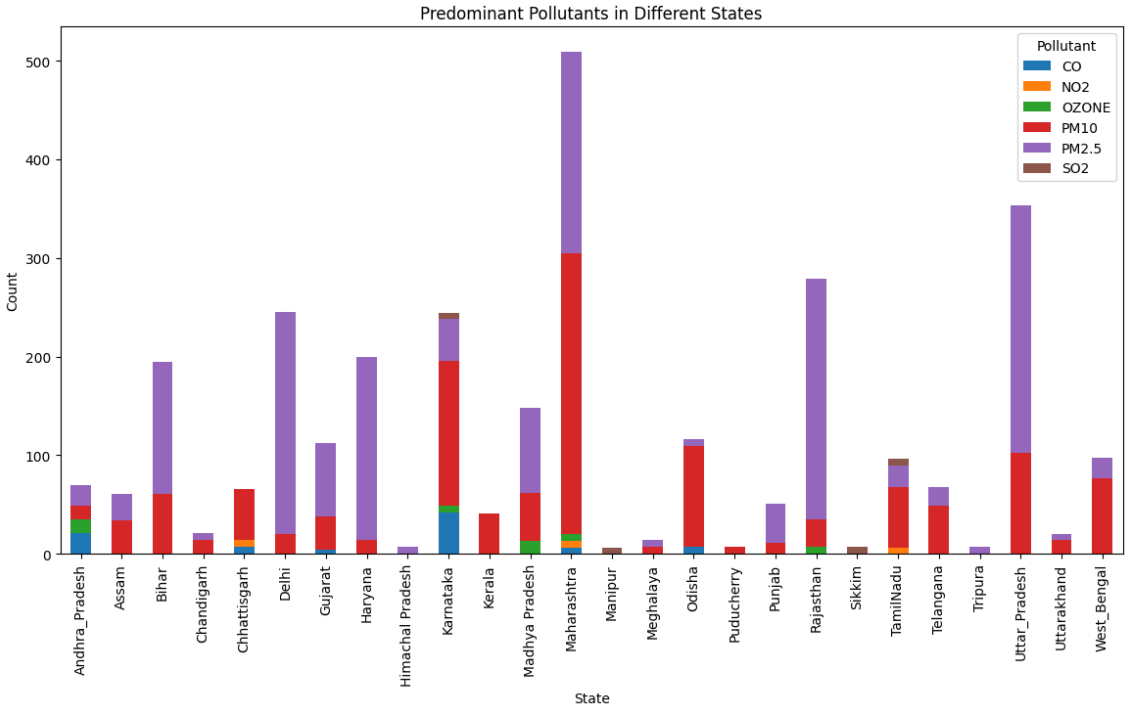

Figure 4: Stacked bar chart of pollutants

Generative AI Statement

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

- (NO) I have used GenAI tools for developing ideas.
- (YES) I have used GenAI tools to assist with research or gathering information.
- (NO) I have used GenAI tools to help me understand key theories and concepts.
- (NO) I have used GenAI tools to identify trends and themes as part of my data
- analysis.
- (YES) I have used GenAI tools to suggest a plan or structure for my assessment.
- (NO) I have used GenAI tools to give me feedback on a draft.
- (NO) I have used GenAI tool to generate image, figures or diagrams.
- (YES) I have used GenAI tools to proofread and correct grammar or spelling errors.
- (NO) I have used GenAI tools to generate citations or references.
- (NO) Other: [please specify]
- I have not used any GenAI tools in preparing this assessment.

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.