


Predicting Air Quality Based on Multiclass Machine Learning Techniques

Nafea Ali Majeed Alhammadi* 

Department of Computer Sciences, Shatt Al-Arab University College, Basrah, Iraq
Nafeaalhamadi@yahoo.com

Abstract

India is among the most polluted nations with severe environmental implications of pollution increase in several basic cities. For the past few years, Indian cities have witnessed an alarming drop in Air Quality (AQ) which is a result of rapid economic growth. The lives of billions of people are deeply affected by Air Pollution (AP) every year. The economic sectors that source poor urban AQ include transportation, agriculture, construction, forestry, and logging, emitting dangerous gases and particulates such as NO₂, NH₃, SO₃, and Benzene. This research attempts to implement four different classifiers independently on the AQ dataset. The four classifiers that are being implemented are the Multiclass Decision Jungle (MDJ), Multiclass Logistic Regression (MLR), Multiclass Decision Forest (MDF), and Multiclass Neural Network (MNN). This project aims to identify the most suitable among the four classification models to build the best model for AQ. The performance of a model is judged by accuracy, precision, and recall. Likewise, in the other three modes, MDF performs best, where it obtained 99.96% accuracy, 98.91% precision, and 99.76% recall.

Keywords: air quality, predicting, machine learning, multiclass logistic regression, multiclass decision jungle, multiclass decision forest, multiclass neural network

Article history: Received:6-3-2024 , Accepted:25-3-2024 , Published: 15-9-2024

This article is open-access under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Air quality (AQ) is particularly poor in large cities, which find it hard to balance the Air Quality Index (AQI) and environmental concerns. Every year, billions of people suffer greatly from Air Pollution (AP). India is one of the most polluted countries in the world, with significant environmental effects from the increasing pollution rise in several big cities [1]. AQ in Indian cities has deteriorated in the last few years as a result of strong economic growth. The sources of poor urban AQ in the economic sectors include transportation, agriculture, construction, forestry, and logging [2]. The relationship between NO₂ pollution and economic growth is robust, with annual increases of up to 4.4 percent [3].

With the increase in India's yearly economic growth, the health problems associated with the pollution caused by poor AQ grow as well [4], [5]. It

is these high pollution levels that cause serious health impacts with Indian estimates for 2010 ranging from 483,000 to 1,267,000 premature deaths every year from outdoor pollution and 748,000-1,254,000 premature deaths due to indoor household pollution. However, their exact size is unknown, and health-related costs are expected to be substantial; an estimate from the World Bank suggests that health-related costs will be 7.7% of national Gross Domestic Production (GDP) in 2013 [6]. In this regard, identification and control of the source of pollution is important. Classification of data by machine learning (ML) algorithms is the only way to be able to minimize the effects of AP on human health [7], [8].

This project aims to build the best classification model to classify AQ based on the dataset from daily AQ across cities in India from 2015 until 2020 using a data mining (DM) approach. This project aims to implement four different classifiers independently on

* Corresponding author: Nafeaalhamadi@yahoo.com

the AQ dataset and determine the best classification model among the four.

This work is divided on the following way. Section 2 reviews all of the related work on the classification model. Section 3 specifies the CRISP-DM methodology applied for the realization of the DM task in addition to the dataset and metrics evaluation. Section 4 presents findings while Section 5 closes with some suggestions for further research.

2. Related Work

The Indian government stated that AP caused 1.1 million deaths in 2015. The rapid expansion of the industrial, power, and transport sectors and planned and unplanned urbanization have all led to the alarming growth of ambient AP in India [4]. Government agencies use the AQI to describe the quality of air at a given location. According to Kumar and Goyal [9], many urban cities worldwide use the AQI for local and regional AQ management.

Most of the research on AQ in India has been conducted through DM approaches. The systematic review of DM and AP by Bellinger et al. [3] emphasized that model selection and evaluation are very important parts of applying ML techniques to real-world systems. The most used DM techniques in the context of AP data are regression, classification, clustering, and association mining. Researchers usually use decision trees (DT), artificial neural networks (ANN), and support vector machines (SVM) [10], [11] as the most common models.

