OPTIMIZATION OF HYPER PARAMETERS IN MACHINE LEARNING TECHNIQUES FOR AIR QUALITY PREDICTIVE ANALYSIS

Basamma Umesh Patil1, Ashoka D. V.2, Ajay Prakash B. V.3*

[1] Department of CSE, SJB Institute of Technology and Research Scholar, CSE Research Center, JSS Academy of Technical Education (Affiliated to VTU), Bengaluru, Karnataka
[2] Department of ISE, JSS Academy of Technical Education (Affiliated to VTU), Bengaluru, Karnataka
[3] Department of CSE, SJB Institute of Technology (Affiliated to VTU), Bengaluru, Karnataka,
India

* Corresponding Author: e-mail: bu.patil25@gmail.com

Abstract: To reduce health related problems due to air pollution, there is a need of effective air quality prediction. In this regard, enhanced AQI (Air Quality Index) prediction machine learning models are proposed. Datasets from different domains like air pollution concentrations and meteorological data are collected and integrated. Machine Learning models such as k-Nearest Neighbors, XGBoost, Support Vector Machine and Decision Tree models have been effectively applied. Optimization of hyper parameters for various machine learning models has been carried out. From obtained results, it is observed that XGBoost gives better results compared to other models with least error rate of 1.6.

Key words: machine learning, meteorological data, pollutant concentration, air quality index, data integration, hyper parameter, air quality prediction.

1. INTRODUCTION

Nowadays, the people have given more focus on health issues due to the increase in air pollution, as it is the main cause of many diseases and deaths in the world. Air pollution can be defined as the pollutants contamination into the air which is highly injurious to human health and the whole planet. It is the leading ecological cause of diseases and premature deaths in the globe nowadays. In 2015, nine million premature deaths were recorded because of the pollution related

diseases [1]. Seven million people are killed every year in the universe by the diseases related to air pollution. From the world health organization statistics, it is revealed that maximum people inhale polluted air [2]. The main causes of air pollution are exhaust from vehicles, industries and factories, power plants, mining and agricultural activities, the burning of fossil fuels, wildfires and volcanoes. Air pollution appears on account of the release of dangerous suspended particulates like carbon monoxide (CO), sulphur dioxide (SO2), particulate matter10 (PM10), particulate matter 2.5 (PM2.5), ozone (O3) in to the atmosphere. The purpose of predicting air quality is to estimate the air pollution level by considering meteorological data and pollutant concentrations data. AQI is the index value used to define the contamination level of the air and to inform people about the air pollution level. To calculate AQI value, the following formula is used.

$$\mathbf{AQI} = \frac{\mathbf{A_{high} - A_{low}}}{\mathbf{P_{high} - P_{low}}}(\mathbf{P - P_{low}}) + \mathbf{A_{low}} \quad (1)$$

Where: P= pollutant concentration; P_{low}= breakpoint of pollutant concentration which is <= P; P_{high}= breakpoint of pollutant concentration which is >= P; A_{low}= breakpoint of index corresponding to P_{low}; A_{high}= breakpoint of index corresponding to P_{high}

In the proposed work, integration of multiple datasets such as air pollutant concentration dataset and meteorological dataset has been done to predict the AQI. Due to variety of data from various sources in various domains, integration of data has become a big challenge. Because these data sets incorporate different modalities, diverse representation, scale, density and distribution. Unlocking the power of knowledge from distinct datasets is the important task in the big data research, which differentiated data mining and big data operations [4].

After studying the prior research outcomes, it has been observed that existing techniques have been employing some artificial neural networks, extreme learning machines, deep learning, natural language processing and spatial interpolation models. However these methods lacked the complete and efficient utilization of the existing air quality related big datasets to get good predictive accuracy model. It is found that current methods have few limitations. Hence, conduction of extensive research on air pollution monitoring is required. As effective integration of meteorological and air pollutant concentration becomes efficient dataset, use of the same with ML models will significantly improve the prediction accuracy.

The work in the paper is structured into following sections. Section II represents the prior work carried on air quality prediction using data integration techniques, section III explores the proposed approaches, section IV outlines about the experiments and results and section V details about conclusion and future work.

