PREDICTION OF AIR QUALITY INDEX USING MACHINE LEARNING

Khushi Sharma [1], Janhavi Diwedi [2], Manisha Tayal [3] , Prachi Sharma [4]
[1,2,3,4] Department of Computer Science and Engineering,
[1,2,3,4] ABES Engineering College, Ghaziabad
[1] khushi.21b1531035@abes.ac.in [2] janhavi.21b1531146@abes.ac.in [3] manisha.21b1531197@abes.ac.in
[4]prachi.21b1531058@abes.ac.in

Abstract— The quality of air, a vital natural resource, has been harmed by economic activities. Predictionof air pollution is not an exception to the constantly growing and permeating influence of machine learning technology in practically every industry. This study uses machine learning techniques to update previous research on predicting air quality index. The most significant papers were chosen using the most well-known databases. After carefully reading those papers, the key characteristics were distilled, servingas a foundation for linking and contrasting them with one another.

INTRODUCTION:

The quality of air, a vital natural resource, has been harmed by economic activities. Urbanization growth also contributes to several issues with various aspects of living, including air quality, transportation, and health care. Nowadays, people are more aware of air pollution because it has a substantial impact on both human health and biological balance. In addition to the effects of hazardous emissions on the environment, air pollution also affects health, productivity at work, and energy efficiency. Both human health and the natural balance are impacted by air pollution. Gas concentrations in the air have a significant impact on human health and can have dangerous consequences. Due to a rise in air pollutants, the amount of rain is also impacted. Since air pollution has several dangerous effects on individuals, it should be examined persistently with the intention of effectively controlling it. Knowing the source, force,and beginning point of air contamination is one method for controlling it.

According to the World Health Organization (WHO), air pollution causes over 1.3 million deaths worldwideeach year. Over the past few decades, there have also been more negative effects such acid rain, global warming, aerosol production, and photochemical smog. Ozone (O3), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Sulfur Dioxide (SO2), and Particulate Matter are the main causes of air pollution (PM). These gases, which are created by burning fossil fuels, wood, industrial boilers, and volcanic eruptions, cannot be seen or noticed. They have the potential to harm people and are mostly responsible for diseases including cancer, birth deformities, and breathing-related issues. Numerous researchers are looking at the underlying pollution-related factors causing COVID-19 pandemics in various countries as a result of the recent rapid spread of COVID-19. The WHO has released evidence that raises concerns about the extent of contamination across the nation. It becomes clear to us that the chance has already passed that we ought to screen the air.

The Air Quality Prediction model aims to work on the concentration of different pollutants like PM2.5, PM10, S02, N02, C0, benzene, toluene, and xylene as well as on the weather conditions that also affects the AQI, or the Air Quality Index, of a region scaling them to a range and defining whether it is healthy, satisfactory, moderate, or unhealthy for the region. After preprocessing the data and appropriately scalingthe data, various machine learning methods are used. The research discussed in this paper focuses on creating AQI prediction models for impending acute air pollution episodes. Investigated are the followingmachine learning (ML) algorithms: decision tree regression, lasso regression, and linear regression models. This study tracks the decline in forecast accuracy across a wider time span. A layperson would not understand the numerical data for the air quality index, which can range from 0 to 300+, even after ithas been predicted. Therefore, utilizing a Quality Check function, the AQI will be divided into 5 categories after being predicted. They are designated as "Healthy, moderate, "Unhealthy," "Very Unhealthy," and "Hazardous."

The research's purpose is accomplished using a variety of machine learning techniques, some of which are mentioned below.

LINEAR REGRESSION:

A machine learning approach called linear regression relies on supervised learning to carry out a regression task. For discovering the relationship between variables and forecasting, linear regression produces a target prediction value based on independent variables. Different regression models are beingexplored, and a list of independent variables is being employed, depending on the relationship between the established and the independent variables: y = mx+c.

