Air pollution modelling using geospatial & machine learning approach

ABSTRACT

We forecast the air quality of India by using machine learning to predict the air quality index of a given area. Air quality index of India is a standard measure used to indicate the pollutant (so2, no2, rspm, spm. etc.) levels over a period. Air quality prediction using machine learning is a project that aims to provide accurate and reliable predictions of air quality in different regions. The project leverages advanced machine learning algorithms to analyse historical data on air quality and predict air quality index. By accurately predicting air quality levels, the project can help individuals and authorities take preventive measures to reduce exposure to pollutants and improve public health. The project utilizes various tools and technologies, including Python and Scikit-Learn to develop a robust and reliable system. Overall, this project has significant potential to positively impact public health and the environment, improving air quality and reducing the negative effects of pollution.

Keywords-AQI, dataset, preprocessing ,outliers, prediction

INTRODUCTION

India, the world's fastest-growing industrial nation, is generating record levels of pollutants, including PM2.5, CO2, and other dangerous air pollutants. According to the Indian air quality standard, pollutants are indexed according to their scale, which shows the amounts of significant pollutants on the atmosphere. The air quality of a given state or nation is a measure of the impact of pollutants on the respected regions. Our atmosphere is contaminated by a variety of atmospheric gases. Every pollutant has a unique index and scales at various levels. The primary contaminants The data can be categorized according to the restrictions using the individual AQI obtained from indexes like (no2, so2, rspm, spm).

We gathered the information from the Indian government database, which includes the concentrations of pollutants at different locations around India. We have created a model that can estimate India's air quality in any location by predicting the air quality index of each available data point in the dataset. We can identify the main pollutants that cause pollution and the areas of India that are most severely impacted by them by forecasting the air quality index. Using a variety of methods, this forecasting model extracts different information from the data to identify areas that are severely impacted within a certain region (cluster). This provides more details and understanding regarding the origin and severity of the pollutants.

LITERATURE REVIEW

Literature Survey of Air Quality Prediction Using Machine Learning

The paper "Air Quality Prediction Using Machine Learning" explores various aspects of air quality monitoring and prediction through machine learning techniques(Raviteja et al., 2024)

Importance of Air Quality Monitoring: The introduction emphasizes the critical need for continuous air quality monitoring due to its significant impact on human health and environmental balance. It highlights that air pollution monitoring is essential for controlling pollution levels effectively and understanding its sources and intensity

- **Machine Learning in Air Quality Prediction**: The paper discusses the growing interest in using machine learning for air quality prediction. It notes that various machine learning models, such as neural networks, decision trees, and support vector machines, have been employed to predict pollutant concentrations. These models utilize a range of input variables, including meteorological data, traffic data, and emission data, to enhance prediction accuracy (Bhattacharya & Shahnawaz, n.d.)
- **Accuracy and Challenges**: The results indicate that machine learning models can achieve high accuracy, with some models reaching up to 90%. However, the performance of these models is heavily reliant on the quality and quantity of input data. Incomplete or low-quality data can lead to inaccurate predictions, and the interpretability of these models remains a significant challenge (Jackulin & Murugavalli, 2022)
- **Incorporating Advanced Techniques**: The paper also suggests incorporating additional data sources, such as satellite imagery and low-cost sensors, to enhance model performance. Addressing uncertainties in input data is another area for future research that could lead to improved accuracy in predictions

Artificial Neural Networks (ANN) have been effectively used to predict particulate matter levels, such as PM2.5, showing promising accuracy. Decision Tree and Random Forest models are popular choices for predicting air quality indices due to their ability to handle complex datasets and produce reliable predictions.(Raviteja et al., 2024)

Support Vector Regression (SVR) also appears frequently in studies, particularly for forecasting pollutants like PM10 and PM2.5, owing to its robustness in regression tasks. Long Short-Term Memory (LSTM) networks, specifically Bidirectional LSTM (BILSTM), are highlighted for their improved performance in air quality forecasting, providing lower error rates such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Lastly, K-Nearest Neighbors (KNN), known for its simplicity and accuracy in classification, is used in the current study to predict air quality based on local pollutant data. Together, these methods reflect the diversity of approaches used in air quality prediction and underline the importance of model selection based on specific pollutants and data characteristic(Raviteja et al., n.d.)

