Juni Khyat
(UGC Care Group I Listed Journal)
ISSN: 2278-4632
Vol-13, Issue-04, March 2023

AN ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR FORECASTING AIR POLLUTION

[#1]**V. SUBBA RAMAIAH,** Assistant Professor,
Department of Computer Science & Engineering,
Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India.

[#2]**Dr. DABBU MURALI,** *Professor,*
Department of Computer Science & Engineering,
Vivekananda Institute Of Technology & Science, Karimnagar, Ts, India.

ABSTRACT: Air pollution is one of the most important environmental concerns that we face today since it poses a threat to people's health as well as the standard of living that they enjoy. Environmental scientists and politicians have come to recognize the importance of conducting research to determine how to forecast future levels of air pollution. Our findings lend credence to the utilization of machine learning as a method for forecasting levels of air pollution, a field that requires additional research. For the purpose of this study, data on weather patterns and measurements of the amount of pollution in the air were gathered over a specific amount of time. Several different regression models were utilized in order to generate forecasts of ozone concentrations in the atmosphere. Models such as Linear Regression, Decision Trees, and Random Forest were among those that were discussed throughout the presentation of the various models.

Keywords: *Machine Learning, Air Pollution, Air dispersion models.*

1. INTRODUCTION

According to the data presented on the Urban Population website, urban areas were home to 56.15 percent of the world's population in the year 2020. According to projections made by the United Nations, by the year 2050, metropolitan regions will be home to 68 percent of the world's inhabitants. As a consequence of these demographic shifts, it is probable that issues relating to public health, transportation, and the quality of the air will become more severe. Air pollution can make it difficult to breathe, has been related to death at an earlier age, and can raise the likelihood that someone will need to be hospitalized for heart and lung problems. A prolonged exposure to air pollution has a detrimental effect on plant leaves, just as it does on human lungs; however, the damage caused to plant leaves is significantly more severe. Dust, PM10 particles with a diameter of less than 10 meters, and PM2.5 particles, which are the most hazardous since their diameter is smaller than PM2.5 microns, are produced when fuel is not completely burned, and process byproducts are also produced. Another key factor that contributes to poor air quality is sulfur dioxide (SO2), which is produced as a byproduct of combustion. Particles having a diameter of less than 10 meters and PM2.5 microns are the most dangerous, followed by PM10 dust and smaller particles. When fossil fuels are burned, three of the most dangerous pollutants are produced. These pollutants include carbon monoxide (CO), nitrogen oxides (NOx), and ozone (O3).

It is necessary to have accurate estimates of the levels of pollution in order to communicate productively with governments and educate the general public about the dangers posed by air pollution. It is common practice to define data on air pollution using a variety of different patterns, including rising and falling trends, seasonality (the variability in time series within a specific time period), cycles (rises and falls that are not specified in time), and erratic movement.

Deep learning can be applied to a variety of problems, including those involving nonlinear, cyclical, seasonal, or sequential dependence between different types of pollutant data. One of these problems is the prediction of air pollution. The structure of deep learning lends itself well to tackling problems of this nature. In contrast to the shallow ANN design, the LSTM approach does an excellent job of retaining dependencies across time as it learns new time series. This is in contrast to the shallow ANN architecture. In contrast to that, this example demonstrates a deep neural network. Because it has its own memory, an LSTM is an excellent choice for solving issues that call for sequence-dependent behavior, such the monitoring of pollutants. To put it another way, the level of pollution that is estimated for each gas is taken from past data in which a pattern of behavior that is comparable was seen.

When developing a deep learning model, one of the many procedures that must be completed is the determination of the hyper parameters for the LSTM model. Because these hyper parameters are handled simultaneously with over fitting, the decision that is taken about them influences the phase of the model that is dedicated to training. In most cases, selecting hyper parameters is a laborious and drawn-out process that takes a lot of time. When attempting to acquire the ideal hyper parameters, the objective is to achieve the lowest possible error rate. It is possible to

ISSN: 2278-4632
Vol-13, Issue-04, March 2023

personalize numerous features of the model, such as the number of layers, the number of cells per layer, the number of units, the batch size, and the kind of activation. Among the other things that may be modified is the batch size.

