WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Air Quality Index Prediction using Machine Learning

Ribadiya Dhruvil Rasikbhai[1]***, Panchal Vedant Rohitbhai***[2]***, Prajapati Pritesh Ashvinbhai***[3]***, Rana Veerpalsinh Bahadursinh***[4]***, Mehta Rohan Sunilbhail***[5]***, Prof. Janki Patel***[6]

[1-6] Sal College Of Engineering, Department Of Engineering, Ahmedabad, Gujarat, India

ABSTRACT

Air pollution is a serious worldwide problem which negatively impacts on human health, climate, and ecosystems. Accurate AQI prediction is necessary for timely warning and environmental policy-making. In this paper, we investigate the use of different machine learning algorithms to forecast AQI based on environmental information like pollutant concentrations (PM2.5, PM10, NO_2, CO, SO_2, O_3), temperature, and humidity. The models are trained and tested on real-time and historical air quality datasets. This study compares the algorithms like Linear Regression, RandomForestRegressor, Support Vector Machine (SVM), XGBoostRegressor, GaussianNB. The outcomes reveal that ensemble-based models, specifically Linear Regression and XGBoostRegressor model offer high prediction accuracy.

Keyword: Air Quality Predication, Data Processing, Feature extraction, Classification, Machine Learning, Support Vector Machine (SVM) etc.

Introduction

Air pollution may be described as a change in air quality that can be defined by the measurements of chemical, biological or physical contaminants in the air. Thus, air pollution refers to the unwanted presence of impurities or the abnormal increase in theratio of some constituents of the atmosphere. It may be divided into 2 parts visible and invisible air pollution. Air pollution results from the occurrence in the atmosphere of poisonous elements, largely brought about by human activities, although at times it may be a result of natural events like volcanic eruptions, dust storms and forest fires, also stripping the air of quality. Air Quality Index (AQI) refers to a measure of air pollution in an area on a number scale. Air pollution is a significant worldwide health hazard, causing 7 million premature deaths every year (WHO). Pollution is rising due to urbanization, industrialization, vehicles, power plants, chemical processes, and some of the other natural processes like agricultural burning, volcanic eruptions, and wildfires. The reason for selecting this topic is air pollution impacts everyone's environment and health. Has AQI helps predict how dirty the air is, so we can take action Current air quality monitoring systems only report current conditions,but can't predict future air quality. We need a reliable way to forecast air quality levels, so we can take proactive steps to reduce pollution, protect public health, and create a sustainable future. AQI index between 0 and 500, utilized to convey how dirty the air is now or is predicted to be.

The AQI incorporates five significant air pollutants which include: particulate matter (PM), ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO), and sulfur dioxide (SO2). This project seeks to design a system capable of forecasting the quality of air we inhale. Through the application of machine learning, wecan look at historical data and weather conditions to predict when air pollution will be high. This can be used by governments and citizens to take measures to minimize pollution and safeguard public health. Our vision is to create a tool that simplifies air quality prediction, makes it accurate, and makes it accessible to all.

LITERATURE SURVEY

Due to uncontrolled growth in deforestation, urbanization, population growth rate, and industrialization leads to the problem of Air quality and, air is getting worse and worse polluted that impact our live. Quality of air is directly connected with human life, trees, and livestock all suffer from the effects of air pollution. Now a days research on air quality prediction has become a hotspot topic. Varous air quality prediction model have been broadly adopted including machine learning, Aritificial Neural Network(ANN) and deep neural network. But exact prediction of air quality is still a major problem.

To deal with extremely dynamic air quality forecasts, a multi-point deep learning model based on convolutional long short term memory (ConvLSTM) is suggested by Mokhtari et al. The authors created a deep learning model to forecast high dynamic air pollution levels. The prediction accuracy of the LSTM and SVR-based models was determined to be 95% and 92.9 percent, respectively.