PROPOSED SYSTEM

There are multiple processes in the suggested machine learning-based air quality prediction system. Gathering historical air quality data from multiple sources, including satellite data and government monitoring stations, is the first stage. In order to prepare the data for use in machine learning models, it is first pre-processed to eliminate any outliers, clean the data, and scale it. Based on their relationship to air quality, pertinent features like traffic patterns, pollution levels, and meteorological data are chosen. Creating machine learning models that can precisely forecast air quality using the chosen features is the next stage. Support vector Classifier, random forests Classifier, XGB Classifier, Logistic Regression , KNN and neural networks are popular machine learning algorithms for predicting air quality.

By imputing missing values, random forest algorithms can deal with missing data and lessen its effect on prediction accuracy. Feature selection, the process of determining the most significant factors influencing air quality, is also made possible by the application of random

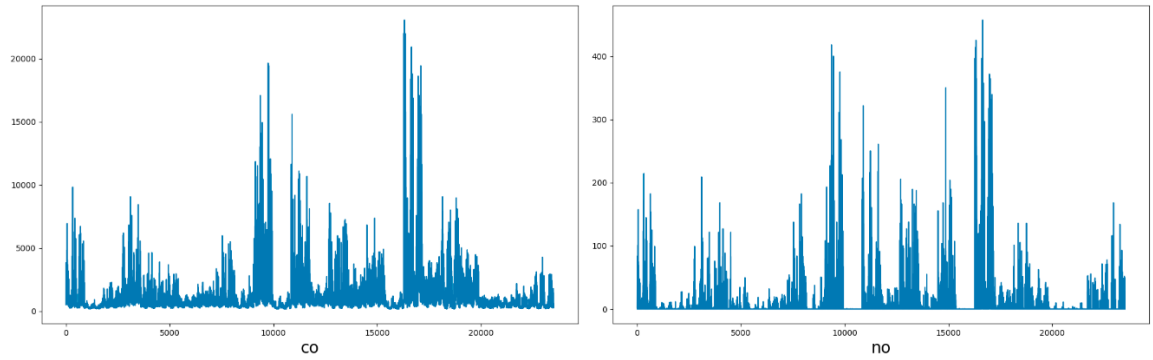
forest and decision tree algorithms. This procedure aids in lowering the model's variable count, which can increase the precision and efficacy of air quality forecasts. Lastly, by examining the random forest can shed light on the variables influencing air quality. In order to guarantee the models' accuracy and dependability in forecasting air quality, they might be trained on pre-processed data and cross-validated.

