International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Comparative Analysis of Air Quality Index Prediction using Machine Learning

S. Lakshmi

Assistant Professor, SRM Institute of Science and Technology
Email: *lakshmi1503[at]gmail.com*

Abstract: ***Quality of air played a major role in the life of human being. Now a days, the amount of air pollution is increased due to so many reasons such as industries, factories and vehicles. Various health issues are created when the quality of air is polluted or degraded. Designing a system to measure the quality of the air is a challenging task, but measuring the air quality is really a boon for our society. It can be used in many ways such as educating the public about the negative effects of low-quality air especially in children. The purpose of this paper is to conduct the comparative analysis of the air quality index by applying various machine learning algorithms.***

Keywords: Air quality index, Machine learning algorithms, pollution and Evaluation

1. Introduction

Air pollution is the contamination of air due to the presents of pollutants which affects our health severely. The air pollution shortens the human life nearly a year according to the study of the environmental engineers. In India and Bangladesh, the air pollution range exceeds the guideline more than ten times. There are many types of pollution such as water pollution, air pollution and soil pollution and so on. Everybody inhales the oxygen through ais, there is a possibility of getting affected immediately by air.

Generally, industries, factories and vehicle cause the outdoor air pollution and emission of gases, smokes and chemicals are the reason for indoor air pollution. There are two different types of pollutants. They are primary pollutants and secondary pollutants.

1.1Primary Pollutants

Primary pollutants are emissions of coal fired power plants, natural gas power plants, biomass burning, forest fires, volcanoes and so on. These primary pollutants are harmful to human beings, animals and plants [1]. Now a days, the emission of primary pollutants has decreased considerably due to the regulations, economic changes and technology [2]. The following is the primary pollutants [3]:

Sulphur oxide (SOX): Sulphur dioxide (SO2) released by burning coal and petroleum. It is released by various factories and plants. When react with Catalyst (NO_2), results in H_2SO_4 causing acid rains that forms the major cause of Air Pollution.

Nitrogen oxide (NOX): Most commonly Nitrogen dioxide (NO2) that is caused by thunderstorm, rise in temperature.

Carbon monoxide (CO): Carbon monoxide is caused by burning of coal and wood. It is released by Vehicles. It is odourless, colorless, toxic gas. It forms a smog in air and thus a primary pollutant in air pollution. Toxic metals – Example are Lead and Mercury

Chlorofluorocarbons (CFC): Chlorofluorocarbons released by air conditioners, refrigerators which react with other gases and damage the Ozone Layer. Therefore, Ultraviolet Rays reach the earth surface and thus cause harms to human beings. Garbage, Sewage and industrial Process also causes Air Pollution. Particles originating from dust storms, forest, volcanoes in the form of solid or liquid causing air pollution.

1.2 Secondary Pollutants

The source of the pollutant reacts with the molecules in the atmosphere. The formation of the secondary pollutants can be formed from many different compounds. Some of the secondary pollutants are smong, ozone, acid deposition and peroxyacyl nitrate.Ground Level Ozone: It is just above the earth surface and forms when Hydrocarbon react with Nitrogen Oxide in the sunlight presence. Acid Rain: When Sulphur dioxide react with nitrogen dioxide, oxygen and water in air thus causing acid rain and fall on ground in dry or wet form.

2. Literature Review

The difference between Primary Pollutants and Secondary Pollutants is Primary Pollutants are those which are released into air directly from Source whereas Secondary Pollutants are those which are formed by reacting with either primary pollutants or with other atmospheric component. [7]. The machine learning algorithms such as Logistic regression and autoregression help in determining the level of PM2.5 [5]. The day wise prediction of pollutant level [4] was removed by various authors further by predicting hourly wise data using different algorithm. Benzene concentration can also account into air pollution and its concentration can be determined with CO [6].

Air pollution causes harm to not only human beings but also animals and plants. It starts with less threating diseases to most dangerous diseases. Doctors advise to wear mask to protect our health. Awareness programs are initiated to understand the real problems due to air pollution. It is the need of an hour to predict the air quality accurately.Various approaches are used for predicting the air quality indextraditionally with lot of mathematical calculations, but the accuracy of the air quality index is still a challenging task.

Machine learning algorithms played a major role in prediction. Air pollution should be avoided or at least considerably reduction is the most required one. It is necessary to measure the air quality index for controlling the air pollution. In this work various machine learning algorithms are implemented for measuring the air quality index and the results are compared to produce the better output.

