International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.148

Air Quality Prediction Using Stacking

Anjali.S
Department of Data Science and Business SystemsSRM IST
SRM Nagar, Kattankulathur, 603203, Tamil Nadu, India.
as8408@srmist.edu.in

SakchiChoudhary
Department of Data Science and Business SystemsSRM IST
SRM Nagar, Kattankulathur, 603203, Tamil Nadu, India.
sc9074@srmist.edu.in

Dr. Kanmani P
Department of Data Science and Business SystemsSRM IST
SRM Nagar, Kattankulathur, 603203, Tamil Nadu, India
kanmanip@srmist.edu.in

***Abstract*—The atmosphere is the most important resource required for the survival of all living beings including plants. The air quality is depleting at a faster pace due to growing industries, construction of houses and the excessive use of vehicles due to the rapid growth of population. Air quality is described as to be the degree which the air is clean in a particular area.As India grows in areas of industrialization and urbanization, the air quality is also slowly depleting due to the emission of harmful gases and particles by industries and vehicles. Every household has at least one car and this also adds up to the problem of depleting air quality in different areas. The aim of this study is to apply a Stacking Ensemble Model in order to make predictions about the air quality of a specific location in India, Stacking Ensemble Model is built on regression models such as Gaussian Process model, Support Vector Regressor Model, Stochastic Gradient Descent Regressor model and the AdaBoostRegressor Model.**

Keywords—Air Quality, Prediction, Gaussian Process Regressor, Support Vector Regressor, Stochastic Gradient Descent Regressor, Stacking Regressor, Ensemble Technique.

I.INTRODUCTION

The earth's atmosphere is comprised of gases and is an essential component of the planet's physical system. Nitrogen and oxygen make up majority of the atmosphere at sea level, accounting for 78% and 21%, respectively. Other gases, such as carbon dioxide, hydrogen, helium, and noble gases, are present in small quantities. The atmosphere serves to protect life on earth by regulating temperature, moderating the differences between day and night, and shielding the planet from harmful UV radiation.

Air pollution is a significant issue for many nations throughout the world, and it significantly contributes to the rise in health issues such as breathing difficulties, heart rhythms, lung dysfunction, the emergence of cardiovascular and respiratory disorders, etc. Particulate matter is made up of very minute liquid and solid particles that are suspended in the atmosphere.

PM2.5, in contrast to other pollutants, is particulate matter made up of particles with a diameter of no more than 2.5 micrometres. These specks are so small that they can get inside the nose, lungs, and the bloodstream.

Urban areas are majorly affected by air pollution as they are surrounded by industries, construction premises and lot of vehicles on the road. Industries release a lot of smoke and harmful gases into the atmosphere that it becomes dangerous for residents living nearby to breathe the air as it will affect their health adversely. Cities are growing in population as they are developing better, and residents are having a better life in cities. Due to this growth, there is a need for more space to build houses and this leads to increase in deforestation. Trees play a key role in purifying the atmosphere but these air purifiers are removed which increases the chance of air pollution. Cars, trucks and motorbikes release a lot of smoke in the air that it becomes difficult to breathe the air.

Predicting the air quality is very important as it will alert us about the possibility that we might be breathing harmful air and we need to be aware of it. It also helps us to understand the need for protecting the air so that we will have clean air to breathe in the near future. Stacking is an ensemble technique that teaches the model how to mix predictions from Base models and meta-models to create a final model with accurate predictions.

II.MOTIVATION

The need of the hour is to be able to breathe good quality air. All parts of the world is facing the problem of air pollution and it needs to be reduced so that there will be safe and clean air to breathe in the future. It is very important to create awareness and the need to treat this natural resource with great care so that it will sustain for the future. We have used a few regression models and to improve their accuracy we have built it with the help of the stacking algorithm.