Methodology

The machine learning method for forecasting Air Quality Index (AQI) entails a few important steps that begin with the retrieval of proper environmental and meteorological data like pollutant concentrations (PM2.5, PM10, NO_2, SO_2, CO, O_3), temperature, humidity, wind speed, and so on. After they are gathered, the data is preprocessed to deal with missing data, delete noise, normalize features, and obtain helpful patterns. If AQI values are not available directly, International Journal on Advanced Computer Theory and Engineering45they can be estimated based on standard pollutant concentration breakpoints specified by environmental agencies. Next, Exploratory Data Analysis (EDA) is conducted to learn about trends, correlations, and feature importance. Relevant features are then chosen based on the insights to train machine learning models. Based on the task, regression models (such as Random Forest Regressor, XGBoost, or LSTM) are applied to forecast continuous AQI values, whereas classification models (e.g., Decision Trees, SVM, or neural networks) are utilized to forecast categorical AQI levels. The chosen model is trained and cross-validated using methods such as cross-validation and optimized for maximum performance. Performance metrics like RMSE, MAE, R^2 score for regression, or accuracy, precision, recall, and F1-score for classification are employed to measure model performance. Lastly, the model is implemented for real-time or scheduled predictions and kept under continuous surveillance to ascertain accuracy, with regular updates in terms of new data to ensure durability over time.

Machine Learning Algorithms

Linear Regression

Linear regression is a supervised machine learning model employed for making predictions of continuous values such as the Air Quality Index (AQI). Linear regression tries to establish a linear relationship between input features like PM2.5, PM10, NO2, etc., and the target, AQI. It learns from historical data by minimizing the difference between the actual and predicted AQI using the least squares method. Once trained, the model will be able to predict AQI from new environmental information. While it is easyto implement and convenient to utilize, linear regression cannot possibly represent sophisticated patterns in air quality data but works well as a baseline model.

