Accuracy of Classifier Models in Presence and Absence of Outliers in a Time-Series Dataset of Air Quality

Pukhraj Rathkanthiwar [1], Karan Badlani [2], Ankita Harkare [3]

Abstract

Classifier models are actively being brought into use in various fields to chart out future possibilities. When it comes to values like the Air Quality Index that directly affect human lives, accuracy is a must. Air pollution claims around 4.2 million lives per year, according to the World Health Organization. Inaccurate predictions in such fields can be detrimental to the environment as well as general public health. However, the accuracy of these classifier models can be affected by how the given data is processed and the presence of outliers. The outliers can considerably alter the results of these classifier models. Hence, the aim of this project was to study the effects of the presence and absence of outliers in a dataset on the prediction accuracy of various classifier models and find the most accurate classification model among the ones that were tested.

DOINumber:10.14704/NQ.2022.20.11.NQ66031 **NeuroQuantology 2022; 20(11): 277-286**

Contents

1

2

3

1 Introduction

1.1 AQI

It's been observed that to deduce & predict outcomes for various real-life situations, Machine learning classifier models are being used (WHO). Predictions in the financial sector & healthcare, weather forecasting, and environmental surveys are its applications to name a few (Deist et al, 2018). The data that is already available is used to train these models and predict the missing or forthcoming values in a dataset. The data is split into training and test after which the classifier model is trained on the training set and is used to test it on the Test part of the dataset (Sethi et al, 2019). By comparing these predictions from the original dataset, we evaluate the accuracy of our machine learning model. Classifier models are basically used in datasets where the target variables lie in a specific domain. For our research, a dataset of air quality and pollution level(4) monitored across different stations in India was procured. Different components of the air, including levels of particulate matter and pollutants, were measured in this dataset. These components included pollutants like particulate matter, nitrogen oxides, sulfur dioxide, carbon oxides, methane, benzene, toluene, xylene, and ozone.

1.2. Issues & Challenges

Predicting changes in air quality can be a daunting task, given that there is an interplay of various physical, chemical and anthropogenic factors that results in the air quality. Any anomaly in even one of these can have a considerable impact on air quality. Good prediction accuracy is of paramount importance in this context. When it comes to predicting values in such datasets that involve various variables each affecting the target variable differently using machine learning models, the data requires pre-processing, without which it might produce inaccurate results. Thus, outliers have to be removed from the dataset. When it comes to planning solutions to improve air quality, accurate predictions can go a long way in dealing with pollution.

1.3. Motivation and Objectives Behind the Study

As already mentioned, classification models can have certain implications in environmental surveys and prediction(5). They can help in establishing a pattern using the available data and deduce future values for a certain variable. This can be monumental when it comes to determining changes in important values such as the Air Quality Index(6). The Air Quality Index directly affects the quality of human life. An unexpected change within the AQI can be detrimental to public health. With the help of classifier models, forthcoming values can be predicted and a holistic change can be planned beforehand. The objectives of this study were to analyze the effect of outliers on the accuracy of some of the most commonly used classifier models and find a reliable classifier in the process. The implications of this study can range from improving the accuracy of existing classifiers and producing more accurate predictions in various fields.The conditional variables viz. the pollutants had some missing values. We trained the dataset using various classifier models and analyzed its results by comparing them to the actual data. The accuracy was measured and documented. The same process was repeated after removing any outliers that presented values beyond the range of the target variable. The differences in accuracy between the two cases were observed. Consequently, the model with the highest accuracy was discussed. Models that showed considerable changes in accuracy were also observed.The paper has been divided into various sections. First off, the resources used for this study were discussed. This included the pre-processing of the dataset and the detection and removal of outliers. Then, the various algorithms that were studied through this project have been discussed. In the end, a detailed comparative analysis of the results was done.