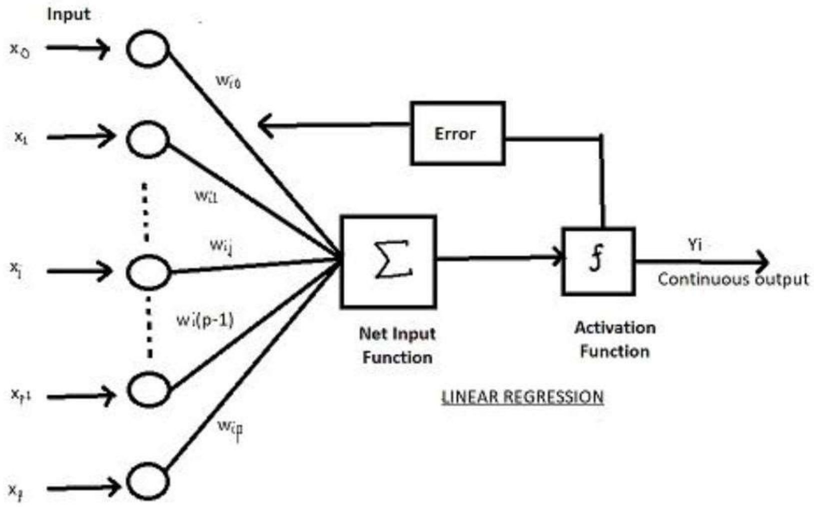

Figure: Illustration of a linear regression algorithm

DECISION TREE REGRESSION:

The Regression on the Decision Tree is both non-linear and non-continuous concept . It represents a function that accepts an attribute values vector as input and outputs a choice. The Supervised Learning group includes decision trees. Regression and classification issues can both be resolved with it. A decision tree decides something by carrying out a series of actions.

RANDOM FOREST REGRESSION:

In contrast to boosting, natural forest is a form of bagging. In a random wood, the trees are running parallel. While the trees are being installed, they are not in contact with one another

It accomplishes its tasks by building a large number of decision trees during the training phase and producing the class that corresponds to each tree's class mode (classification) or average prediction (regression). A random forest is a met estimator that aggregates numerous decision trees and makes someuseful improvements. It integrates the results of many forecasts.

A particular percentage of the total number of functions at each node that can be split on (known as the hyper parameter). Every tree uses a random sample from the data set from which it separates, adding stillanother element of randomization that avoids overfitting.

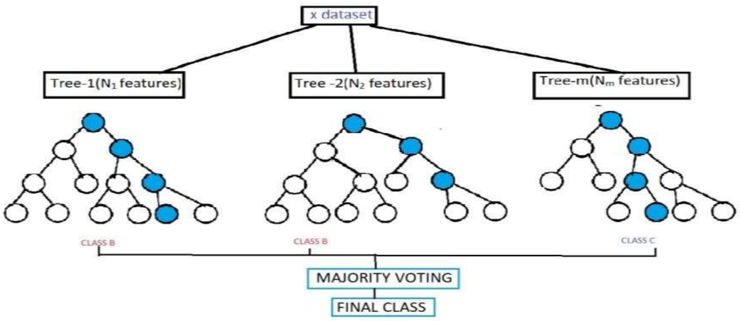

Figure: Illustration of a random forest algorithm

LASSO REGRESSION:

Shrinkage is used in the linear regression method known as lasso regression. When data values shrink toward a middle value, such as the mean, this is called shrinkage. The lasso technique promotes straight forward, minimal.

LITERATURE REVIEW:

Prediction of air pollution is not exempt from the sectors where machine learning technologies are havinga significant increase in influence and penetration. This study uses machine learning techniques to updateprevious research on predicting air pollution within the framework of smart cities, based on sensor data.The most pertinent papers were chosen using the most widely used databases and the corresponding filters. After carefully reading those papers, the key characteristics were distilled, serving as a foundation for linking and contrasting them with one another. C we can draw the following conclusions: 1) The authors currently use complex and advanced methodologies rather than simple machine learning techniques. 2) In terms of a case study, China was the best-performing nation. 3) The primary prediction target was particulate matter with a diameter of 2.5 micrometers. 4) In 41% of publications, authors made predictions for the following day. 5) 66% of studies used data with an hourly rate. 6) 49% of papers used open data, and since 2016 there has been an increase in this trend. 7) It's crucial to consider the externalaspects like the weather, regional characteristics, and temporal features for effective air quality forecast.

