WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Predicting Air Quality Index using Machine Learning.

Suru Sirisha

B. Tech Student, GMR Institute of Technology, Rajam,532127, India

ABSTRACT

Air pollution is increasingly becoming a problem in the world. It caused by cars, industrial activities, and even by burning of fossil fuels and coal. Air pollution is one the serious threat to the environment that can harm human health, especially in cities.The air contains some of the harmful pollutants, such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO). India's rapid Industrialization, urbanization, and burning of fossil fuel have witnessed a sharp drop in the air quality factor in the country. This paper explores how the Stacked model can be used in India to accurately predict the air quality index classification labels such as Good, Moderate ,Unhealthy, etc. The XGBoost model was used as the base model, and that was combined with an SVC model which acts as final estimator in the setup of stacked classifier. The hybrid pipeline is designed to enhance the predictability by providing reliable AQI classifications leading to better environmental monitoring and awareness about public health. The dataset is taken from Kaggle repository which has the air quality information

Keywords: Air pollution, Air Quality Index (AQI), Stacked model, Support vector Classifier, XGBoost algorithm, Machine learning.

Introduction

Air pollution consists of chemicals or particles in the air that can harm the health of humans, animals, and plants. It also damages buildings. Pollutants in the air take many forms. They can be gases, solid particles, or liquid droplets. Pollution enters the Earth's atmosphere in many different ways.Most air pollution is created by humans, taking the form of emissions from factories, cars or planes. Second-hand cigarette smoke is also considered air pollution. These man-made sources of pollution are called Anthropogenic sources. Some types of air pollution, such as smoke from wildfires or ash from volcanoes, occur naturally. These are called natural sources. Air pollution is most common in large cities where emissions from many different sources are concentrated. Sometimes, mountains or tall buildings prevent air pollution from spreading out. This air pollution often appears as a cloud making the air murky. It is called smog. The word "smog" comes from combining the words "smoke" and "fog''.To predict air quality, machine learning algorithms are used. These algorithms analyze data from various pollutants and provide forecasts based on patterns and trends. Common machine learning models include decision tree ,XGBoost etc. XGBoost, in particular, is favored for its high accuracy in predicting air quality. These models use input data such as pollutant levels (e.g., PM2.5, NO2, CO) to assess the air quality and categorize it into different levels, which helps individuals to take appropriate actions.

AQI Categories and Health Concerns:

Good: Air quality is good. The health risk is minimal, or there is no risk at all.

Moderate: Air quality is moderate. Sensitivity groups-including children, the elderly, and persons with respiratory or heart problems-individuals may experience some minor health effects.

Unhealthy for sensitive groups: The air quality is unhealthy for sensitive groups. They may experience more significant health effects.

Unhealthy: Air quality is unhealthy. Everyone may begin to experience health effects.

Very unhealthy: Air quality is very unhealthy. Health effects may be serious for everyone.

Hazardous: Hazardous levels of air quality. All people affected with major health problems.

Literature Review

The study focuses on predicting air pollution in Bengaluru, it is a severe problem due to urbanization, industrialization, burning fossil fuels and affecting human health. It uses a DECISION TREE algorithm to predict air quality based on historical data from the past year and analyzing pollutants levels such as pm2.5,pm10,co,nox,so2 and the dataset is taken from official website of the government of Karnataka to train this model. The model achieved a mea squared error of 25.67 and an R²score of 0.54 [1]. This study introduces an optimized model using GREY WOLF OPTIMIZATION WITH DECISION TREE algorithm for accurate AQI predictions. It achieves high accuracy rates with maximum performances in some of the cities in India and they have mentioned the future research aims to integrate deep learning for improved accuracy [2].

The project develops a machine learning algorithm, specifically LINEAR REGRESSION, to predict hourly pollution levels, focusing on PM2.5,NO2,SO2. In this paper the dataset is taken from the Delhi government's website, detailing pollutants and their concentrations across various locations and the model gets the precision of 96% for the AQI [3]. The paper discusses air quality has a significant impact on human health, especially in children and the study aims to build a model using SVR to predict air quality levels based on historical data. The support vector regression model was used to forecast pollutant levels and gets the accuracy of 93.4% [4].

The paper demonstrates a method called GA-KELM. The suggested technique consists of applying genetic algorithms to a kernel extreme learning machine in order to enhance the predictions of air quality levels. Compared with traditional models, the model offers lower RMSE and MSE values, with a high correlation coefficient R^2 [5]. In this study, the machine learning models from the domains of SARIMA, SVM, and LSTM are utilized to predict AQI in the city of Ahmedabad. Techniques used are quite diverse, including outlier handling along with missing values, among others [6]. But air pollution has the chance of causing harmful effects to human health. Some studies have produced predictions and estimation of particulate matter levels by employing machine learning techniques namely ARIMA and Facebook Prophet methods [7].

