MACHINE LEARNING TECHNIQUES FOR AIR QUALITY PREDICTION

Rajesh Bhaskar Pawar

Keraleeya Samajam's Model College, Khambalpada Road, Thakurli, Dombivli(East), Kanchangaon Maharashtra.

Abstract:

Air pollution poses a major threat to the society. Monitoring the quality of air is the key to assure public health. Naïve and empirical methods for predicting air pollution is not accurate. This work proposes to develop an air pollution prediction system that can predict the quality of air using machine learning techniques. A review is done to study the different machine learning approaches conducted by researchers. An architecture is developed which takes in input data and follows different stages like pre-processing, learning and evaluation. A model can be developed using the training data. The model is then tested using the test data. Air pollution can then be predicted on the test data to know how accurately the model predicts. Various techniques like Logistic Regression, Random Forest, Support Vector Machine and Neural Networks are discussed for constructing the model to be used for prediction.

Keywords:

Machine learning, Support Vector Machine, Neural Networks, Logistic Regression

1 INTRODUCTION

It is observed that 33% of Indian population lives in cities [12]. Urbanization in cities have led to degradation of the atmosphere and environment leading to poor health introduced by noise pollution, water pollution, air pollution and several other problems related to waste disposal etc. One of the major threats is air pollution as every human breathes air. Air pollution is the result of introduction of harmful and toxic gases into the atmosphere. Inhaling the polluted air leads to premature death, high blood pressure, heart problems, malnutrition etc. At least 140 million people breathe air more over the WHO safe limit with India as one among the cities to record highest annual level of air pollution [12]. Not only to human beings, effects of air pollution affects the environment leading to acid rain, change in climatic condition, damaged vegetation, acid rain etc. Studies conducted by WHO reveal that 9 out of 10 people in the world breathe polluted air [13].

Automobiles are one of the major contributors of air pollution. From 79073 number of registered vehicles in 2009, it's reached to approximately 2 lakhs in 2018 [14]. Hydrocarbons formed during partial burning of the fuel, like Carbon monoxide (CO) released due to incomplete combustion, Nitrogen Oxides (NOx) produced as a result of the reaction between nitrogen in the air and oxygen inside the engine at high temperature, particulate matter PM10 and PM2.5 which are particles of very small size, sulfur oxide (SOx) formed by burning fuel are the major outcomes of pollution by automobiles [15].

Air Quality Index (AQI) is a term used to communicate how polluted the air currently is or how polluted it can become. Corresponding to different national air quality standards, different countries or regions have set their own pollutants to consider. India calculates its AQI based on eight pollutants namely PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb. The air quality is then categorized as Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe. Consider the Figure 1.1 for detailed description.

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
When the AQI is in this range:	*..air quality conditions are:*	*...as symbolized by this color:*
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Fig 1.1: AQI values and its health concerns [16]

Furthermore, the AQI chart of India is represented in the Figure 1.2 which indicates that several states of India are heavily polluted.

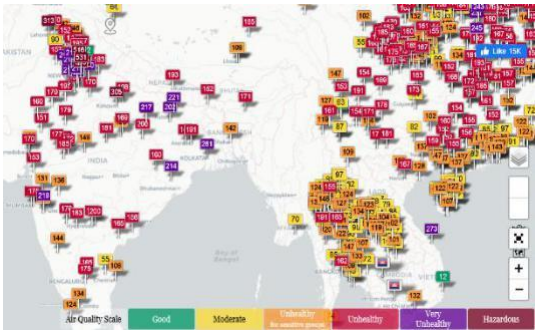

Fig 1.2: AQI chart of India [16]

This work proposes to develop a system that can monitor the pollutants of a particular region and predict if the air is polluted using machine learning techniques.

A. *Machine Learning*

Machine learning is the subset or application of Artificial Intelligence (AI) that enables computers with the capability to perform a specific task automatically without explicitly programmed into it. It is an emerging and exciting technology used in real time analysis and prediction. Like humans it gives the computers the ability to learn. It relies on patterns and inferences to perform a task. It accesses data and uses it to learn for themselves and helps accurately predict future based on the examples provided. This allows computers to learn automatically without intervention by humans. Machine learning uses algorithms that can take in input data and uses statistical analysis to predict the output. The algorithms in machine learning builds a mathematical model based on the "training data" to enable predictions without explicitly programmed. The machine learns from the data provided to it, and its performance improves using its past experiences. Machine Learning are classified as supervised learning and unsupervised learning.

