Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

ENHANCED STACKING ENSEMBLE MODEL IN PREDICTIVE ANALYTICS OF ENVIRONMENTAL SENSOR DATA

J. Angelin Jebamalar
PhD Research Scholar,School of Computing Sciences,
Vels Institute of Science, Technology and Advanced
Studies(VELS-VISTAS),Pallavaram, Chennai, India

Dr. T.Kamalakannan
Professor, School of Computing Sciences,
Vels Institute of Science, Technology and Advanced
Studies(VELS-VISTAS), Pallavaram, Chennai, India

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395755

***Abstract*— Predictions based on existing data are an effective way to protect human lives. It plays a vital role in taking precautionary measures and minimize the degree of damage. In India, air pollution is a major issue causing several health problems like respiratory difficulties, lung cancer and even cardiopulmonary deaths. Air is contaminated by various pollutants, among which Particulate Matter($PM_{2.5}$) is known to be the toxic particles smaller than 2.5 micrometers in diameter. This paper focuses on prediction of $PM_{2.5}$ concentration level in the air using an enhanced stacking ensemble machine learning model. The experimental outcome indicates our proposed model performs better comparative to other ensemble models.**

Keywords—$PM_{2.5}$ prediction; Outlier; Boosting regression algorithms; Stacking ensemble

I. INTRODUCTION

Air pollution is one of the main environmental issues contributing to global warming and has a major effect on human health, leading to [15] premature death, heart problems and even lung cancer. The Air Quality Index (AQI) is a measure based on the concentration of many pollutants in the atmosphere to define the level of air quality. $PM_{2.5}$, PM_{10}, sulphurdioxide, carbon monoxide, nitrogen dioxide, and ozone are the most common pollutants. Among which, $PM_{2.5}$ is the most dangerous pollutant containing fine particles of diameter smaller than 2.5 micrometers. These fine particles can enter easily into the bloodstream and causes severe health hazards[24]. Fig. 1.shows $PM_{2.5}$ is the highest pollutant.

Recent studies reveals that air particles have found their way directly to the placenta through lungs and can reach the foetus of the mother.

Regression algorithms [22] of machine learning play an important role in data extraction and find hidden knowledge helpful in predictive analysis. Recent researches focused on machine learning ensemble models[18] for more accurate prediction. This paper focuses on enhanced stacked ensembling technique for the prediction of $PM_{2.5}$ concentration levels in air.

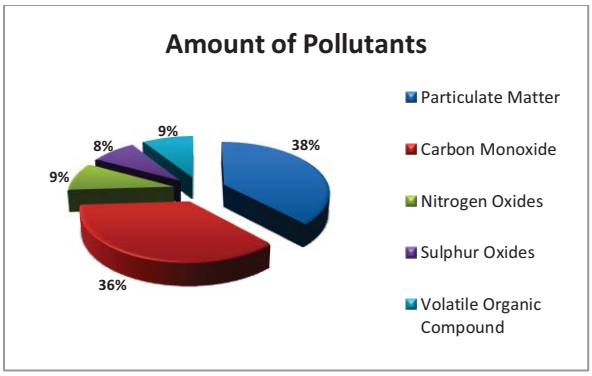

Fig. 1. Major air pollutants

II. RELATED WORK

Several machine learning approaches [2] have been suggested based on simple regression models to solve air pollution prediction problems. In [1] the author compared four machine learning algorithm for predicting air quality and found with random forest model a high precision is achieved. The limitation is that the random forest performs well with classification problems only. Authors in [26] used K-means clustering for air pollution prediction with sensor data but the anomaly detection is unexplored.LSTM model [6] is also used for $PM_{2.5}$ prediction but it always results in overfitting of model. Zhu[17] have used regularization and optimization technique for air quality prediction but the dataset used is of

limited generality. Zhou, H. Jiang[18] suggested a model using a combination of several neural networks to forecast air quality and concluded that the random forest model showed better results but the dataset is too small for multi layer network. Ian Dia[11] used an heterogeneous ensemble model that integrated multiple machine learning algorithms and predictor variables to estimate daily $PM_{2.5}$. The author in [29] predicted only the ozone layer factor using deep learning concepts. Hui Liu [14] proposed an ensemble forecasting model that uses the Hampel identifier to detect and correct the outliers in the original series and has been shown in the research that the model proposed in the study has better accuracy and wider applicability compared to the current models. In [28] authors developed ensemble model for $PM_{2.5}$ prediction using LASSO,Adaboost and XGB as base learners were used but our model enhanced stacking ensemble achieves higher accuracy.

III. METHODOLOGY

The flow of prediction of the proposed Stacked boost model is shown in the Fig. 2.The process starts with preprocessing the data and then the preprocessed dataset is used to train the boosting regression models namely Random forest,GBM and XGBoost as base learners.