III.RELATED WORK

The researchers employed two machine learning models, that's ANN and GPR to forecast the quality of air in six Indian cities. The R-values obtained for ANN and GPR were 0.96 and 0.98, respectively. The GPR model outperformed the ANN model with lower RMSE, MAPE, and MAE values of 21.40, 7.89%, and 13.58, respectively. The results suggest that GPR is a promising approach for predicting air quality in Indian cities [1].An exploratory data analysis is conducted to reveal new insights, identify concealed trends, and identify harmful substances that may affect air quality directly. The authors employed five different models namely, KNN, GNB,SVM, Random Forest and XGBoost to predict the air

DOI: rp-9788770040723.148
quality and concluded that Out of all the models tested, the XGBoost model had the highest performance, achieving an R-squared score of 0.83. This result was better than any other model evaluated.[2].A comparison analysis on how air quality has affected regional and national lockdowns. After conducting the research, it was discovered that during the nationwide lockdown, air quality improved by 33% in industrial areas, 41% in commercial areas, and 15% in residential zones. However, during the regional lockdown, these rates increased to 53%, 46%, and 43%, respectively [3]. A new method was proposed that employs both the Genetic Algorithmand LSTM deep learning algorithm. The results of this method were highly accurate, with an RMSE of 9.58 when forecasting air quality. [4]. Forecasting the air quality can be achieved by combiningLSTMand GRU deep learning algorithms. The effectiveness of this LSTM-GRU fusion was evaluated by comparing it with various other models such as KNN, Linear Regression, SVM, LSTM, and GRU.After comparing their performances, the hybrid model was found to be more effective than the independent models, producing a mean absolute error (MAE) value of 36.11 and an R-squared (R2) value of 0.84. [5]. An analysis of pollutant emissions in China, using a roadmap of carbon neutrality and evolution of clean air policies. The researchers utilized an air quality model to simulate the levels of O3and PM2.5 pollutants at the regional and national levels for three specific years: 2030 (the year of carbon peak), 2035 (in line with the "Beautiful China 2035" initiative), and 2060 (the year of carbon neutrality).The study found that in each of these years, emissions of main PM2, NOx, SO2 and VOCs are expected to decrease by 44%, 42%, 42%, and 28%, respectively [6].A study on the levels of four major pollutants, namely PM2.5, PM10, O3 and NO2 in Baghdad before the lockdown withfour partial intervals and total lockdown from March to July 2020. It was found that during the first partial and whole lockdown from March to mid of April,concentrations of PM2.5, PM10, NO2 decreased by 8%,15% and 6% respectively whereas concentrations of O3 increased by 13% when compared to the levels before lockdown [7]. To achieve accurate PM2.5 value predictions, a framework that combines cloud-based and edge-based methods was proposed. The framework was evaluated using real-world data from air quality sensors in Calgary, Canada, with both original and pre-processed data used to assess prediction quality. Results indicate that the framework improved prediction accuracy by an average of 40.18% on Mean Absolute Percentage Error [8]. A novel approach called CT-LSTM was proposed, which combines chi-square testwith theLSTMmodel to improve prediction accuracy. The results showed that this approach achieved a 93.7% accuracy rate. [9] A novel technique to determine the level of contaminated air present in a particular area. The method involves comparing four distinct architectural designs and incorporating weather data into the photos to enhance their classification accuracy. To overcome the problem of class imbalance, the suggested method uses generative adversarial networks and data augmentation techniques. The experiment showed that the proposed approach was able to achieve a strong accuracy of about 0.88 [10].The variations in the quality of air in New York that resulted from COVID- 19 shutdown measures. The research study was conducted to measure the concentration levels of two distinct pollutants, namely NO2 and PM2.5, in the air. The study required the gathering of daily information from 15 central surveillance sites located in the five boroughs from January to May for the years 2015 to 2020. The scientists employed a linear time lag model to evaluate pollutant levels in the present period against those observed in 2020.The findings revealed there was no significant difference between the two years; however, a reduction in concentration of PM2.5 by 36% and concentration of NO2 by 51% was observed shortly after the shutdown [11]. A new model that combines LSTM with data from neraby sources of pollution and individual health profiles has been proposed. The model was integrated into a tool known as My Air Quality Index (MyAQI), which was tested in a real-life scenario in Melbourne Urban Area. The outcomes from the MyAQI tool demonstrated precision levels of approximately 90-96%. [12]. Researchers developed a new deep learning method called Aggregated LSTM (ALSTM) using the LSTM architecture. They evaluated the performance of the proposed model against SVR and GBTR through several experiments. The findings indicate that the ALSTM model significantly outperformed the other models, achieving a higher accuracy in prediction, as evidenced by the lowerRMSE value of 3.94. [13]. A proposal was made to represent air quality data in a three-dimensional format and introduce a comprehensive deep learning model that employs spatiotemporal collaborative approach. CNN and LSTM was combined to forecast regional air quality, considering various aspects of air quality. When compared with other neural networks, the proposed model demonstrated superior adaptability to these aspects, resulting in more precise air quality predictions [14]. A potential solution to air quality forecasting involves utilizing deep learning techniques to estimate hourly levels of air pollutants such as ozone, PM2.5, and sulphur dioxide. In this study, a CNN model was employed, and the outcomes demonstrated promising performance in air quality forecasting, with a RMSE of 6.07 for the training data. [15]. A novel approach utilizing a transferred bi-directional long short-term memory (TL-BLSTM) model was proposed to predict the quality of air. The effectiveness of this framework was evaluated through a case study conducted in Guangdong, China. Comparative analysis with commonly used machine learning algorithms revealed that the TL-BLSTM model yielded better results specifically for higher temporal resolutions. [16]. A novel method was developed using the LightGBM model to forecast PM2.5 levels for 35 air quality monitoring stations in Beijing. The results of the study demonstrate that the proposed model outperforms other existing methods. [17].