The studies employed a variety of machine learning approaches, including neural networks, regression, ensemble, hybrid, regularization and optimization in one study, and multinomial naive bayes and multinomial logistic regression in another. Due to the diverse data sets and temporal granularities utilizedin the analysis of the studies, it is generally very challenging to compare the results. They should be created and tested using the same datasets utilizing all the mentioned above ways. The outcomes could then be contrasted on a comparable and fair basis.

The quality of air, a vital natural resource, has been affected by economic activity. The ability to anticipateoccurrences of poor air quality has been extensively studied, but the majority of these studies are constrained by the lack of longitudinal data, making it challenging to take seasonal and other factors intoconsideration. On the basis of an 11-year dataset gathered by Taiwan's Environmental Protection Administration, several prediction models have been created (EPA). For predicting the level of the air quality index (AQI), machine learning techniques such adaptive boosting (AdaBoost), artificial neural networks (ANN), random forests, stacking ensembles, and support vector machines (SVM) show promise.The stacking ensemble consistently outperforms AdaBoost and random forest for R 2 and RMSE, but AdaBoost performs best for MAE, according to a series of studies employing datasets for three different areas to gain the greatest prediction performance from each.

The use of artificial intelligence techniques yields encouraging outcomes for AQI forecasting. This study used information gathered over an 11-year period by Taiwan's EPA and CWB. Those three areas are (North: Zhongli, Central: Taiwan's Changhua and Fengshan(South: Fengshan) were taken into consideration, as well as other locations known for having poor air quality all year round. Stacking ensemble and AdaBoost provide the best performance of target predictions based on three separate datasets, with good results for R 2. To be more precise, AdaBoost produces the best MAE results, whereasthe stacking ensemble produces the best RMSE results. All findings indicate that SVM produces the worst outcomes of all investigated approaches and only offers useful outcomes for 1-h predictions. The outcomes also demonstrate that the two machine learning techniquesused in this study—AdaBoost and stacking ensemble—can beat well-known techniques from the literature including SVM, random forest, and ANN. To put it another way, AdaBoost and stacking ensemblecan be thought of as fresh and better options for AQI forecast.

This study also shows that there are regional differences in Taiwan's prediction performance. The best findings are for the Fengshan AQI prediction (Southern Taiwan), where performance decay with increasedtime step is less noticeable than in Zhongli (North) and Changhua (central). Additionally, 1-hour, 8-hour, and 24-hour forecast 95% confidence intervals are calculated, accordingly. The 95% C.I. can give the decision-maker a better reference than the single value estimate.

To maintain excellent air quality, the air quality monitoring framework calculates various air pollutants atvarious locations. In the present context, it is the main problem. The introduction of hazardous gases into the environment from businesses, vehicle exhaust, and other sources taints the air. These days, air pollution has reached elementary levels and has gone beyond the standards for air quality set by the government in many large urban locations. It has a huge impact on a person's soundness. Due to advances in machine learning technology, it is now possible to predict toxins based on historical data. In this research, we describe a device that can take current poisons and, with the help of earlier poisons, executea computation based on ML to predict the information of contaminations in the future. The discovered data is retained inside the Excel sheet for later analysis. These sensors are used on the Arduino Uno platform to collect data about pollution.

By using unique formulas like direct relapse, Decision Tree, and Random Forest, we forecast the air qualitylist. We concluded from the results that the Random Forest approach provides a superior expectation ofthe list of air quality.

