Air Quality Prediction using Ensemble Machine Learning Algorithm

P. Durga Devi, M. Chandrakala
Assistant Professor, Dept. of ECE,
Mahatma Gandhi Institute of Technology,
Hyderabad, Telangana, India.

***Abstract*—Air is the fundamental element that is essential for living organisms. Such an essential element is getting polluted day by day. It is important to get an update on the quality of the air in our surroundings. This is possible by measuring Air Quality Index(AQI).This paper presents the performance evaluation of machine learning algorithms, (1) linear regression; (2) k-Nearest Neighbor regression;(3) decision tree regression, and (4) the combination of the three mentioned models as an ensemble learning model or stacking method; to predict AQI. Six air pollutants, Ozone (O_3), Carbon monoxide (CO), Benzene, Sulphur Dioxide (SO_2), Oxides of Nitrogen (NO_x), and Particulate Matter (PM2.5) levels for the years 2017to 2020 in Hyderabad, India, were used to train the regression models to predict the AQI. Among the four, stacking method got more accuracy with MAE = 8.271, RMSE = 10.818, and validation score = 0.939.**

***Keywords*—Air Quality Index; stacking; regression;**

I. INTRODUCTION

Different kinds of vehicles or engines that burn fossil fuels have been the major contributors to CO and Benzene in the outdoor air. CO is a colorless, odorless gas that can be released when something is burnt. As Benzene is found in petrol and oil, can be released due to vehicles. Stratospheric ozone that occurs naturally in the upper atmosphere protects us from harmful ultraviolet rays. But Tropospheric ozone that exists at the ground level is a harmful air pollutant.Ground-level ozone will be generated due to the chemical reactions that occur between oxides of nitrogen (NOx) and volatile organic compounds.It all happens when the pollutants that are emitted from vehicles, power plants, industries, and other sources react chemically in the presence of sunlight. This ground-level ozone arises breathing problems, asthma to people who are activeoutdoors especially children(https://www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution).

Symptoms of carbon monoxide poisoningcause headaches, faintness, exhaustion, vomiting, andpossiblyloss of consciousness. Oxygen deficiency, anemia, and cardiac diseases are observed as a result of oxygen loss when the competitive binding of carbon monoxide is effective [1]. Benzene can cause loss of white blood cells which leads to low immunity. It also causes anemia by reducing the production of red blood cells [2]. Klemm et al., have investigated the correlation between mortality and air pollution for two years period in Atlanta [3]. They found a significant increase in deaths due to respiratory problems and cancer with an increase of PM 2.5 levels. It is also observed that people above 65 years of age were getting affected by CO levels in the air.

A stacking approach with SVR and k-NN as base models have got more accuracy in the prediction of super-resolution images and also achieved high consistency with human visual judgments [4]. A greedy stacking method for various biomedical disciplines with different data sets got more accuracy in prediction compared to linear, genetic algorithm stacking, and a brute force approach [5]. Wolpert and Macready have investigated some bootstrap algorithms and derived that the stacking method has shown improved performance [6]. Dragomir [7] has performed the k-NN regression method to predict the quality of air in Ploiesti and obtained zero prediction error for 19 out of 29 instances.

Shishegaran et al., have used four prediction models that are (1)Auto-Regressive Integrate Moving Average, (2) Principal Component Regression, (3) combination of 1& 2, and (4) combination of model 1 with Gene Expression Programming to estimate daily AQI in Tehran, Iran [8]. Among these, the last model which is the non-linear ensemble regression model has shown the best results.

In the present work, we have compared four regression models viz., linear regression, k-NN regression, decision tree regression,and stacking method in estimation of the air quality index (AQI) based on six air pollutants.

II. DATA AND METHODS

A. Study Area

Hyderabad is the capital city of Telangana state, India. It is one of the industrially fast-growing metropolitan cities in the country, with a population of around 10 million (worldpopulationreview.com). This paper focused on predicting the air quality index of six areas in Hyderabad where Continuous Ambient Air Quality Monitoring Stations (CAAQMS) stations are located.

B. Data Collection

For the present work, data related to AQI and air pollutants of six CAAQM stations of Hyderabad, which are located at HCU, Sanathnagar, Zoopark, Pashamylaram, Bollaram, and ICRISAT, collected from the official website (https://tspcb.cgg.gov.in/Pages/Envdata.aspx) of the pollution control board of Telangana state for the period of four years, i.e., from January 2017 to December 2020.