2 Resources Used

For this project, we obtained a publicly available time-series dataset of Air quality from the Central Pollution Control Board (7). This dataset is a collection of data procured from different

Station IDs spread across the whole country.These stations monitor the quality of air and detect the presence of pollutants in the air. They are also able to calculate the levels of pollution in the air and produce an automatic calculation that goes on to become the Air Quality Index. This calculation is made by taking into consideration various elements in the air like particulate matter (PM2.5 and PM10 particles) and other substances. Particulate matter can go unchecked in the human body and damage the lungs and internal organs, resulting in decreased function. Other substances include chemical compounds like ammonia and oxides like nitrous oxide, nitric oxide, carbon monoxide, sulphur dioxide and even ozone. In tandem with nitrogen oxides, Ammonia contributes to nitrogen pollution in the environment. Atmospheric ozone is detrimental to plants and other ecosystems, as it obstructs the process of photosynthesis in them. Carbon monoxide is a greenhouse gas that is one of the leading contributors to global warming. Cyclic compounds like xylene, benzene and toluene are known to degrade the soil and water quality. They can be acutely harmful to aquatic life. Unmonitored exposure to them can cause anaemia, irritation and nausea in humans. The AQI bucket refers to a range of values in a domain that helps in deciding the possible effects of air pollution on the quality of human life. For example, if the AQI lies between 0-50, the health impacts of the air in that specific region can be considered minimal. If the same lies between 401-500, then it can severely impact the general health of the public(8). Given above is a short section of the time-series data set used for this project.

Since a classification model cannot be used with a dataset without an index column, Pandas Libraries were used to convert the date and time columns into index columns (9).

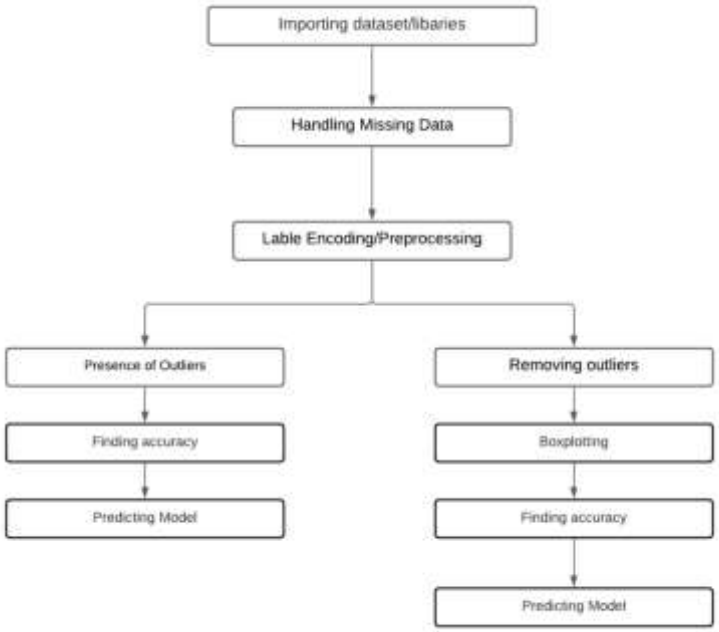

Fig. 1 Flowchart of Methodology.

Table 1

Impact of AQI on Health of people

AQI	REMARK	POSSIBLE HEALTH IMPACTS
0-50	Good	Minimal Impact
51-100	Satisfactory	Minor Breathing Discmort to sensitive
101-200	Moderate	Breathing discomfort to people with lung and heart diseases
201-300	Poor	Breathing discomfort to most people
301-400	Very Poor	Development of Respiratory Illness
401-500	Severe	Affects healthy people and seriously impacts those with existing diseases

2.1. Handling Missing Data

The missing data in this dataset was managed using the Numpy library. For columns containing integer data, the mean of all the values in that column was taken and put into the missing values. For the columns containing categorical data, the value that was seen most frequently was used to fill the missing values in the dataset.

2.2. Methodology

Once the missing values were found, the data was then pre-processed using the SKLearn library. To train the dataset with an algorithm, the dataset needed to be in a certain format. With this process, the data became much more accessible and usable for the training models.

2.2.1. Label Encoding

Label encoding refers to converting categorical data into a machine-readable numerical form. Alphabetic data is not suitable for processing and needs pre-processing to become easily accessible to the algorithms. Once pre-processed, the data becomes legible for the machines. (10)

Label encoding works by giving object type data a certain numeric value. This value is given on the basis of alphabetical order. For example, if it's a column of days in a week, then the assignment of values will begin on Friday and end on Wednesday (11).

