Proceedings of the Third International Conference on Smart Electronics and Communication (ICOSEC 2022)
IEEE Xplore Part Number: CFP22V90-ART; ISBN: 978-1-6654-9764-0

Prediction Analysis using Random Forest Algorithms to Forecast the Air Pollution Level in a Particular Location

Puli Dilliswar Reddy1
Research Scholar,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, TamilNadu, India, 602105.
pulidilliswarreddy17@saveetha.com

L.Rama Parvathy
Project Guide, Corresponding Author,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, TamilNadu, India, 602105.
ramaparvathy1.sse@saveetha.com

2022 3rd International Conference on Smart Electronics and Communication (ICOSEC) | 978-1-6654-9764-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOSEC54921.2022.9952138

Abstract: To forecast the degree of air pollution in a specific area of a region using techniques Innovative Random Forest against Naive Bayes. Two groups of algorithms are Random Forest and Naive Bayes. The technique was developed and tested on a 32516-record dataset. In a programming experiment, each approach was iterated N=5 times to identify different levels of air pollution. The threshold value is 0.05 percent, and the confidence interval is 95%. The G-power test is around 80% effective. When compared to Naive Bayes, the innovative Random Forest method (98.26%) offers higher accuracy (97.32%).Random forest has the highest accuracy in comparison to the Naive Bayes algorithm. Significance value for accuracy is 0.056(p>0.05), Precision 0.02(p<0.05) and recall 0.01(p<0.05) based on 2-tail analysis. Conclusion: Random Forest has improved performance when compared to Naive Bayes in the forecast of pollution in air.

Keywords: Machine learning, Innovative Random Forest algorithm, Naive Bayes algorithm, Air pollution, Prediction, Recall

I. INTRODUCTION

As populations increase in size, so does the transportation infrastructure that is dependent on fossil fuels. The increase in vehicle use causes an increase in traffic-related pollution emissions. Because air pollutants have a negative impact on human health, urban pollution might become a big issue in both developed and developing countries. Numerous ailments, such as carcinoma, are also induced by numerous air contaminants. Pollution may also cause other major environmental issues, such as air pollution and, as a result, the gas effect. For example, SO2 and NO2 are the leading sources of air pollution, but CO2 and N2O are the leading causes of the gas effect.The total number of publications published on this topic in at least 2 databases in the last five years is 25. In this experiment [1]authors found that the one-hour dataset had better resultsthanthefive minute dataset, regardless of the algorithms used. In this study [2], accuracy obtained is 70%.

These results demonstrate that accuracy may be improved with adequate training and control of class imbalance in data([3]). This study has obtained an accuracy of 80%. Various machine algorithm strategies were used to this data to forecast emission rate, and a comparison study was performed. In this paper [4], has attained the accuracy of 75.5%. The authors of this research suggested a deep machine learning approach for predicting air pollution. Based on the above analysis [5] is the best study because the model obtained 80% of accuracy. As the approach results in less accuracy rate, we use the machine learning domain to predict and replace less accuracy rate with better accuracy percentage using the random forest algorithm. Though there is a lot of research about the project, some of the papers yet failed to show good accuracy rates by their methods. This limitation made us take on this project and our goal is to get better accuracy. By comparing the Random Forest method to the naïve bayes algorithm, the performance analysis in classification with accuracy of forecasting the air pollution level in a specific location was improved. We engaged with numerous writers from our universities, which allowed us to finish the assignment quickly and accurately[6].However there is a ton of examination about the undertaking, a portion of the papers yet neglected to show great precision rates by their strategies. This impediment made us take on this undertaking and our objective is to improve precision. To further develop execution investigation in grouping with precision of prediction of air pollution.

II. MATERIALS AND METHODS

The proposed work is being researched in the DBMS Lab at the CSE department of SIMATS' Saveetha School of

Engineering. Using clincalc.com, the sample size was calculated to be 5 per group, with an 80 percent G power, a threshold value of 0.05 percent, and a confidence interval of 95 percent[7].For size computation, the mean and standard deviation were computed using existing research. This paper utilises two groups in total. The first group employs an established method, naive Bayes, whereas the second employs a novel algorithm, Random Forest.

A. Naive Bayes algorithm (NBA)

NBA is an administered algorithm used for learning classification which works on principle, i.e. that every pair of features being classified is independent of every other.