2. RELATED WORK

Due to health concerns, prediction and prevention of air pollution is very important and research is underway to measure the air quality. LightGBM, histogram-based model and sliding window mechanisms have been proposed in [5] to predict quality of air. The model was built to fuse historical data, predictive data and meteorological data together to find the PM2.5 concentration in the upcoming 24 hours. Predicator framework (spatial and temporal) is proposed in [7] for air quality detection using linear regression and artificial neural network (ANN).The datasets used are meteorological data (weather forecast), AQI and wind speed. The accuracy of the model was found to be 75%. Regression models have been proposed in [8] to forecast air quality. Air quality prediction data, which consists of pollutant concentrations of Delhi and Houston city, was used for training and testing the model. For achieving more accuracy, computational complexity can be increased.

An air quality prediction method by developing Long Short Term Memory (LSTM) along with k Nearest Neighbors (k-NN) has been proposed in [9]. Datasets used are pollutant concentration values of CO, O3, SO2, PM10 and PM2.5 and site details including site name, site latitude and longitude. However the method failed to predict air quality considering the meteorological and geographical factors. Air quality prediction approach using deep neural network (DNN) is proposed in [10]. Spatial transformation has been done on the dataset consisting of spatial air quality to convert it into AQI. Then AQI and other datasets are given as input to the deep neural network to forecast the pollution in air. Multisensory space time data fusion highlighting improvement of PM2.5 concentration is proposed in [11]. The regression model's r2 value was 0.89. A model for road traffic speed prediction using K- nearest neighbor Euclidian distance is proposed in [12]. The datasets used are social media tweet sensors, CCTV speed sensors and car trajectory sensors collected from map. Limitation of the work is, it needs to be extended for real time large traffic model.

Several research works have been carried out on air quality prediction by integrating different data sets from multiple domains. The proposed research work focuses on the different data sets from multiple domains of Bengaluru city.

3. PROPOSED METHODOLOGY

Proposed methodology consists of mainly seven stages such as data loading from multiple sources, data integration from multiple sources, data pre-processing, data splitting, building machine learning models and optimization of hyper parameters, predict AQI and evaluate the models using SMAPE, MSE and MAE. The pictorial representation of the proposed architecture is given in figure 1.

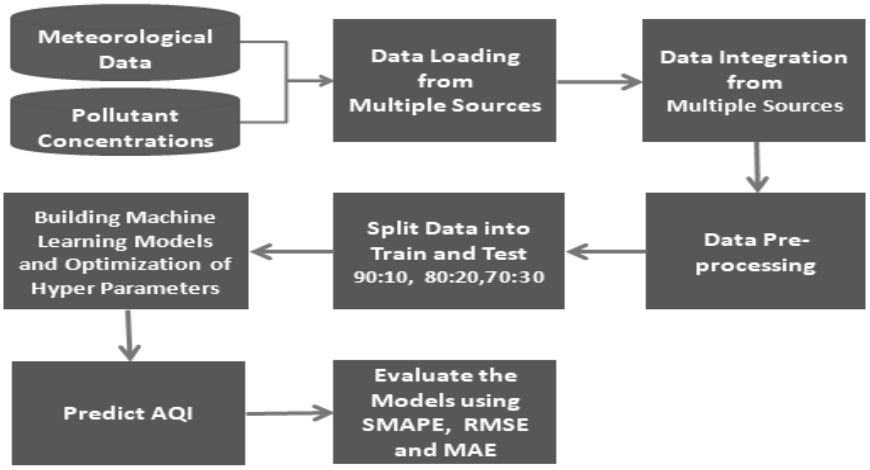

Fig 1. Proposed architecture for predicting AQI

3.1. Data Loading from Multiple Sources

It is the first and most important stage where the much needed data for the air quality prediction is required. The data collected for air quality prediction are pollutant concentration data [13] and meteorological data [14]. Meteorological data is collected from world weather online website and air pollutant concentrations data is collected from pollution control board. Table 1 depicts five days meteorological dataset sample, which consists of temperature, humidity, pressure and wind speed. Table 2 depicts pollutant concentration dataset which consist of PM2.5, PM10, 03, NO2, SO2 and CO in μg/m3 which can be used to calculate AQI.