In our dataset, we employed this process for the AQI bucket column and the station ID column. AQI bucket uses different ranges of pollution. These ranges are then replaced by numerical data.

2.2.2. Training and Testing Sets

The dataset was split into X and Y variables. X consisted of all conditional variables in the dataset, like the amount of various pollutants in the air. On the basis of these variables, the target variable is predicted. Y consisted of the target variables, that is the AQI bucket. The dataset was further split into training and testing datasets. Training datasets are used to train the machine based on available input. The test datasets are corresponding data that are used for evaluating the trained model. Training and testing sets were created for both X and Y variables. x_train is a feature used for training the model by using conditional variables as

input. y_train refers to the target variables that correspond to those of x_train. The test datasets viz. x_test and y_test are the features and their corresponding labels for further evaluating the trained model. (12) For this study, we split the dataset into training and testing sets in the ratio of 3:1. The random state was set to zero so that the values of these sets are not randomized when an algorithm is applied.

The ratio of training and test sets according to the total dataset (%).

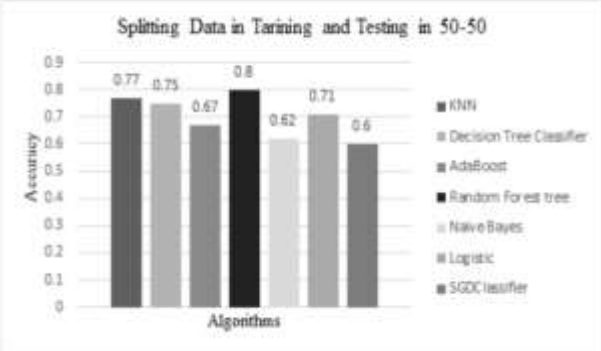

Fig. 2 Bar graph of Splitting data 50-50

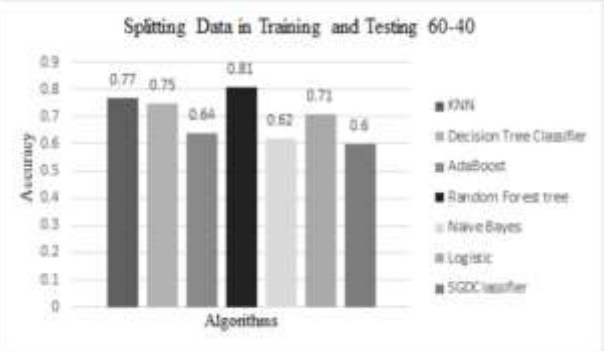

Fig. 3 Bar graph of Splitting data 60-40

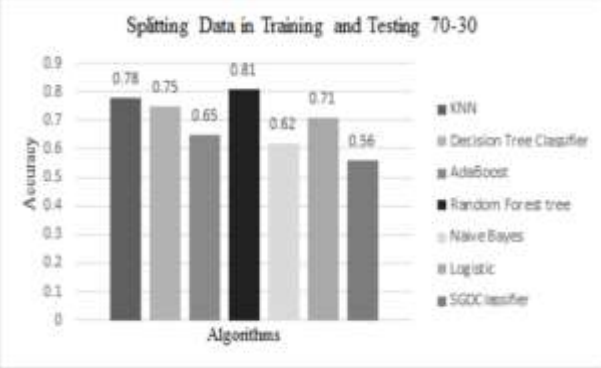

Fig. 4 Bar graph of Splitting data 70-30

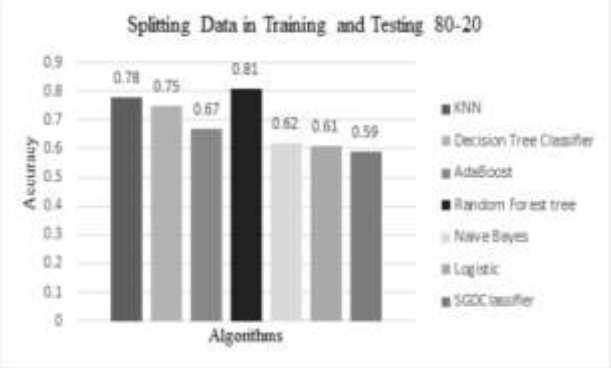

Fig. 5 Bar graph of Splitting data 80-20

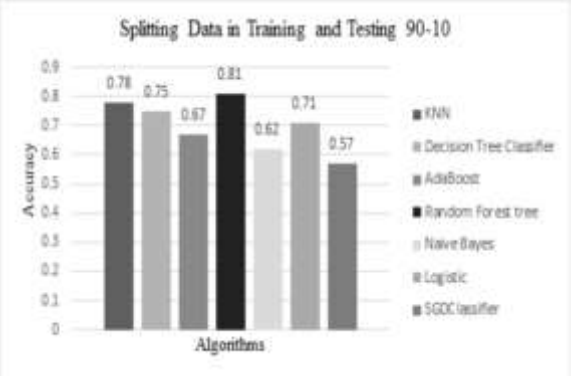

Fig. 6 Bar graph of Splitting data 90-10

2.2.3. StandardScaler