This work aims to find a near-ideal model for predicting air pollution using the Metaheuristics method, which is known for its ability to locate near-ideal solutions in large areas; to solve the problem of selecting hyper parameters; and to predict air pollution. Given the importance of the window size and the number of LSTM units in the LSTM inputs, this work aims to find a near-ideal model for predicting air pollution using the Metaheuristics method.

Within the scope of this investigation, our objective was to improve the performance of the LSTM model by utilizing the Genetic Algorithm (GA), a well-known optimization technique. After tomorrow, the LSTM model will be trained with GA to estimate the concentrations of a number of air pollutants (PM2.5, PM10, CO, and NOx), as well as the appropriate window size and LSTM unit (PM2.5, PM10, CO, NOx).

2. USING MACHINE LEARNING MODELS

LINEAR REGRESSION:

Linear regression, to put it another way, is a technique for completing regressions that is used in supervised machine learning. Data analysis employing this technique. This procedure is used to do the necessary regressions. Linear regression, which generates a value for target prediction based on the independent variables, is widely used for establishing causality between the variables and for making predictions. Predictions can be useful in establishing the connection between the variables. This method sees regular use for the aforementioned purpose. In this piece, we'll look at one of the more common uses of linear regression. The list of independent variables to be explored and the type of regression model to be used will be determined by the nature of the link that exists between the variables that are regarded to be dependent and those that are considered to be independent. Both choices will be determined by the nature of the link.

$y = mx+c$

As will be illustrated in the next statement, the equation just stated is representative of both the input training data (x) and the output labels (y) (input parameter).

During the prediction-generation phase of model training, the variable x is used to make predictions about the variable y, and the best solution is the line that produces the best match.

c = intercept m = slope of line

Building the most accurate possible fit line requires optimizing both the m esteem and the c esteem to their maximum values. Our model will function in this way and be able to predict the estimation of y for the information estimation of x when we put it to use for anticipating. When we finally put our model to use for forecasting, we will find this to be the case. We will confirm this when we put our model to use for forecasting in the future.

DECISION TREE:

Due to its inability to produce continuous results, the Regression on the Decision Tree cannot be used as a linear model. Results in a linear model must be continuous. This occurs because it is essential for linear models to produce continuous results. It's the label given to a process that accepts a set of attribute values (or "attribute vector") and returns a logical conclusion. This function takes its input from a list of attribute values.

Supervised learning strategies include the decision tree algorithm, which is more often known as the decision tree. This is what we mean when we talk about "classical" learning. It is likely to be useful in solving problems associated with regression and classification. A decision tree will eventually reach a conclusion after carrying out a sequence of actions in a prescribed order.

RANDOM FOREST:

Increasing the quantity of bags harvested is more common in natural forest management than improving productivity. In random forests, tree development tends to occur in parallel, perpendicular lines that all point in the same direction. Throughout the entire planting procedure, the individual trees will not come into contact with one another because they will be kept at a great distance from one another.

In order to predict the correct target class, it builds a forest of decision trees during training and then uses the class mode, a classification or the average forecast of those trees. This is performed in an effort to foretell the identity of the target class (regression). When multiple decision trees are combined into one "random forest," a more accurate mean can be calculated. Its moniker comes from the haphazard nature of the forest's layout. Named thus because the locations of the trees were changed (i.e., it takes the average of many predictions). The fraction of a node's total functionalities that can be isolated for use with just that node is limited. This proportion varies from one network node to the next (known as the hyper parameter). In order to prevent overfitting, we randomly select data points from the whole initial data set to create the splits in each tree. For the tree to provide a better representation of the data, this is done. By increasing the amount of randomization already in place, this helps to sidestep the issue that was found.

ARTIFICIAL NEURAL NETWORKS:

When considering information processing, the Artificial Neural Network (ANN) technology stands out because of its foundation in biological brain networks. This is because it computes data, considers various possibilities, and draws

ISSN: 2278-4632
Vol-13, Issue-04, March 2023

conclusions in a way that is similar to the way a human mind might.