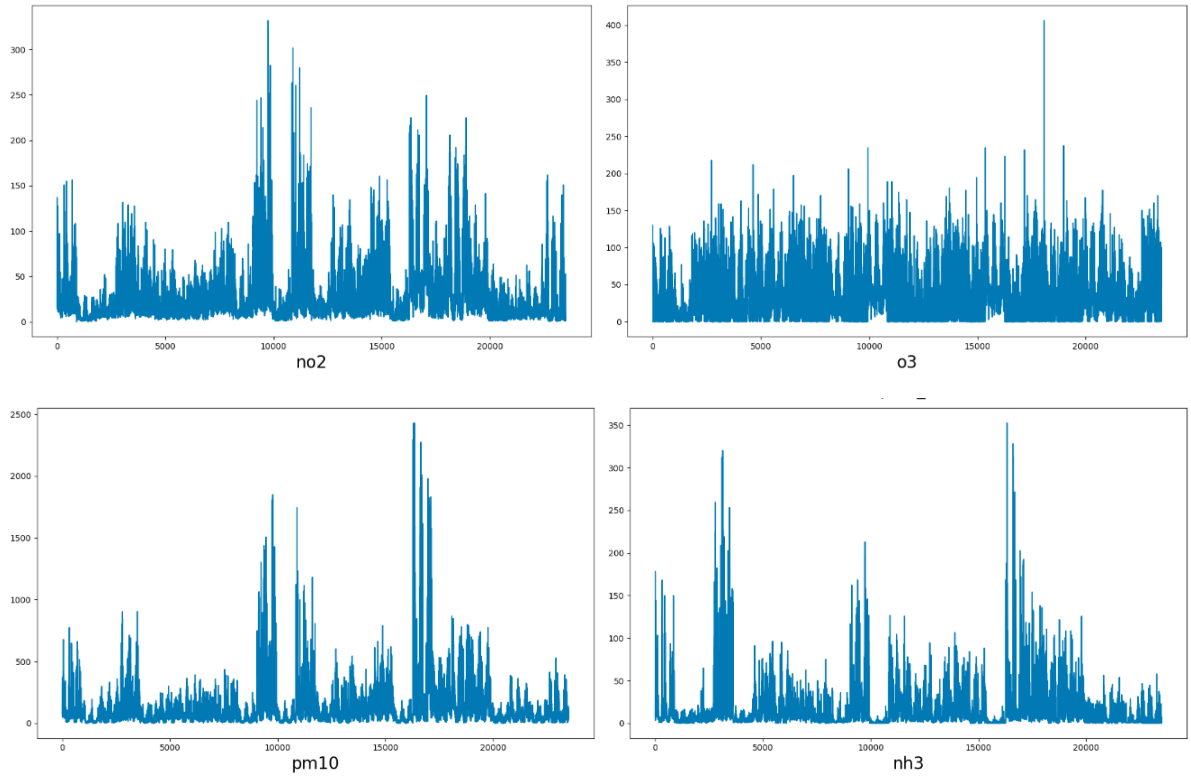
2. IMPLEMENTATION

2.1 DATA COLLECTION:

The first step in data collection is to identify the sources of data. There are several sources of air quality data, including government agencies, private organizations, and research institutions. The most reliable source of air quality data is government agencies, which collect and report data regularly. These agencies use various types of instruments to measure the levels of air pollutants in the atmosphere. Once you have identified the sources of data, the next step is to collect the relevant data. The data should include environmental factors that affect air quality, such as temperature, humidity, wind speed, and other factors. The dataset should also contain the corresponding AQI values for each data point. It is important to ensure that the data is of high quality and is collected using standardized methods to ensure consistency and accuracy.

EXPLORATORY DATA ANALYSIS



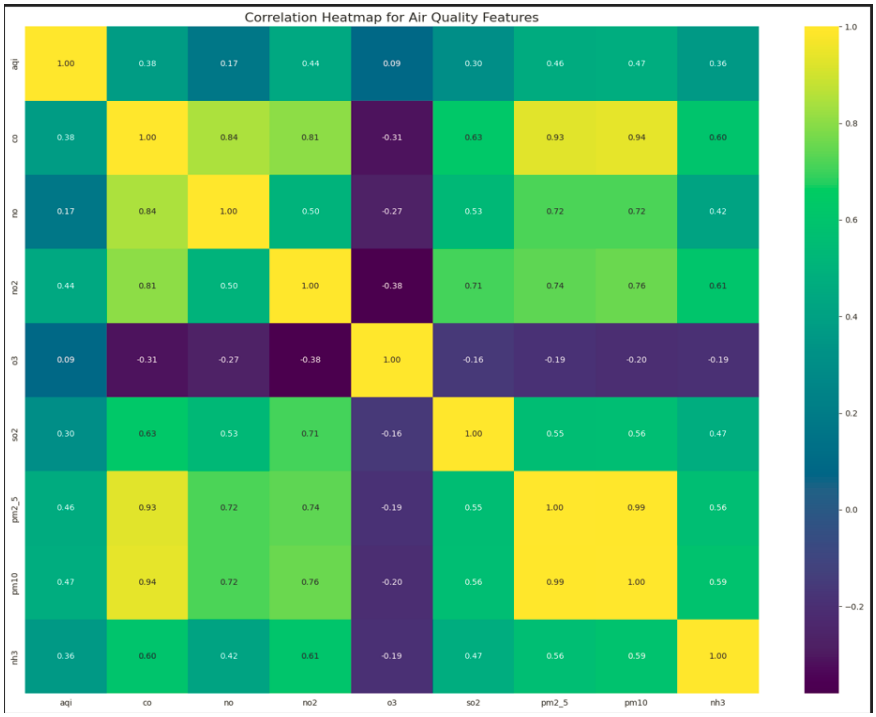


2.2 PREPROCESSING

Data preprocessing is an important step in preparing the data for analysis. It involves transforming the raw 3 data into a format that can be easily analysed by machine learning algorithms. The following are the common steps in data preprocessing:

Data Cleaning: Data cleaning involves removing or fixing any missing or incorrect data points in the dataset. This is important because missing or incorrect data can affect the accuracy of the predictions.

