Grenze International Journal of Engineering and Technology, Jan Issue

Homogenous Ensemble Learning for Air Quality Index Prediction

Shrabani Medhi[1], Rituraj Boruah[2], Sagar Pratim Baruah[3] and Hiranya Kumar Das[4]
[1]Girijananda Chowdhury University, dept. of Computer Science and Engineering, Guwahati, India
Email: shrabani_cse@gcuniversity.ac.in
[2-4]Girijananda Chowdhury Institute of Management and Technology, dept. of Computer Science and Engineering, Guwahati, India
Email: shrabani_cse@gcuniversity.ac.in, spbaruah000@gmail.com, hiranyagimt22@gmail.com

***Abstract*— Air pollution is a global issue of concern. Accurate prediction of the air quality index (AQI) plays a crucial role in understanding pollution patterns and implementing effective mitigation strategies. This research paper presents a comprehensive study on the application of ensemble learning techniques for AQI prediction, aiming to enhance prediction accuracy and provide insights into the driving factors behind pollution levels. The primary objective is to conduct a comparative analysis of homogenous ensemble method, namely, bagging in conjunction with individual base models, such as multiple linear regression, support vector regression, decision trees, and artificial neural networks. In total we have created 4 non-ensembled models and 4 ensembled models. Additionally, the research investigates the importance of different features in AQI prediction models, aiming to identify the key factors influencing air pollution. Hyperparameter tuning is done to optimize the models.**

***Index Terms*— Air Quality Index Prediction, bagging, ensemble learning, machine learning, air pollution**

I. INTRODUCTION

Air pollution is a well-known global issue, particularly in urban areas[1]. The increasing number of vehicles, industries and population growth in cities contributes to the traffic volume and, ultimately, the exhaust gases emissions. AQI (Air Quality Index) is the index for reporting the daily air quality. AQI for India is shown in Table I[2]. AQI is calculated using the formula below in Eq. (1)

$$AQI = \frac{[Ihigh - Ilow]}{[BPhigh - BPlow]} * (CP - BPlow) + Ilow \qquad (1)$$

where, AQI = Air Quality Index, CP = Pollutant Concentration, BPhigh = Concentration breakpoint that is >=CP, BPlow = Concentration breakpoint that is <CP, Ihigh= AQI value corresponding to BP, Ilow = AQI value corresponding to BPlow.

TABLE I. AQI INDEX VALUES

	Category of AQI					
	Good	Satisfactory	Moderate	Poor	Very Poor	Severe
AQI Value	0 - 50	51 - 100	101 - 200	201 - 300	301 - 400	401 - 500

Grenze ID: 01.GIJET.10.1.64

TABLE II. LITERATURE REVIEW

SL no	Author((s) and Year	Algorithm applied	Evaluation metrics used	Results
1	F. Matloob *et al 2021* [9]	Random Forest, boosting, bagging, stacking, voting, extra trees	AUC, accuracy, F-measure, Recall, Precision, MCC	A thorough review is done by analyzing the most related research papers. It was proved that ensemble learning performed better than machine learning.
3	Ali Ezzat 1, Min Wu 2, Xiao-Li Li 3, Chee-Keong Kwoh 2017 [10]	SVM, RF, PCA, LDA, ReliefF and KNN	prediction accuracy, precision, recall, AUC-ROC, RMSE.	It shows EnsemKRR achieving the highest AUC (94.3%) for predicting drug-target interactions
4	Yassaman Kazemi 1, Seyed Abolghasem Mirroshandel 2017 [11]	Bayesian model, DT, ANN, and Rule-based classifiers	accuracy, precision, recall, F1 score, and AUC-ROC	It provides a novel way to study stone disease by deciphering the complex interaction among different biological variables
5	Xueheng Qiu, Le Zhang, Ye Ren, P. N. Suganthan and Gehan Amaratunga 2014[12]	Boosting, Stacking, Bagging, Bayesian model averaging	MSE, MAE, RMSE, R-squared and MAPE	The ensemble model outperformed other machine learning models and traditional regression models in terms of accuracy and robustness.
6	Achin Jain, Francesco Smarra and Rahul Mangharam 2017[13]	Regression Tree (RT) algorithm and Ensemble Learning (EL) algorithm	MSE, MAE, R-squared, and CPI	Able to track a reference trajectory while minimizing the control effort. To be used in real-world applications, as it is able to provide accurate and efficient control of the HVAC system.
7	Akpona Okujeni, Sebastian van der Linden, Stefan Suess and Patrick Hostert 2017[14]	SRV, Bagging, Boosting, Synthetic Data Generation	MAE, RMSE, R-squared and R2	The proposed approach outperforms other regression models such as linear regression and random forest regression, as measured by several evaluation metrics. Can effectively quantify urban land cover from remote sensing data.
8	Nikhil Pachauri & Chang Wook Ahn 2022[15]	Random Forest algorithm, principal component analysis, and cross-validation	CV-RMSE, CV-MAE, and CV-R^2	Higher Coefficient of Determination (R-squared), Lower Mean Absolute Error (MAE), and Lower Root Mean Square Error (RMSE). can effectively capture the nonlinear relationships between the input variables (building characteristics) and output variables (heating and cooling loads) and provide accurate predictions.