In the study by Aditya et al. [1], logistic regression (LR) and autoregression (AR) were reviewed in this research paper for modeling use. Based on the system's function, classify whether the air is polluted (0) or not polluted (1). The results found that LR is best suited among all. According to Aditya et al. [1], the AQI is classified into six different categories: "good, satisfactory, moderate, poor, very poor, and severe," based on the value of the AQI. The attributes being mined in previous research include Particulate Matter (2.5 & 10), Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide, Nitrogen oxide, Ozone, and temperature.

The project is explained in the next section based on the related works. A recent study by Jagtap and Babbar [8] compared two classification models: Random Forest (RA) and ANN. The results showed that a RA has a 91.06% accuracy more than an ANN, which has an accuracy of 87.5%.

In the study by Shashi and Sanjay [12], the AQI was calculated using a linear regression model, referring to the aggregate of the highest AQI in that region of India. The pollutant's max-sub index calculates the AQI for that specific area. Apart from that, the Nave Forecast approach splits data 85:15 between test and training datasets to spot huge seasonal changes and trends in AP. In another study by Taneja et al. [13] LR and ANN were compared to study and anticipate existing trends in AP in Delhi and analyze time series data to study the data's behavior.

3. Materials and Methods

This project compares the performance of Multiclass Decision Jungle (MDJ), Multiclass Decision Forest (MDF), Multiclass Logistic Regression (MLR), and Multiclass Neural Network (MNN) individually in the classification of AQ using the daily AQ dataset across cities in India.

The research project was built upon the Cross-Industry Standard Process for DM (CRISP-DM). The six (6) phases of CRISP-DM that have been carried out in this project. The details of each phase are in the next sub-sections, and the classification methodology is illustrated in Fig. 1.

Business Understanding: The first step is to look at the dataset from a business perspective. Upon reviewing the dataset of daily AQ across cities in India, which consists of AQI Bucket, the research objective of classifying the AQ was set. The primary objective of this project is to develop the most accurate classification model for AQ.

Data Understanding: The second phase involves better grasping the data. The dataset contains 25,531 instances and 16 attributes, with all data in numeric features except for the AQI Bucket, which is in string format.

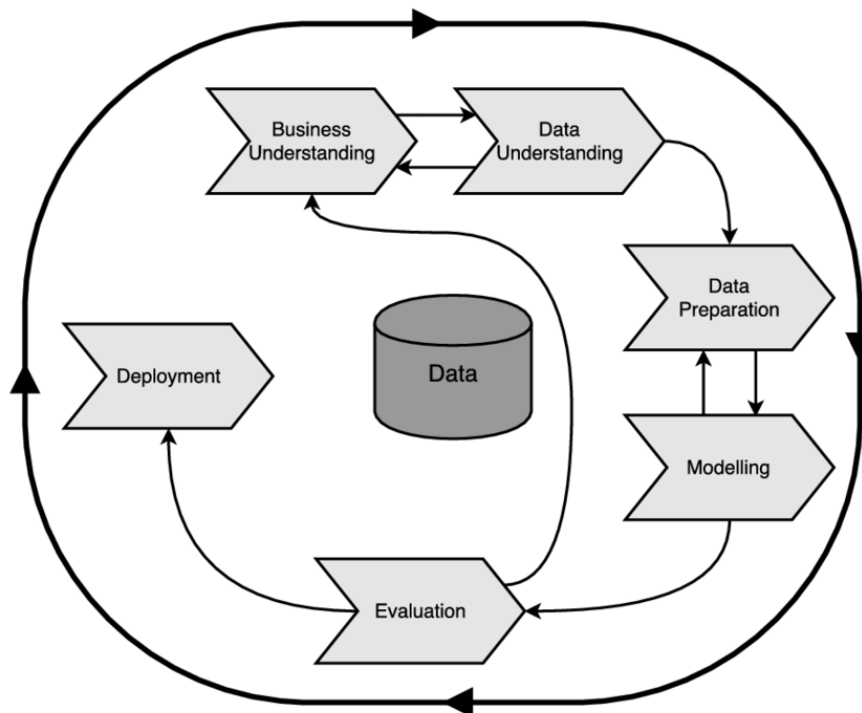


Fig. 1 The CRISP-DM

Data preparation: This stage involves manipulating raw data to come up with the final dataset used in the model. In the next phase, the models to be constructed require the identification and preparation of the variables or attributes in the dataset. A classification model was developed for the AQI Bucket dataset. Data cleaning will occur at this stage to remove the missing data, confirm that the data is of good quality and prevent the inconsistencies.

