AQI Prediction Using Machine Learning

1st Raisa Tarannum
dept. name of organization (of Aff.)
Department of Computer Science, Brac University
Dhaka, Bangladesh

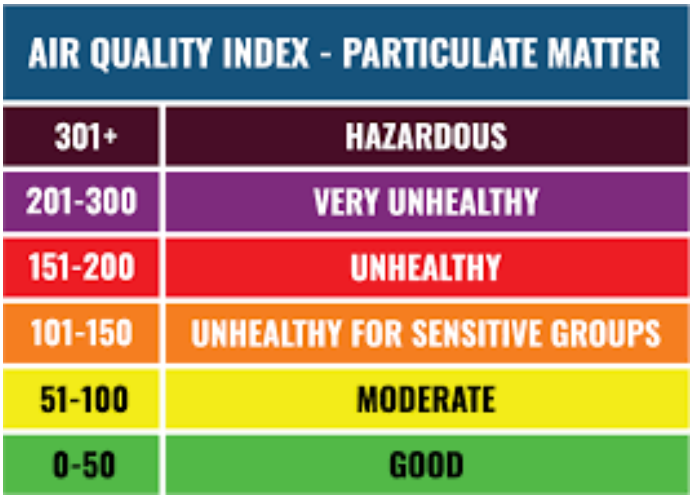
raisa.tarannum@g.bracu.ac.bd

***Abstract*—With the ongoing rise in population and number of vehicles, many environmental issues all over the world have grown significantly. One such concern is that of increased air pollution. With Industrial processes continuously releasing gases that are detrimental to the environment and lives of all living beings, proper prediction of air quality is imperative to minimize the harmful effects on health. This literature review centres around the different methods and techniques that is used for forecasting concentrations of pollutants and to predict and model the Air Quality Index, as a means to reduce pollution in an area.**

***Index Terms*—AQI, air pollution, forecasting**

I. Introduction

As the quality of air deteriorates, air pollution continues to become even more of a health concern. This is where the air quality index is used to measure how polluted the air is. Normally, the AQI is measured by six levels of air pollution where the higher the level or AQI, the greater the concentration of pollution in the air. The levels of AQI ranges from 0 to 50 for perfect, 51 to 100 for moderate, and so on, till its hazardous when over 301+. This is why it is extremely important for people to understand the importance of the air quality around them to ensure a healthy and cleaner environment.



The AQI can be calculated using the equations separately based on the number of parameters used. Generally, to calculate AQI, at least three parameters need to be considered, of which, at least one needs to either be PM2.5 or PM10. Then, depending on what basis the AQI is being calculated, for example using CO, PM2.5 and SO2, the sub-index for each of these parameters need to be calculated separately. From there, the least favourable sub-index represents the AQI.

It is said that 17 percent of cities in high-income nations have air quality that is lower than the global air quality guidelines set up by WHO. Not only that, it is also seen that less than about 1 percent of the cities of low-income countries fall within the WHO guidelines. That is why it is of utmost importance that studies are done to find the most effective ways to reduce the pollution in air.

II. Related Works

This literature reviews looks at a few different research papers conducted on this topic.

Mayuresh Mohan Londhe (2021) selected AQI data of 13 different Indian cities that are high in air pollution. These cities include Ahmedabad, Amritsar, Bengaluru, Chandigarh, Chennai, Delhi, Gurugram, Hyderabad, Jaipur, Kolkata, Lucknow, Mumbai, Vishakhapatnam. The author also collected data from the stations across Bihar, Delhi, and Maharastra and Pradesh state. Their prime objective was to predict AQI using these algorithms while also exploring dimension reduction. In the preprocessing stage, they used the fill funtion to make up for the null data, with the assumption that there were not any drastic changes as these were data from back to back days. They used supervised learning algorithms to build and train three regression and two classification models. Since all the variables in the data were continuous type, they applied Principal Component Regression and Partial Least Squares Regression for reducing dimension. For the Stations dataset, the dependent variable was a categorical data which led them to using classification techniques, K Nearest Neighbor and Multinomial Logistic Regression for AQI prediction. The author found that Partial Least Squares Regression model gave the most accurate results incase of lowest RMSE. Their research was well thought-out and may help to improve AQI prediction.