Table 1. Meteorological Data

Sl. No	***Date***	***Temp (in cel)***	***Humidity (in %)***	***Pressure (in mb)***	***Wind speed (in kmph)***
1	*2019-07-01*	*26*	*63*	*1006*	*25*
2	*2019-07-02*	*27*	*61*	*1006*	*28*
3	*2019-07-03*	*26*	*62*	*1007*	*28*
4	*2019-07-04*	*27*	*62*	*1007*	*27*
5	*2019-07-05*	*26*	*63*	*1007*	*30*

Table 2. Air pollutant concentrations

Sl. No	***Date***	***PM 2.5***	***PM 10***	***O3***	***NO2***	***SO2***	***CO***
1	*2019-07-01*	*48*	*49*	*34*	*10*	*1*	*5*
2	*2019-07-02*	*43*	*48*	*30*	*9*	*1*	*6*
3	*2019-07-03*	*34*	*38*	*22*	*9*	*1*	*4*
4	*2019-07-04*	*33*	*40*	*27*	*9*	*1*	*4*
5	*2019-07-05*	*36*	*35*	*27*	*9*	*1*	*4*

3.2. Data integration from multiple sources

It is the process of fusing data from different sources of different domains to create more reliable, precise and helpful knowledge than produced by any individual source. Usage of a single dataset cannot result in the accurate prediction of air quality. Hence data integration is performed by combining both the data sets such as meteorological and pollutant concentrations as shown in table 3. Six months dataset, based on the 24 hours average of hourly readings, from July 2019 to December 2019 of Bengaluru city has been collected.

Table 3. Integrated data

Date	***Temp***	***Humidity***	***Pressure***	***Wind speed***	***PM 2.5***	***PM 10***	***O3***	***NO2***	***SO2***	***CO***
2019-07-01	*26*	*63*	*1006*	*25*	*48*	*49*	*34*	*10*	*1*	*5*
2019-07-02	*27*	*61*	*1006*	*28*	*43*	*48*	*30*	*9*	*1*	*6*
2019-07-03	*26*	*62*	*1007*	*28*	*34*	*38*	*22*	*9*	*1*	*4*
2019-07-04	*27*	*62*	*1007*	*27*	*33*	*40*	*27*	*9*	*1*	*4*
2019-07-05	*26*	*63*	*1007*	*30*	*36*	*35*	*27*	*9*	*1*	*4*

3.3. Prescreened Process

It is one of the essential steps in data science and machine learning to handle the missing values in the dataset. The purpose is to remove or filter out the unwanted and noisy data from the data set. In the dataset, all the empty strings are replaced by null values. After dropping the rows with null values, the prescreened process and fused data has been saved into a csv file.

3.4. Data Splitting

To use datasets in machine learning models, it should be first divided into train and test set. Using train set, model is trained. To find the accuracy of the model, test set is used. Three scenarios are used for the proposed machine learning models such as 90:10, 80:20 and 70:30 for train and test purpose. Different machine learning models comparison has been done.

3.5. Building Machine Learning Models and Optimization of Hyper Parameters

From survey, it is observed that few of the machine learning models such as artificial neural networks, deep learning, natural language processing and spatial interpolation models were used for AQI prediction. Still more algorithms need to be explored. In this regard we are implementing k-nearest neighbors algorithm (k-NN), Decision tree, Support Vector Machine and XGBoost algorithms. Optimization of hyper parameters is carried out to obtain the best results of machine learning techniques. The performance metrics such as SMAPE, RMSE and MAE values of proposed machine learning models were calculated after optimizing the parameters required by the models and as a result of this we could find better results.

3.6. Optimization of Hyper Parameters

Different tuneable parameters are used in various machine learning algorithms. The parameter values are varied to obtain the best results. The proposed algorithm for Optimizing Hyper Parameters (OHP) in various machine learning techniques for air quality predictive analysis is given below.

Algorithm OHP: Optimizing *Hyper Parameters for Machine Learning Techniques*

Input: Multi source data from meteorological and pollutant concentrations
Output: Prediction of Air Quality Index (AQI)
Method:

M ← Meteorological dataset, **P**← Pollutant concentration
Let M = {m_1, m_2, m_3, ……. m_n} and P = {p_1, p_2, p_3, ……. p_n}. m_i and p_i represent independent variables. **A** ←AQI dependent variable to be predicted
Perform data fusion of M and P dataset

$$M \bigcup P = \{m_i \bigcup p_i \mid m \in M, p \in P\}$$

$$\mathrm{D} \Leftarrow M \bigcup P,\ D \leftarrow Integrated\ Dataset$$

Pre-Processing **D** by eliminating redundant values, missing values.