Modeling: In this step, four (4) classification models have been selected and implemented, which are MDJ, MLR, MDF, and MNN.

Evaluation: This phase is geared towards evaluating the models established during the modeling phase. Evaluation is also needed to determine whether the data models have resolved the problems identified during the business understanding phase, as checking the model's performance. The evaluation metrics used in this project are Accuracy, Precision, and Recall.

Development: This stage is the one when the execution of the DM process applies the created models and requires the outcomes to determine the project's goal and objective.

The project was carried out using the R ML Studio with a 5-fold validation method for training and testing. The dataset was uploaded to the R ML studio and prepared for data preparation. Then, the implementation of method setting (MDJ, MLR, MDF, and MNN) and the classification process to train the model. Lastly, the evaluation of the model was conducted independently. This project is very important to classify the AQ so that interventions can be made to reduce AP and reduce the risk of health problems faced by the people in India.

3.1 Dataset

This project utilized the daily AQ data in all India's cities and it covers the years from 2015 to 2020 [11]. The data set contains 29,531 instances having 16 features which are city, date, Particulate matter 2.5 (PM2.5), Particulate matter 10 (PM10), Nitric Oxide (NO), Nitric Dioxide (NO2), Any Nitric Oxide (NOX), Ammonia (NH3), Carbon Monoxide (CO), The data set was collected from the Central Pollution Control Board (CPCB), an official body of India government [13]. The snippet of the dataset is presented in Table 1.

Table 1. Daily Air Quality across cities in India (2015-2020)

City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI Bucke
Amaravati	1/1/2018	59.03	108.83	3.45	23.36	15.23	18.03	0.09	26.05	87.55	0.23	3.2	0.09	120	Moderate
Amaravati	1/2/2018	61.88	109.82	3.87	23.14	15.45	19.57	0.09	12.81	81.38	0.2	3.84	0.04	132	Moderate
Amaravati	1/3/2018	71.57	137.48	5.01	39.99	25.36	25.61	0.09	16.82	80.3	0.24	6.55	0.11	136	Moderate
Amaravati	1/4/2018	92.44	174.75	5.84	53.79	33.36	19.45	0.11	16.93	71.09	0.3	7.23	0.23	182	Moderate
Amaravati	1/5/2018	92.9	162.69	6.38	56.39	35.19	21.02	0.08	23.11	68.79	0.29	7.88	0.2	222	Poor
Amaravati	1/6/2018	85.15	151.97	11.05	61.55	41.72	18.71	0.07	11.86	90.8	0.26	7.14	0.16	230	Poor
Amaravati	1/7/2018	81.47	145.68	6.45	46.9	30.18	20.18	0.07	9.63	96.49	0.25	7.26	0.15	224	Poor
Amaravati	1/19/2018	64.98	116.14	6.98	45.4	29.83	21.32	0.06	8.83	50.28	0.18	6.79	0.14	111	Moderate
Amaravati	1/20/2018	75.46	132.84	9.56	56.53	37.84	21.88	0.13	30.48	51.98	0.2	6.58	0.14	137	Moderate
Amaravati	1/21/2018	81.58	141.85	4.99	38.15	24.35	19.33	0.05	11.04	59.76	0.16	5.83	0.07	172	Moderate
Amaravati	1/22/2018	78.27	142.86	5.98	45.36	29.01	19.12	0.06	10.32	58.53	0.16	5.57	0.06	160	Moderate
Amaravati	1/23/2018	79.01	138.14	6.16	45.43	29.17	20.29	0.06	10.18	58.04	0.18	5.76	0.07	163	Moderate
Delhi	1/17/2015	196.46	356.2	42.18	53.04	80.18	125.45	8.87	5.13	25.77	9.92	17.59	3.36	360	Very Poor
Delhi	1/18/2015	201.51	359.75	45.51	53.77	75.83	120.45	8.99	5.43	23.22	9	20.95	4.03	370	Very Poor
Delhi	1/19/2015	183.35	305.13	27.32	39.88	55.67	148.45	9.02	4.21	25.34	5.57	12.93	3.62	362	Very Poor
Delhi	1/20/2015	165.63	257.04	16.47	34.23	37.68	146.31	8.03	4.11	22.69	3.42	6.76	3.28	340	Very Poor
Delhi	1/21/2015	159.54	235.27	12.27	22.56	33.02	63.23	9.01	7.05	18.56	4.1	8.26	2.72	338	Very Poor
Delhi	1/22/2015	143.68	183.89	14.75	21.82	34.88	39.65	10.58	8.64	6.94	4.15	9.54	2.78	332	Very Poor