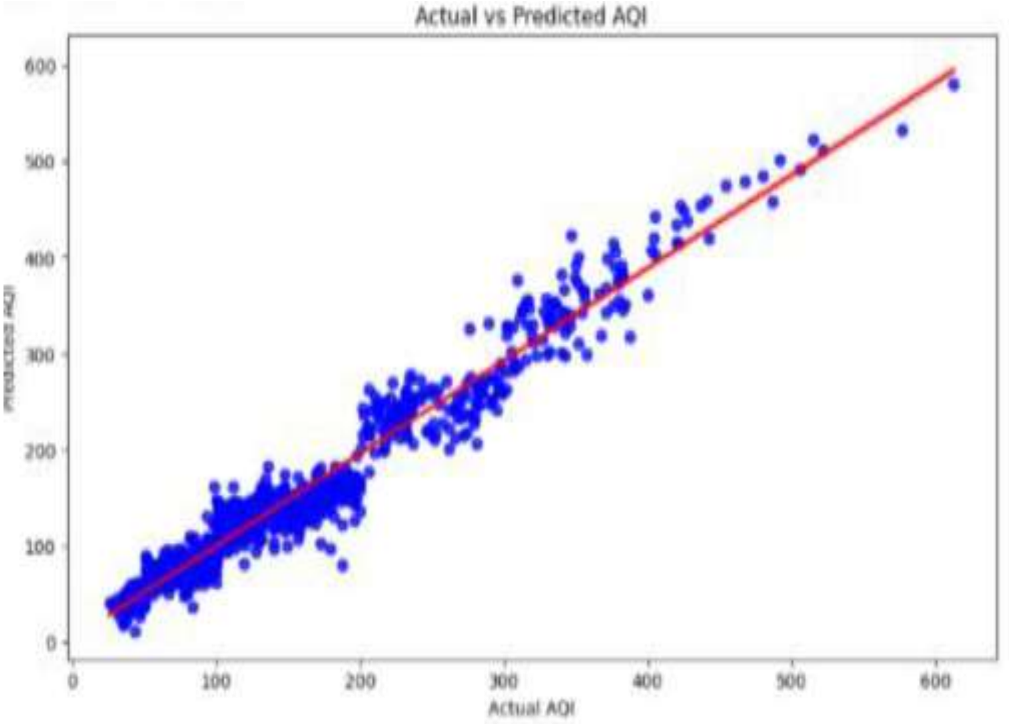

Fig. 1 Linear Regression

Random Forest

Random Forest is a machine learning ensemble algorithm that performs very well in predicting continuous values such as the Air Quality Index (AQI). It achieves this by training multiple decision trees and then averaging their predictions to enhance accuracy and minimize overfitting. In predicting AQI, inputs such as PM2.5, PM10, NO2, CO, and other environmental factors are utilized to train the model. Each tree in the forest is trained on a randomsubset of the data, which allows it to capture complicated, non-linear relationships within air quality patterns.

Random Forest is stable, capable of dealing with missing data, and tends to produce high prediction accuracy and is therefore a reliable option for AQI prediction.

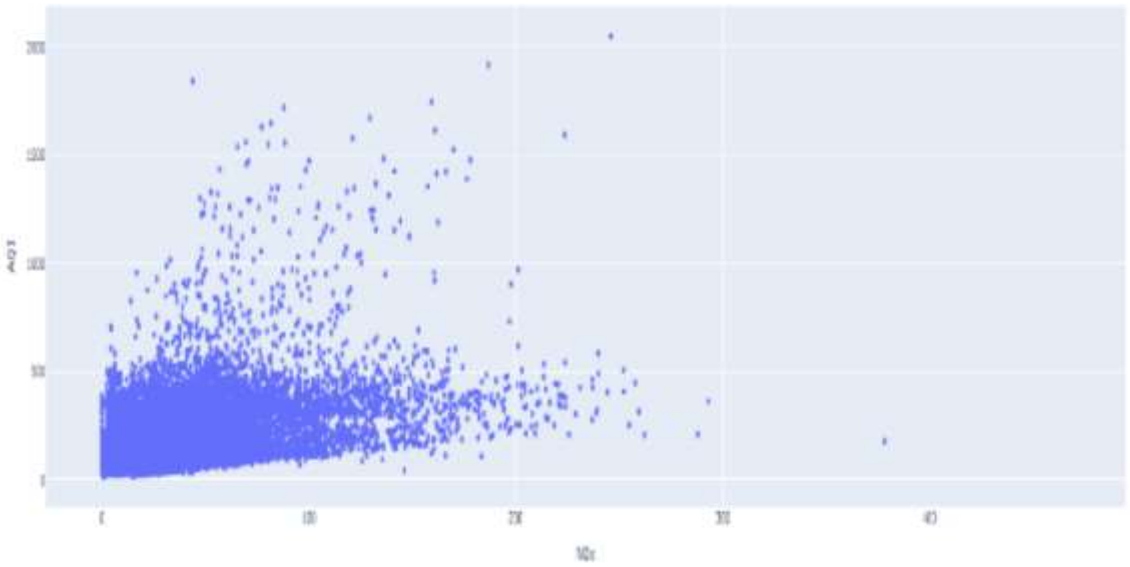

Fig. 2 Random Forest

Support Vector Regression (SVR)

Support Vector Machine regressor is an efficient supervised learning algorithm for making predictions of continuous values such as the Air Quality Index (AQI). SVM regressor searches for a hyperplane in the high-dimensional feature space that minimizes theerror while keeping the error within some threshold (epsilon) and the margin between actual and predicted values as large as possible. In AQI forecasting, input variables like PM2.5, PM10, NO2, and other contaminants are utilized to train the model to learn patterns and trends in air quality. The SVM regressor can identify non-linear relationships with kernel functions and is thus ideal for intricate environmental data. Though it needs more computational power than linear regression, it tends to offer greater accuracy for AQI forecasting when appropriately tuned.