The SVR has been used in California for the important air pollutants: CO, NO2, SO2, and PM2.5 as well as the AQI. Data techniques have been applied to enhance the predictiveness of SVR. The performance of the SVR model is highly accurate, as 94.1% of categories of AQI were correctly classified. Accurate forecasting of air quality is an important public health issue. It helps give timely warnings and ways to control emission [8]. The paper refers to a superior decision tree algorithm used to predict air quality called C4.5 with 94.5% accuracy. The model of a decision tree is more accurate and can predict the quality of air reliably. According to the paper the better model works for the classification and prediction with a large amount of data from air quality [9].

The paper illustrates various algorithms applied in the prediction of air quality such as Random Forest (RF), Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), Decision Trees (DT), and Long Short-Term Memory (LSTM) networks in air quality prediction. It also emphasizes that some of the key weather conditions including relative temperature, relative humidity, and wind speed are significant elements in the construction of air quality prediction models [10]. This paper explores different machine learning techniques for predicting the Air Quality Index (AQI) of New Delhi, Bangalore, Kolkata, and Hyderabad support vector regression (SVR), random forest regression (RFR), and CatBoost regression (CR). The research applies the synthetic minority oversampling technique -smote for the data to be balanced and accurate. The paper aims to study how best AQI can be predicted for climate control and to raise the alarm among people regarding hazardous health effects due to air pollution [11].

Six years of air pollution data from 23 Indian cities will enable enhancement in the monitoring effort. The analysis finds that AQI is mainly affected by the following pollutants: PM10, PM2.5, CO, NO2, and SO2. The main contributors among these are the PM10 and PM2.5. Five machine learning models were used. Gaussian Naive Bayes had the best accuracy, while Support Vector Machine had the worst. XGBoost did the best job of predicting real values of AQI. There was a dramatic fall in 2020 of levels of air pollution by nearly all pollutants, caused by the strict lockdowns in place during the COVID-19 pandemic [12]. It considers different models of machine learning to predict the Air Quality Index (AQI) and Air Quality Grade (AQG). Based on six pollutants that got data from 2014 to 2019, the Stacked Model did better than the single models, with a score of 0.973 for predicting AQI and an accuracy of 0.970 for AQG. This presents the fact that the Stacked Model is strong and very accurate in all measures. The stacked machine learning model would help reduce the problems of single models. This would lead to better and more accurate predictions for both AQI and AQG. It seems to offer promise in managing air quality [13].

The document discusses a number of machine learning algorithms that can predict how much SO2 is in the air. It emphasizes watching air pollutants for public health and the health of the environment. Time series analysis in AR and ARIMA models is used to predict SO2 levels. This explains how predictive models can make sense of changes in air quality. Future research should focus on attributes like pm2.5 and AQI, suggesting ongoing model updates and enhancements [14]. The review talks about different machine learning methods used to predict air quality. These methods include Linear Regression, Decision Tree, Random Forest, Neural Networks, and Support Vector Machine (SVM). Data from places like Kaggle is split into training and testing sets, but changes in the atmosphere make predictions harder, especially for pollutants like PM2.5. Several pollutants, including CO2 and NO2, greatly contribute to air pollution, which can cause health problems like lung disease and cancer. Of all the methods tested, Neural Networks work best, especially with continuous data, but there are still problems with sensor quality that affect real-time predictions [15].

Methodology

Data Collection and Preprocessing:

The dataset is taken from kaggle It consisted of extraction of air pollution data in the experimental dataset. Concentrations of various kinds of pollutants such as CO, NO, NO2, O3, SO2, PM2.5, PM10, and NH3 along with date-wise temporal information constituted this dataset. The column for date was converted into a datetime format, so year, month, and day were obtained as other features. This original column of date was removed to ensure proper pre-processing. In addition, the categorical features like the city were encoded using one-hot encoding into a binary format to accommodate machine learning. All pre-processing ensured that the data set was clean and properly structured for input to the XGBoost classifier.