Various machine learning techniques has been used for AQI values prediction[3]–[5]. Still ensemble learning techniques has shown lots of potential in the field of air quality prediction due to their ability to combine multiple models for improved accuracy and prediction[6]. Inclusion of several base models provide a basis for comparison. These simple models include various linear models, support vector regression, decision trees, and neural networks, etc. Each base model is trained using the training data and tuning is done using appropriate hyper parameters. Various visualization techniques can be used to visualize the data[7].

Robust evaluation testing is performed to evaluate the performance of the prediction models. Metrics include Mean Absolute Error (MAE), Mean Square Error (MSE), and coefficient of determination (R^2 score) and Root Mean Square Error (RMSE)[8]. These measurements help to gather information about the model's accuracy, precision, and fitness quality. The literature review, methodology and results and discussion are explained in the below sections.

II. LITERATURE REVIEW

The literature review is given in Table II.

III. METHODOLOGY

The workflow diagram of the paper is given in Fig.1.

A. Data Source

The dataset contains Continuous Ambient Air Quality Monitoring Station (CAAQMS) data for Guwahati city which is collected from collected from Pollution Control Board, Assam located in Bamunimaidam. AQI value 1 hour ahead is predicted.

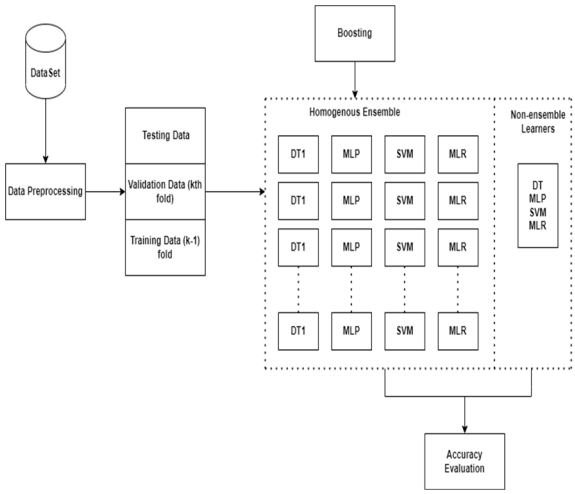

Figure 1. Workflow diagram of the systm

B. Data Preprocessing

The effective construction of forecasting models is dependent on the most important criterion which is data quality and representativity. Data preparation phase is responsible for influencing the capacity of a machine learning algorithm to generalize. In this paper, missing data imputation, eliminating or changing outlier observations, data transformation, and feature engineering is done. Interquartile range (IQR) is used to identify abnormalities. IQR calculates the first quartile (25 percent) and third quartile (75 percent) of the data range. Box plots helps to give a visual representation of the data distribution. It has a line representing the median and a box representing the interquartile range (IQR) of the data. Whiskers range from the box to various minimum and maximum values, typically 1.5 times the IQR. Depending on the particular characteristics of the data set attributes are modified using appropriate statistical methods. In machine learning, it is important to first determine whether the data is static or non-stationary. Non-stationary data exhibit patterns, seasonality, or statistical features that change over time. Checking the station is important because many machine learning algorithms, especially those based on time analysis, assume stability in the data. Unstable data can lead to inaccurate estimates and unstable models. Gaussian distribution, time plots and summary statistics is used to check the stationarity of data.