Supervised Learning: In supervised learning, input and output is provided to the computer along with the training data. The model is trained on the data set which is labeled (input and output). The accuracy of predicted data is analyzed during training. This makes the system to learn and map input to its output. It is used to develop predictive model based on both input and output data. Categories of supervised leaning are Classification and Regression. Classification is the way in which the algorithm classifies the input data into one of several predefined classes. It is used to predict categorical results that fit in the predefined labels. Regression predicts a continuous value as output using the training data. Various algorithms used in supervised learning are support vector machine, neural network, linear regression, logistic regression, random forest and decision trees.

Unsupervised Learning: In unsupervised learning, the machine works on its own with unlabeled data to discover information without any guidance. Here the machine groups the data based on similarities and differences between the elements without any training. Categories of unsupervised learning are clustering and association. In clustering it finds a structure or a pattern in a collection of uncategorized data. Association rules helps to establish associations amongst data objects in databases. Various algorithms used in unsupervised learning are K means clustering, Hidden Markov Models, neural networks, fuzzy c-means etc.

B. *Importance of Machine Learning Techniques for Air Pollution Prediction*

Air Pollution prediction mechanism using machine learning techniques is a useful approach to better predict the quality of air accurately than naive methods. Machine learning and statistical algorithm helps to train a model using collected or historical data. It is very important to know the quality of air we breathe. As air is everywhere, the system can be used to calculate the quality of air at any region. The system built using historical data is called the training data model. The model is then tested with the test data. In this paper; Section 2 presents related work done by researchers. Section 3 describes the proposed system. Section 4 concludes the proposed work.

2 LITERATURE SURVEY

In this chapter, a brief review of the implementation and study work done by several researchers are presented along with the pollutant parameters and machine learning techniques used by them. Prediction of air pollution is studied and implemented by several researchers. Fuzzy logic approaches are proposed and implemented by several researchers to predict air pollution but these are created with limited data. Empirical models are also proposed and implemented by several researchers but since the models are more

dependent on experience than logic it may result in poor performance compared to other models. Simulation models are also widely explored but are based on theories from chemistry and physics but performance may vary based on the scale at which they are simulated. Machine learning approaches have not been fully explored but are widely used to analyse and predict based on statistics, probability and mathematical approaches. Hamed Karimaian et al. [1] in their work has implemented multiple additive regression trees (MART), deep feedforward neural network (DFNN) and a new hybrid model based on long short-term memory (LSTM) to predict the concentration of PM2.5. As per the designed model LSTM technique achieved the best results compared to others. Yi-Ting Tsai et al.
[2] in their work has implemented an approach to forecast PM2.5 concentration using RNN (Recurrent Neural Network) with LSTM (Long Short- Term Memory). Their result showed that the technique can effectively forecast the value of PM2.5. Ziyue Guan et al. [3] in their work has implemented a prediction model using various machine learning techniques like Linear Regression, Artificial Neural Network and Long-Term Short-Term Memory Recurrent Neural network. They have calculated the accuracy of the models using these techniques to predict the pollutant PM2.5. Aditya CR et al.
[4] in their work has predicted the pollutant PM 2.5 to determine the quality of air by constantly keeping a check on the its level in the atmosphere. Logistic regression is employed to detect if the air is polluted or not. Auto regression is employed to predict future PM2.5 values based on previous PM2.5 values. Yong Tian et al. [5] proposed a new framework to evaluate the quality of air in airports. The framework is composed of a combination of the standard assessment procedure and machine learning methods to predict AQI (Air Quality Index). Ghaemi Z et al. [6] has implemented a system to predict the AQI using spatio-temporal LaSVM based online algorithm. The system is evaluated by comparing the predicted values with traditional SVM to show that LaSVM showed outstanding results in prediction than traditional SVM. Rubal, Dinesh Kumar [7] in their work has implemented an air pollution prediction system using Random Forest method. This method is compared with existing methods where Bayesian network and multi-label classifier are used for the estimation. AQI is predicted with more accurate values and provide precise forecasting information. Cole Brokamp et al. [8] in their work had trained a model to accurately predict PM2.5 using random forest method in an urban area. They have used satellite, meteorologic, atmospheric, and land-use data to predict PM 2.5. W.C. Leong et al. [9] has implemented an air pollution model to predict API (Air Pollution Index) using Support vector machines. There are several parameters affecting the performance of the support vector machine model: penalty factor (C), regularization parameter (ε) and the type of kernel function used. The author has used only kernel functions model parameters. Jasleen Kaur Sethi and Mamta Mittal
[10] has worked on air pollution prediction technique using Decision trees from classification and Support Vector Regression from regression for prediction Air Quality Index. They found that from all supervised learning techniques, Decision tree and Support Vector Machines are more effective in air pollution predicition. Milica Arsić et al. [11] has predicted Ozone concentration in the vicinity of the city of Zrenjanin (Serbia). Multiple linear regression analysis (MLRA) and artificial neural networks (ANNs) were used to detect the ozone concentration. Their model revealed that Artificial Neural Network are more efficient than Multilinear regression to predict ozone concentration.