Pseudocode of Naive Bayes
Input: Determine various training and test data
Output: Determine calculated accuracy

1. X_new =rfe.transform(X).
2. X_train, X_test, y_train, y_test = train_test_split(X_new, y, test_size = 0.25)
3. nb.fit(X_train, y_train)
4. y_pred = nb.predict(X_test)
5. mse=mean_squared_error(y_test, y_pred)
6. mae=mean_absolute_error(y_test, y_pred)
7. r2=r2_score(y_test, y_pred)
8. Return accuracy

Figure 1 represents the illustration of the Naive Bayes algorithm. It explains that the first step is to initialize the dataset and pre-processes it. After feature selection, the algorithm measures the performance by training and testing the dataset.

B. Random forest algorithm

Pseudocode:
Input: Determine various training and test data
Output: Determine calculated accuracy

1. X_new =rfe.transform(X)
2. X_train, X_test, y_train, y_test = train_test_split(X_new, y, test_size = 0.25)
3. rf.fit(X_train, y_train)
4. y_pred = rf.predict(X_test)
5. mse=mean_squared_error(y_test, y_pred)
6. mae=mean_absolute_error(y_test, y_pred)
7. r2=r2_score(y_test, y_pred)
8. Return accuracy

Figure 2 represents the flow diagram of Random Forest algorithm. It explains that the first step is to initialize the dataset and pre-processes it. After feature selection, the performance of the algorithm is measured by training and evaluating the dataset. An Intel i5 processor, 1 TB hard drive, and 8GB RAM are used in the hardware setup. Software Configuration is the use of the Windows 10 operating system and the execution of CoLab and Microsoft Office.Initially, import the dataset as per classifiers and investigate the information to sort out what they resemble. Later, the data is pre-processed by splitting data into attributes and labels. After splitting data, training has been applied to data with algorithms. The algorithms predicted the expected results. Finally, evaluation has been done for results. The attributes in the dataset are Country, State, City, Place, Last update, Avg, Max, Min, Pollutants.Table 1 is the data collected from 5 different sample datasets. In Table 1, accuracy and test size were compared with two classifiers. Dataset used in this paper has 250 rows and 9 columns. Attributes give information such as avg, max, min etc. Table 1 represents data collection of Naive bayes and Random Forest with sample size five (N=5). Random Forest has higher accuracy than Naive Bayes algorithm and it has parameters like MAE, MSE, R SQUARED, RMSE, Precision, Recall and Accuracy [8].

The statistical software used for implementation is IBM SPSS version 21. The independent variables of the data are protocol type, service, flag such as dependent variables in the data are the accuracy that is considered in this task. Independent samples T-test is carried out in this work.

III. RESULTS

From Table 1, Data collection of Naive Bayes and Random Forest with sample size five (N=5) is given. Accuracy Table where the accuracy of Random Forest is about 99.15 % and Naive Bayes is about 97.60%. The accuracy varies with the number of decimals in the test. The algorithm's accuracy fluctuates as the test size is changed arbitrarily. Table 2 shows the results of a group statistics study of two groups.The mean accuracy of NB and RF is 92.71 and 94.17 respectively. The mean precision of NB and RF is 84.40 and 92.60. The mean recall of NB and RF is 79 and 88.

Table I: Accuracy analysis of Naive Bayes algorithm and Random Forest algorithm

S. No	Naive bayes		Random forest	
	Test Size	Accuracy(%)	Test Size	Accuracy (%)
1	0.25	92.60	0.25	94.70
2	0.25	91.26	0.25	95.93
3	0.25	88.60	0.25	93.60

4	0.25	90.40	0.25	98.20
5	0.25	91.20	0.25	96.60

Table II : GroupStatisticsfor with Standard Error and Mean for NB (1.0768[redacted] 92.9240) and RF (0.73674[redacted] 94.720).

Statistics of two Groups				
Collections	**K**	**AVG**	**STD.DEV.**	**S.E.M.**
ACCURACY				
NB	5	91.49	3.44733	1.5416
RF	5	96.02	2.92774	1.3093
PRECISION				
NB	5	88.03	3.0902	1.3820
RF	5	92.66	1.7260	.7719
RECALL				
NB	5	76.14	3.0573	1.3672
RF	5	84.31	1.8003	.8051

From Fig. 1, Flow diagram of Naive Bayes algorithm is illustrated. Architecture for predicting accuracy which consists of the steps that include in the procedure for the prediction of Air pollution level Accuracy. It consists of steps which are initialize Data Collection, pre-processing, Feature selection applying NB algorithm, Performance measures as given in Fig. 2.