C. Feature Selection

In order to find the correlation between features and hence obtain the desired number of attributes correlation matrix is used. From Fig. 2 it is observed that the correlation is not very high. So, all the attributes are considered in the paper. It is also important to find the skewness in data. The machine learning algorithms underperform if skewness is found. It is observed that Rainfall (RF) attribute has skewness. Logarithmic transformation is applied to deal with the skewness.

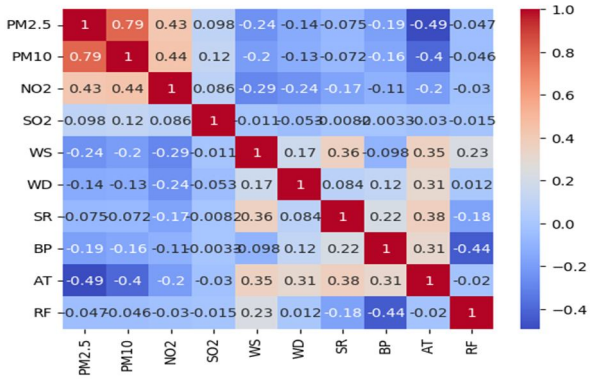

Figure 2.Correlation matrix

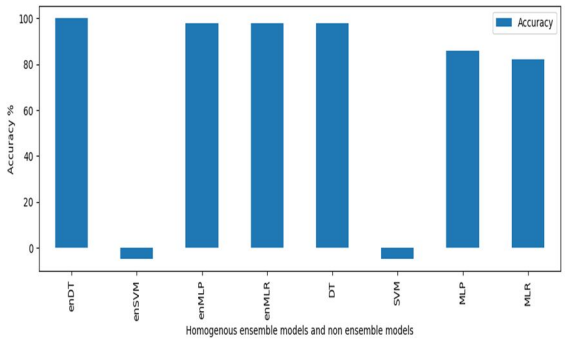

Figure 3.Accuracy of models

D. Model Training

The details of the models is given in Table III. The dataset is divided into training and test sets to evaluate the performance of the model. 70% of the data is used for training and 30% of the data is used for testing. Cross validation technique, namely, k-fold cross validation is used to test the model and evaluate its generalizability. Boosting as the ensembled technique is used. Here future predictions are enhanced and improved by learning the mistakes of the past predictions. Weak learners are arranged in a sequential fashion so each learner learns from the mistakes of the preceding learner in the sequence ultimately resulting in a better predictive model. [16] Numerous subsets, called bags, are created from the original training dataset by bootstrapping technique. Replacement is permissible. Bags help to obtain a non-discriminatory idea of the complete set. The size of the bags is less than the original dataset. Bagging is used to decrease the variance of the models. It uses DT[17] along with boosted gradient in order to improve the gradient and performance of the model. The base learner parameters were as follows: ANN [18]uses Multi -Layer Perceptron with three hidden layers: 5, 5 and 10 nodes respectively. Rectified Linear Unit (ReLU) is used as the activation function and adam is used as optimizer. Maximum iteration was set to 2000. For SVM, RBF kernel is used and regularization of 100 is used. For DT, criterion=MSE is used. 1-100 estimators is used for homogenous ensembling.To evaluate the models MAE, MSE, R2 score and RMSE are used[19].

IV. Results And Discussion

In Table IV the performance of homogenous ensembled models, namely, enMLR, enDT, enSVM and enMLP and non-ensembled models, namely MLR, DT, SVM and MLP is shown based on error metrics. It is observed that for only 1 estimator enMLP performed best out of all the models. As the number of estimators increased, enDT outperformed all other models. enDT performed best when number of estimators ranged from 10 to 100. From this it can be concluded that ensembled decision tree is dependent on the number of estimators. enSVM performed worst out of all the models. It is also observed that its performance is not dependent on the number of estimators used. The performance of enMLP and enMLR is almost same. Their performance improved when the number of estimators were increased.