Summarizing the techniques used by various researchers to predict the air pollution it can be inferred that pollutants vary based on the location. Major pollutant that contribute to air pollution is PM2.5. Other pollutants include CO2, NO2, temperature, humidity, smog etc. It can also be inferred that Support Vector Machine, Random Forest and Neural networks are more preferred to predict air pollution over other techniques for better accuracy in the prediction model. Based on the analysis in the literature review, in the proposed system, a data set is collected and the above three: Support Vector Machine, Random Forest and Neural Networks are planned to be implemented and tested, to check using which model the accuracy is the best.

Table 2.1 Summary of literature review

Author	Monitored air pollutants	Machine learning Techniques
Hamed Karimaian et al. 2018 [1]	• PM2.5 • T(Temperature) • P(Surface level pressure) • W(Wind) • RH(Relative Humidity)	• Multiple additive regression trees (MART) • Deep feedforward neural network (DFNN) • Hybrid model Long short-term memory (LSTM)
Yi-Ting Tsai et al. IEEE. 2018 [2]	• SO2(sulphur dioxide), CO(Carbon monoxide), O3(Ozone level), PM10,PM2.5 • NOX, NO, NO2 • Temperature, Rainfall • Humidity, Wind speed	• RNN (Recurrent Neural Network) • LSTM (Long Short- Term Memory).
Ziyue Guan et al. IEEE 2018 [3]	• Latitude and Longitude • Wind, Rainfall, Temperature • PM2.5, Humidity • Sound Traffic	• Linear Regression • Artificial Neural Network • LSTM (Long Short- Term Memory)
Aditya C R et al. IJETT 2018 [4]	• Temperature, Wind, Dewpoint • Pressure, PM2.5	• Logistic Regression
Z. Ghaemi et al. 2018 [6]	• SO2 (sulphur dioxide) • CO(Carbon monoxide) • NO2 (Nitrogen dioxide) • PM10, O3(Ozone level)	• LaSVM (Spatio-Temporal Support Vector Machines)
Rubal et al. ICCIDS 2018 [7]	• SO2 (sulphur dioxide) • CO(Carbon mon,oxide) • NO2 (Nitrogen dioxide) • PM2.5, PM10, O3(Ozone level), C6H6(Benzene)	• Random Forest
Cole Brokamp et al. 2018 [8]	• Aerosols, PM2.5, Location • Weather, Temperature • Location Greenspace	• Random Forest
W.C. Leong et al. 2018 [9]	• SO2 (sulphur dioxide) • CO (Carbon monoxide) • NO2 (Nitrogen dioxide) • PM10, O3(Ozone level)	• Support Vector Machines
Yong Tian et al. 2019 [5]	• NOx (Nitrogen oxides) • CO(Carbon monoxide) • CO2(Carbon dioxide) • SO2 (sulphur dioxide)	• Support Vector machine • Random Forest • Logistic Regression
Jasleen Kaur Sethi et al. 2019 [10]	• SO2 (sulphur dioxide) • CO(Carbon monoxide) • NO2 (Nitrogen dioxide) • PM2.5, O3(Ozone level)	• Support Vector Machines • Decision Trees
Milica Arsić et al. 2019 [11]	• SO2(Sulphur Dioxide) • CO(Carbon Monoxide) • H2S(Hydrogen sulphide) • NO, NO2, NOx(Nitrogen oxides), PM10, Benzene, ethylbenzene • Toluene, m,p-Xylene, o-Xylene	• Multilinear Regression • Artificial Neural network

Table 2.2 Summary of Machine Learning technique used.