K.Kumar, B.P.Pande (2022) attempted AQI prediction using machine learning techniques since it is a topic that is often overlooked in India. They selected data from 23 individual Indian cities starting from January 2015 to July 2020. The authors analysed PM2.5, PM10, CO, ozone level, along with some other major contributors to air pollution. For data

handling, they used median values to fill the missing data standardized the data using a normalisation process. The dataset was split into a 75 percent training and 25 percent testing subset. They presented evaluation of the ml models against one another, both with and without SMOTE resampling technique. The models applied consist of KNN, SVM, GNB, RF and XGBoost. The results showed that the XGBoost model obtained highest accuracy with SVM producing the least accurate results. All in all, it was seen that all the models showed improvement in every assessment when SMOTE resampling technique was used. This study took into account data over a span of 6 years which helps better understand the AQI.

Elia Georgiana Dragomir (2010) used the K Nearest Neighbor Technique in this study for air pollution forecasting. The author focuses on this algorithm taking it as a classification problem. They used data from June 2019, containing data from over a span of 29 days, for the experiment and applied the KNN method to predict the value of air quality index. Sulphur dioxide, nitrogen monoxide, nitrogen dioxide, carbon monoxide along with the ozone level were the features selected for this experiment. The whole thing was carried out using a data mining software known as Weka. Experiment results showed a rate of 65.51 percent for correct prediction of the AQI and that there is a strong correlation between the parameters used. The training dataset used for this study took into account only 29 days of data which is not the best way to evaluate the AQI. Using long periods of data may help produce better results.

Suraya Ghazali, Lokman Hakim Ismail, Johor made use of Artificial Neural Networks to create an AQI predicting model. The parameters selected for this study include carbon monoxide, sulphur dioxide, nitrogen dioxide as well as nitric oxide. The temperature, humidity and air velocity were also taken into account. They evaluated the model's performance using Mean Square Error and R square. The results concluded that a model with a network structure of 7-20-4 performed best.

R. Gunasekaran, K. Kumaraswamy, P.P. Chandrasekaran, R. Elanchezhian (2012) conducted a research on the air quality of Salem City in Tamil Nadu. The primary goal of the study was to identify the status and trends of some of the major contributors to air pollution such as nitrogen oxides, sulphur dioxide, PM10 and PM100. The data used here is from the month of April 2010 to March 2011 and it showed a rise in PM10 and concluded that the issue may only get worse if not brought under control.

Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu (2018) focused on PM2.5 in their study where they made use of machine learning algorithms. They used logistic regression and autoregression model to predict the value of PM2.5 in a certain city, based on previous data.

Huixiang Liu,Qing Li,Dongbing Yu, Yu Gu (2019) used Support Vector Regression and Random Forest Regression to build and train models for AQI prediction in Beijing prediction. It was seen that the SVR model did a better job accurately predicting AQI.

III. Methodology

A. Data collection and pre-processing

The data used for the project is found through Kaggle. The data set contains air quality data from Indian cities and has attributes needed for AQI calculation such as PM2.5, PM10, SO2, NO2, CO, O3, etc.

Some of the data was missing so I dropped the rows will null values in the pre-processing stage. I also dropped some of the columns, such as, 'Date','Benzene','Toluene','Xylene' since they were not needed for model training. Initially the dataset had about 25000 datapoints, and the final data set had around 11500 datapoints.

The categorical data were all converted to numerical data for the models' convenience. For model training, the data was split into a test size of 0.3. Here, 30 percent of the data was for testing and the rest 70 percent were for training. I also set random set=2 for better seeding.

The data was scaled using MinMax and Standard Scalar to make it easy for the models to understand it.

B. Techniques

Since the models are meant to predict AQI index and classify it to level accordingly, this is considered a classification problem. The features in the dataset are both categorical as well as numerical. I used Decision Tree, KNN and SVM for AQI prediction and compare the accuracy.