XGBOOST REGRESSION:

Without the Boost technique, it would be impossible to construct highly accurate supervised regression models. Boost is shorthand for "ensemble learning," a broad notion that includes several specialized areas. In order to provide a single prediction, this method requires the training and combination of a large number of models, here collectively referred to as base learners. The combined name for these models is "basis learners."

LASSO REGRESSION:

Moreover, lasso regression, a form of linear regression based on shrinkage, can be utilized. Shrinkage refers to the process through which data values are reduced to converge on a mean value. This procedure can also be viewed as an effort to bring the values into harmony with one another. The lasso method is effective for building models that don't have a lot of moving parts or a particularly intricate structure.

3. IMPLEMENTATION

Figure 1 presents a block diagram of the air pollution prediction module; in this section, you will find a thorough description of each stage that is illustrated in that diagram. You can find it here.

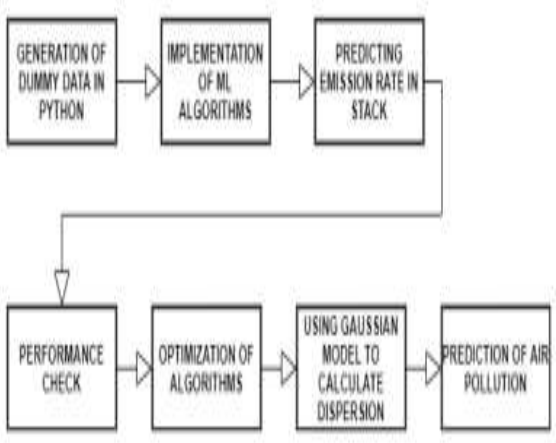

Fig. 1. Block Diagram of Air Pollution Prediction module

Generation of dummy data in python: In compliance with the criteria issued by India's Central Pollution Control Board, a Python script is used to produce a total of one thousand data samples connected to pollution (CPCB). The erroneous data were not selected entirely at random; rather, they were tied to parameters such as latitude and longitude, in addition to the speed and direction of the wind. It was vital to incorporate the ideas and suggestions of experienced professionals from the relevant field into the process of teaching the machine. This was done in order to guarantee that the training would be both effective and efficient. This was done in order to guarantee that the machine would be educated in the proper manner. It was found that the accuracy of the forecast may be significantly improved by include other meteorological aspects, such as the time of day, the season, and the length of the year. This was a discovery that was made.

Implementation of Machine Learning (ML) algorithms: In order to provide estimations that were as accurate as possible with regard to wind speed and Q-emission rate, methods from the field of machine learning were used to the data that was created. In order to accomplish this objective, a training set consisting of 80% of the data and a test set consisting of 20% of the data were generated. Both sets were used in the analysis. After that, both sets were evaluated in light of one another. During the course of the analysis, both sets were taken into account.

Performance check: Following the completion of that step, an accurate measurement of the total amount of emissions that had been predicted to occur was carried out. We estimated the mean square error for each of the several outcomes that could occur in order to provide concrete evidence that the precision of our predictions cannot be called into doubt.

Optimization of algorithms: We were able to successfully reduce the amount of error that was built into the projections by improving the functionality of a wide variety of analytical methodologies. This allowed us to make more accurate predictions. After giving each of these considerations the close attention they deserved, the algorithm that ended up proving to be the most successful was chosen.

Prediction of Air Pollution Dispersion: By simulating the spread of the contamination using a Gaussian distribution, it was possible to determine how far the contamination had spread and precisely how extensive it was. This was done in order to determine the exact scope of the contamination in its various forms. We came to the conclusion that the Gaussian dispersion model would be the best one to utilize because it would need the fewest resources, both in terms of time and effort. This led us to the conclusion that this model would be the optimum one to employ. Because of this, we arrived at the decision that employing it would be the best option available to us.