Paper	Machine learning techniques						
	Linear Regression	MultiLinear Regression	Logistic regression	SVM	Neural Networks	Random Forest	Decision Trees
Jasleen Kaur Sethi et al. 2019 [10]				✓			✓
Milica Arsić et al. 2019 [11]		✓	✓		✓		
Hamed Karimaian et al. 2018 [1]					✓		✓
Yi-Ting Tsai et al. IEEE. 2018 [2]					✓		
Ziyue Guan et al. IEEE 2018 [3]	✓				✓		
Aditya C R et al. IJETT 2018 [4]			✓				
Z. Ghaemi et al. 2018 [6]				✓			
Rubal et al. ICCIDS 2018 [7]						✓	
Cole Brokamp et al. 2018 [8]						✓	
Yong Tian et al. 2019 [5]			✓	✓		✓	
W.C. Leong et al. 2018 [9]					✓		

Table 2.1 gives a summary of the literature review done by several authors in machine learning for air pollution prediction. Table 2.2 gives a summary of the techniques used by various researchers to predict the air pollution. From table 2.1 it can be inferred that pollutants vary based on the location. Major pollutant that contribute to air pollution is PM2.5. Other pollutants include CO2, NO2, temperature, humidity, smog etc. From Table 2.2 it can be inferred that Support Vector Machine, Random Forest and Neural networks are more preferred to predict air pollution over other techniques for better accuracy in the prediction model.

Based on the analysis in the literature review, in the proposed system, a data set is collected and the above four: Logistic Regression, Support Vector Machine, Random Forest and Neural Networks are planned to be implemented and tested, to check using which model the accuracy is the best.

3. PROPOSED SYSTEM

Data collected from different data sets are subject to errors. This can be due to duplicate data, misspelled data, typographical errors etc. In the proposed system the input data or raw data are collected and labelled. The input data undergoes different pre-processing stages namely feature extraction and scaling, feature selection, dimensionality reduction and sampling. This processed data will be given to next phase for learning algorithms. In this phase, different algorithms are used to learn the data and model selection, cross validation, performance metrics and optimization is done. Based on the selected algorithm the model is built on the test data. Once the model is finalized, the model is tested on test data, accuracy is analysed and finally the air pollution is predicted. The complete proposed architecture is represented in Figure 3.1.

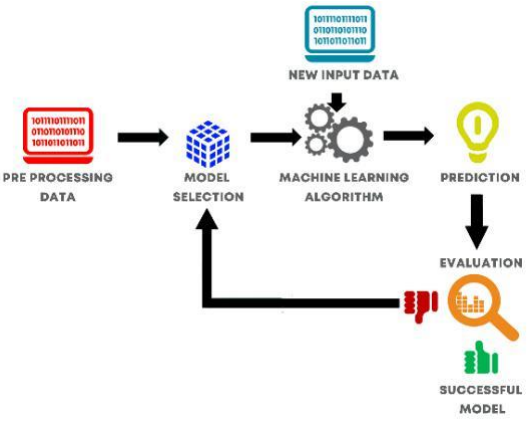

Fig 3.1: Architecture of the proposed system

A. Pre-processing data

Pre-processing is an important step which cannot be discarded. Raw data collected are susceptible to missing values, noisy data, incomplete data, inconsistent data and outlier data. So, it is important for these data to be processed. In pre-processing the collected dataset is divided into two parts, one for testing and other for training purpose. The dataset is labelled. The pre-processing stage involves feature extraction and scaling, feature selection, dimensionality reduction and sampling. Pre-processing helps in cleaning data to be made fit for further processing.

Feature extraction: In air pollution prediction system, feature extraction is the process of transforming the raw data from the data set to a more meaningful and useful information that can be further processed on. It helps to create a new, smaller set of the original data set that captures the most useful and meaningful information.

Feature selection: Feature selection is used to filter irrelevant or redundant features from the air pollution dataset. Not all columns contain useful data for predicting air pollution, hence a subset of the original dataset is considered.

Dimensionality reduction: Dimensions refer to the number of features (i.e. input parameters) in your dataset. Mostly all features of the original dataset are not required for air pollution prediction. More dimensions can lead to over fitting or under fitting of the models. Dimensionality reduction is done with the help of feature extraction and feature scaling to reduce redundant and irrelevant features in the dataset.