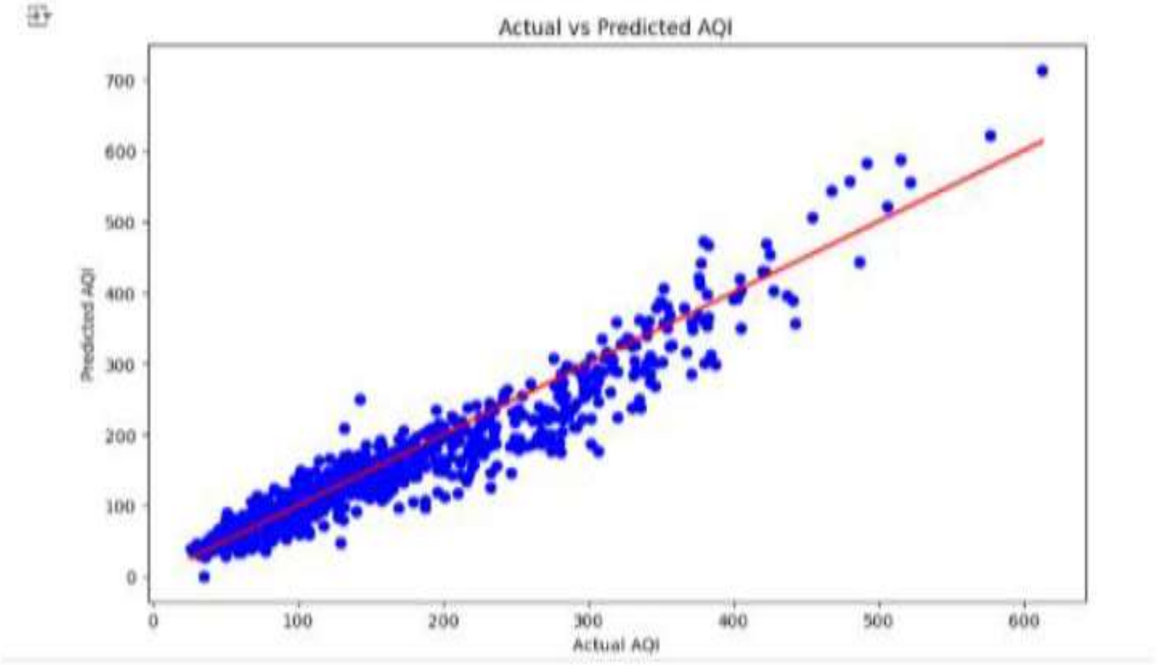

Fig. 3Support Vector Regression (SVR)

XGBoost

XGBoost(Extreme Gradient Boosting) is a fast and effective machine learning algorithm commonly employed for the prediction of continuous values such as the Air Quality Index (AQI). It performs by creating an ensemble of decision trees sequentially, with each subsequent tree aimed at correcting the mistakes made by the preceding ones, optimizing the model through gradient descent. In predicting AQI, XGBoost employs characteristics such as PM2.5, PM10, NO2, CO, and weather to learn intricate patterns in the data. XGBoost manages missingdata, overfitting, and big data effectively using methods such as regularization and parallel computing. Because of its efficiency in accuracy and speed, XGBoost is most commonly used for accurate and consistent AQI forecasting.