for i in n:
  do
    $D = \{d_1, d_2, d_3, \ldots\ldots\ldots\ldots d_n\}$
    if $d_i == null$ *then*
      $E\,\mathrm{lim}inate\ d_i$
    end if
end for
 # Generating Hyper Parameters Value for Optimization of ML Techniques
$HP \leftarrow Hyper\ Parameters\{\mathrm{HP_{XGBoost}}, \mathrm{HP}_{DT}, \mathrm{HP}_{SVM}, \mathrm{HP}_{KNN}\}$
$\mathrm{HP_{XGBoost}} \leftarrow$ {learning rate (l_r),depth (d), gamma (γ)}
$\mathrm{HP}_{DT} \leftarrow$ {depth (d), samples split (s_s), features (f), and leaf nodes (l_n)}
$\mathrm{HP}_{SVM} \leftarrow$ {C, degree (d_e), kernel (k_e), gamma (γ), epsilon (ε)}
$\mathrm{HP}_{KNN} \leftarrow$ {n neighbors (n_h), weights (w_i), algorithm (alg), metric (m_e), p}
for each split ratio $\in \{70:30, 80:20, 90:10\}$ do
  $x \leftarrow train\ ratio, y \leftarrow test\ ratio$
  for each $ML_i \in \{XGBoost, DT, SVM, KNN\}$ $HP_i \in \{\mathrm{HP_{XGBoost}}, \mathrm{HP}_{DT}, \mathrm{HP}_{SVM}, \mathrm{HP}_{KNN}\}$
    start with default values of selected ML_i *with* HP_i
    Generate the HP_i *value and* $\mathrm{Pr}edict\ AQI$
    Measure $RMSE, SMAPE, MAE$
      if $Errorrate < Threshold$ *then*
        select the HP_i *value for optimal prediction* $u\sin g\ \ ML_i$
      end if
  end for
end for

3.7. Predict AQI

Table 4 depicts the AQI predicted by proposed machine learning models such as k-nearest neighbor, decision tree, support vector machine and XGBoost for the new test data set.

Table 4. AQI prediction with new test dataset.

Temp	Humid	Pressure	Wind speed	PM 2.5	PM 10	O3	NO2	SO2	CO	AQI Prediction			
										KNN	DT	SVM	XGB
26	63	1006	25	48	49	34	10	1	5	144	137	139	136
27	61	1006	28	43	48	30	9	1	6	146	151	148	149
26	62	1007	28	34	38	22	9	1	4	133	125	130	125
27	62	1007	27	33	40	27	9	1	4	133	125	132	125
26	63	1007	30	36	35	27	9	1	4	132	125	128	125

3.8. Evaluate the models using SMAPE, RMSE and MAE

To validate the performance of models, the evaluation metrics considered are Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The metrics measure accuracy based on error percentage and can be defined as follows.

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \tag{2}$$

Here, A_t denotes actual value and F_t denotes predicted value

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{M}(\text{Predicted value}_k - \text{Actual value}_k)^2}{M}} \tag{3}$$

Here, i is a variable and M is number of observations.

$$\text{MAE} = \frac{1}{n}\sum|t - \hat{p}| \tag{4}$$

In equation 4, n is the data points count, t is the actual output, $\hat{p}$ is predicted output and $t - \hat{p}$ is the absolute value of the residual.

4. EXPERIMENTS AND RESULTS

In order to conduct experiments on proposed machine learning models, python language is used for implementation. Experiments are conducted in Jupiter notebook with six months dataset from July 2019 to December 2019. First exploratory data analysis (EDA) is applied to understand the structure of data. As shown in the figure 2, histogram graph has been drawn to show the count of AQI across its range. As we can clearly see most of the AQI lies between 120 and 180.

This indicates that most of the AQI value belongs to moderately polluted category. As depicted in figure 3, heat map graph has been drawn to indicate the positive and negative correlation of both meteorological data and pollutant concentrations data.