Table III. Details About Models

Model Name	**Base Learners**	**Amalgamation Method**	**No of estimators**
DT	Decision Tree	--	--
MLP	Multi-Layer Perceptron	--	--
SVM	Support Vector Machine	--	--
MLR	Multiple Linear Regression	--	--
enDT	Decision Tree	Bagging	1-100
enMLP	Multi-Layer Perceptron	Bagging	1-100
enSVM	Support Vector Machine	Bagging	1-100
enMLR	Multiple Linear Regression	Bagging	1-100

Table V shows the error metrics of non-ensemble machine learning models. It is observed that out of all the models MLP (Multi-Layer Perceptron) model performed best. SVM (Support Vector Machine) model performed worst. Performance of MLR (Multiple Linear Regression) model and DT (Decision Tree) model were average. Overall

if the homogenous ensemble and non-ensemble models are compared, it can be concluded that homogenous model performs better.

TABLE IV. ERROR METRICS OF HOMOGENOUS ENSEMBLE MODELS

Models	No. of estimators	MAE	MSE	RMSE	R^2
enDT	1	0.053	0.042	0.256	0.815
enSVM	1	0.367	0.326	0.683	-0.593
enMLP	1	0.128	0.113	0.318	0.898
enMLR	1	0.051	0.047	0.286	0.825
enDT	20	0.000	0.001	0.000	1.000
enSVM	20	0.369	0.327	0.623	-0.528
enMLP	20	0.113	0.109	0.318	0.984
enMLR	20	0.018	0.017	0.013	0.981
enDT	50	0.000	0.000	0.000	1.000
enSVM	50	0.360	0.316	0.632	-0.518
enMLP	50	0.115	0.122	0.381	0.985
enMLR	50	0.012	0.023	0.033	0.983
enDT	100	0.000	0.000	0.000	1.000
enSVM	100	0.378	0.321	0.643	-0.528
enMLP	100	0.127	0.117	0.391	0.987

V. CONCLUSION

A rigorous comparison is performed between homogenous ensemble learning models using bagging with decision tree, multi-layer perceptron, multiple linear regression and support vector machine as the base learners and non-ensemble models to predict the AQI index in Guwahati city. Total of 1-100 estimators were tested. From the results it is observed that ensemble models outperformed non-ensemble models. Among ensemble models enDT performed best and enSVM performed worst. The performance of enMLP and enMLR was average. In total 8 models are analyzed. In future prediction using heterogenous ensemble models can be performed. The accuracy score of all the models is given in Fig. 4.

ACKNOWLEDGEMENT

First and foremost, we would like thank almighty GOD for everything that happens to us and makes us patient, dedicated and courageous to our work. We are very grateful to Pollution Control Board, Assam located in Bamunimaidan for providing us the necessary dataset without which it was not possible for us to carry out our research. Finally, we wish to warmly thank our institution "Girijananda Chowdhury Institute of Management and Technology-Guwahati" for encouraging us to do research work in our field.

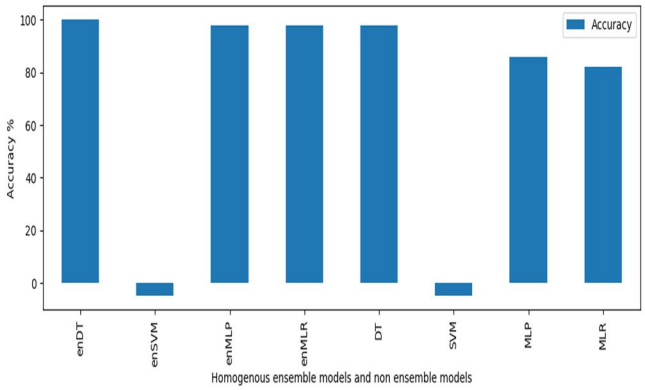

Figure. 4. Accuracy of models

TABLE V. ERROR METRICS OF NON ENSEMBLE MODELS

Non ensemble models	MAE	MSE	RMSE	R^2
MLP	0.023	1.8744e-05	0.0026	0.983
SVM	0.785	0.0089	0.1784	-0.059
MLR	0.002	5.4552e-06	0.0034	0.868
DT	0.086	9.2318e-06	0.0089	0.829

REFERENCES

[1] N. Barman and S. Gokhale, "Urban black carbon-source apportionment, emissions and long-range transport over the Brahmaputra River Valley," Science of the Total Environment, vol. 693, p. 133577, 2019.