PROPOSED METHODOLOGY:

The steps in this study's methodology are data collection, preprocessing, feature selection, time windowing, and model development. The open-source data mining platform will be used to build all the machine learning models used in this study. Data was collected in order to train the algorithm to detect air quality. The intended attribute set for the data set included CO, SO2, O3, NO2, NO, benzene, toluene,and xylene. The system is trained using the meteorological data information set parameters of temperature, wind speed, humidity, and wind direction. The research's purpose is accomplished using a variety of machine learning techniques, but the best machine learning method, which yields the most accurate results, is **random forest regression.**

Random Forest Regression:

In contrast to boosting, natural forest is a form of bagging. In a random wood, the trees are running parallel. While the trees are being installed, they are not in contact with one another

It accomplishes its tasks by building a large number of decision trees during the training phase and producing the class that corresponds to each tree's class mode (classification) or average prediction (regression). A random forest is a met estimator that aggregates numerous decision trees and makes someuseful improvements. It integrates the results of many forecasts.

A particular percentage of the total number of functions at each node that can be split on (known as the hyper parameter). Every tree uses a random sample from the data set from which it separates, adding stillanother element of randomization that avoids overfitting.

The Random Forest is a collective term for ensemble methods utilizing tree-type classifiers, where is a produced classifier {h(x, θk), k=1 ,...,} where θk is an independent, identically distributed random vector, x is an input pattern. It generates several trees using recursive partitioning and then aggregates the outcomes. Using a bootstrap sample of the training data, each tree is individually built by first dividing theparameter set into many parts based on one of the parameters, then repeating the process for each part.

The unlabeled data are added to each decision tree after it has been created. The estimated probability of the AQI level I with each tree is denoted by p(ci)., where T is the number of decision trees as previouslydiscussed, defines the final probability of the AQI level i p'(ci) in the random forest:

$$P'(Ci) = 1/T \quad {}_{K=1}\Sigma^{\mathbf{1}'} P(Ci)$$

$$p'(c_i) = \frac{1}{T}\sum_{k=1}^{T} p(c_i)$$

The equation which determines the result is given by:

$$C'(i) = Max(p'(ci))$$

RESULT:

All the algorithms—random forest, decision tree, linear regression, AdaBoost, ANN perform well but the random forest regression achieves marginally greater accuracy.

Random Forest is a classifier that uses many decision trees on different subsets of the input dataset and averages the results to increase the dataset's predicted accuracy.

A very well supervised machine learning technique utilized in classification and regression challenges is Random Forest. This algorithm's ability to handle a dataset with continuous variables, like in the case of regression, is one of its key characteristics.

In comparison to other algorithms, it requires less training time. Even with the enormous dataset, it operates effectively and predicts the outcome with a high degree of accuracy. When a significant amountof data is absent, accuracy can still be maintained.

TYPES OF ALGORITHMS	PREDICTED PROBABILITY OF VARIOUS PARAMETERS								
	CO	O3	SO2	NO2	Benzene	Xylene	Toluene	PM 2.5	PM 10
LINEAR REGRESSION	0.05	0.07	0.04	0.11	0.02	0.08	0.06	0.04	0.03
DECISION TREE REGRESSION	0.59	0.65	0.57	0.60	0.63	0.76	0.54	0.80	0.57
RANDOM FOREST REGRESSION	0.80	0.75	0.72	0.71	0.86	0.89	0.69	0.89	0.78

Figure: Table showing predicted probability of various parameters using different machine learning algorithms