IV. Experiment

A. About the Dataset

The data set contains information on different locations, recorded on a daily basis. Columns include station code,

DOI: rp-9788770040723.148
sampling date, state, location, agency, type, SO2, NO2, RSPM, SPM, location monitoring station, PM2.5, and date.

B. Methodology

First, the required libraries are imported. Then, all the unnecessary columns, including agency, station code, date, sample date, and location monitoring station, are removed from the dataset. The values for Sulphur Dioxide Index, Nitrogen Dioxide Index, Respirable Suspended Particle Matter Index, and Suspended Particle Matter Index are then calculated. The Air Quality Index (AQI) of each data value is then determined.

Secondly, we define Gaussian Process Regressor, Support Vector Machine Regressor, Stochastic Gradient Descent Regressor as the Basis Estimator and Elastic Net as the final estimator to fit into our stacking model.

Next, we define the Stacking regressor using the base estimator and the MLP regressor as the final estimator. As the MLP Regressor in this instance, we have selected the Ada boost Regressor.

In conclusion, the ensemble technique is utilized to train both the base estimators and final estimators. We then use this technique to make predictions and calculate the R-Squared values for both the training and testing datasets.

We have used the following formula to calculate the Index of a particular pollutant:

$$Ip = [IHi – ILo / BPHi – BPLo] (Cp – BPLo) + ILo$$

where,

Ip = index of pollutant p
Cp = truncated concentration of pollutant p
BPHi = concentration breakpoint of pollutant i.e. greater than or equal to Cp
BPLo = concentration breakpoint of pollutant i.e. less than or equal to Cp
IHi = AQI value corresponding to BPHi
ILo = AQI value corresponding to BPLo

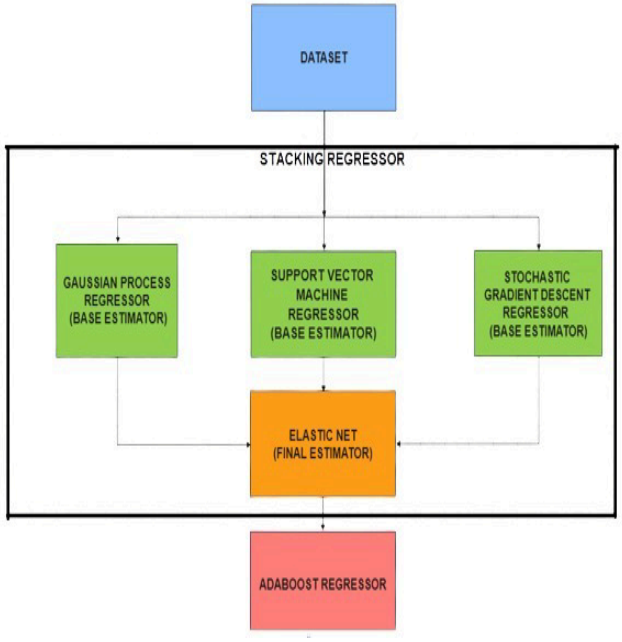

Fig. 1 Methodological Structure of Stacking Regressor

C. Model Architecture

The model is constructed with 5 different models and each model plays an important role in predicting the final outcome. The models used are:

1. Gaussian Process Model:

A widely used probabilistic supervised machine learning model for regression and classification tasks is the Gaussian process model. A Gaussian processes regression (GPR) model that incorporates previous knowledge (kernels) and provides uncertainty measurements for forecasts can create predictions. The regression model is utilised for extrapolation, interpolation, and forecasting of data. They are employed in a wide range of industries, including banking, robotics, and biology.

2. Support Vector Regressor Model:

In order to predict discrete values, a type of supervised learning called support vector regression utilizes a method similar to SVMs. The key principle of SVR is to find the hyperplane or line that best fits the data points. This approach is especially useful because it is capable of handling outliers, noise, and high-dimensional data. Various industries, such as finance, medicine, and energy consumption, have successfully utilized SVR for their predictive needs.

Another model used for regression is the Stochastic Gradient Descent Regressor. This model works by iteratively adjusting the parameters to minimize the loss function, allowing it to optimize the model for better predictions.

3. Stochastic Gradient Descent

DOI: rp-9788770040723.148

Stochastic Gradient Descent Regressor is a popular optimization algorithm that solves regression problems. It is a type of gradient descent algorithm that works well with large datasets as it updates the model parameters iteratively based on gradient loss function that is implemented on small subsets of training data called mini-batches. It is widely used in Natural Language Processing (NLP), Image and video processing, Marketing, etc.

4. ElasticNet Regression Model:

ElasticNet Regression Model, is a machine learning model that combines feature elimination from the Lasso regression model and feature coefficient reduction from the Ridge regression model to regularize regression models. This model is frequently employed to avoid overfitting regression models. It is very useful when there are many features in a dataset as it will automatically perform feature selection. It is majorly used in fields like Healthcare, Marketing, Bioinformatics and many more.

5. AdaBoostRegressor Model:

AdaBoostRegressor is an ensemble model widely used to solve regression problems. The basic idea behind AdaBoostRegressor is that it combines several weak learner algorithms and form a strong learner algorithm. In each iteration, it adds a new weak learner to the ensemble by giving more weight to the incorrectly predicted values from the previous iteration. It has uses in domains like finance, engineering, healthcare,etc.

To combine the predictions of Gaussian Process model, Support Vector Regressor model, Stochastic Gradient Descent Regressor model and ElasticNet model (base estimators), we used an ensemble technique known as Stacking Regressor that performs aggregation with AdaBoostRegressor (MLP Regressor). Stacking Regressor is mainly used to combine multiple regression models to improve their prediction accuracy.

This Stacking Regressor model takes two important parameters that is the base estimator and MLP Regressor. The base estimator takes the list of models that we will be aggregatingto get the final prediction. The MLP Regressor is the model which we will use as to do the aggregation process.

V. Results

The stacking model was built with five different models and the metric used to analyze the model's performance was the R-squared metric. The R-squared value is a statistical measure used to evaluate how well the model has fit the observed data. Its value always ranges between 0 and 1, where 0 indicates that the model is not fit well and 1 indicating otherwise. R-squared value can be calculated with the formula as given below:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where *RSS* is the sum of squares of residuals and *TSS* is the total sum of squares.

This stacking model showed a R-squared value of 0.96 for the train data and a R-squared value of 0.93 for test data, this indicates that the stacking model was able to fit the observed data points well and it is able to predict the values well.

VI. Conclusion

It is very important to understand the need to safeguard and preserve the air resource that we have at hand. India is growing its economy everyday with the gradual decrease of air quality due to need of industries, shelter to home the growing population and provide better facilities to cater to their needs such as schools, hospitals, etc. It is very important to analyze how the air quality will be in the near future and this analysis will help us to deliver solutions to build a better future for the generations to come. The Stacking algorithm used in this research is an ensemble model that aggregated a few profound regression models and it is found that the model is performing well with R-squared value of 0.96 for train data and 0.93 for test data.