Real-world data can be too bulky for algorithms to process effectively. For that purpose, a simple Feature Scaling tool, StandardScaler has been used to restrict the data to a certain range. The effectiveness with which an algorithm

processes the data is greatly increased with this process. StandardScaler standardizes the data into such a range that the mean of the distribution is always 0 and the standard deviation is always 1. This scaling has been applied to the X_train and X_test datasets. After the process, the data was segregated into a range from -3 to 3.

2.2.4. Outlier Detection and Elimination

Outliers are values that lie distinctly far from other values in a given dataset. For example, a value that is larger or infinitesimal compared to the other values in a dataset is called an outlier. Outliers can be detrimental to the algorithm's prediction accuracy (13). A single outlier can throw off the algorithm remarkably and affect the results. For a categorical time-series dataset like this one, outliers can cause inaccurate and implausible results. Hence, removing outliers becomes mandatory. Outliers can be detected by various methods. For this dataset, the Boxplot method was used to detect them.

The Boxplot method creates a range consisting of quartiles and inner quartiles. With the help of these, a lower limit and an upper limit are determined. Any values lying beyond these lower and upper limits are considered outliers.

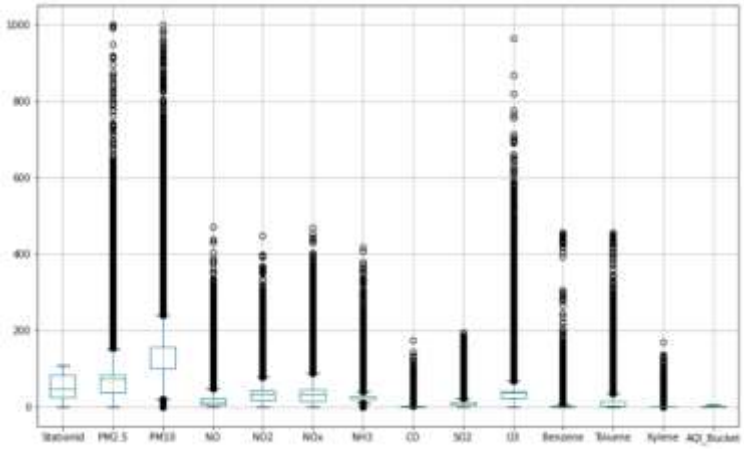

Fig. no 7 Construction of Boxplot

Thus, any values lying beyond the lower and upper limits thus created are eliminated. For that process, the given steps were used:

1. The lower bound was found using the formula: lower_bound= (q1-1.5 x IQR)
2. The upper bound was found using the formula: upper_bound= (q3 +1.5 x IQR)

With the help of these formulae, a condition was created which detected the outliers and replaced them with NaN (Not a Number).