The Pollution Control Board of Telangana has been arranged CAAQM stations that can monitor air quality automatically every 15 minutes.For the present study, monthly average data of AQI (unitless) and six pollutants, O3 (µg/m3), CO (mg/m3), Benzene (µg/m3), SO2 (µg/m3),

NOx (µg/m3), and PM2.5 (µg/m3) have been considered to estimate AQI. These predicted values of AQI were compared with measured AQI.

C. Methods

1) Multiple Linear regression

It is the basic method of regression that has more than one independent variable. In the present work, six air pollutants are taken as independent variables to estimate the AQI value. The equation of prediction is given as,

$$D = b + m_1 i_1 + m_2 i_2 + m_3 i_3 + m_4 i_4 + m_5 i_5 + m_6 i_6 \quad (1)$$

Where,D is the predicted dependent variable (AQI)b is the intercept.

$m_1, m_2, m_3, m_4, m_5,$ and m_6 are the estimated regression coefficients

i_1, i_2, i_3, i_4, i_5 and i_6 are the independent variables.

2) k-Nearest Neighbor regression

To predict the values of new data points, the k-NN algorithm uses feature similarity. This means that a value is assigned to the new point based on how closely it approaches the points in the training set. The Euclidean distance between the new point and each training point is calculated.The closest k data points are selected based on the distance. The mean value of all the k data points is the predicted value of the new data point. The error varies concerning the k value.

3) Decision Tree Regression

In this method, the data points get split using a binary tree.There are parent nodes and child nodes or leaf nodes. Generally, there are three elements 1) selection of splits; 2) the decisions regarding when to terminate the node or continuation of splitting; and 3) assignment of each leaf node to a group.The impurity of nodes reduces with best splits[9]. That means homogeneityincreases at every child node. Similarly, the best node will be selected for a new data point and the mean of all data points in that node is the predicted value of the new data point.

It is crucial to find the best tree split that reduces the impurity of child nodes. For this, it is required to compute variance reductionbecause, for the regression, variance is the measure of impurity. The higher the variance reduction value, the lower the impurity.

The variance reduction can be calculated as follows:

$$\text{Variance Reduction} = \text{var (parent node)} - \sum W_i \text{ var (child}_i) \quad (2)$$

where, W_i is called the weight of ithnode that is the ratio ofthe size of the child node to the size ofthe parent node.

The split which has more variance reduction is the better split. That means there is a significantdifference between the child nodes.

4) TheStacking regression model

In the present work, AQI was predicted using the stacking method which has shown better performance compared to the individual regression techniques. Stacking is an ensemble machine learning algorithm.Generally, in the stacking method, two or more base regression models are used for prediction at the first level. In the second level, a meta-regression model is used to get the final prediction by combining the predictions of all base regression models. The workflow of the stacking method is shown in Fig. 1.

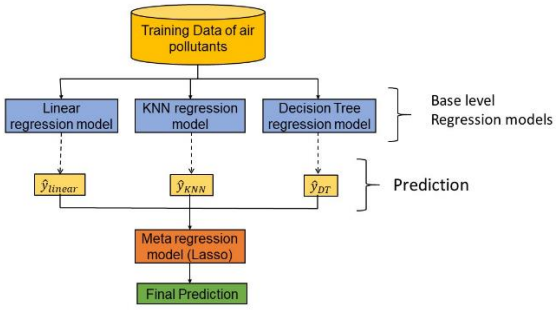

Fig. 1. *Workflow of stacking method of regression*

In the presentstacking or ensemble learning method, we have used multiple linear regression, k-NN, and decision tree regression methods as base models and the lasso method as the meta-model.At the first level, the value of the dependent variable (AQI) was predicted using the three base models. All these values were given as input to the meta regressor in the second level. This meta regressor gave the final predicted value of AQI.

III. RESULTS AND DISCUSSION

The correlation matrix of six air pollutants concerned with AQI is given in Table 1. It is observed that CO, NOx, and PM2.5 have a significant correlation with AQI compared to other parameters.