Kostandina Veljanovska1 et al., [8] compared unsupervised neural network algorithms in which output is notknown with supervised algorithm K-NearestNeighbour, Support vector machine, DecisionTrees. Savita Vivek Mohurle et. al., [9] predicted pm2 and pm10level using Fuzzy logic. MaryamAljanabi et al., [10] depicted the ozone layerdepending upon the Temperature, humidity,wind speed and wind direction.

3. Dataset

Air quality dataset is available in Kaggle. It contains the daily and hourly data from various stations in India. It contains the detail from 2015 to 2020.The dataset consists of the details from 2015 to 2020. AQI bucket is created with poor, good and satisfactory. First of all the dataset is cleaned by replacing all the missing details and incorrect information in the dataset. Then the dataset is divided into four such as New Delhi, Hyderabad, Bangalore and Kolkata. The sample dataset is given in Table-1.

Table 1: Sample dataset for Ahmedabad

City	PM2.5	NO	NO2	NOx	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
Ahmedabad	94.52	24.39	32.66	52.61	24.4	67.4	111.33	0.24	0.01	7.67	514	Severe
Ahmedabad	136	43.48	42.08	84.57	43.5	75.2	102.7	0.4	0.04	25.9	782	Severe
Ahmedabad	178.3	54.56	35.31	72.8	54.6	55	107.38	0.46	0.06	35.6	914	Severe
Ahmedabad	139.7	30.61	28.4	56.73	30.6	33.8	73.6	0.17	0.03	11.9	660	Severe
Ahmedabad	80.65	2.37	22.83	24	2.37	25.7	47.3	0	0	0	294	Poor
Ahmedabad	58.36	2.6	21.39	23.31	2.6	32.7	53.54	0	0	0	149	Moderate
Ahmedabad	79.29	1.16	26.94	26.83	1.16	67.4	59.3	0	0	0	190	Moderate
Ahmedabad	88.7	7.29	31.32	37.73	7.29	80.1	44.76	0	0	0	247	Poor
Ahmedabad	74.28	8.92	27.3	33.42	8.92	54.3	47.42	0	0	0.27	379	Very Poor
Ahmedabad	113.9	4.32	24.27	26.86	4.32	48.7	39.94	0.02	0	3.55	341	Very Poor
Ahmedabad	105.4	1.41	18.21	18.75	1.41	35.9	56.15	0	0	3.31	256	Poor
Ahmedabad	66.52	6.34	23.8	28.24	6.34	66.6	53.14	9.7	9.63	16.5	388	Very Poor
Ahmedabad	65.04	14.19	30.1	44.21	14.2	65.9	31.88	7.72	17.46	2.51	288	Poor
Ahmedabad	103.4	18.18	39.56	57.33	18.2	80.4	40.11	11.29	24.35	3.35	510	Severe

4. Materials and Methods

Machine learning algorithms are used for prediction of air quality index. The following algorithms are used in this work for predicting the air quality index.

1) K-Nearest Neighbour
   It is a non-parametric supervised algorithm used to take the k closest training examples from the air quality index dataset. For a given distance and a k value, the algorithm calculates the distance between the data point and the training dataset for selecting the k nearest ones.
2) Naïve Bayes
   It is two stage classification algorithms. In the learning stage, train the model is done on the known dataset. In the evaluation stage, the performance testing can be done for multiple applications. Basically, the prediction is done based on the bayes theorem.
3) Support vector Regression
   It is a supervised learning algorithm. It is used to explore the relationship between one or more predictor variables and the dependent variable. The basics of SVR is to map the input data to a random high dimensional feature space.[11]
4) Random Forest Regression
   It is a supervised learning algorithm. Here the decision tree is created by taking the samples from the training dataset and the regression line is created.
5) XG Boost
   It stands for extreme gradient boosting. It is popular and widely used machine learning algorithm. Missing values can be handled efficiently in this model and it has buit-in support for parallel processing. It is also highly customizable and allows to make changes in the parameters of the model which leads to optimization.

The following steps are implemented to produce the results:

1) Data set selection is the first step which is taken from Kaggle and download the air quality index prediction dataset as a CSV fille.
2) Data pre-processing is the first step for cleaning the dataset.The air quality dataset is filtered and cleaned by removing missing values, null values and unwanted values for the important cities in India such as Delhi, Kolkata and so on to measure the AQI .
3) Dataset is normalized by using the Standard Scalar from Scikit learn library.
4) Spitting the dataset into training and testing phase in the ratio of 80:20 and the training is given by taking the random samples
5) Apply various machine learning algorithms and predict the air quality index then the results are tabulated.
6) Evaluation can be done for the machine learning models.The evaluated metrics such as MSE. RMSE, MAE are used to measure the accuracyand the values are tabulated.
7) Find out the highest accuracy value of the machine learning algorithm and declare that the machine learning algorithm suits to this problem and the overall performance measure is calculated.