3 Methodology

Seven algorithms including decision tree, and AdaBoost, were used. Four of these models are Ensemble learning algorithms. Ensemble learning algorithms are divided into two parts: bagging and boosting. In bagging, multiple models of an algorithm are trained with the original dataset. On the basis of the results of these various models, a final answer is generated. In the process of boosting, various subsidiary models are created. These models reiterate the previous one to produce more accurate predictions. In simple terms, the first sub-model will train the dataset and produce some results. However, these results may or may not be accurate. Any wrongly classified data is given more weight in the dataset, thus increasing its probability of landing in the next subsidiary model's training dataset. This process continues until n number of sub-models. Once all the models have produced results, they are tested against the Y dataset i.e. the target variable. The results obtained are observed and

a final answer is generated on the basis of the majority.

Seven algorithms are used in this:

- K-Nearest Neighbour
- Decision Tree
- Random Forest
- Adaboost
- Naive Bayes
- Logistic Regression
- SGD classification

4 Result and Discussion

After processing the data using the aforementioned algorithms, the results were seen as shown above in Fig. As mentioned, the dataset was trained at first without removing the outliers. In that case, the Random Forest Tree Classifier model displayed the highest prediction accuracy of 0.81 per prediction. The second highest accuracy was achieved by the K-Nearest Neighbor classifier - an accuracy of 0.78. This was followed by the Decision Tree Classifier and Adaboost with 0.75 and 0.65 accuracies per prediction respectively. The poorest accuracies observed were those of the Naive Bayes Classifier and SGD Classifier which provided accuracies of 0.62 and 0.56 per prediction respectively. Consequently, outliers were removed and the whole dataset was trained by all the algorithms again. In this case, Random Forest's accuracy underwent a negligible change. It still maintained the first position and resulted in 0.81 accuracy per prediction. K-Nearest Neighbor came second, with a slightly increased accuracy of 0.79 per prediction. The Decision Tree Classifier came up with an unchanged accuracy of 0.75. It was followed by AdaBoost which came up with a considerable increase in accuracy, going from 0.65 to 0.67. Naive Bayes experienced a similar hike in accuracy going from 0.62 to 0.64. Logistic Regression# Classifier was the only model that underwent a decrease in accuracy after the removal of the outliers. Its accuracy went from 0.71 to 0.63. SGD Classifier also displayed a slight increase, going from 0.56 to 0.58. The data in Figures

3,4,5,6 & 7 point out that the different ratios of training and testing sets also resulted in variations in results. The Decision Tree and the Random Forest Tree model remained unaffected by these changes and delivered maximum accuracy. However, other classifier models were subject to noticeable changes.

Table 2
Accuracies in presence and absence of outliers

Algorithm Name	Presence of Outlier	Absence of Outlier
K-Nearest Neighbour	0.78	0.79
Decision Tree	0.75	0.75
Random Forest	0.81	0.81
Adaboost	0.65	0.67
Naive Bayes	0.62	0.64
Logistic Regression	0.71	0.63
SGD classification	0.56	0.58

5 Conclusion

In the paper, a comparative study of the accuracies of different classifier models in the presence and absence of outliers has been presented. A time-series dataset was trained with various classifier models for the study. Air pollution has risen considerably in the past few decades and has already doled out several global risks. Machine learning can be an effective tool to predict environmental factors with respect to pollution. With the help of machine learning, changes in environmental factors such as Air Quality Index and even Climate Change can be predicted. Hence, this study was focused on performing a comparison of algorithms that offer the best accuracy for an Air Quality Index dataset. It was found that the highest accuracy was delivered by the Random Forest Tree classifier algorithm which gave an accuracy percentage of approximately 81% in both cases. K-Nearest Neighbor and Logistic Regressor Classifier models gave moderate accuracy. It was also observed that Naive Bayes Classifier and SGD Classifiers algorithms had the poorest accuracy amongst all the ones that were studied. The time-series dataset used was procured from Central Pollution Control Board, India. The proposed research in this study is an effective way to compare these classifying algorithms. However, depending on the type of dataset, the performance of these models might vary. The inferences of this study can have various implications in meteorology and pollution control in today's world. This work can also be collaborated by the study made by (Harkare et al, 2018) and (Mahajan et al, 2020) for fuel adulteration systems.