3.2 Data preparation

Before the implementation of the method set, the dataset goes through the data pre-processing to ensure that the data is of good quality and reduces the inconsistencies. The dataset consists of 25,531 instances with 16 features but a high number of missing values, as shown in Table 2. The entire row was removed for the missing values, and after the data cleaning process, 6236 instances were proceeded into the next process.

Table 2. Missing values according to features

Features	Missing Values
Particulate matter 2.5	4598
Particulate matter 10	11140
Nitric Oxide	3582
Nitric Dioxide	3585
Any Nitric Oxide	4185
Ammonia	10328
Carbon Monoxide	2059
Sulphur Dioxide	3854
Ozone	4022
Benzene	5623
Toluene	8041
Xylene	18109
Air Quality Index (AQI)	4681
AQI Bucket	4681

3.3 Algorithms

In this project, the classification algorithms used are Multiclass Decision Jungle (MDJ), Multiclass Logistic Regression (MLR), Multiclass Neural Network (MNN), and Multiclass Decision Forest (MDF). In each section below, there is an explanation for each method [14], [15], [16], [17]:

- **MDJ:** Decision jungles are relatively recent innovations to the family of decision forests. A choice-driven acyclic graph (DAG) collection constitutes a jungle of decisions. Although a decision DAG has more generalization performance owing to its ability to merge trees, the increase in training time makes it have a lower memory footprint when compared to a decision tree. Non-linear decision boundaries can also be represented by nonparametric models called decision jungles. They combine the stages of feature selection and classifier and have a strong immunity to noisy features.

- **MNN:** An ANN consists of connected layers. The inputs form the input layer, and an acyclic weighted graph represents the output layer with nodes and edges. The input, hidden, and output layers can be embedded. A single, or at most, a few

hidden layers can effectively solve most of the predictive tasks. The number of hidden layers is not always large. However, when many hidden layers are utilized in the calculations of MNN, this architecture has been found to be very powerful in complex tasks like image or speech recognition. The layers below are semantic layers of levels getting deeper and deeper. The model learned the input-output relationship from an input data during the training. The architecture is broken down away from the input layer towards the hidden layer and then the output layer. The nodes in one layer are related to nodes on another layer using weighted edges.

- **MLR:** There is a statistical prediction method known as LR that has been widely used to predict whether a given result will occur, and this technique has been found to be very helpful in the case of categorization jobs. The method estimates the probability of an occurrence based on data fitting to a logistic function. A set of possible outcomes in multiclass logistic regression can be predicted using the classifier. Assume that there are three or more classes for the output probabilities in multiclass or multinomial logistic regression.

- **MDF:** Decision forest is an ensemble model that, based on the tagged data, learning creates a series of DT in a relatively short time. The decision forest algorithm is a classification approach that stems from ensemble learning. The approach is to generate many DT and vote for the biggest part of the output class. The tree yields a non-normalized histogram of label frequencies in the classification decision forest. It is a case of aggregation. This aggregation technique combines these histograms and normalizes the resulting outcome to calculate the probabilities for a given label. Higher predict confidence trees are favored in the ensemble final.

