Air Quality Index Prediction using Machine Learning Techniques

Priyanka Maan[1], Bhoomi Gupta[2] and Deepika Bansal[3]

[1]Department of Computer Science & Engineering, Faculty of Engineering & Technology, SRM University, Delhi-NCR, Haryana, 131029, India.
[2,3]Department of Information Technology Maharaja Agrasen Institute of Technology
New Delhi 110086, India
Corresponding Author. Tel: 8512811975, E-mail: priyankamaan@srmuniversity.ac.in

Abstract. *Without air, humanity could not possibly survive. Air quality is negatively affected by continuous changes in almost every area of modern human civilization, posing a threat to sustainable development. By integrating sustainable practices into air quality monitoring and forecasting, we can aim to minimize the generation of risky contaminants and reduce the total environmental footprint of transportation, business, and home approaches. In comparison to standard techniques, machine gaining knowledge based totally prediction technology studying techniques have been shown to the best equipment for studying such contemporary risks. To analyze and expect air satisfaction, the prevailing take a look at appears at six years' really worth of air pollution facts from 23 Indian towns. The dataset has undergone thorough preprocessing, and correlation analysis was used to identify essential properties.*

An exploratory data analysis is conducted to determine the pollutants that might have a direct influence on the air quality index as well as to find patterns that are hidden in the dataset. Almost all contaminants have significantly decreased in the epidemic year of 2020. Four models of machine learning are employed to evaluate air quality while resampling is utilized to address the problem of data imbalance. The outputs of models are contrasted with established standard. The Support Vector Machine model is the least accurate. The most accurate model is the Gaussian Naive Bayes one. The known performance metrics is used to evaluate and compare these model's performances. The winning one was the XGBoost one, which also exhibited the highest degree of linearity between predicted and the data that was observed.

Keywords: Air Quality Index, Support Vector Machine, Random Forest, Machine Learning, Deep Learning.

1 INTRODUCTION

Air is the only thing that keeps humans alive. For our well-being, its quality needs to be monitored and comprehended. Millions of individuals worldwide have physiological problems and respiratory mortality as a result of air pollution. Scientific research indicates that the single biggest threat to the environment is the pollution of air. As a result of fast industrialization and the harmful emissions of gas it produces, population levels have significantly expanded. The quality of the air is seriously harming our health. Environmental and Public Health in Hindawi being contaminated by hazardous substances. The air quality has significantly decreased as a result of this unmanaged pollution.

A numerical index called the AQI is used to quantify and communicate air pollution levels. The AQI value that is high denotes extremely polluted air, which has a gravely detrimental effect on health. The AQI can be used to monitor the quality of air in real time. In our area, many weather stations have also recorded AQI data daily and hourly. To use this data in the proposed study, they will be mined and harvested.

2 LITERATURE SURVEY

The authors forecasted air quality using machine learning using Google Street View data at numerous places in California. He concentrated on the regions that had the missing data. For each neighborhood in a city, a web application to predict air quality was developed by the author [1].

The author attempted to use the Support Vector Regression (SVR) machine learning technique to predict the amounts of air pollutants and particles in California. The authors created a novel method for simulating hourly air pollution [2]. Few authors evaluated how effectively six ML classifiers predicted AQI of Taiwan using data spanning the previous 11 years [3].

In [4], the authors examined twenty distinct literary works which was based on the contaminants they researched, the machine learning methods were employed, and calculated their performances. Several studies employed meteorological data, including temperature, humidity, and wind speed, as found by the author to accurately predict pollution levels. They found that compared to other well-known ML techniques, boosting models and neural network (NN) strategies performed better. The authors found that the air pollutants concentration was considerably influenced by temperature, humidity, direction of wind, and speed of wind [5]. The RF approach exhibited the fewest classification errors, according to the supervised machine learning methods the authors used to estimate AQI.

Bhalgat predicted the air's SO2 content in Maharashtra, India, by applying machine learning. The authors concluded that due to their excessive pollution, this Indian region has some cities that need immediate medical attention [6].

The authors presented a very fascinating study in 2018 on their survey of the international research on pollution of air and respiratory health that has been peer review [7]. From the Scopus database, the authors took 3635 records from publications between 1990 and 2017. They noticed that the number of articles significantly increased between 2007 and 2017. After observing active nations, organisations, journals, writers, and international partnerships in the field, the authors stated that there is a great deal of interest in the relationship between respiratory health and air pollution research. They recommended gathering public opinion on investments in green technologies and the reduction of outdoor air pollution. The issue of AQI prediction was referred to be multi-task learning challenge by the authors [8] [9].