4. RESULTS AND DISCUISION

We used Jupiter notebook, which we found to be an extremely helpful tool, to write the code for a number of different machine learning algorithms. For this particular attempt, the Python programming language was decided to be the best option. The plot that follows demonstrates that all of the features that were used in the process of producing the prediction are related with one another, which indicates that they can be used in the process of training the model. This is demonstrated by the fact that all of the features are shown to be related with one another. This is indicated by the fact that all of the characteristics are revealed to be

ISSN: 2278-4632
Vol-13, Issue-04, March 2023

associated with one another. This demonstrates that this is the case. This was demonstrated by the fact that all of the qualities share a connection with one another in some fashion.

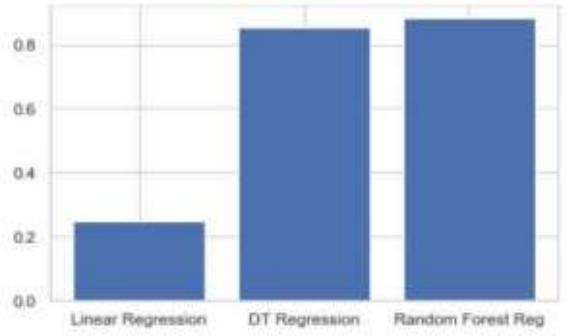

Fig. 2. SO2 prediction probability

The accuracy of predictions made with regard to sulfur dioxide and the forecasts contained inside

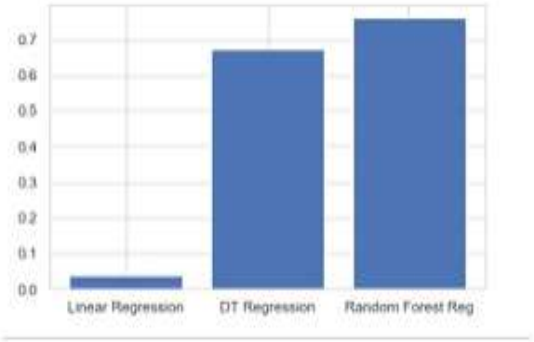

Fig. 3. CO prediction probability

In order to determine how accurate the CO prediction is, it is possible to do the following calculations:

Type of Algorithm	Prediction Probability of CO
Linear Regression	0.02
Decision Tree	0.61
Random Forest regression	0.79

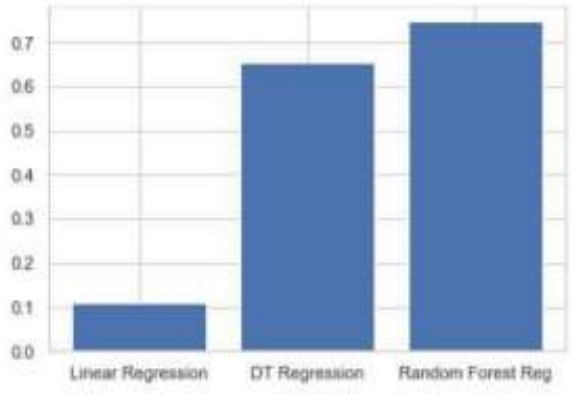

Fig. 4. O3 prediction probability

It is possible to quantify the degree of precision attained in the forecasting of O3 using the following formula:

Type of Algorithm	Prediction Probability of O3
Linear Regression	0.09
Decision Tree	0.62
Random Forest regression	0.79

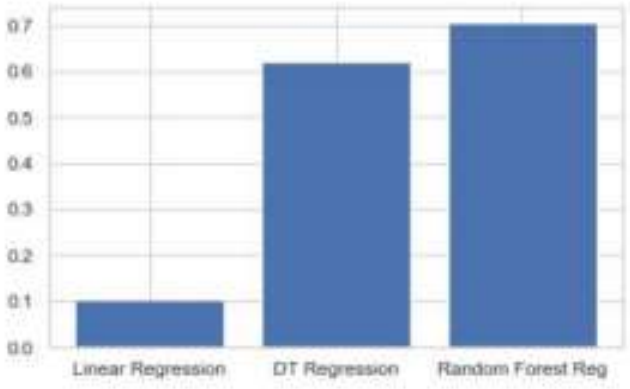

Fig. 5. NO2 prediction probability

By putting the following strategy into action, you will be able to ascertain the degree of precision that the NO2 forecast is likely to have.