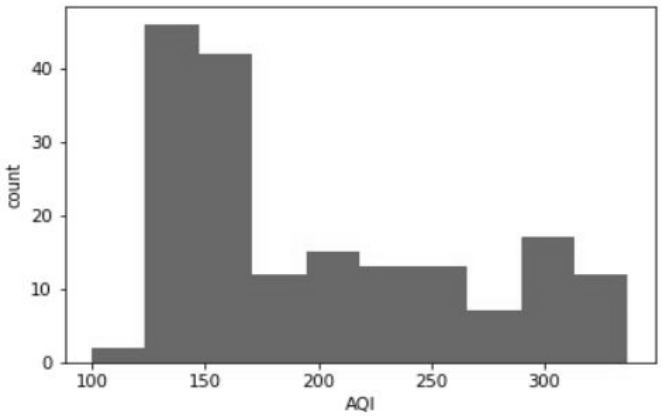

Fig 2. AQI count

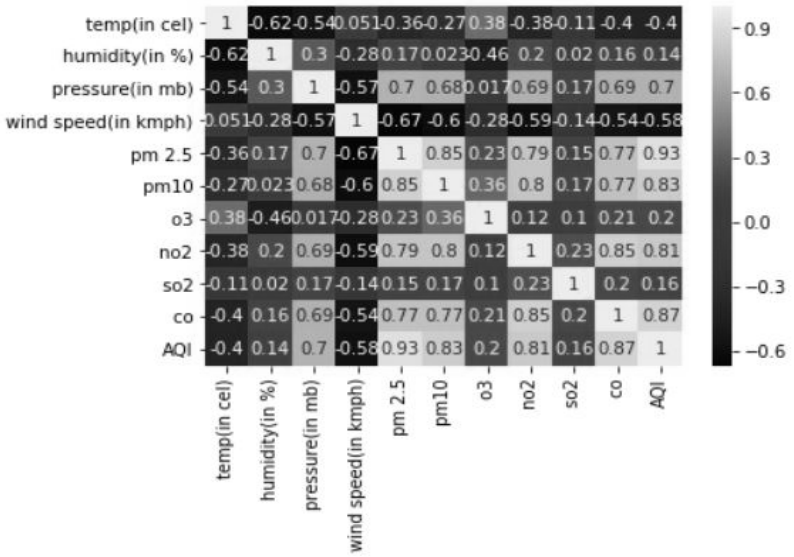

Fig 3. Heat map of the interrelated features

Figure 4 depicts the proposed model's predicted fitting effect on the test data. This graph indicates the variations of AQI in three different regions. It is plotted using October 2019 data sets for different regions of Bengaluru city. Here we can clearly see Silkboard of Bengaluru is more polluted than other two regions named Jayanagar and Hebbal.

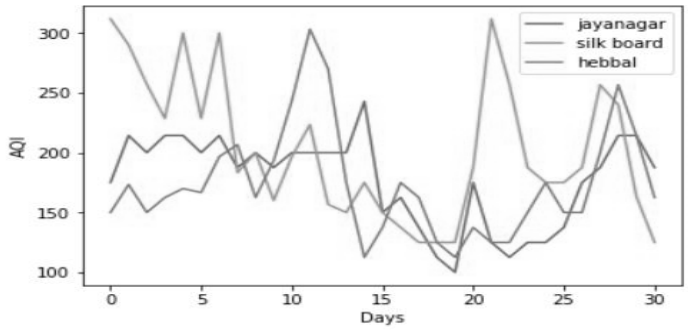

Fig 4. Fitting curve

After EDA proposed machine learning models are implemented with three training and testing scenarios namely 90:10, 80:20 and 70:30. In order to explore on different tuning parameters of proposed machine learning models, we have conducted extensive research and applied optimization technique.

The tuneable parameters used in each machine learning models are given in the table 5. As shown in table 6, the tuneable parameter values are varied and the results are highlighted under the evaluation metrics in XGBoost. Optimization of hyper parameters such as ratio, learning rate, max depth, gamma and subsample is performed with suitable values and results of evaluation metrics such as SMAPE, RMSE and MAE values are observed.

Table 5. Tuneable parameters list.

Sl. No.	***ML Model***	***Tuneable parameters used***
1.	*XGBoost*	*learning_rate, max_depth, gamma, subsample*
2.	*Decision tree*	*max_depth, min_samples_split and max_leaf_nodes*
3.	*Support Vector Machine*	*degree, kernel, gamma, epsilon*
4.	*k- nearest neighbour*	*n_neighbors, algorithm, metric*

Table 6. XGBoost results under the evaluation metrics.