References

[1] R.S.Suri, A.K. Jain, N.R. Kapoor, A. Kumar, H.C. Arora, K. Kumar, and H. Jahangir,"Air Quality Prediction-A Study Using Neural Network Based Approach," Journal of Soft Computing in Civil Engineering, vol. 7, no. 1, pp.93-113, 2023.

[2] K. Kumar, and B.P. Pande, "Air pollution prediction with machine learning: A case study of Indian cities," International Journal of Environmental Science and Technology, pp.1-16, 2022.

[3] P. Das, I. Mandal, S. Pal, S. Mahato, S. Talukdar, and S. Debanshi, "Comparing air quality during nationwide and regional lockdown in Mumbai Metropolitan City of India," Geocarto International, pp.1-26, 2022.

[4] G.I. Drewil, and R.J. Al-Bahadili,"Air pollution prediction using LSTM deep learning and metaheuristicsalgorithms," Measurement: Sensors, vol. 24, p.100546, 2022.

[5] N. Sarkar, R. Gupta, P.K. Keserwani, and M.C.Govil,"Air Quality Index prediction using an effective hybrid deep learning model," Environmental Pollution, vol. 315, p.120404, 2022.

[6] Dhanabalan, S. S., Sitharthan, R., Madurakavi, K., Thirumurugan, A., Rajesh, M., Avaninathan, S. R., & Carrasco, M. F. (2022). Flexible compact system for wearable health monitoring applications.Computers and Electrical Engineering, 102, 108130.

[7] B.M.Hashim, S.K.Al-Naseri, A. Al-Maliki,and N. Al-Ansari, "Impact of COVID-19 lockdown on NO2, O3, PM2. 5 and PM10 concentrations and assessing air quality changes in Baghdad," Iraq. Science of the Total Environment, vol. 754, p.141978, 2021.

[8] S. Ojagh, F. Cauteruccio, G. Terracina, and S.H. Liang, "Enhanced air quality prediction by edge-based spatiotemporal data preprocessing," Computers & Electrical Engineering, vol. 96, p.107572, 2021.

[9] J.Wang, J.Li, X. Wang, J. Wang,and M.Huang,"Air quality prediction using CT-LSTM," Neural Computing and Applications, vol. 33, pp.4779-4792, 2021.

[10] J.Kalajdjieski,E. Zdravevski, R. Corizzo, P. Lameski, S. Kalajdziski, I.M.Pires, N.M. Garcia, and V. Trajkovik, "Air pollution prediction with multi-modal data and deep neural networks," Remote Sensing, vol. 12, no. 24, p.4142, 2020.

[11] S.Zangari, D.T.Hill, A.T.Charette, and J.E.Mirowsky,"Air quality changes in New York City during the COVID-19 pandemic," Science of the Total Environment, vol. 742, p.140496, 2020.

DOI: rp-9788770040723.148
[12] D. Schürholz, S. Kubler, and A. Zaslavsky, “Artificial intelligence-enabled context-aware air quality prediction for smart cities,” Journal of Cleaner Production, vol. 271, p.121941, 2020.

[13] Y.S. Chang, H.T. Chiao, S. Abimannan, Y.P. Huang, Y.T. Tsai, and K.M. Lin, “An LSTM-based aggregated model for air pollution forecasting,” Atmospheric Pollution Research, vol. 11, no. 8, pp.1451-1463, 2020.

[14] G.Zhao, G.Huang, H.He, H. He,and J.Ren,“Regional spatiotemporal collaborative prediction model for air quality,” IEEE Access, vol. 7, pp.134903-134919, 2019.

[15] Y. Mao,and S.Lee,“Deep convolutional neural network for air quality prediction,” In Journal of Physics: Conference Series, IOP Publishing,vol. 1302, no. 3, p. 032046, August2019.

[16] Pazhani. A, A. J., Gunasekaran, P., Shanmuganathan, V., Lim, S., Madasamy, K., Manoharan, R., &Verma, A. (2022).Peer–Peer Communication Using Novel Slice Handover Algorithm for 5G Wireless Networks.Journal of Sensor and Actuator Networks, 11(4), 82.

[17] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang, and L. Huang, “A predictive data feature exploration-based air quality prediction approach,” IEEE Access, vol. 7, pp.30732-30743,2019.