Sampling: Sampling of the data is done to select a subset or divide the entire dataset into different subsets. Datasets of air pollution is divided as training data and test data. Maximum of the data in the dataset is considered as training data. Model is built using the training data. The accuracy of the model is later tested using the test data.

B. Model Selection

In this step a model can be built using different machine learning techniques. Several steps of learning are model selection, cross validation, performance metrics and hyper parameter optimization. Model Selection techniques are discussed based on the inferences drawn from the literature review (Section 2).

Logistic Regression:

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. It predicts the output of a categorical dependent variable. The outcome must be a categorical or discrete value like Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic regression shown in Figure 3.2 transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. In the proposed system, to predict air pollution, binary logistic regression or multi logistic regression can be used. In order to map predicted values to probabilities, sigmoid function is used. The function maps any real value into another value between 0 and 1.

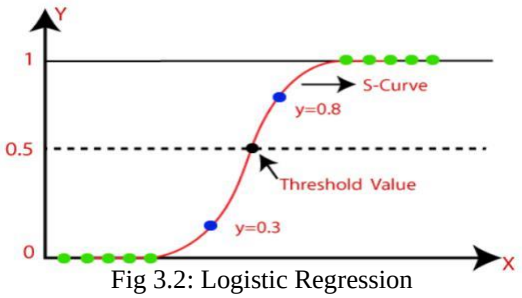

Fig 3.2: Logistic Regression

$S(z) = 1/(1+ e^{-z})$ ……(3.1)

The sigmoid function (logistic function) is,

- s(z) = output between 0 and 1 (probability estimate)
- z = input to the function (algorithm's prediction e.g. mx + b)
- e = base of natural log

The Logistic regression equation can be derived from Linear Regression equation. Mathematical representation of a straight line is given in equation 3.2.

$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots\ldots\ldots + b_nx_n$ …….(3.2)

In Logistic Regression y can have only values between 0 and 1 only. Hence dividing the equation 3.2 by (1-y) as we get the value as shown:

$y/(1-y)$; 0 for y=0, and ∞ for y=1 …….(3.3)

Taking log of equation 3.3 it gives

$\log[y/(1-y)] = b_0 + b_1x_1 + b_2x_2 +$

$b_3x_3+\ldots\ldots\ldots\ldots+ b_nx_n$ ……..(3.4)

Equation 3.4 is the final equation for Logistic Regression.

Random Forest: Random Forest is another standard machine learning method implemented using decision trees. It follows a divide-and-conquer approach to improve the performance of the system. In the tree the data input is entered at the top and the data traverses down the tree and gets confined to smaller sets. A random forest is a classifier consisting of a collection of tree-structured classifiers:

$h(x, \Theta k), k=1,2,\ldots..$ …….(3.5)

Where in equation (3.5), {Θk} are identically distributed independent random vectors and each tree casts a unit vote for the popular class at input x [17]. Random Forests sometimes have high accuracy prediction and can handle more features due to the embedded feature selection in the model generation process. When the number of features is more, it is better to use a higher number of regression trees. Random Forests are sufficiently robust to noisy data, but the biological inter-predictability of Random Forests is limited. Depending upon the features selected in the pre-processing phase several decision trees can be formed to get a better prediction of the air pollution data.

Support Vector Machine: Support Vector Machines exist in different forms, linear and non-linear. A support vector machine is a supervised classifier. In the usual context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable as shown in Figure 3.3. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset perfectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line".

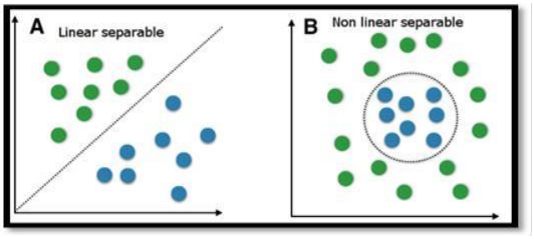

Fig 3.3: Support Vector Machine

The best line is found by maximizing the distance to the nearest points of both classes in the training set. The maximization of this distance can be converted to an equivalent minimization problem, which is easier to solve. The data points on the maximal margin lines are called the support vectors. Most often datasets are not nicely distributed such that the classes can be separated by a line or higher order function. Real datasets contain random errors or noise which creates a less clean dataset. Although it is possible to create a model that perfectly separates the data, it is not desirable, because such models are over-fitting on the training data. Over fitting is caused by incorporating the random errors or noise in the model. Therefore, the model is not generic, and makes significantly more errors on other datasets. Creating simpler models keeps the model from over-fitting. The complexity of the model has to be balanced between fitting on the training data and being generic. This can be achieved by allowing models which can make errors. SVM can make some errors to avoid over-fitting. It tries to minimize the number of errors that will be made. Support vector machines classifiers are applied in many applications. They are very popular in recent research. This popularity is due to the good overall empirical performance. Comparing the naive Bayes and the SVM classifier, the SVM has been applied the most