3.4 Evaluation Metrics

The evaluation metrics in the project include accuracy, precision, and recall. They include the parameters of TP: True Positive; TN; True Negative; FP: False Positive; FN: False Negative. The description of each evaluation metric is explained as follows [14], [15], [16], [17]:

- **Accuracy.** The accuracy is the true positive to all evaluated by the classifier cases ratio. If the precision of the classifier is acceptable, it is possible to use it for classification of the new data tuples, for which the class label is unknown. The equation of determining accuracy is provided in Eq.1.

$$\text{ACCURACY (A)} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (1)$$

Where: TP: True Positive; TN; True Negative; FP: False Positive; FN: False Negative

- **Precision.** Precision is the proportion of true results and overall positive results. The formula for scheming precision is shown in Eq. 2.

$$\text{PRECISION (P)} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (2)$$

Where: TP: True Positive; FP: False Positive

- **Recall.** Recall is the fraction of all correct results returned by the model. The formula for scheming recall is shown in Eq. 3.

$$\text{RECALL (R)} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (3)$$

Where: TP: True Positive; FN: False Negative

4. Results and Discussion

The applied models begin with data preparation, with the dataset being the daily AQ of cities in India. The datasets were subjected to attribute selection with sixteen (16) features (city, date, Particulate matter 2.5 (PM2.5), Particulate matter 10 (PM10), Nitric Oxide (NO), Nitric Dioxide (NO2), Any Nitric Oxide (NOX), Ammonia (NH3), Carbon Monoxide (CO), After that, the data cleaning process is executed to eliminate all the missing values and to have a quality data set for model testing. The data is then split into training and testing, a 5-fold cross-validation method (30:70, 40:60, 50:50, 60:40, and 70:30). Each algorithm (MDJ, MLR, MDF, and MNN) will conduct the same process independently. The project intends to evaluate the effectiveness of four different algorithms in the AQ dataset is using the evaluation metrics. The results data is presented in Table 3.

Table 3. The classification performance of the four algorithms

Data split (%)	Algorithms	Accuracy	Precision	Recall
30:70	MDJ	0.998625	0.980639	0.988456
	MLR	0.915846	0.793096	0.568439
	MDF	0.998702	0.978026	0.992305
	MNN	0.971669	0.898168	0.930298
40:60	MDJ	0.99902	0.985009	0.992865
	MLR	0.919205	0.7971	0.596062
	MDF	0.999198	0.985595	0.992784
	MNN	0.976127	0.878104	0.931515
50:50	MDJ	0.999252	0.988787	0.996076
	MLR	0.921531	0.782437	0.61893
	MDF	0.999679	0.98913	0.997685
	MNN	0.978405	0.906597	0.949065
60:40	MDJ	0.999599	0.987805	0.997143
	MLR	0.924218	0.823728	0.613711
	MDF	0.999599	0.987805	0.997143
	MNN	0.980754	0.94095	0.909184
70:30	MDJ	0.999466	0.986111	0.996269
	MLR	0.926243	0.818756	0.62198
	MDF	0.999466	0.986111	0.996269
	MNN	0.980581	0.953448	0.931678

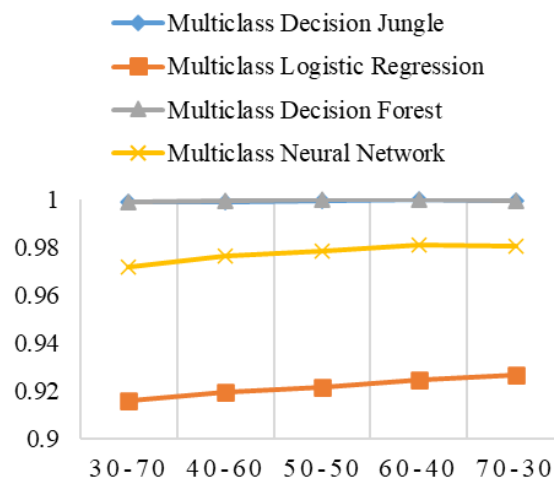


Fig. 2 The variation between the classification accuracy results

The highest accuracy is at 0.999679 by MDF, with data split (training and testing) into 50%. Meanwhile, the lowest accuracy was found

in MLR at 0.915846, with the data split into 30% training and 70% testing as displayed in Fig. 2.