Machine learning Model:

Machine learning is one of the components of artificial intelligence, which allows systems to learn from data in order to make predictions or decisions without direct programming. There are basically three types: supervised learning, unsupervised learning, and reinforcement learning. XGBoost is a super powerful machine learning algorithm that can be used on both classification and regression tasks. It applies several weak learners, mostly decision trees, in producing a strong predictive model. XGBoost uses gradient boosting to iteratively improve predictions and regularization in order to prevent overfitting and the XGBoost model is used as base model in the stacked classifier. The meta model is a Support Vector Classifier or SVC which is a machine learning algorithm that is used specifically for classification tasks. It is basically an algorithm that looks for the best decision boundary that helps differentiate between classes that should be separated by maximizing their margin.

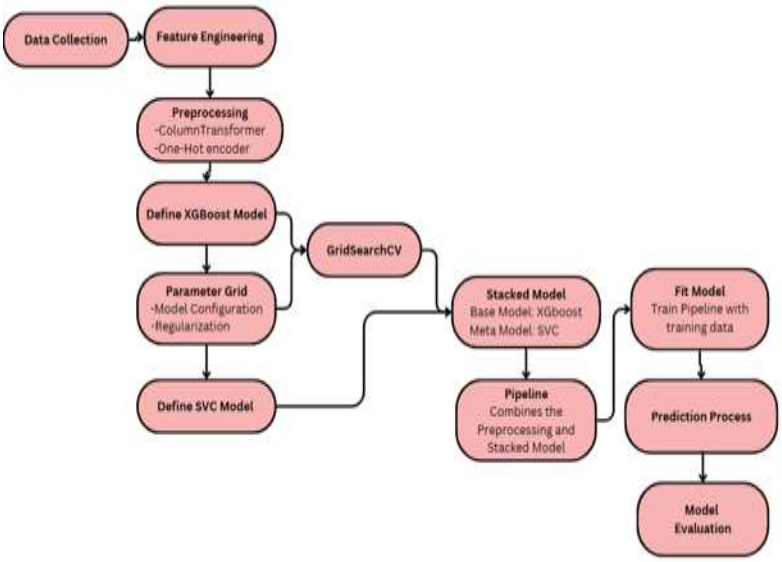


Figure 1: Flowchart of AQI Prediction with Stacked Model

XGBoost Model:

The chosen model for training was the XGBoost classifier, a powerful gradient-boosting model that was well known to handle structured data. For the optimization of model performance, the process of hyperparameter tuning included GridSearchCV. A set of hyperparameters-the number of estimators, the learning rate, and maximum depth-were tried in combinations to determine which one best predicted the air quality. Cross-validation was used, in which GridSearchCV had already trained each configuration on a training data subset to avoid overfitting and improve the model's generalization across data partitions.The XgBoost is performed as base model in stacked model.

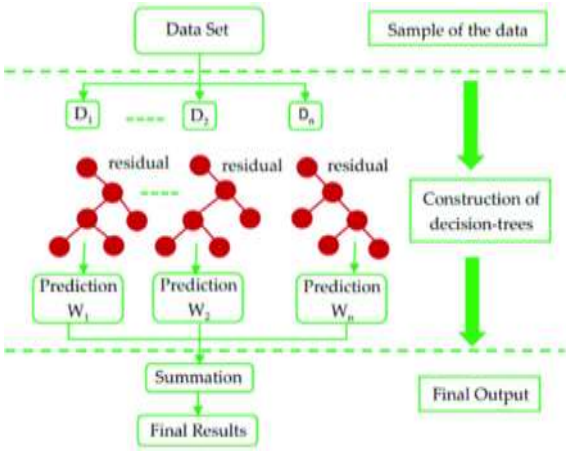


Figure 2: workflow of xgboost

Support Vector Classifier (SVC):

This code uses the SVC as the last estimator in a model of stacking. Kernelized support vector machines with an RBF (Radial Basis Function) kernel help discover nonlinear relations in data. The function model SVC works by seeking an optimal decision boundary that maximizes the margin between classes. The SVM acts as a final estimator that basically tends to utilise collective information from the earlier-based models in order to produce a final decision. Thus, in this layered approach, the model can learn with much more perspectives with improvement towards the better accuracy of the predictive model.

Stacked Model:

It uses StackingClassifier to build a stacked model in which it combines predictions from multiple models in this case (XGBoost and SVC). The base model is XGBoost that learns features within training data, and the prediction of the base model in turn acts as an input to the SVC final estimator. The layered approach likely enhances the accuracy of the predictions by making the best of the strength built into every model wherein SVC can be regarded as a meta-model to generate final predictions depending on the outputs from XGBoost.