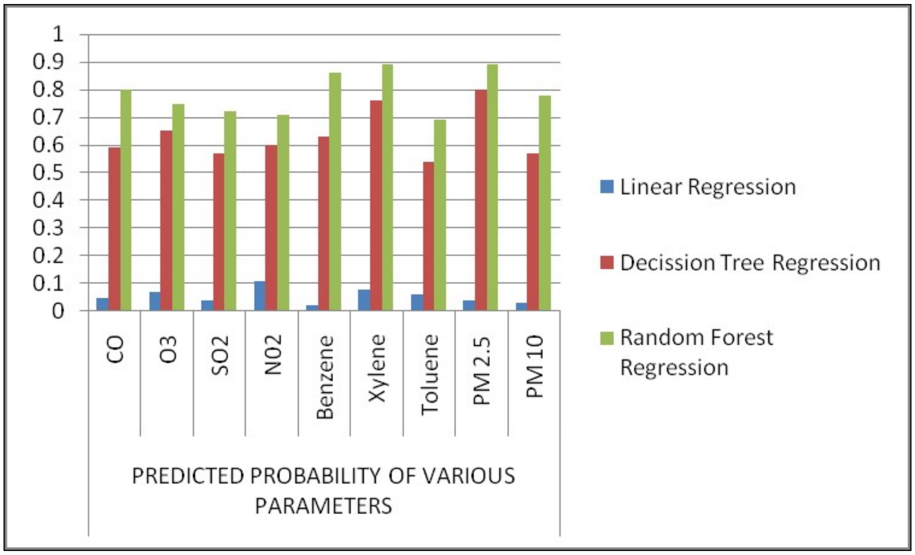

Figure: Predicted Probability of Various Parameters

CONCLUSION:

By examining previous papers, the goal of this work is to provide an overall impression of the current approaches to the idea of air quality prediction. The key discovery was that a dataset of additional elements that influence air quality should be included in addition to air quality data to improve the accuracy of air quality prediction. Due to the diverse data sets and temporal granularities utilised in the analysis of the studies, it is generally very challenging to compare the results. An exhaustive work is suggested as future effort. The outcomes could then be contrasted on a comparable and fair basis. With the use of special computations like linear, decision tree, and random forest, we forecast the air quality Index. We concluded from the results that the Random Forest algorithm provides a more accurate prediction of the air quality Index. In Random Forest Regression Less time is needed for training. It performs well and makes most accurate predictions about the outcome. Accuracy can be kept even whena sizable amount of data is unavailable.

REFERENCES:

1. Kennedy Okokpujie, Etinosa Noma-Osaghae, Odusami Modupe, Samuel John, and Oluga Oluwatosin, “A SMART AIR POLLUTION MONITORING SYSTEM,” International Journal of Civil Engineering and Technology (IJCIET), vol. 9, no. 9, pp. 799–809, Sep. 2018.

2. D. Zhu, C. Cai, T. Yang, and X. Zhou, “A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization,” Big Data and Cognitive Computing, vol. 2, no. 1, p. 5, Mar. 2018.

3. A. Masih, “Machine learning algorithms in air quality modeling,” Global Journal of Environmental Science and Management, vol. 5, no. 4, pp. 515–534, 2019.

4. Campbell-Lendrum, D., & Prüss-Ustün, A. (2018). Climate change, air pollution and non-communicable diseases. Bulletin of The World Health Organization, 97(2),160-161.

5. Random Forest Regression in Python - GeeksforGeeks. GeeksforGeeks. (2020). Retrieved 22 July 2020.

6. Decision Tree Regression — scikit-learn 0.23.1 documentation. Scikit-learn.org. (2020). Retrieved 22July 2020.

7. Rokach, Lior; Maimon, O. (2008). Data mining with decision trees : theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

8. Kostandina Veljanovska and Angel Dimoski, “Air Quality Index Prediction Using Simple Machine Learning Algorithms,” International Journal of Emerging Trends & Technology in Computer Science, vol. 7, no. 1, 2018.

9. Yi, X.; Zhang, J.; Wang, Z.; Li, T.; Zheng, Y. Deep Distributed Fusion Network for Air Quality Prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, UK, 19–23 August 2018; pp. 965–973.

10. Veljanovska, K.; Dimoski, A. Air Quality Index Prediction Using Simple Machine Learning Algorithms. Int. J. Emerg. Trends Technol. Comput. Sci. 2018, 7, 25–30.

11. Ghorani-Azam, A.; Riahi-Zanjani, B.; Balali-Mood, M. Effects of Air Pollution on Human Health and Practical Measures for Prevention in Iran. J. Res. Med. Sci. 2016, 21, 1–12.

12. Rocca, J. Ensemble Methods: Bagging, Boostingand Stacking. Available online: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (accessed on 23 April 2019).

13. Awad, M.; Khanna, R. Support Vector Regression. In Ecient Learning Machines; Apress: Berkeley, CA, USA, 2015.