Type of Algorithm	Prediction Probability of NO2
Linear Regression	0.1
Decision Tree	0.64
Random Forest regression	0.701

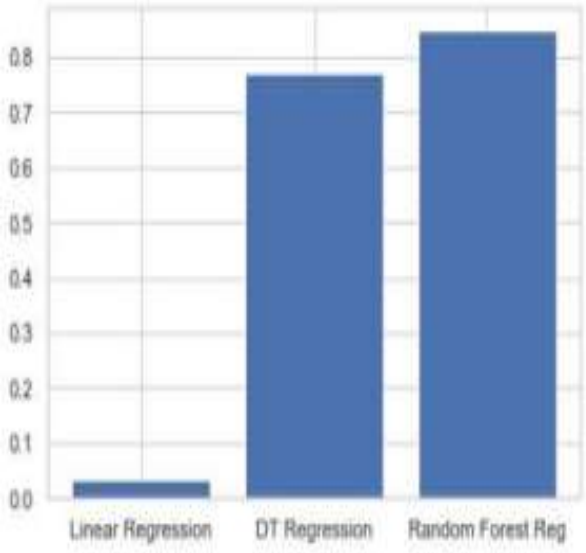

Fig. 6. PM2.5 prediction probability

If you carry out the calculations that are described in the following equation, you will be able to determine the mean square coefficient that was utilized for the PM2.5 forecast. This will be the case if you are successful.

ISSN: 2278-4632
Vol-13, Issue-04, March 2023

Type of Algorithm	Prediction Probability of PM2.5
Linear Regression	0.02
Decision Tree	0.75
Random Forest regression	0.86
Type of Algorithm	**Prediction Probability of PM210**
Linear Regression	0.02
Decision Tree	0.61
Random Forest regression	0.79