Ratio	***learning_ rate***	***max_ depth***	***Gamma***	***Subsample***	***SMAPE***	***RMSE***	***MAE***
90/10	*0.2*	*4*	*0.2*	*0.6*	*2.7*	*8.0*	*4.8*
90/10	*0.1*	*5*	*0.2*	*0.7*	*1.9*	*7.0*	*3.3*
90/10	*0.1*	*5*	*0.1*	*0.9*	*2.0*	*7.6*	*3.8*
80/20	*0.1*	*5*	*0.1*	*0.5*	*2.0*	*6.8*	*4.0*
80/20	*0.2*	*4*	*0.2*	*0.6*	*2.2*	*7.9*	*4.7*
80/20	*0.1*	*5*	*0.1*	*0.9*	*1.6*	*5.4*	*3.2*
70/30	*0.1*	*5*	*0.1*	*0.5*	*2.6*	*8.4*	*5.3*
70/30	*0.1*	*5*	*0.2*	*0.7*	*2.3*	*8.0*	*4.9*
70/30	*0.1*	*5*	*0.1*	*0.9*	*2.2*	*7.9*	*4.7*

As depicted in table 7, Decision tree results are highlighted. Parameters such as ratio, max depth, min samples split and max leaf nodes values are varied and the values of performance metrics SMAPE, RMSE and MAE are observed. Results are highlighted under the evaluation metrics in support vector machine as shown in table 8. Parameters such as ratio, degree, kernel, gamma and epsilon are modulated and performance metrics such as SMAPE, RMSE and MAE values are observed.

Table 7. Decision tree results under the evaluation metrics.

Ratio	***max_ depth***	***min_samples _split***	***max_leaf _nodes***	***SMAPE***	***RMSE***	***MAE***
90/10	*9*	*9*	*9*	*3.4*	*8.5*	*5.5*
90/10	*10*	*10*	*10*	*2.8*	*9.6*	*5.4*
90/10	*12*	*12*	*12*	*3.9*	*11.1*	*6.8*
80/20	*9*	*9*	*9*	*4.9*	*17.3*	*10.7*
80/20	*10*	*10*	*10*	*4.7*	*14.7*	*9.5*
80/20	*10*	*10*	*9*	*4.9*	*17.3*	*10.7*
70/30	*9*	*9*	*9*	*4.6*	*13.7*	*9.3*
70/30	*10*	*10*	*10*	*2.8*	*9.6*	*5.4*
70/30	*10*	*10*	*9*	*4.6*	*13.7*	*9.3*

Table 8. Support vector machine results under the evaluation metrics.

Ratio	***Degree***	***Kernel***	***Gamma***	***Epsilon***	***SMAPE***	***RMSE***	***MAE***
90/10	*3*	*Rbf*	*0.001*	*0.1*	*3.4*	*7.6*	*5.3*
90/10	*4*	*Linear*	*0.0001*	*0.2*	*7.2*	*14.0*	*11.8*
90/10	*2*	*linear*	*10*	*0.001*	*7.3*	*14.0*	*11.9*
80/20	*3*	*Rbf*	*0.0001*	*0.1*	*5.9*	*14.6*	*10.3*
80/20	*3*	*Rbf*	*0.001*	*0.1*	*2.7*	*6.4*	*4.3*
80/20	*4*	*linear*	*0.0001*	*0.001*	*6.8*	*16.9*	*13.0*
70/30	*3*	*Rbf*	*0.0001*	*0.1*	*6.2*	*15.5*	*11.2*
70/30	*3*	*Rbf*	*0.001*	*0.1*	*3.0*	*7.2*	*5.1*
70/30	*2*	*linear*	*10*	*0.001*	*7.2*	*18.3*	*14.3*

Table 9 shows the results that are highlighted under the evaluation metrics in k- nearest neighbor. Optimization of parameters such as ratio, n neighbors, algorithm and metric values is performed with suitable values and the SMAPE, RMSE and MAE values are observed.

Table 9. k- nearest neighbor results under the evaluation metrics.