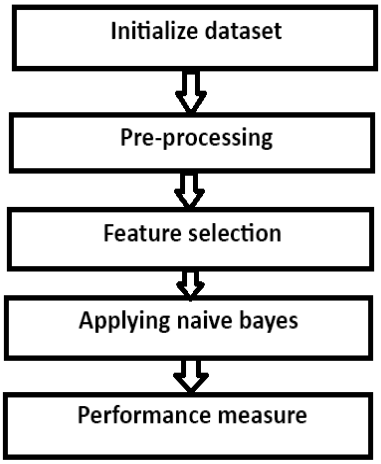

Fig. 1. Flow illustration of Naive Bayes.

From Fig. 2, Flow illustration of Random Forest algorithm is illustrated. Architecture for predicting accuracy which consists of the steps that include in the procedure for the prediction of Air pollution level Accuracy. It consists of steps which are initialize Data Collection, pre-processing, Feature selection, Applying RF algorithm, Performance measures as given in Fig. 1.

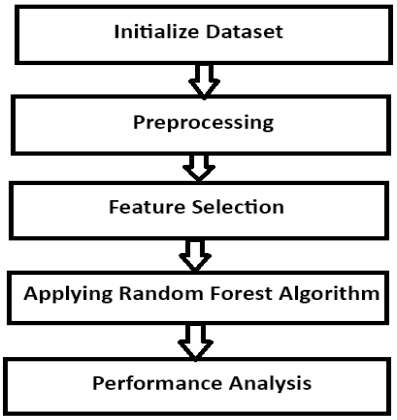

Fig. 2. Flow chart of random forest algorithm.

From Fig. 3, Simple Bar Mean of Accuracy by (RF, NB) is a bar chart that compares the mean accuracy of the Random Forest Algorithm (99.26) with the Naive Bayes Algorithm (97.32). Proposed Random Forest algorithm achieved better performance than Naive Bayes.

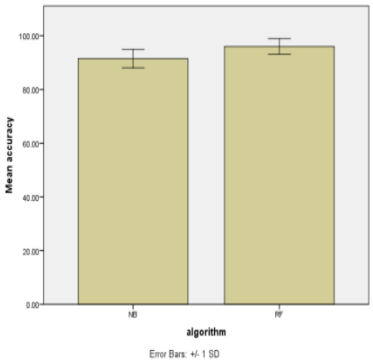

Fig. 3. Comparison of Random Forest and Naive Bayes in terms of mean accuracy.

Table 3 shows Independent Samples Test analysis of Random Forest Algorithm and Naive Bayes Algorithm. Significance value for accuracy is 0.056($p>0.05$), Precision 0.02($p<0.05$) and recall 0.01($p<0.05$) based on 2-tail analysis. When compared against other algorithms, the suggested Random Forest Algorithm performed better than the Naive Bayes Algorithm.

Table III: T-Test for Independent Sample and Comparison of the accuracy of air pollution prediction using Random Forest with naive bayes algorithms. NB is meaningfully lesser than the RF algorithm ($p < 0.05$).

Variance				t-test -Variance					95% Confidence - difference	
		X	∑.	x	diff	∑(2-tailed)	Mean diff.	S.E.D.	MIN	MAX
Accuracy	Assumed	0.039	0.448	-2.230	8	0.056	-3.468	1.55540	-7.05476	.11876
	Not Assumed			-2.230	7.057	0.057	-3.468	1.55540	-7.07201	.13601
Precision	Assumed	.147	.312	-4.362	8	0.02	-6.200	1.42127	-9.47745	-2.92255
	Not Assumed			-4.362	7.827	0.03	-6.200	1.42147	-9.49007	-2.90993
Recall	Assumed	7.509	0.487	-3.266	8	0.011	-4.00	1.22474	-6.82427	-1.17573
	Not Assumed			-3.266	7.965	0.011	-4.00	1.22474	-6.82645	-1.17355