5. Results and Discussion

The evaluation metrics are discussed in detail in this session. The Mean Square Error is a parameter which is used to measure the closeness of the data points with the fitted lines. If the mean square error is 0 means the model is a perfect model and produces 100% accuracy. Table-2a and 2b shows the results of the machine learning algorithm in terms of accuracy, precision, recall and F1-score for training as well as testing phase. Precision is the ratio of true positive and all the positive values. Accuracy is the ratio of the correctly labelled data with the entire dataset.

$$\text{Precision} = \frac{True\ positive}{True\ positive + False\ positive}$$

$$\text{Recall} = \frac{True\ positive}{True\ positive + False\ Negative}$$

$$\text{Accuracy} = \frac{True\ positive + True\ Negative}{True\ Positive + False\ positive + True\ Negative + False\ Negative}$$

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table -3 shows the performance of the various machine learning algorithms with the evaluation parameters. The XGBoost algorithm performed very well when we compare with other algorithms. The naïve Baye's model got the good score for R^2 error prediction.

Table 2 (a): Comparison results of the training set

Model	Accuracy	Precision	Recall	F1-Score
KNN	88	93	89	95
Naïve Bayes	84	90	93	87
SVM	80	89	92	87
Random Forest	87	92	87	90
XGBoost	90	94	94	90

Table 2 (b): Comparison results of the testing set

Model	Accuracy	Precision	Recall	F1-Score
KNN	84	91	84	93
Naïve Bayes	82	87	88	91
SVM	76	90	89	82
Random Forest	85	91	90	89
XGBoost	89	94	93	90

Table 3: Results of Various ML algorithm for AQI Prediction

Models	MAE	RMSE	RMSLE	R^2
KNN	0.812	4.123	0.612	0.412
Naïve Bayes	0.521	3.214	0.223	0.393
SVM	0.632	3.811	0.112	0.612
Random Forest	0.630	2.122	0.211	0.642
XGBoost	0.481	1.006	0.1.45	0.794

6. Conclusion

Due to the dynamic nature of the environment, variation in time and pollution, the prediction of the air quality index is becoming a challenging task to the researchers. Consistent monitoring of the air quality leads us to avoid negative impacts to some extend by giving proper alarming to the society. The results of the models for both training and testing are given in terms of the standard metrics like accuracy, precision, recall and F1-score. The XGBoost model performs well in both training and testing dataset. The future enhancement of this research work is planned to implement deep learning algorithms for measuring the air quality index. It is also planned to implement ensemble techniques to get more accurate results.

References

[1] BC Air Quality. (Accessed July 23, 2015). Pollutants: An Introduction [Online], Available: http://www.bcairquality.ca/101/pollutants-intro.html
[2] NASA. 2014. New NASA images highlight U.S. Air Quality Improvement [Online], Available: https://www.nasa.gov/content/goddard/new-nasa-images-highlight-us-air-quality-improvement/#.VbEhrrNVikp
[3] LSC Atmospheric Sciences. (Accessed July 23, 2015). Primary Pollutants [Online], Available: http://apollo.lsc.vsc.edu/classes/met130/notes/chapter18/primary.html
[4] Kostandina Veljanovska1 & Angel Dimoski2, Air Quality Index Prediction Using Simple Machine Learning Algorithms, 2018, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
[5] Aditya C R, Chandana R Deshmukh , Nayana D K and Praveen Gandhi Vidyavastu , Detection and Prediction of Air Pollution using Machine Learning Models,2018, International Journal of Engineering Trends and Technology (IJETT)
[6] Arwa Shawabkeh , Feda Al-Beqain , Ali Rodan, Maher Salem, Benzene Air Pollution Monitoring Model using ANN and SVM, 2018, IEEE]
[7] J. Angelin Jebamalar& A. Sasi Kumar, PM2.5 Prediction using Machine Learning Hybrid Model for Smart Health,2019, International Journal of Engineering and Advanced Technology (IJEAT)
[8] Kostandina Veljanovska1 & Angel Dimoski2, Air Quality Index Prediction Using Simple Machine Learning Algorithms, 2018, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).
[9] Savita Vivek Mohurle, Dr. Richa Purohit and Manisha Patil,A study of fuzzy clustering concept for measuring air pollution index,2018, International Journal of Advanced Science and Research
[10] Maryam Aljanabi, Mohammad Shkoukani and Mohammad Hijjawi, Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman,Jordan,2020, International Journal of Automation and Computing.
[11] W. Wang and Z. Xu, "A heuristic training for support vector regression," Neurocomputing, vol. 61, pp. 259–275, 2004.