This code evaluates the performance of a hybrid stacking model that combines XGBoost and SVM for the AQI prediction task. Preprocessing is performed by doing one-hot encoding on the 'city' column, but it retains the levels of the pollutants as continuous features. GridSearchCV was performed to optimize the parameters of XGBoost such as max_depth, number of estimators to further improve the performance of the model. SVM, with its fixed parameters, was the final estimator in the stacking setup. With the data split into training and test data, the pipeline is fitted on the training data. Finally, predictions on the test data are compared with the actual values to obtain accuracy computed using `accuracy_score`, showing the model's effectiveness.

Key Metrics for Classification Models:

$$Accuracy = \frac{Number\ Of\ correct\ predictions}{Total\ Number\ of\ predictions}$$

$$Precision = \frac{True\ positives}{True\ positives + False\ positives}$$

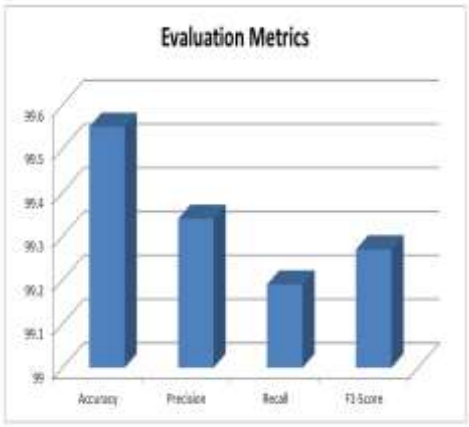
$$Recall = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$F1\ score = \frac{(Precision * Recall)}{(Precision + Recall)}$$

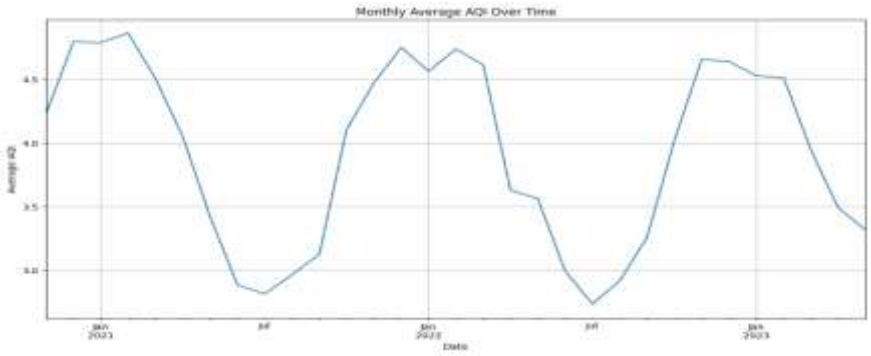
Accuracy measures the overall proportion of correct predictions. Precision and Recall assess how well the model identifies true positives, with precision focusing on false positives and recall on false negatives.F1-score balances precision and recall, especially useful for imbalanced datasets.

Results and Discussion

The hybrid stacking model with XGBoost and SVC demonstrated outstanding accuracy at 99.55% in the AQI prediction. With a very low false positive rate, precision was at 99.34%. High AQI instance identification capability was also observed with a recall of 99.19%. Its F1 score, representing the balance between precision and recall, was thus at 99.27%. These results show that the proposed model is quite good in the prediction of the levels of the AQI, generalized towards unseen data. The misclassification was minimal on the confusion matrix, thereby supporting the robustness and effectiveness of the model for AQI prediction.

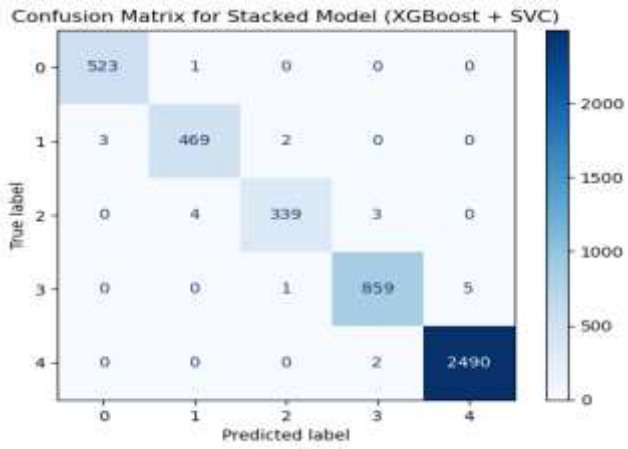


Monthly Average AQI Time Series Plot :