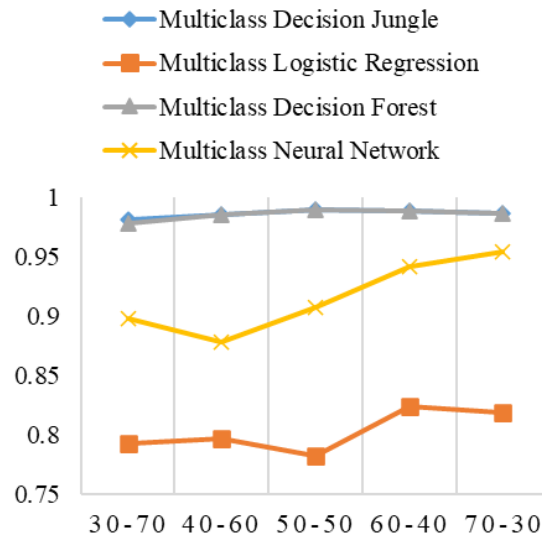


Fig. 3 The variation between the classification precision results

The highest precision is at 0.98913 by MDF, with data split (training and testing) into 50%. Meanwhile, the lowest accuracy was found in

MLR at 0.782437, with the data split into 50% training and 50% testing as displayed in Fig. 3.

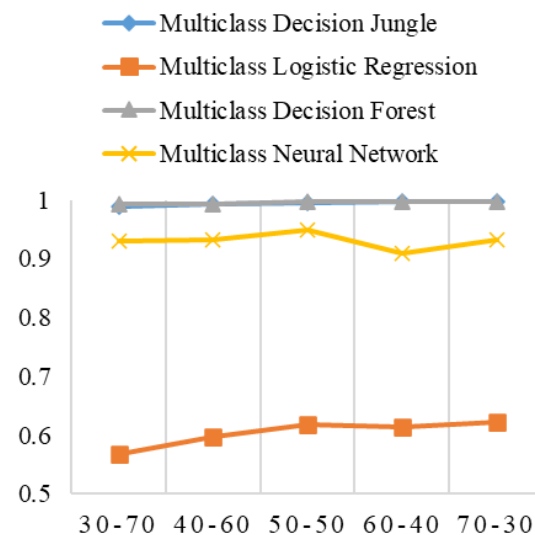


Fig. 4 The variation between the classification recall results

The highest recall is at 0.997685 by MDF, with data split (training and testing) into 50%. Meanwhile, the lowest accuracy was found in MLR at 0.568439, with the data split into 30% training and 70% testing as shown in Fig. 4.

The results showed that MDF, with the data divided into 50% for each of the training and testing, has the highest accuracy, precision, and recall compared to the other three methods (MDJ, MLR,

and MNN). The experiment concluded that the performance of MDF is the best in classifying AQ, considering all evaluation metrics. The lowest performance is MLR, with the lowest accuracy (0.915846), precision (0.782437), and recall (0.568439).

The dataset includes categorical (e.g., AQI Bucket) and continuous values (e.g., the value of NO₂ and SO₂) which are well suited to MDF's

capabilities in capturing non-linear decision boundaries and in efficient training and prediction computation and memory usage. In addition, this approach can also achieve combined feature selection and classification and is resistant to noisy features.

5. Conclusions and Future Work

The forecasting of AQ is crucial for human health and environment protection. Poor AQ, indicated by excessive amounts of pollutants such as particulate matter, nitrogen dioxide, and ozone, poses a substantial danger to health and results in exacerbations of respiratory and cardiovascular diseases and increased mortality. AQ forecasting enables the authorities to put measures in place to prevent people from exposure, such as health advisories, regulation of industrial emissions, and changes in transport policies. What is more, this means that people can plan their outside activities and do them in such a way that they will not be exposed personally to bad air pollutants. In addition, tracking the development of AQ trends enables us to evaluate the efficiency of pollution control measures and creates foundation for policy-making to provide pure and healthy air for nowadays and coming generations. So many findings have been brought in by the AQ data set in India in indicating economic development and the health status of the people in India. This project deploys four ML algorithms, which are MDJ, MNN, MLR, and MDF to perform classification tasks. The highest accuracy is at 0.999679 by MDF, with data split into 50% for each of the training and testing.