IV. DISCUSSIONS

Random forest algorithm (99.154%) has better accuracy when comparedwithNaiveBayes (97.32%). As a consequence, in terms of accuracy, the Random forest exceeds the previous algorithms that were altered and converted to the best method. There is a statistically significant difference in accuracy between group statistics and the independent sample T-test. In this work[9], the probabilistic model is utilized to uncover the qualities of urban communities regarding various boundaries. Much of the time it's impractical to visit settings and get the qualities without any problem. In [10] the accuracy obtained in this experiment is 85.6%. The authors forecasted the air quality record using a variety of machine learning algorithms, including Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). Based on the results, the scientists concluded that the Random Forest predicts the air quality index more accurately.In[7], the accuracy attained in this experiment is 51.21%. In this paper, a progression of strategies for consolidating classifiers along with certain recently publicly released boosting calculations were investigated to anticipate the not so distant future's air quality expectation. In[3].In this experiment, the accuracy is 91.62%. A machine learning strategy for anticipating air quality indices for smart cities is proposed in this work. The proposed technique for characterising the air quality record using machine learning-based prediction models was successfully deployed in [1]. The key constraint of this study is that there are relatively few attributes in the dataset that can predict accuracy percentage for novel classification. Accuracy can be enhanced in the future by increasing the number of independent and dependent variables.

V. CONCLUSION

In this research, within our limitations, innovative Random forest (99.154%) has appeared to be better accuracy in comparison to the Naive Bayes (97.32%) with significance value for accuracy is 0.056($p>0.05$), Precision 0.02($p<0.05$) and recall 0.01($p<0.05$) based on 2-tail analysis.

REFERENCES

[1] Amado, Timothy M., and Jennifer C. Dela Cruz. 2018. "Development of Machine Learning-Based Predictive Models for Air Quality Monitoring and Characterization." TENCON 2018 - 2018 IEEE Region 10 Conference. https://doi.org/10.1109/tencon.2018.8650518.

[2] Mahalingam, Usha, Kirthiga Elangovan, Himanshu Dobhal, Chocko Valliappa, Sindhu Shrestha, and Giriprasad Kedam. 2019. "A Machine Learning Model for Air Quality Prediction for Smart Cities." 2019 International Conference on Wireless

Communications Signal Processing and Networking (WiSPNET). https://doi.org/10.1109/wispnet45539.2019.9032734.

[3] Mahanta, Soubhik, T. Ramakrishnudu, Rajat Raj Jha, and Niraj Tailor. 2019. "Urban Air Quality Prediction Using Regression Analysis." TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON). https://doi.org/10.1109/tencon.2019.8929517.

[4] Andri M Kristijansson and Tyr Aegisson, "Survey on Technique and User Profiling in Unsupervised Machine Learning Method", Journal of Machines and Computing, vol.2, no.1, pp. 009-016, January 2022. doi: 10.53759/7669/jmc202202002

[5] Pasupuleti, Venkat Rao, Uhasri, Pavan Kalyan, Srikanth, and Hari Kiran Reddy. 2020. "Air Quality Prediction Of Data Log By Machine Learning." 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). https://doi.org/10.1109/icaccs48705.2020.9074431.

[6] Reddy, M. Ramana, and M. Ramana Reddy. 2020. "IoT BASED AIR AND SOUND POLLUTION MONITIORING SYSTEM USING MACHINE LEARNING ALGORITHMS." Journal of ISMAC. https://doi.org/10.36548/jismac.2020.1.002.

[7] Simu, Shreyas, Varsha Turkar, Rohit Martires, Vranda Asolkar, Swizel Monteiro, Vaylon Fernandes, and Vassant Salgaoncary. 2020. "Air Pollution Prediction Using Machine Learning." 2020 IEEE Bombay Section Signature Conference (IBSSC). https://doi.org/10.1109/ibssc51096.2020.9332184.

[8] Xayasouk, Thanongsak, and Hwamin Lee. 2018. "AIR POLLUTION PREDICTION SYSTEM USING DEEP LEARNING." Air Pollution XXVI. https://doi.org/10.2495/air180071.

[9] Zheng, Hong, Yunhui Cheng, and Haibin Li. 2020. "Investigation of Model Ensemble for Fine-Grained Air Quality Prediction." China Communications. https://doi.org/10.23919/j.cc.2020.07.015.