The output of first level base learners is then combined and is used to train the second level model LightGBM which act as meta estimator for obtaining better accuracy.

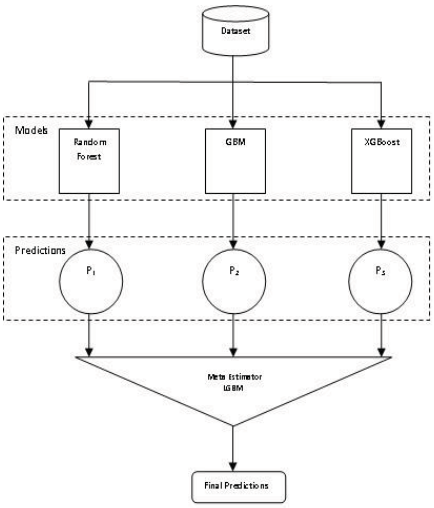

Fig. 2. Stacked Boost Model

A. Pre-Processing

The dataset taken for the research is taken from kaggle which contains air quality sensor data captured from monitoring station in Delhi city from 2017 to 2019 using pollutant sensor nodes. The dataset includes 13 independent attributes contributing to the final estimation of target variable. In order to manage the missing values, irrelevant data pre-processing steps are used and thus suitable for implementing machine learning algorithms to achieve the high accuracy.

B. Cleaning

The dataset contains sensor data of pollutants may have outliers or anomalies[17] which are extreme or unusual values. Fig. 3 shows the outliers of the attributes.These were removed using z-score method for improved accuracy of result.

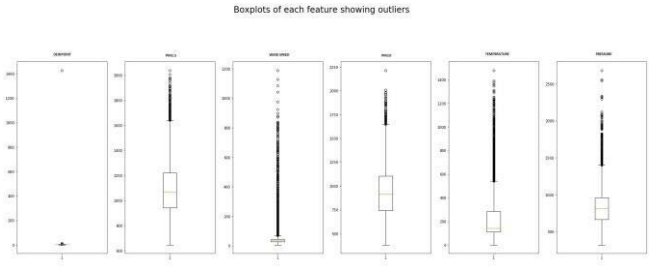

Fig. 3. Features with outliers

In this research, z-score method is used for detecting and removing outliers. The z_score value indicates the deviation of datapoint far from mean and that flags outlier.

$$z = |x-m(x)|/s(x) \quad (1)$$

where z=z_score
m=mean
s = standard deviation

C. Estimation Models

1) Random forest regressor

Random forest is an ensemble model that does assumptions based on a series of base models. The difference of the two values is used to test the efficiency of the regression model.Initially,the $PM_{2.5}$ estimation was forecasted using the regression models Random forest, Adaboost, GBM, XGB, Light GBM models. As Light GBM model prediction result is comparatively better this model is used as meta regressor in our proposed Stacked Boost

This can be expressed as

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots f_k(x) \quad (2)$$

where f_k represents k number of regressors.Each base model is built from a sample drawn with replacement from the training set.

2) GBM

Gradient Boosting or GBM is another ensemble regression machine learning algorithm that incorporates a set of weak learners to build a strong learner. i.e In order to produce the final predictions, it combines the predictions from different decision trees. In addition, the errors produced by the previous trees are taken into account in each new tree. Thus,each successive decision tree, is therefore based on the mistakes of the previous trees.

3) *XGBOOST*

An advanced implementation of gradient-boosted decision trees is XGBoost for extreme gradient boosting. This model works by iteratively adding models in which the error of previous model are rectified by the next model ensembled till the training set is predicted correctly. XGBoost is a high power predictive model,thereby increasing the efficiency of the model.

4) *Stacked Boost Model*

Stacking is an ensemble technique that combines multiple predictions generated by using different learning algorithms on a single dataset. This technique consists of two phases, with a set of base-level regressors as the first phase regressors and meta-level regressor in the second phase, which learns by combining the outputs from the base-level regressors.

Our model uses stacking of random forest regressor,GBM and XGBoost regressor as the base level regressors and LightGBM as meta regressors that selects first level regression algorithms and the output predictions of first level regressors are fed into second level regressor as input then the meta regressor is fitted using the training set data. Optimization is based on the minimization of least square errors.

IV. Experiment

The research dataset is taken from kaggle contribute to the estimate of PM2.5 concentration level in air. Fig. 4 Shows the correlation of attributes using Heatmap which helps to perform feature of importance with target variable.Heatmap is plotted using the matplot library.

The dataset was splitted as 80% of training set and 20% of test set. The training set trains the model and the prediction efficiency of model is determined by the test set data. The same dataset is used by the other models for prediction and the accuracy is thus compared with the proposed model.