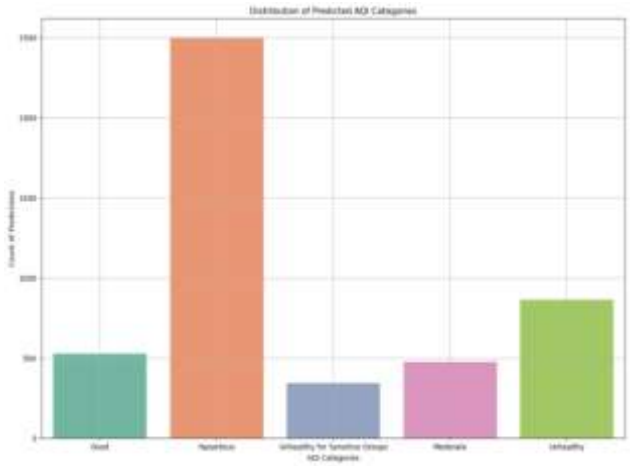
The above time series plot is done whole dataset and The date column was made the index for time-series analysis, and average monthly AQI was computed in order to monitor long-term trend in air quality. Plotting the time series revealed seasonal variations and perhaps shifts in the AQI values over time, thus offering an overview of changes in air quality that might affect model predictions.

Confusion Matrix:



The confusion matrix clearly shows how well the model performed; it measures, at the actual level, the number of correct as well as incorrect predictions that took place for each class. Diagonal elements refer to correct predictions whereas off-diagonal elements indicate misclassifications.

Distribution of Predicted AQI Categories:



The bar chart below gives the distribution of the predicted categories of the Air Quality Index. "Hazardous" holds the highest count of predictions, reasonably well above any other category. "Good" and "Unhealthy" categories hold a moderate number, while "Moderate" and "Unhealthy for Sensitive Groups" hold the lowest counts, that is, poor air quality predictions are on the higher side.

Conclusion

This work demonstrates the effectiveness of a hybrid stacking model combining XGBoost and SVC for AQI prediction with 99.55% of accuracy. In future work , the regression models can be used instead of classifiers to predict AQI as a continuous percentage, allowing for more detailed AQI level predictions. Techniques like XGBoost Regression or Support Vector Regression would be suitable for this approach.Such improvements will further enhance AQI forecasting, providing insights useful in public health planning and policy measures.

References

[1] Naveen, S., Upamanyu, M. S., Chakki, K., Chandan, M., & Hariprasad, P. (2023, July). Air Quality Prediction Based on Decision Tree Using Machine Learning. In 2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES) (pp. 1-6). IEEE.

[2] Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. Scientific Reports, 14(1), 6795.

[3] Singh, J. K., & Goel, A. K. (2021, March). Prediction of air pollution by using machine learning algorithm. In 2021 7th International conference on advanced computing and communication Systems (ICACCS) (Vol. 1, pp. 1345-1349). IEEE

[4] Bhattacharya, S., & Shahnawaz, S. (2021). Using machine learning to predict air quality index in new delhi. arXiv preprint arXiv:2112.05753.

[5] Liu, C., Pan, G., Song, D., & Wei, H. (2023). Air quality index forecasting via genetic algorithm-based improved extreme learning machine. IEEE Access.

[6] Maltare, N. N., & Vahora, S. (2023). Air Quality Index prediction using machine learning for Ahmedabad city. Digital Chemical Engineering, 7, 100093.

[7] Gladkova, E., & Saychenko, L. (2022). Applying machine learning techniques in air quality prediction. Transportation Research Procedia, 63, 1999-2006.

[8] Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. Complexity, 2020(1), 8049504.

[9] Wang, Y., & Kong, T. (2019). Air quality predictive modeling based on an improved decision tree in a weather-smart grid. IEEE Access, 7, 172892-172901.

[10] Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. Artificial Intelligence Review, 56(9), 10031-10066.

[11] Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of air quality index using machine learning techniques: a comparative analysis. Journal of Environmental and Public Health, 2023(1), 4916267.

[12] Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology, 20(5), 5333-5348.

[13]Aram, S. A., Nketiah, E. A., Saalidong, B. M., Wang, H., Afitiri, A. R., Akoto, A. B., & Lartey, P. O. (2024). Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. International Journal of Environmental Science and Technology, 21(2), 1345-1360.

[14] Bhalgat, P., Pitale, S., & Bhoite, S. (2019). Air quality prediction using machine learning algorithms. International Journal of Computer Applications Technology and Research, 8(9), 367-370

[15] Madan, T., Sagar, S., & Virmani, D. (2020, December). Air quality prediction using machine learning algorithms–a review. In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 140-145). IEEE.