References

1. World Health Organization official site
2. Deist, Timo M., et al. "Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers." *Medical physics* 45.7 (2018): 3449-3459.
3. Sethi, Jasleen Kaur, and Mamta Mittal. "A new feature selection method based on machine learning technique for air quality dataset." *Journal of Statistics and Management Systems* 22.4 (2019): 697-705.
4. Sharma, Arjun, et al. "Estimation of Air Quality Index from Seasonal Trends Using Deep Neural Network." *International Conference on Artificial Neural Networks*. Springer, Cham, 2018
5. Singh, Monika, et al. "Prediction of Pollutant Oxide of Nitrogen Component in Delhi City using Artificial Neural Network." https://cpcb.nic.in/
6. Jassim, Majeed S., and Gulnur Coskuner. "Assessment of spatial variations of particulate matter (PM 10 and PM 2.5) in Bahrain identified by air quality index (AQI)." *Arabian Journal of Geosciences* 10.1 (2017): 19.
7. McKinney, Wes. "pandas: a foundational Python library for data analysis and statistics." *Python for high performance and scientific computing* 14.9 (2011): 1-9.
8. McKinney, Wes. "Data structures for statistical computing in python." *Proceedings of the 9th Python in Science Conference*. Vol. 445. 2010.
9. McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.
10. Potdar, Kedar, Taher S. Pardawala, and Chinmay D. Pai. "A comparative study of categorical variable encoding techniques for neural network classifiers." *International journal of computer applications* 175.4 (2017): 7-9.
11. Uçar, Muhammed Kürşad, et al. "The effect of training and testing process on machine learning in biomedical datasets." *Mathematical Problems in Engineering* 2020 (2020).
12. Lebanon, Guy. *Riemannian geometry and statistical machine learning*. Carnegie Mellon University, 2005.
13. Williamson, David F., Robert A. Parker, and Juliette S. Kendrick. "The box plot: a simple visual method to interpret data." *Annals of internal medicine* 110.11 (1989): 916-921.
14. Sarkar, Manish, and Tze-Yun Leong. "Application of K-nearest neighbours

algorithm on breast cancer diagnosis problem." *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000.
15. Freund, Yoav, and Llew Mason. "The alternating decision tree learning algorithm." *icml*. Vol. 99. 1999.
16. Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., ... & Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, *151*, 147-160.
17. S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458.
18. Harkare, Ankita H. "Analytical Study for Development of Fuel Adulteration Detection System." (2018): 4327-4333.
19. Mahajan, Harsh, et al. "Study and Development of Fuel Adulteration Detection System." *Helix* 10.04 (2020): 181-185.

Biography of Authors

Pukhraj Prashantraj Rathkanthiwar pursuing his Bachelor of Engineering (Electronics and Communication) from Shri Ramdeobaba College of Engineering and Management, Nagpur, India (2019-2023). He is working in the field of machine learning.
He had completed his internship at Dassault Systemes.
Email: rathkanthiwarpp@rknec.edu

Karan Rajkumar Badlani pursuing his Bachelor of Engineering (Electronics and Communication) from Shri Ramdeobaba College of Engineering and Management, Nagpur, India (2019-2023). He is interested in the field of Data Science and Machine learning and its algorithms.
He had completed an internship in a startup where his role was based on Data Analytics.
Email: badlanikr@rknec.edu

Ankita Hitesh Harkare pursued her Bachelor of Engineering (Electronics and Communication) and Masters in Engineering (VLSI Design) from Shri Ramdeobaba College of Engineering and Management, Nagpur, India in 2009 and 2014 respectively. She is currently pursuing her PhD from Indian Institute of Information Technology, Nagpur, India. She worked as a System Engineer at Tata Consultancy Services after completion of graduation. She ventured into the field of teaching and research due to her dedication towards academia. She is currently working as Assistant Professor at Shri Ramdeobaba College of Engineering and Management, Nagpur, India. She has attended and presented papers at 8 National and International Conferences and published 11 Journal papers till date. She has 3 published Patent, 3 Copyright, and 1 Book Chapter in Springer Series to her credit. She has reviewed many peer reviewed journals including those of Elsevier Publications. She has delivered expert lectures in the field of VLSI Design and conducted workshops related to same. Her Current Research Interest are Embedded VLSI Design and Antenna Design. She is a life member of ISTE and IAENG.

Email: harkareah@rknec.edu