TABLE 1: CORRELATION COEFFICIENTS OF SIX POLLUTANTS AND AQI TO EACH OTHER

	O3	**CO**	**Benzene**	**SO2**	**NOx**	**PM2.5**	**AQI**
O3	1	0.50	0.15	0.38	0.16	0.36	0.4
CO		1	0.32	0.13	0.6	0.62	0.64
Benzene			1	-0.01	0.36	0.2	0.21
SO2				1	-0.01	0.27	0.3
NOx					1	0.64	0.64
PM2.5						1	0.96
AQI							1

A. Performance evaluation

The statistical metrics such as mean absolute error (MAE) and root mean square error (RMSE) had been chosen to evaluate the performance of three individual regression techniques and the stacking method. The metrics are described as following:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|T_i - P_i| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i - P_i)^2} \quad (4)$$

Where nis the total number of samples, T is the measured value, and P is the predicted value of AQI.

The MAE, RMSE, and validation scores of all regression models are presented in table 2.

TABLE 2: PERFORMANCE METRICS OF BASE LEARNING AND STACKING MODELS

Model	MAE	RMSE	Validation Score
Linear Regression	10.072	12.197	0.916
k-NN Regression	8.757	12.07	0.921
Decision Tree Regression	9.164	12.835	0.873
Stacking	8.271	10.818	0.939

The validation scores of the four regression models are shown in Fig. 2. The stacking model has got best validation score which is 0.939.

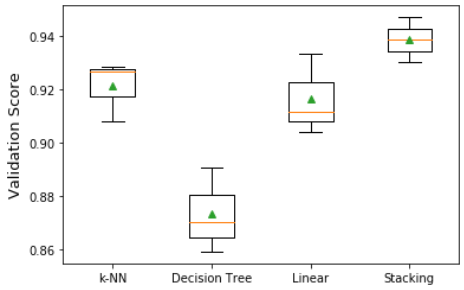

Fig. 2. *Validation scores of the four regression models*

The lowest values of MAE, RMSE, and the highest value of validation score represent the best regression method.

IV. CONCLUSION

With the present study, it is concluded that the stacking method has shown better performance compared with the three regression methods; linear regression, k-NN regression, and Decision tree regression alone. The ensemble learning method has got highest validation score and lowest error, RMSE. By stacking the regression models, we have achieved about a 10% reduction in RMSE.

ACKNOWLEDGMENT

As the data for the present study was taken from TSPCB official website, the authors would like to thank them.

REFERENCES

[1] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, “Environmental and Health Impacts of Air Pollution: A Review,” Front. Public Heal., vol. 8, no. pp. 1–13,February, 2020, doi: 10.3389/fpubh.2020.00014.

[2] C. Jia, S. Batterman, and C. Godwin, “VOCs in industrial, urban and suburban neighborhoods-Part 2: Factors affecting indoor and outdoor concentrations,” Atmos. Environ., vol. 42, no. 9, pp. 2101–2116, 2008, doi: 10.1016/j.atmosenv.2007.11.047.

[3] R. J. Klemm, F. W. Lipfert, R. E. Wyzga, and C. Gust, “Daily mortality and air pollution in Atlanta: Two years of data from ARIES,” Inhal. Toxicol., vol. 16, no. SUPPL. 1, pp. 131–141, 2004, doi: 10.1080/08958370490443213.

[4] K. Zhang, D. Zhu, J. Li, X. Gao, F. Gao, and J. Lu, “Learning stacking regression for no-reference super-resolution image quality assessment,” Signal Processing, vol. 178, 2021, doi: 10.1016/j.sigpro.2020.107771.

[5] C. F. Kurz, W. Maier, and C. Rink, “A greedy stacking algorithm for model ensembling and domain weighting,” BMC Res. Notes, vol. 13, no. 1, pp. 1–6, 2020, doi: 10.1186/s13104-020-4931-7.

[6] D. Wolpert and W. Macready, “Combining stacking with bagging to improve a learning algorithm,” St. Fe Inst. Tech. Rep., no. pp. 1–28, October,1996, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.9933&rep=rep1&type=pdf.

[7] E. G. Dragomir, “Air Quality Index Prediction using K-Nearest Neighbor Technique,” Ser. Mat. - Informatică - Fiz., vol. LXII, no. 1, pp. 103–108, 2010,http://bmif.unde.ro/docs/20101/pdf_final_12EDragomir.pdf.

[8] A. Shishegaran, M. Saeedi, A. Kumar, and H. Ghiasinejad, “Prediction of air quality in Tehran by developing the nonlinear ensemble model,” J. Clean. Prod., vol. 259, p. 120825, 2020, doi: 10.1016/j.jclepro.2020.120825.

[9] Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.