Feature Selection: Feature selection is the process of selecting the relevant features that will be used in the prediction model. In air quality prediction, the features include temperature, humidity, wind speed, and other factors that affect air quality.



Feature Scaling: Feature scaling involves scaling the features to a similar range, typically between 0 and 1, so that the model can learn effectively. This is important because some features may have a larger range than others, and this can affect the performance of the machine learning algorithm.

Data Splitting: Data splitting involves dividing the dataset into training and testing sets. The training set is used to train the machine learning algorithm, while the testing set is used to evaluate the performance of the model.

Data Encoding: Data encoding involves transforming categorical data into numerical data for machine learning algorithms to process. This is important because machine learning algorithms can only process numerical data.

SMOTE : In our project, we applied Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the dataset. SMOTE generates synthetic samples for the minority class by interpolating between existing instances, effectively increasing its representation without merely duplicating samples. This approach helps the model to learn more effectively by providing a balanced view of both classes, improving overall classification performance and reducing bias toward the majority class. By applying SMOTE to the preprocessed dataset, we

aimed to achieve better prediction accuracy and generalization across imbalanced data.

2.3 FEATURE EXTRACTION

In preprocessing the focus is on feature selection to improve the performance of the model. Among all the available features, the ones with highest as the most important are being selected. In this model NO2, SO2, PM2.5, CO are selected as the most important features that impact on the decision. This avoids the overfitting problem by avoiding or reducing unwanted or partially relevant features in the dataset. The availability of many features may arise the problem of curse of dimensionality, that may in turn reduce the efficiency of the model by a greater extent Feature selection is the process of selecting the relevant features that will be used in the prediction model. In air quality prediction, the features include temperature, humidity, wind speed, and other factors that affect air quality Feature scaling involves scaling the features to a similar range, typically between 0 and 1, so that the model can learn effectively. This is important because some features may have a larger range than others, and this can affect the performance of the machine learning algorithm.

2.4 TRANING AND BUILDING MODEL

After completing data collection and preprocessing. the next step in air quality prediction is to train and build machine learning models. In this module, we will discuss the process of model training and building using four different algorithms: Logistic Regression, KNN Classifier, Support vector classifier, Random Forest, XGB classifier and ANN

KNN Classifier KNN (K-Nearest Neighbors) is a classification algorithm that determines the class of a new data point based on the classes of the k-nearest neighbors in the training set. In air quality prediction, KNN can be used to predict the AQI for a new combination of environmental factors by finding the k-nearest neighbors in the training set and determining the

average AQI for those neighbors. The model is trained using the preprocessed dataset. The training data is used to determine the optimal value of k and to calculate the distances between the data points.

Random Forest Classifier Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions. In air quality prediction, Random Forest can be used to predict the AQI for a new combination of environmental factors by combining the predictions of multiple decision trees. The model is trained using the preprocessed dataset. The training data is used to construct multiple decision trees using different subsets of the features and the data points.

Support Vector classifier SVC chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine. The model is trained using the preprocessed dataset. The training data is used to construct multiple decision trees using different subsets of the features and the data points.

XGB Classifier: The XGBoost Classifier, short for Extreme Gradient Boosting, is an ensemble technique that builds a sequence of decision trees. Each new tree attempts to correct the errors of the previous ones by emphasizing incorrectly classified data points. This model uses gradient descent to minimize the loss function, making it highly efficient and often more accurate than traditional boosting methods. The model is trained using the preprocessed dataset, where each tree iteration progressively improves the overall prediction accuracy.

K-Nearest Neighbors (KNN): The KNN algorithm classifies new data points based on the 'k' closest examples in the training set. By calculating distances (typically using Euclidean distance), KNN finds similar cases, making it a memory-based algorithm. The model is trained using the preprocessed dataset, but it does not explicitly learn an internal model. Instead, it stores all data points and compares them during the prediction phase, categorizing a point based on the majority class of its nearest neighbors.

Logistic Regression: Logistic Regression is a statistical model that classifies data by estimating the probability that a given input belongs to a certain category. It applies the logistic function to ensure that the output probabilities are bounded between 0 and 1. This model learns weights for each feature during training on the preprocessed dataset, optimizing them to minimize the error in predictions by using a cost function. It is particularly suited for binary classification tasks, where it assigns data points to one of two classes.