[2] S. Sharma, "What is Air Quality Index (AQI) & How Is It Calculated ?," Prana Air, Sep. 27, 2021. https://www.pranaair.com/blog/what-is-air-quality-index-aqi-and-its-calculation/ (accessed Mar. 23, 2022).

[3] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 437–446.

[4] M. Johnson, V. Isakov, J. S. Touma, S. Mukerjee, and H. Özkaynak, "Evaluation of land-use regression models used to predict air quality concentrations in an urban area," Atmospheric Environment, vol. 44, no. 30, pp. 3660–3668, 2010.

[5] I. O. Alade, M. A. Abd Rahman, and T. A. Saleh, "Predicting the specific heat capacity of alumina/ethylene glycol nanofluids using support vector regression model optimized with Bayesian algorithm," Solar Energy, vol. 183, pp. 74–82, 2019.

[6] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," Journal of Big Data, vol. 7, no. 1, pp. 1–40, 2020.

[7] S. Medhi and M. Gogoi, "Visualization and Analysis of COVID-19 Impact on PM2. 5 Concentration in Guwahati city," in 2021 International Conference on Computational Performance Evaluation (ComPE), IEEE, 2021, pp. 012–016.

[8] W. Wang and Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," in IOP conference series: materials science and engineering, IOP Publishing, 2018, p. 012049.

[9] F. Matloob et al., "Software defect prediction using ensemble learning: A systematic literature review," IEEE Access, vol. 9, pp. 98754–98771, 2021.

[10] A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, "Drug-target interaction prediction using ensemble learning and dimensionality reduction," Methods, vol. 129, pp. 81–88, 2017.

[11] Y. Kazemi and S. A. Mirroshandel, "A novel method for predicting kidney stone type using ensemble learning," Artificial intelligence in medicine, vol. 84, pp. 117–126, 2018.

[12] R. K. S. Gautam and E. A. Doegar, "An ensemble approach for intrusion detection system using machine learning algorithms," in 2018 8th International conference on cloud computing, data science & engineering (confluence), IEEE, 2018, pp. 14–15.

[13] A. Jain, F. Smarra, and R. Mangharam, "Data predictive control using regression trees and ensemble learning," in 2017 IEEE 56th annual conference on decision and control (CDC), IEEE, 2017, pp. 4446–4451.

[14] A. Okujeni, S. van der Linden, S. Suess, and P. Hostert, "Ensemble learning from synthetically mixed training data for quantifying urban land cover with support vector regression," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 4, pp. 1640–1650, 2016.

[15] N. Pachauri and C. W. Ahn, "Regression tree ensemble learning-based prediction of the heating and cooling loads of residential buildings," in Building Simulation, Springer, 2022, pp. 2003–2017.

[16] P. Pintelas and I. E. Livieris, "Special issue on ensemble learning and applications," Algorithms, vol. 13, no. 6. MDPI, p. 140, 2020.

[17] I. S. Amiri, O. A. Akanbi, and E. Fazeldehkordi, A Machine-Learning Approach to Phishing Detection and Defense.Syngress, 2014.

[18] K. E. Koech, "Softmax Activation Function — How It Actually Works," Medium, Nov. 18, 2021. https://towardsdatascience.com/softmax-activation-function-how-it-actually-works-d292d335bd78 (accessed Feb. 17, 2023).

[19] Y. Liu, "Error awareness by lower and upper bounds in ensemble learning," International Journal of Pattern Recognition and Artificial Intelligence, vol. 30, no. 09, p. 1660003, 2016.