Ratio	n_neighbors	Algorithm	Metric	SMAPE	RMSE	MAE
90/10	5	Auto	Minkowski	6.9	15.6	11.3
90/10	7	ball_tree	Manhattan	6.9	14.8	11.1
90/10	5	kd_tree	Manhattan	6.9	15.6	11.3
80/20	5	Auto	Minkowski	6.6	16.6	11.7
80/20	6	Brute	Euclidean	6.5	16.6	11.5
80/20	7	kd_tree	Manhattan	6.1	15.6	10.9
70/30	5	Auto	Minkowski	6.1	15.6	11.1
70/30	5	Brute	Euclidean	6.5	16.7	11.3
70/30	5	kd_tree	Manhattan	5.8	14.8	10.4
70/30	5	ball_tree	Manhattan	6.3	15.8	11.4

As shown in figure 5, we compare SMAPE, RMSE and MAE values of proposed machine learning models. XGBoost has given good results for predicting the air quality index with SMAPE value of 1.6, RMSE value of 5.4 and MAE value of 3.2. As it has very less SMAPE, RMSE and MAE values compared to other machine learning techniques, can be stated as a better accuracy model. The XGBoost model uses boosting technique to merge multiple weak classifiers and make them powerful for better prediction. We found 50% improvement in the performance metrics of XGBoost compared to other models.

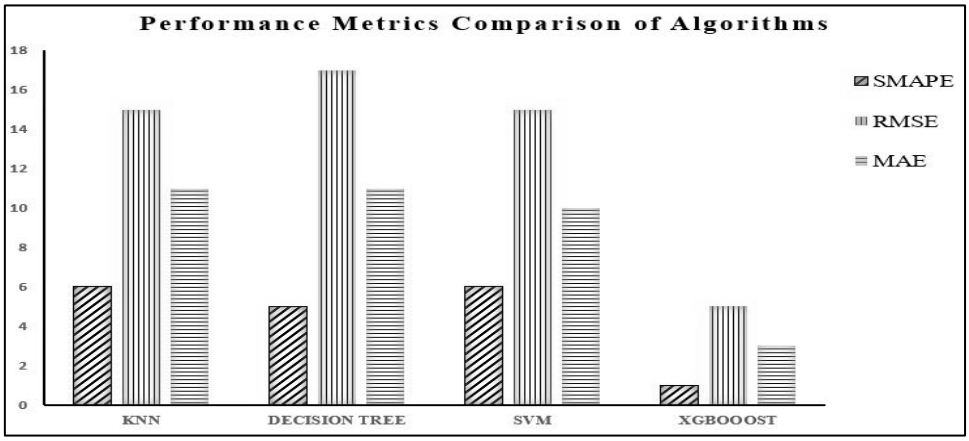

Fig 5. Performance metrics comparison of different algorithms

5. CONCLUSION AND FUTURE WORK

The proposed approach predicts the air quality index by integrating different data sets namely air pollutant concentrations and meteorological data. In this

proposed work, different machine learning models like xgboost, decision tree, support vector machine and k- nearest neighbor were evaluated for the given data set. Optimization of hyper parameters for different machine learning models has been carried out with three different train and test scenarios such as 70:30, 80:20 and 90:10. The performance metrics such as SMAPE, RMSE and MAE were used for evaluating the models. The comparison was made by varying the parameters of machine learning models and results were recorded for the same. Xgboost model gave good results for predicting the air quality index with SMAPE value of 1.6 which was best compared with k-nearest neighbour, decision tree and support vector machine. This proposed work can be enhanced to find AQI by using different data sets like grid node data and weather forecast data. This work can be scalable to other regions data sets also. The integration of datasets can be used for health risk assessment, environmental management and source localization of pollution.

ACKNOWLEDGEMENT

This research was supported by Visvesvaraya Technological University, Jnana Sangama, Belagavi.

REFERENCES

[1] Acosta NJR, Fuller R., Landrigan P. J. *The Lancet Commission on pollution and health*. 2017, Available at: https://pubmed.ncbi.nlm.nih.gov/29056410/ (visited on 09.11.2020)

[2] World Health Organizatio. *Air pollution details*. 2018, Available at: https://www.who.int/health-topics/air-pollution#tab=tab_1 (visited on 15.11.2020)

[3] Global weather and climate center. *Weather and climate data*. 2017, Available at: https://www.globalweatherclimatecentenr.com/air-quality-topics/limitations-of-epas-air-quality-index-credit-physics-today (visited on 29.09.2020)

[4] Yu Zheng. Methodologies for Cross-Domain Data Fusion: An Overview. *IEEE Transactions on Big Data*. ISSN: 2332-7790, Vol. 1, No. 1, 2015, pp. 16-34, DOI:10.1109/TBDATA.2015.2465959.