Decision Tree Classifier can be applied on both classification, as well as regression problems. It is a supervised learning algorithm with a tree-like structure where each of the internal nodes denotes the features of a dataset and the branches describe the outcome of the test, while the leaf nodes contain a class label.

Support Vector Machines are another type of supervised learning model that is often used for both classification and regression problems, as well as for detecting outliers. SVM is a super fast classifier, however it has a higher training time when working with massive datasets.

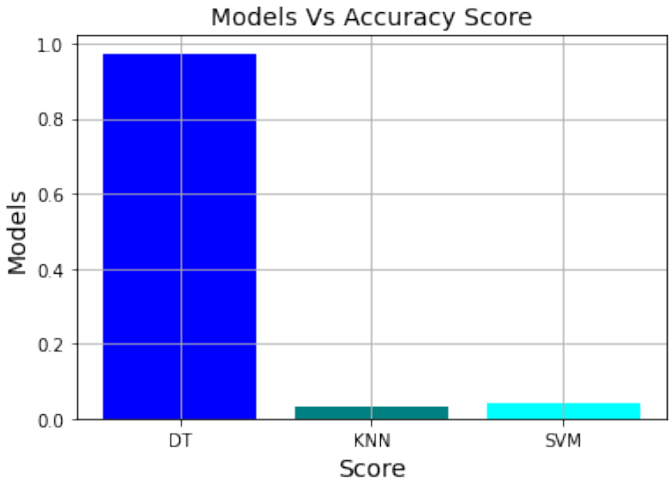
Lastly, the final model used in this study is the K-Nearest Neighbor algorithm which is a very simple supervised learning technique. Much like the other algorithms, this too can be used for both regression as well as classification, however it is mostly used for the latter. The way it works is by deciding on a value k for which is nearest to the data point that is being classified. To put it simply, it classifies a data point based on similarity, meaning it assigns the new data to one class or another.

I also used K-Fold Cross Validation to improve the model prediction.

IV. Results

As mentioned above, SVM took a longer time to train the model since the data was large. The results showed that the Decision Tree model had a high accuracy of 98 percent whereas KNN and SVM had very low accuracy. That means that Decision Tree model made a lot of correct predictions

whereas KNN and SVM models barely made any correct predictions.



V. Discussion

The noticeable difference in accuracy between the models lead me to think there was some sort of error in training and testing the KNN and SVM models.

VI. Conclusion

THe methods used above are some of the most common ones used by researchers to predict AQI. Although the decision tree gave me highly accurate results, the same cannot be said for the other two models. I hope to work more on the models to find the issue.I would also like to work on more models like LASSO, logistic regression and random classifier for air pollution forecasting.

VII. References

References

[1] Mayuresh Mohan Londhe,"Data Mining and Machine Learning Approach for Air Quality Index Prediction",International Journal of Engineering and Applied Physics (IJEAP) (2021) 2737-8071..

[2] K. Kumar, B. P. Pande,"Air pollution prediction with machine learning: a case study of Indian cities",International Journal of Environmental Science and Technology (2022).

[3] Elia Georgiana Dragomir,"Air Quality Index Prediction using K-Nearest Neighbor Technique",Universităţii Petrol – Gaze din Ploieşti (2010) 103 - 108.

[4] Suraya Ghazali, Lokman Hakim Ismail, Johor,"Air Quality Prediction Using Artificial Neural Network", UTHM Institutional Repository.

[5] R. Gunasekaran, K. Kumaraswamy, P.P. Chandrasekaran, R. Elanchezhian, "MONITORING OF AMBIENT AIR QUALITY IN SALEM CITY, TAMIL NADU", International Journal of Current Research. (2012) pp.275-280.

[6] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, "Detection and Prediction of Air Pollution using Machine Learning Models", International Journal of Engineering Trends and Technology (IJETT). (May2018).

[7] Huixiang Liu,Qing Li,Dongbing Yu, Yu Gu,"Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms", Appl. Sci.(2019).

[8] https://www.kaggle.com/code/melvin97n/air-pollution-india-forecast/notebook