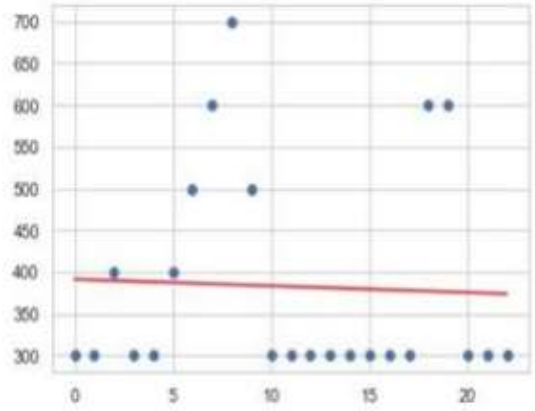

Fig. 7. Linear regression fitted curve for CO

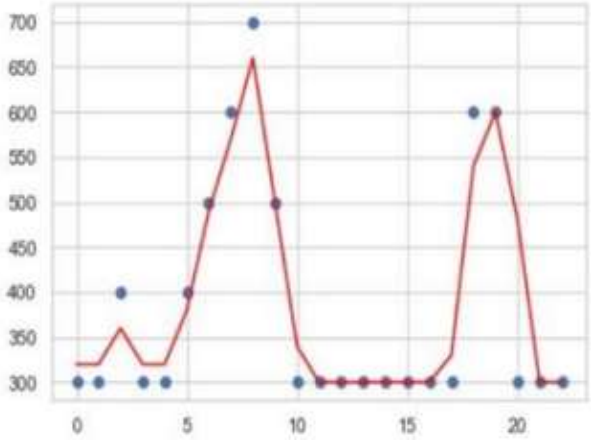

Fig. 8. Random forest fitted curve for CO

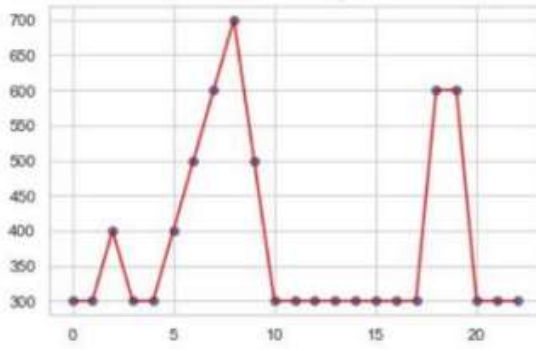

Fig. 9. Decision tree fitted curve for CO

5. CONCLUSION

The purpose of this study is to determine whether or not it is possible to utilize machine learning algorithms for the purpose of predicting levels of air pollution. The purpose of this inquiry is to acquire additional knowledge concerning this subject. In addition to this, the research analyzes the manner in which contaminants disperse across a region, as well as the concentrations of these contaminants at different distances from the initial source of the contamination. Python was used inside of the Spyder IDE to carry out the implementation of the Gaussian air dispersion model that was necessary for the air dispersion module. Spyder is an integrated development environment. This was essential in order to complete the module. Python was applied because there was a requirement for its utilization, hence Python was utilized. In order to offer an accurate estimation of the levels of air pollution, the data were analyzed using five different machine learning algorithms. This was done so that the estimates could be as accurate as possible. These methods are referred to by their respective names, which include Random Forest, Multi-layer Perception, K-Nearest Neighbor, Support Vector Regression, and Multi-linear Regression. After applying each of these approaches to the data, a comparison of the results was carried out between the various approaches. According to the results of the research, using the multi-layer Perception technique results in the smallest amount of mean squared error when compared to using any of the other options that are at your disposal. If additional research were to be conducted, various distinct models of how air disperses might be compared and contrasted in order to increase one's ability to predict how far pollution will travel in the atmosphere. This could be accomplished by studying how air disperses.

REFERENCES

[1]. https://en.wikipedia.org/wiki/Air_quality_index
[2]. 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.
[3]. 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.
[4]. W. W. Nazaroff and C. J. Weschler, “Cleaning products and air fresheners: exposure to primary and secondary air pollutants,” Atmospheric environment, vol. 38, no. 18, pp. 2841–2865, 2004.
[5]. F. Ministry of Environment and G. o. I. Climate Change, “Environment laws (amendment) bill, 2015,” 2015.
[6]. C. C. United Nations, “Paris agreementl, 2015,” 2015.

ISSN: 2278-4632
Vol-13, Issue-04, March 2023

[7]. Z. Yang and J. Wang, "A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction," Environmental research, vol. 158, pp. 105–117, 2017.

[8]. J. 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, no. 7569, pp. 367– 371, Sep. 2015. Wolrd Health Organization (WHO), "Ambient air pollution: A global assessment of exposure and burden of disease," 2016. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/250141/9789241 511353-eng.pdf. [Accessed: 26-Jul-2019].

[9]. P. Eastwood and T. Gupta, Air Pollution and Control. Singapore: Springer Singapore, 2018.

[10]. "Ministry of Environment - EEAA > Topics > Air > Air Quality > Air Quality Forecast." [Online]. Available: http://www.eeaa.gov.eg/enus/topics/air/airquality/airqualityforecast.aspx. [Accessed: 17- Feb-2019].

[11]. United Nations Population Division, The world bank, Urban population (% of total population). https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?end=2020&name_desc=false&start=1960&view=chart, 2020 accessed Sep. 25, 2021.

[12]. UN, "United Nations. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html, 2018 accessed Sep. 25, 2021.

[13]. Mrs. A. GnanaSoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd), Akshaya A.C. "Indian Air Quality Prediction And Analysis Using Machine Learning". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue).

[14]. Suhasini V. Kottur , Dr. S. S. Mantha. "An Integrated Model Using Artificial Neural Network

[15]. United Nations Population Division, The world bank, Urban population (% of total population). https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?end=2020&name_desc=false&start=1960&view=chart, 2020 accessed Sep. 25, 2021.

[16]. UN, "United Nations. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html, 2018 accessed Sep. 25, 2021.