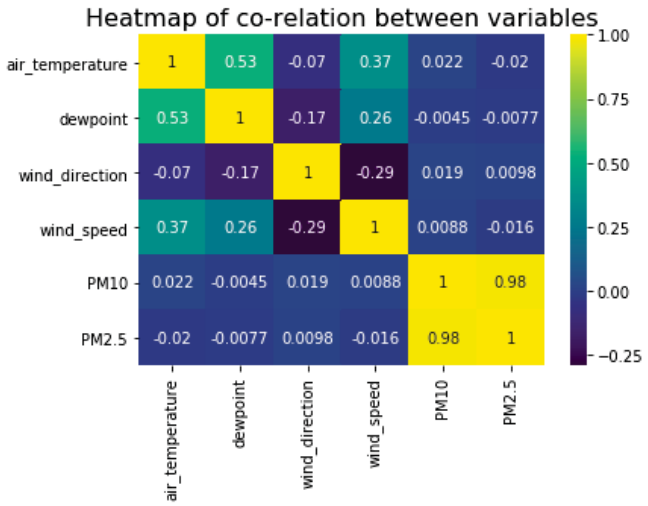

Fig. 4. Heatmap of interrelated features

V. Implementation

The experiments are performed in the jupyter IDE and coded in python 3.0. The scikit-learn libraries were also used to make predictions.For the implementation of our stacking boosting ensemble model with three base regressors random forest ,GBM and XGBoost.At the next level LightGBM is used as meta regressors.The meta regressor does the final prediction by combining the prediction of base level regressors.

VI. Results

The following metrics, Root Mean Squared Error(RMSE),Mean Absolute Error(MAE) and R- Squared are used to evaluate the regression model.

Mean Square Error is used to calculate Root Mean Squared Error which is the square root of Mean Square Error(MSE), where MSE is the difference between actual observed and predicted values of model.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \quad (3)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad (4)$$

Better the model with lesser value of RMSE.Perfect fit is indicated by zero value,since the of difference value 0 indicates the predicted and actual value are the same.

Where y_i is the actual value

$\hat{y}$ is the predicted value and N is the number of data points.

The Mean Absolute Error represents the difference between original and predicted values from the absolute difference of entire data.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad (5)$$

Another metric is R-square used to define the accuracy based on the regression line produced.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (6)$$

Table 1.1 Comparison of Metrics of Stacked Boost Model with other Regression Models

Model Name	MAE	RMSE	R-Square
Random Forest	0.6124	0.1144	0.2033
AdaBoost	0.6075	0.1321	0.2062
GBM	0.5942	0.1206	0.2067
XGB	0.5911	0.1232	0.2276
Light GBM	0.5863	0.1128	0.2293
STBoost	0.5768	0.1048	0.2346

i.e.From the fitted regression line the deviation is calculated for the final results which forms the R-Squared metric.The performance of the model is better with higher the values.

Therefore lower values of RMSE,MSE and MAE cause higher accuracy in the regression model. However, a higher value of R square is considered desirable.

Therefore lower values of RMSE,MSE and MAE cause higher accuracy in the regression model. However, a higher value of R square is considered desirable.

Table 1.2 Accuracy of Regression Models

Regression Models	Accuracy
Random Forest	0.967128631
AdaBoost	0.973290343
GBM	0.985684595
XGB	0.974326053
Light GBM	0.986516156
STBoost	0.995218792

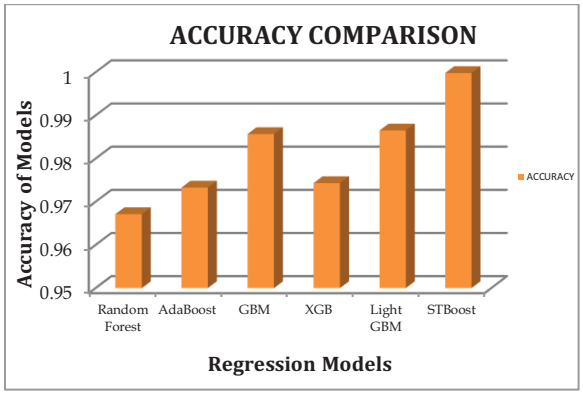

Figure 5. Accuracy of Models

VII. Conclusion

This study performs prediction of $PM_{2.5}$ using stacked boosting ensemble model that defines uncertainties of IoT acquired sensor data using z-score optimization technique and focuses on the concept of predicting $PM_{2.5}$ from environmental sensor data streams.This approach stacks multiple boosting algorithms such as Random forest algorithm, XGBoost regression and GBM regression as the base level regressor and Light GBM as the meta regressor thus improves the overall prediction accuracy.The results from the table and chart shows that the optimal solution is achieved by Stacked Boosting ensemble model with 0.1048 Root Mean Square Error(RMSE) value and Accuracy score of 99.52.