[5] Ying Zhang, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao, Rongrong Zhang, Qingqing Wang, Linyan Huang. A Predictive Data Feature Exploration-Based Air Quality Prediction Approach. *IEEE Access*, ISSN: 2169-3536, Vol. 7. 2019, pp. 30732-30743, DOI:10.1109/ACCESS.2019.2897754.

[6] X. Zhang, M. He, B. Shao, C. Ren. Physical-social fusion to assist public services in the war against air pollution in China. *IEEE 14th International Conference on Industrial Informatics (INDIN),* 19-21 July, 2016, Poitiers, France,

ISBN: 978-1-5090-2870-2, IEEE, pp. 916-920, DOI: 10.1109/INDIN.2016.7819292.

[7] Yu Zheng, Xiuwen Yi Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, Tianrui Li. Forecasting Fine-Grained Air Quality Based on Big Data. KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 10 August, 2015, Sydney, NSW, Australia. ISBN:978-1-4503-3664-2/15/08, ACM, pp. 2267-2276, http://dx.doi.org/10.1145/2783258.2788573.

[8] S. S. Ganesh, S. H. Modali, S. R. Palreddy, P. Arulmozhivarman. Forecasting air quality index using regression models: A case study on Delhi and Houston. *International Conference on Trends in Electronics and Informatics (ICEI),* 11-12 May 2017, Tirunelveli, India, ISBN: 978-1-5090-4257-9, IEEE, pp. 248-254, DOI: 10.1109/ICOEI.2017.8300926.

[9] Zepeng Qin, Chen Cen, Xu Guo, Prediction of Air Quality Based on KNN-LSTM. *Journal of Physics: Conference Series*, Vol. 1237. Issue 4, 2019, ISBN: 1237 042030, IOPScience Publishing, pp. 1-6, doi:10.1088/1742-6596/1237/4/042030.

[10] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, Yu Zheng. Deep Distributed Fusion Network for Air Quality Prediction. *Proceedings of the 24th SIGKDD Conference on Knowledge Discovery and Data Mining*, 19 July, 2018, New York, NY, USA, ISBN: 9781450355520, Association for Computing Machinery, pp. 965-973, https://doi.org/10.1145/3219819.3219822.

[11] Yuan-Chien Lin, Wan-Ju Chi, Yong-Qing Lin. The improvement of spatial-temporal resolution of PM2.5 estimation based on micro-air quality sensors by using data fusion technique. *Environment International, ISSN 0160-4120,* Vol. 134. 105305, 2020, pp.1-18. https://doi.org/10.1016/j.envint.2019.105305.

[12] L. Lin, J. Li, F. Chen, J. Ye , J. Huai. Road Traffic Speed Prediction: A Probabilistic Model Fusing Multi-Source Data. *IEEE Transactions on Knowledge and Data Engineering,* Vol. 30, No. 7, 2018, pp. 1310-1323, DOI: 10.1109/TKDE.2017.2718525.

[13] Bengaluru Air Pollution. *Real-time Air Quality Index*. 2019, Available at: https://aqicn.org/city/bangalore/. (visited on 19.06.2020).

[14] World Weather Online. *Bengaluru Weather averages*, 2019, Available at: https://www.worldweatheronline.com. (visited on 25.06.2020)

Information about the authors:

Basamma Umesh Patil – Assistant Professor, Department of Computer Science and Engineering, SJBIT, Bengaluru, India. Area of Scientific research: Machine Learning, Big Data Analytics, Internet of Things and Computer Networks.

Dr. D V Ashoka – Dean Research and Professor, Department of Information Science and Engineering, JSSATE, Bangalore. His fields of interest are Requirement Engineering, Big Data Analytics, Software Architecture, Computer Networks and Machine Learning.

Dr. Ajay Prakash B V – Associate Professor, Department of Computer Science and Engineering, SJBIT, Bengaluru, India. Area of Scientific research: Machine Learning, Big Data Analytics and Software Engineering.

Manuscript received on 02 June 2021