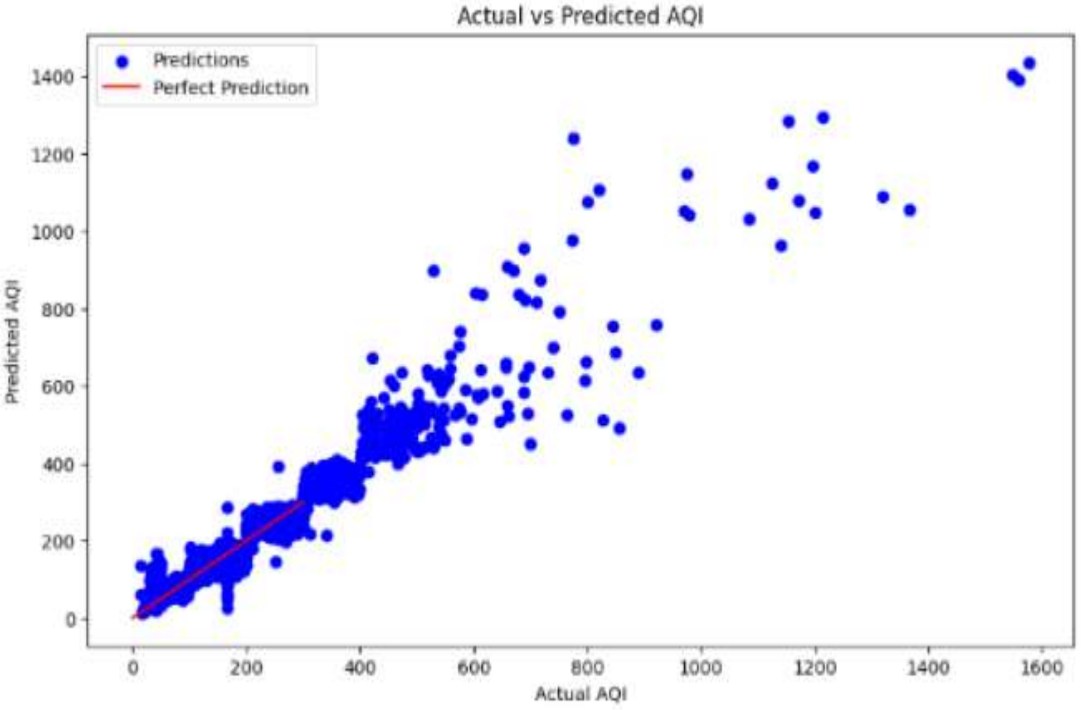

Fig. 4 XGBoost

GaussianNB

Gaussian Naive Bayes (Gaussian NB) is a probabilistic machine learning model that is generally employed for classification, but can be modified for regression-type tasks like Air Quality Index (AQI) prediction by classifying AQI into various levels (e.g., good, moderate, unhealthy). It operates on the basis of Bayes' Theorem under the assumption that the input features—like PM2.5, PM10, NO2, and CO—are normally distributed and independent of one another. The algorithm computes the probability of every AQI category using the input features and predicts the most probable category. Gaussian NB is efficient, easy, and does well even with small data when the independence assumption is marginally broken. It is less efficient for predicting the exact AQI value but can be beneficial in classifying air quality into health-related classes.

Result

The output of machine learning (ML) based air quality index (AQI) prediction normally entails the forecasting of concentrations of different air pollutants like particulate matter (PM2.5, PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide(CO), and ozone (O3). Prediction is carried out using models trained with past air quality records comprising environmental variables like temperature, humidity, wind speed, and location.For instance, a machine learning model such as a Random Forest, Support Vector Machine (SVM), or deep learning model may be trained to forecast AQI values. The model takes input features such as time of day, meteorological data, and past levels of pollutants to project future AQI levels. The result is usually a forecasted AQI value, usually categorized into levels like "Good," "Moderate," "Unhealthy," etc., to guide people on the air quality of a specific location. The model's performance is measured based on metrics such as Mean Squared Error (MSE) or R-squared, depending on whether it is a regression or classification problem. By making precise AQI predictions, these machine learning algorithms assist with decision-making on public health, environmental policy, and air quality control.