VIII. FUTURE SCOPE

Future work is to increase the novel technique efficiency using other ensembling methods.Also the other factors that affects the air pollution can be studied.Further the ensembling technique can be extended to detect and remove outliers.

The future scope can be further extended to forecast air quality based on images using neural network convolutional deep learning concepts.

References

[1] Saba Ameer, Munam Ali Shah, AbidKhan,"Comparative analysis of machine learning techniques for predicting air quality in smart cities",IEEE 2019.

[2] C. A. Keller, M. J. Evans, J. N. Kutz, and S. Pawson, "Machine learning and air quality modeling," *IEEE ,2017 pp*. 4570-4576.

[3] Gongbo Chen, Shanshan,"A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information",Elsevier 2018, pp.52-60.

[4] Yi Lin ,Long,"Air quality forecasting based on cloud model granulation",Springer,2018.

[5] SachitMahajan,Liu,"Improving the accuracy and efficiency of PM2.5 Forecast Service Using Cluster-Based Hybrid Neural Network Model",IEEE,2017.

[6] Yi-Chung Chen ,Dong-Chi Li," Selection of key features for PM2.5 prediction using a wavelet model and RBF-LSTM" ,Springer 2020.

[7] Unjin Pak,JunM."Deep learning-based PM2.5 prediction considering the spatiotemporal correlations",Elsevier,2020.

[8] K. A. Delic,"On resilience of iot systems: The internet of things"vol.1,February, 2016.

[9] Y.Xing,Xu,Shi,Y.X. Lian,"The impact of PM2.5 on the human respiratory system",Journal of Thoracic Disease, vol. 8,pp.69–74, January 2016.

[10] Ping Wei Soh, Jia Wei Chang "Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations"IEEE,May 2018.

[11] Qian Dia, Heresh Amini "An Ensemble-based Model of PM2.5 Concentration across the Contiguous United States with High Spatiotemporal Resolution" Elsevier,2018

[12] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", ACM, August 2016.

[13] Fi-John Chang,Li-Chiu Chang,"Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques",Elsevier 2020.

[14] Hui Liu ,Yinan Xu , Chao Chen " Improved pollution forecasting hybrid algorithms based on the ensemble method" , Elsevier,2019,pp .473-476.

[15] 7 million premature deaths annually linked to air pollution." [Online]. Available: https://www.who.int/phe/eNews_63.pdf

[16] Asgari, Marjan, Mahdi ,"Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster", ACM 2017, pp.89-93.

[17] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization"

[18] Q. Zhou, H. Jiang, J.Wang, and J. Zhou, "A hybrid model for PM2:5 forecasting based on ensemble empirical mode decomposition and a general regression neural network,",Oct,2014,pp. 264-274

[19] Mahmudul Hasan , Md. Milon Islam ," Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches", Elsevier 2019

[20] Debry, E. Mallet, V. "Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM10", Atmos. Environ. 2014,pp.71–84.

[21] Wu, Q.; Lin, H. "A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors", Sci. Total. Environ. 2019, pp.808–821.

[22] Doreswamy,Harishkumar,Ibrahim,"Forecasting Air Pollution Particulate Matter ($PM_{2.5}$) Using Machine Learning Regression Models", Elsevier,2020.

[23] Zhou, Qingping, et al.,"A hybridmodel for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network" Sci. Total Environ.,pp.264–274.

[24] Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer, "The contribution of outdoor air pollution sources to premature mortality on a global scale," *Nature*, vol. 525, 69, p. 367, 2015.

[25] Z. Niu, S. Shi, J. Sun, and X. He, "A survey of outlier detection methodologies and their applications," in *Proc. Int. Conf. Artif. Intell. Comput.*I*ntell.* Berlin, Germany: Springer, 2011, pp. 380_387.

[26] G. R. Kingsy, R. Manimegalai, D. M. S. Geetha, S. Rajathi, K. Usha,and B. N. Raabiathul, "Air pollution analysis using enhanced K-means clustering algorithm for real time sensor data," in *Proc. IEEE Region Conf.,* Nov. 2016, pp. 1945_1949.

[27] T. G. Dietterich: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1-15. Springer, 2000.

[28] Binxu Zhai ,Jianguo Chen, "Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China", Elsevier,2018,pp.644-658.

[29] A. B. Ishak, M. B. Daoud, and A. Trabelsi, ``Ozone concentration forecastingusing statistical learning approaches," *J. Mater. Environ. Sci.*, vol. 8, no. 12, pp. 4532_4543, 2017.

[30] Hui Liu, Chao Chen "Prediction Of Outdoor PM2.5 Concentrations Based On A Three-Stage Hybrid Neural Network Model"Elsevier 2020.