In [10], they sought to develop a model that linked air pollution and traffic density. The author claims that gathering this kind of traffic data may be done on a budget and that adding climatic information would increase the data's accuracy. The hybrid model outperformed all others, and according to the authors, morning time data has the highest accuracy [11].

3 METHODOLOGY

3.1 Data Preprocessing

The first and most crucial condition for the development of effective ML models is data quality. Preprocessing methods contribute to the reduction of data noise, which eventually accelerates processing and expands the use of ML algorithms. Data extraction and monitoring applications have two of the most common errors which are outliers and missing data. The data preparation step involves modifying or deleting outlier data, filling out data that is not a number (NAN), and performing other operations on data.

Many statistics may be missing for a variety of reasons, such as a station that has the capability to view data but not the means to record it. The proposed flowchart of the model is shown in Fig. 1.

At the first stage, Air Pollution dataset is taken and fed into the preprocessing stage followed by normalization of data through feature selection. Then the data is explored and divided into training and testing sets. Then we apply the Machine learning algorithm on the training dataset to train the model. After the training phase, the test

data set is fed into the model for validation. As a final result, the AQI of the test data set is predicted by the designed model on which further evaluation is done and comparative analysis has been carried out with the state-of-the-art methods.

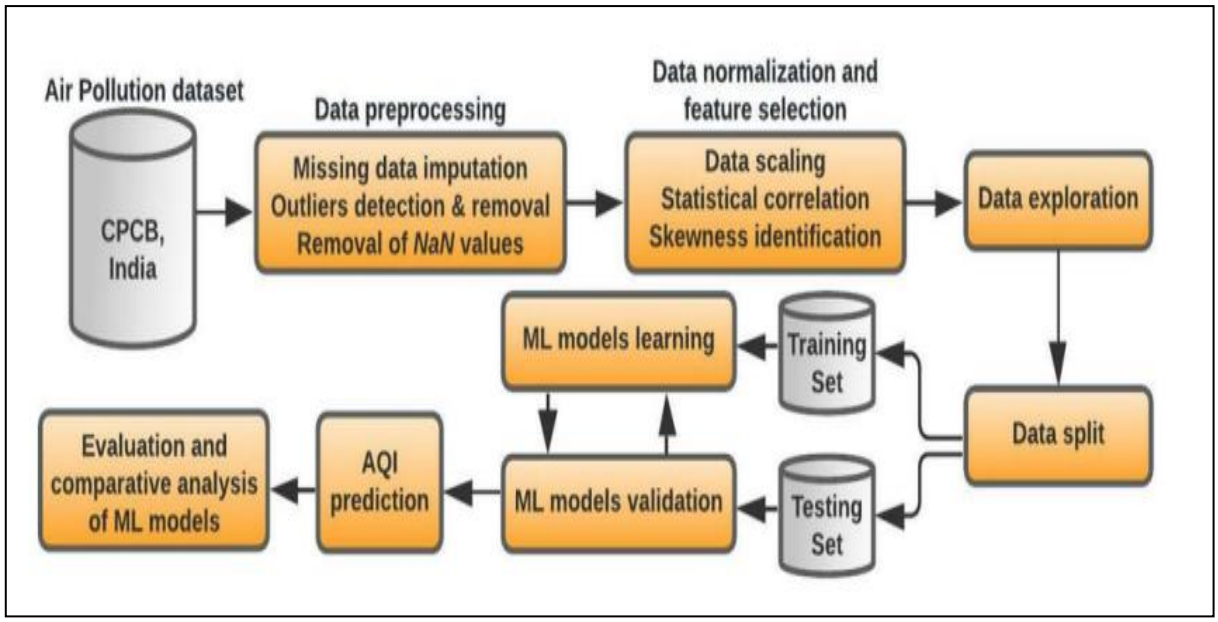

Fig. 1 Proposed model flowchart

3.2 Feature Selection

Government organizations utilize the specific metric, AQI, in the CPCB dataset under investigation to enlighten the public about the quality of air and to train forecasters. The National Ambient Air Quality Standards classify AQI into six categories: acceptable (0-50), tolerable (51-100), moderate (101-200), poor (201-300), extremely poor (301-400), and severe (401-500). Lowering the number of input variables, according to experts in the area leads to lowering of the computing cost of modeling and boosts prediction accuracy. In the current work, the quantity of contaminants (input variable) required to build a predictive model has been determined using a correlation-based feature selection technique.

Each set of both the target and input variables exhibits a correlation when features are chosen using algorithms based on statistical correlations. The variables are further investigated if they strongly correlate with the target variables. Since many machine learning algorithms are prone to outliers, it is imperative to identify any characteristic in the input dataset that deviates from the overall trend of the data. The outliers in the current dataset were discovered using a correlation-based statistical outliers identification method. A correlation investigation between the AQI characteristic and features of other contaminants was used to identify significant aspects.