The highest precision is at 0.98913, and the highest recall is at 0.997685, which the MDF made. Meanwhile, the lowest accuracy was found in MLR at 0.915846, with the data split into 30% training and 70% testing. In future work, it is recommended to have more reliable data with less missing value to make the data scalable, more precise, and of higher quality. It is hoped that this model can be used to help other countries and cities classify the AQ to monitor good AQ, and interventions can be taken into consideration to prevent health problems.

References

- [1] Aditya, C. R., Deshmukh, C., K, Nayana, K., & Gandhi, P., & Vidyav. (2018). Detection and Prediction of Air Pollution using Machine Learning Models. *International Journal of Engineering Trends and Technology*. 59. 204-207.
- [2] Beig, G., Sahu, S. K., Singh, V., Tikle, S., Sobhana, S. B., Gargeva, P., Ramakrishna, K., Rathod, A., Murthy, B.S., (2019). Objective evaluation of stubble emission of North India and quantifying its impact on air quality of Delhi. *Science of the Total Environment*, 136126.
- [3] Bellinger, C., Jabbar, M. S. M., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC public health*, 17(1), 1-19.
- [4] Gordon, T., Balakrishnan, K., Dey, S., Rajagopalan, S., Thornburg, J., Thurston, G., Agrawal, A., Collman, G., Guleria, R., Limaye, S., Salvi, S., Kilaru, Va., Nadadur, S., (2018). Air pollution health research priorities for India: Perspectives of the Indo-U.S. Communities of Researchers. *Environment International*, 119, 100–108.
- [5] Gurjar, B. R., & Ajay S. N., (2015). "Indian megacities as localities of environmental vulnerability from air quality perspective." *Journal of Smart Cities* 1.1
- [6] Harishkumar, K S. (2018). A Survey of Air Pollution Studies Using Data Mining Techniques in Smart City.
- [7] Hilboll, A., Richter, A., & Burrows, J. P. (2017). NO₂ pollution over India observed from space—the impact of rapid economic growth, and a recent decline. *Atmospheric Chemistry and Physics Discussions*, 1-18.
- [8] Jagtap, I., & Babbar, N. (2021). Predicting Air Pollutant using Data Mining and Machine Learning Algorithms.
- [9] Kumar, A. & Goyal, P., (2011). Forecasting of air quality in Delhi using principal component regression technique. *Atmospheric Pollution Research*, 2(4), 436–444.
- [10] Misra, P., & Takeuchi, W. (2015). Assessing Impact of Economic Activities on Urban Air Quality in India by Nightlight and Atmospheric Measurement Datasets. In *37th Asian Conf. Remote Sens.*
- [11] Sharma, R., Kumar, R., Sharma, D., Kumar, S., Le Hoang, Priyadarshini, I., Pham, B., Thai, T.

- B., Dieu, R., Sakshi, (2019). Inferring air pollution from air quality index by different geographical areas: case study in India. *Air Quality, Atmosphere & Health*.
- [12] Shashi, B. & Sanjay, T., (2020). AQI Predication Using Temporal Data Mining.
- [13] Taneja, S., Sharma, N., Oberoi, K., & Navoria, Y. (2016). "Predicting trends in air pollution in Delhi using data mining." *Information Processing (II CIP), 2016 1st India International Conference on. IEEE, 2016*
- [14] Aram, S. A., Nketiah, E. A., Saalidong, B. M., Wang, H., Afitiri, A. R., Akoto, A. B., & Lartey, P. O. (2024). Machine learning-based prediction of air quality index and air quality grade: A comparative analysis. *International Journal of Environmental Science and Technology, 21(2)*, 1345-1360.
- [15] Dhanalakshmi, M., & Radha, V. (2022). Discretized Linear Regression and Multiclass Support Vector Based Air Pollution Forecasting Technique. *International Journal of Engineering Trends and Technology, 70(11)*, 315-323.
- [16] Jayaraj, M. (2020, July). Air quality monitoring and disease prediction using IoT and machine learning. In *International Conference on Internet of Things and Connected Technologies* (pp. 18-32). Cham: Springer International Publishing.
- [17] Sani, S. H., Shopon, M., & Rakib, S. H. (2021, April). Air Quality Index Prediction Using Azure IoT & Machine Learning for Smart Cities. In *Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2020* (pp. 721-733). Singapore: Springer Singapore.