MODEL	Training Accuracy	Testing Accuracy
ANN	85%	86%
Logistics Regression	75%	76%
Random Forest	100%	95%
XGB Classifier	94%	85%
Support Vector Classifier	51%	51%
KNN	94%	91%

PREDICTION

This is the final step in the air quality prediction process, which involves using the best-performing model to predict the AQI for new data points. The process of making predictions begins with collecting 8 new data on the environmental factors that affect air quality. Once the new data has been collected, it is preprocessed using the same preprocessing steps. This includes data cleaning, feature selection, feature scaling, data splitting, and data encoding to ensure that the new data is suitable for machine learning analysis. After the new data has been preprocessed, the best-performing model can be used to predict the AQI for the new data points. The predicted AQI values can then be used to determine whether the air quality is good or poor. It is important to note that the accuracy of the predictions depends on the quality of the data and the performance of the machine learning algorithm used to predict the AQI.

RESULT AND DISCUSSION

Air quality prediction using machine learning is a widely researched area due to its potential for mitigating the health and environmental effects of air pollution. Several machine learning models have been used to predict air quality, including neural networks, decision trees, and support vector machines. These models use a range of input variables, such as meteorological data, traffic data, and emission data, to predict pollutant concentrations in the atmosphere. Studies have shown that machine learning models can accurately predict air quality, with some models achieving up to 92% accuracy. However, the performance of these models is highly dependent on the quality and quantity of the input data. Models trained on incomplete or low-quality data may produce inaccurate predictions. In addition, the interpretability of machine learning models remains a challenge in air quality prediction. As machine learning models are often regarded as black boxes, it can be difficult to understand how the models arrive at their predictions. This lack of transparency can make it challenging to identify the causes of air pollution and develop effective mitigation strategies. Overall, air quality prediction using machine learning has shown promising results, but further research is needed to improve the accuracy and interpretability of these models. By combining machine learning with traditional air quality monitoring techniques, it may be possible to better understand the sources and impacts of air pollution and develop effective strategies for reducing its effects on human health and the environment

CONCLUSION AND FUTURE ENHANCEMET

In conclusion, air quality prediction using machine learning has shown potential for accurately predicting air pollution concentrations. However, the performance of these models is highly dependent on the quality and quantity of the input data, and the interpretability of these models remains a challenge. Therefore, further research is needed to enhance the accuracy and interpretability of machine learning models for air quality prediction. Future research could focus on improving the quality and quantity of input data used for air quality prediction. This could involve incorporating more sources of data, such as satellite imagery or data from

low-cost air quality sensors. Additionally, research could focus on developing methods to account for uncertainty in input data, which could improve the accuracy of machine learning models. Another important area for future research is improving the interpretability of machine learning models. This could involve developing methods for explaining the predictions made by machine learning models, such as feature importance analysis or local interpretability methods. Finally, machine learning models for air quality prediction could be integrated with traditional air quality monitoring techniques to provide a more comprehensive understanding of air pollution. This could involve combining machine learning models with ground-based air quality monitoring stations or integrating them with mobile air quality monitoring platforms, such as drones or vehicles. Overall, air quality prediction using machine learning 6 has shown promising results, and further research could lead to improved prediction accuracy and more effective strategies for mitigating the health and environmental effects of air pollution

FUTURE ENHANCEMENT The Future enhancements deals with collecting data from the IOT Device which is built to get the required pollutants of our location. With the help of other Advance machine learning algorithms, the project can predict the future quality of air like for next hours, days and week

REFERENCES

Bhattacharya, S., & Shahnawaz, S. (n.d.). *Using Machine Learning to Predict Air Quality Index in New Delhi*.

Jackulin, C., & Murugavalli, S. (2022). A comprehensive review on detection of plant disease using machine learning and deep learning approaches. *Measurement: Sensors*, *24*, 100441. https://doi.org/10.1016/j.measen.2022.100441

Raviteja, B., Tejaswini, P., & Reddy, U. S. (n.d.). *Air Quality Prediction Using Machine Learning*.

Raviteja, B., Tejaswini, P., & Reddy, U. S. (2024). *Air Quality Prediction Using Machine Learning*. *6*(2).