3.3 Exploratory Data Analysis

In order to know about the numerous patterns that are hidden within the dataset, this section of study focuses on data analysis and exploration. Exploratory data analysis is the initial step in the field of data analytics, which is implemented prior the utilization of any ML models. The concerned significant issues are being examined as a result: (a) assessing the current state of air pollution and its trends over the previous six years, or from 2015 to 2020; (b) analyzing the top six most polluted municipalities and their average AQI parameters, as well as the distribution of pollutants in the air; and (c) determining the top four contaminants that are particularly contributing to the rise in AQI numbers.

Pollutants involved directly in increasing AQI values
The AQI and numerous pollutants have been correlated, and pollutants that have correlation coefficient values more than the threshold value of 0.5, indicating a highly positive link, have been discovered.

4 RESULTS

The plan of the experiment and the empirical evaluation used to anticipate AQI values based on airborne pollutants are covered in this section. Before ML models are assessed, the dataset is split into two sections: screening (25%) and training (75%). On the cloud platform known as Google Collab Pro, Python programs have been executed.

The dataset is then investigated to determine the value of the AQI concerning those contaminants that significantly contribute to increasing the AQI value. Table 1 shows comparison of model results.

Table 1 Comparison of model results

Model	Accuracy	Precision	Recall	F1-score	Prediction time (in seconds)
KNN	85	92	85	94	0.018
GNB	83	88	89	92	0.016
SVM	78	91	90	83	0.027
RF	86	92	91	90	0.023
XGBoost	90	96	95	91	0.041

An AQI timeline graph is displayed over a few particular contaminants that are directly responsible for elevated AQI readings. Winter pollution is more severe due to seasonal fluctuations in PM2.5 and PM10 pollution levels than summer pollution levels. While the level of O3 remained constant from 2018 to 2020, the level of SO2 started to increase after 2018. Levels of BTX2 exhibit a similar tendency as well. Almost all pollutants, except CO, have seasonal fluctuations. Observe how India's pollution levels decline from June to August. It can be a result of the Indian subcontinent experiencing its first monsoon at this time.

Between March and April, BTX levels significantly decreased; from May to September, they somewhat increased; and from October to December, they significantly increased. Given that 2020's median findings are lower as compared to those for previous years, it's possible that the year's pollution level was significantly lower. Human and industrial activities were forbidden.

5 CONCLUSION

It can be challenging to anticipate air quality because of the changing environment, unpredictability, and range of contaminants present at different times and places. The AQI forecast in India, however, received little consideration from scientists. The train and test subsets of the dataset are divided by a 75-25% ratio, respectively. It contrasts the usage of the SMOTE resampling technique with ML-based AQI prediction. The outcomes of ML models are shown with respect to common measures such as precision, accuracy, recall, and F1-Score for both the train and test groups. In contrast to SVM model, which had the least accuracy, the model using XGBoost had the highest accuracy for both train and test sets.

The XGBoost model performs best overall because it produces the best outcomes during both the training and testing periods. The RF model exercised using SMOTE performed rather well for the training phase. However, throughout the testing phase, nearly all ML models demonstrated improvements.

6 REFERENCES

[1] Alade, I. O., Abd Rahman, M. A., & Saleh, T. A. (2019). Predicting the specific heat capacity of alumina/ethylene glycol nanofluids using support vector regression model optimized with Bayesian algorithm. *Solar Energy*, 183, 74-82.

[2] Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC public health*, 17, 1-19.

[3] 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.

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

[5] Gopalakrishnan, V. (2021). Hyperlocal air quality prediction using machine learning. *Towards data science.*

[6] 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.

[7] Patil, R. M., Dinde, H. T., Powar, S. K., & Ganeshkhind, P. M. (2020). A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms. *Int J Innov Sci Res Technol*, 5(8), 1148-1152.

[8] Johansson, C., Zhang, Z., Engardt, M., Stafoggia, M., & Ma, X. (2023). Improving 3-day deterministic air pollution forecasts using machine learning algorithms. *Atmospheric Chemistry and Physics Discussions*, 2023, 1-52.

[9] Sonawani, S., & Patil, K. (2024). Air quality measurement, prediction and warning using transfer learning based IOT system for ambient assisted living. *International Journal of Pervasive Computing and Communications*, 20(1), 38-55.

[10] Van, N. H., Van Thanh, P., Tran, D. N., & Tran, D. T. (2023). A new model of air quality prediction using lightweight machine learning. *International Journal of Environmental Science and Technology*, 20(3), 2983-2994.

[11] 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.