Neural Networks: These are used to model or simulate the distribution, functions or mappings among variables as modules of a dynamic system associated with a learning rule or a learning algorithm. The modules here simulate neurons in nervous system and hence Neural Networks represented in Figure 3.4 collectively refers to the neuron simulators and their synapses simulating interconnections between these modules in different layers. The defining aspect of a Neural Network is the function implemented at each neuron and the learning algorithm for the dynamic weights assigned to the interconnections among neurons. What makes Neural Network stand apart is its ability to simulate human thought process coupled with continuous learning, growth and evolution. It is also capable of handling large number of parameters and large set of data with noise and yet achieves high accuracy.

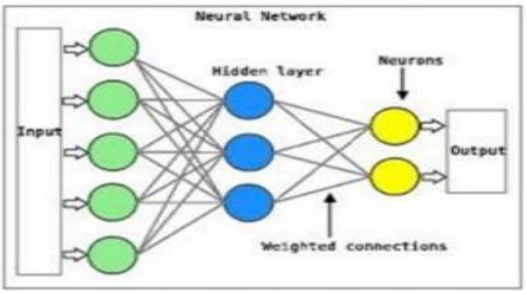

Fig 3.4: A three-layer neural network perception [18]

Neural Network is built by stacking together multiple neurons in layers to produce a final output as shown in Figure 3.4. First layer is the input layer and the last is the output layer. All the layers in between are called hidden layers. Each neuron has an activation function. Some of the popular Activation functions are Sigmoid, ReLU, tanh etc. The parameters of the network are the

weights and biases of each layer. The goal of the neural network is to learn the network parameters such that the predicted outcome is the same as the ground truth. Back-propagation along loss-function is used to learn the network parameters.

C. Prediction

On working with real-world data of air pollution prediction, we have to define our own labels to train the model based on supervised learning model. The test data is inputted in the model and tested if it makes the correct predictions with our sample data. Correct predictions indicate the data is outputted in the correct label.

D. Evaluation

Accuracy in air pollution prediction is number of correct predictions made by the model over all kind's predictions made with respect to the confusion matrix. Accuracy is a good measure when the target variable classes in the data are nearly balanced. Model is evaluated and the accuracy of the system is determined on the test data. Accuracy is the proportion of correctly classified profiles or instance. To evaluate the proposed system, following metrics are used: precision, recall, F1 score and support. For air pollution prediction system, confusion matrix is used to check the performance of the classification model on a set of data values for which true values are known. Figure 3.4 gives the confusion matrix with classification metrics.

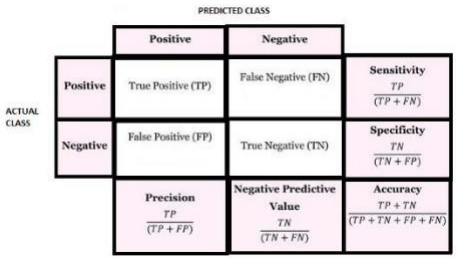

Fig 3.4: Confusion matrix with classification metrics

True positives are the cases when the actual class of the data set was True and the predicted is also True. True negatives are the cases when the actual class of the data set was False and the predicted is also False. False positives are the cases when the actual class of the data set was False and the predicted is True. False is because the model has predicted incorrectly and positive because the class predicted was a positive one. False negatives are the cases when the actual class of the data set was True and the predicted is False. False is because the model has predicted incorrectly and negative because the class predicted was a negative one. Specificity is defined as True Negative rate or Recall. It's the measure of negative sets labeled as negative by the model. The value of specificity should be higher. Precision gives the correctness of the positive predictions. The precision is the ratio of correctly predicted positive results to the total predicted positive results as given in equation 3.6.

Precision = True Positive / (True Positive + False Positive) ….(3.6)

Accuracy is the proportion of the total predictions that are correct. Recall (Sensitivity) is the ratio of correctly predicted positive results to all the results in the actual class i.e True Positives and False Negatives as given in equation 3.7.