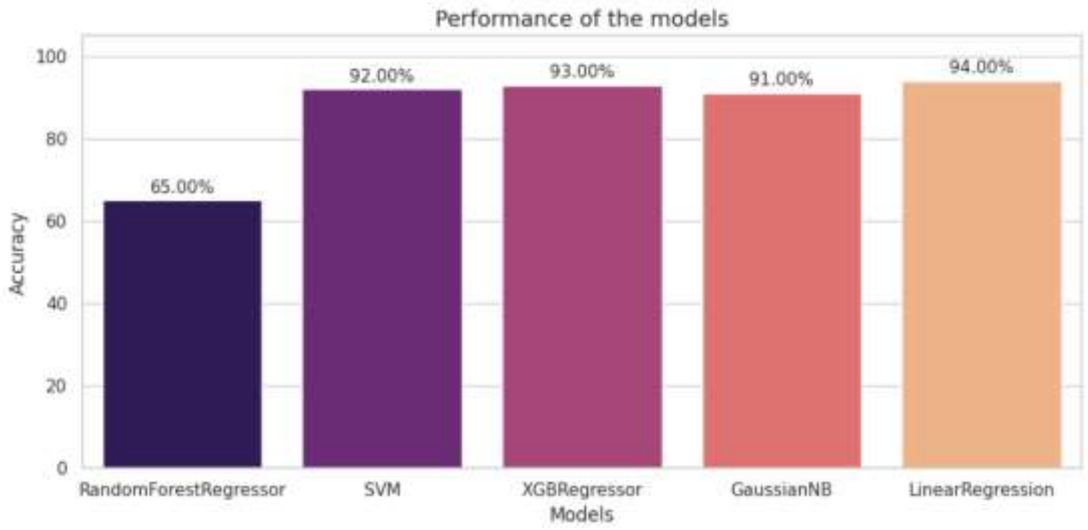

Fig.6 Comparision of all Algorithms

Future Scope

The future potential of air quality index (AQI) forecasting with machine learning is bright, with increasing capabilities in data acquisition through sensors and IoT devices having a positive effect on the accuracy of models. Hyper-local and real-time predictions will improve with time, and interventions in hotspot regions can be done promptly. Machine learning algorithms may incorporate more heterogeneous data, including socio-economic and behavioral factors, to deliver rich information on pollution trends. Moreover, AQI forecasting could be a critical component of smart city technology, allowing for the best traffic and industrial operations in real-time, based on pollution predictions. With advancements in AI methods, the use of explainable AI (XAI) might enhance transparency, boosting confidence in the forecasts and supporting better public health and environmental policy.

Conclusion

In summary, machine learning has vast potential in air quality index (AQI) prediction, making more precise and timely predictions that can be used to reduce the effects of air pollution. ML models, using historical and real-time data, allow for better insight into patterns of pollution and their impact on public health. Predictions inform decision-making in urban planning, environmental policy, and public health interventions. As technology continues to evolve, machine learning algorithms will become finer and provide local and more accurate AQI forecasts. Overall, using machine learning for AQI forecasting will be an important enabler in designing smarter, healthier cities and enhanced environmental sustainability.

References

- Y. Zhang et al., “A Predictive Data Feature Exploration-Based Air Quality Prediction Approach,” IEEE Access, vol. 7, pp. 30732–30743, 2019, doi: 10.1109/ACCESS.2019.2897754.
- Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. B. Valarmathi, G. Arulkumaran (2023). Journal of Environmental and Public Health. Compares SVR, Random Forest, and CatBoost regression for multiple Indian cities; addresses imbalance with SMOTE.
- An air quality index prediction model based on CNN-ILSTM. Wang, Jingyang, Li, Xiaolei, Jin, Lukai, ... (2022). Scientific Reports. Proposes CNN-ILSTM (CNN + improved LSTM) model, comparing with SVR, Random Forest, MLP, etc.
- Kamaljeet Kaur Sidhu et al. (2024). arXiv preprint. Uses multiple ML and deep learning models (Random Forest, XGBoost, SVM, LSTM) to predict AQI across monitoring stations; also looks at contributor like stubble burning.
- Mokhtari, W. Bechkit, H. Rivano, and M. R. Yaici, "Uncertainty-Aware Deep Learning Architectures for Highly Dynamic Air Quality Prediction," IEEE Access, vol. 9, pp. 14765-14778, 2021, 10.1109/ACCESS.2021.3052429.
- Air Quality Dataset. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set/data.
- Q. Liu, B. Cui, and Z. Liu. Air quality class prediction using machine learning methods based on