Recall = True Positive/ (True Positive + False Negative) ….(3.7)

F-1 score can be considered as a weighted harmonic mean of the precision and recall as given in equation 3.8.

F1 Score = 2*(Recall * Precision) / (Recall + Precision) ...…….(3.8)

The support is the number of occurrences of each class in test data. It's the number of true responses that lie in that class.

4. CONCLUSION

Air pollution prediction systems can be used to predict the quality of air we breathe. The proposed system aims to build a model using machine learning technique. In most of the prediction systems, researchers have implemented various machine learning techniques like logistic regression, linear regression, SVM, Random Forest, Decision Tree. Supervised classification algorithms are more preferred for prediction to gain better accuracy in the prediction. The proposed work aims to develop models based on different machine learning techniques to get maximum accuracy in predicting the level of PM2.5 in the atmosphere, which is an important index to determine the AQI.

REFERENCES

[1] Hamed Karimian, Qi Li, Chunlin Wu, Yanlin Qi, Yuqin Mo, Gong Chen, Xianfeng Zhang, Sonali Sachdeva, in "Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations" in Aerosol and Air Quality Research, 19: 1400–1410, 2019 ISSN: 1680-8584 print / 2071-1409 online doi: 10.4209/aaqr.2018.12.0450.

[2] Yi-Ting Tsai, Yu-Ren,Zeng, Yue-Shan Chang in "Air pollution forecasting using RNN with LSTM" 2018 IEEE 16th Int. Conf. on Dependable, Autonomic & Secure Comp., 16th Int. Conf. on Pervasive Intelligence &Comp., 4th Int. Conf. on Big Data Intelligence & Comp., and 3rd Cyber Sci. & Tech. Cong.

[3] Ziyue Guan, Richard O. Sinnott, in "Prediction of Air Pollution through Machine

Learning Approaches on the cloud" 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT).

[4] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, "Detection and Prediction of Air Pollution using Machine Learning Models" in International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018.

[5] Yong Tian, Weifang Huang, Bojia Ye andMinhao Yang, Hindawi, "A New Air Quality Prediction Framework for Airports Developed with a Hybrid Supervised Learning Method"
.

[6] Z. Ghaemi & A. Alimohammadi & M. Farnaghi Environ Monit Assess, "LaSVM-based big data learning system for dynamic prediction of air pollution in Tehran" (2018) 190: 300 https://doi.org/10.1007/s10661-018-6659-6.

[7] Rubal, Dinesh Kumar "Evolving Differential evolution method with random forest for prediction of Air Pollution" in International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

[8] Cole Brokamp, Roman Jandarov, Monir Hossain, and Patrick Ryan, "Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model" article in Environmental Science and Technology.

[9] W.C. Leong, R.O. Kelani, Z., "Prediction of air pollution index (API) using support vector machine (SVM)" Ahmad Journal of Environmental Chemical Engineering.

[10] Jasleen Kaur Sethi, Mamta Mittal, "Ambient Air Quality Estimation using Supervised Learning Techniques" in EAI Endorsed Transactions on Scalable Information Systems article.

[11] Milica Arsić, Ivan Mihajlović, Djordje Nikolić, Živan Živković & Marija Panić, "Prediction of Ozone Concentration in Ambient Air Using Multilinear Regression and the Artificial Neural Networks Methods" ISSN: 0191-9512 (Print) 1547-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/bose20.

[12] Indian Population information available-http://worldpopulationreview.com/countries/indiapopulation/cities/ [Online].

[13] Pollution in Indian cities available: https://www.bbc.com/news/world-asia-india43972155 [Online].

[14] Ministry of Statistics and Program implementation http://mospi.nic.in/statisticalyear-book-india/2018/189 [Online].

[15] Pollution due to automobiles available: http://www.yourarticlelibrary.com/environment/automobile-pollution-sources-effects-andcontrol-of-automobile-pollution/9984 [Online]

[16] https://airnow.gov/index.cfm?action=aqibasics.aqi [Online]

[17] Atherosclerotic Plaque Characterization Methods Based on Coronary Imaging, 2017 [Online].

[18] Air Pollution – Monitoring, Modelling and Health, Edited by Mukesh Khare p. cm. ISBN 978-953-51-0424-7, InTech Janeza Trdine 9, 51000 Rijeka, Croatia [Online]