ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 14, Issue 1, 2020

Using Machine Learning Algorithms to Analyze and Forecast Air Quality

[1]B.Sai teja,[2]M.Muthukumaran,[3]C.Manikandan
[1,2,3]Assistant professor,
department of CSE,
Chadalawada Engineering College,India

***Abstract*—** The application of machine learning algorithms, including artificial neural networks (ANN), support vector machines (SVM), and random forests (RF), has been covered in this research. One of the most significant environmental problems facing the globe today is air pollution. It is a serious risk to both the environment and human health. Every day, the quality of the air in cities gets worse. The quality of the soil and water is also impacted by air pollution. This study discusses the analysis and prediction of air quality using machine learning methods. We can estimate the levels of air pollution using machine learning algorithms, enabling people to take preemptive action to reduce their exposure to it.

Keywords*—Air Pollution; Algorithms; Machine learning; Prediction; Artificial Neural Networks*

I. INTRODUCTION

Around the world, industrialization and urbanization have been primarily blamed for environmental pollution. Across the globe, one of the most important problems facing cities is air pollution. In order to shield individuals from respiratory illnesses, it is critical to keep an eye on the quality of the air. The lack of air quality monitor stations in cities due to high construction and maintenance costs is one of the main issues that the suggested approach addresses. The data gathered, which includes the concentration of air pollutants and metrological parameters like temperature, pressure, relative humidity, and air humidity, can be used to anticipate air pollution. To forecast air quality, machine learning techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forests (RF) have been used. To increase accuracy and decrease error %, a prediction model was put forth to enhance the forecast. Primary and secondary pollutants are the two general categories of pollutants. In addition to sulfur dioxide emitted by factories and carbon monoxide gas from automobile exhaust, primary pollutants are typically created during a volcanic eruption. Regarding secondary contaminants, they do not originate from direct emissions. Instead, they are created in the atmosphere through interactions or reactions with main pollutants. Ground-level ozone is one of the most significant instances of a secondary pollutant [1]. Air quality predictions have been made using neural networks (ANNs).

II. IMPLEMENTATION

1. Data Collection:

Data Collection is the process of collecting and measuring information from a variety of sources. It must be collected and stored in a way that makes sense for the problem at hand. The dataset includes a concentration of pollutants and meteorological factors. The total attributes in the dataset are twelve:Temperature, CH4(Methane),CO(Carbon Monoxide),NMHC (Non Methane Hydro-Carbons), NO (Nitrogen Monoxide), NO2(Nitrogen Dioxide), NOx (Nitrogen Oxides),O3(Ozone), PM10 (Particulate Matter),PM2.5, RH (Relative Humidity), and SO2 (Sulfur Dioxide)

2. Data Preprocessing

A dataset can be viewed as a gathering of data objects, which are frequently also called a record, points, vectors, patterns, events, cases, samples, observations, or entities. Data is always taken from multiple and different sources which are normally not too well-grounded and that too in different formats. It is simply impracticable to expect that the data will be perfect. There may be a hindrance due to human error, constraints for measuring devices, or flaws in the data collection process.

A. Cleaning

It is very much customary to have missing values in the dataset. It may have happened during data collection. To solve this problem the rows with the missing data are eliminated. Object type is converted into numeric type because it is easy for a model to understand numerical inputs.

B. Attribute Selection

The new attribute is selected from the given set of attributes. The attributes which majorly contribute to air pollution and the row-wise highest value is considered as Air Quality Index.

C. Normalization

It means scaling the data values in the specified range. Air Quality Index is a number used to communicate to the public how polluted the air is currently. The Air quality index (AQI) increases, an increasingly large percentage of pollution as shown in Figure1.TheAirQualityIndex(AQI)value is scaled between the specified range from 0 to 4.

Predictive analytics uses historical and pre-processed data to predict future events. Historical data is used to build a model that captures important trends. This model is then used on the current data to predict. Machine learning algorithms are used for the

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 14, Issue 1, 2020

analysis and prediction of air quality.

AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+

Figure 1: Air Quality Index (AQI) range

D. Formatting Convert from one file format (xlxs) to another file format (CSV file).

3. Data Visualization:

Data visualization is the graphical representation of information and data and it plays an important role in the portrayal of both small-scale and large-scale data. Graphical elements like charts, graphs, and maps, data visualization tools provide an approachable way to see and fathom trends, outliers, and patterns in data.

III. ALGORITHMS/METHODOLOGY

1. Artificial Neural Network

An artificial neuron network (ANN) is a computational algorithm based on the structure and functions of biological neural networks which is predicated upon the collection of units called Artificial neurons. ANNs have three layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to the hidden layer which in turn sends the output neurons to the output layer. The flowchart of Artificial Neural Network is shown in Figure 2. In Python, Keras is an easy-to-use, open-source neural network library written. It proficiently runs on top of the Tensor flow. It is very easy to build complex models and recapitulate quickly.

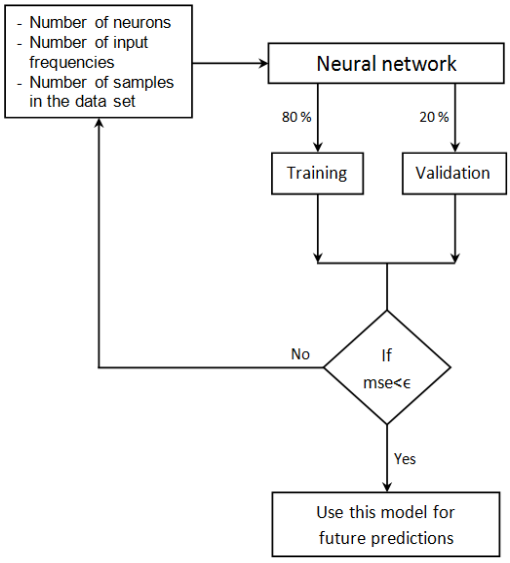


Figure 2: Flow chart of Artificial Neural Networks

A. Defining the Model

An elementary model is defined in the Sequential class and it's a linear stack of layers. It is uncomplicated and the easiest way to build a model in Keras. It permits you to build a model layer by layer. It has the 'add()' function to add layers to our model. Every neuron in each layer is connected to every neuron of the following layer. There are four hidden layers with32,32,16,8 neurons respectively. The number of input dimensions is twelve. There are five neurons in the output layers. The activation function used are ReLu (Rectified Linear Unit) and sigmoid. ReLU is zero for all negative inputs. Mathematically,it is stipulated as y = max(0, x). Visually, it looks as shown in Figure 3.

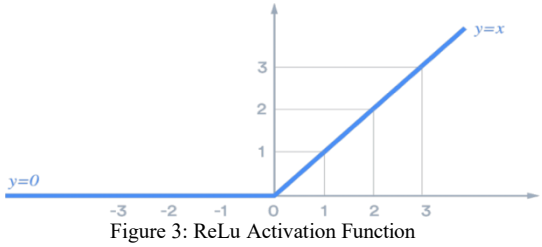


Figure 3: ReLu Activation Function

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 14, Issue 1, 2020

The sigmoid function returns an actual output value between 0 and1for any input value.It will return zero if the input is very large and small. Visually, it looks as shown in Figure 4.

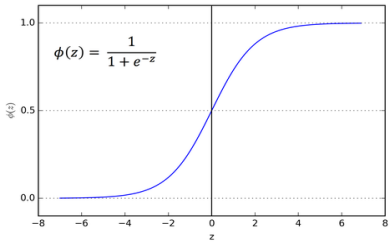


Figure 4: Sigmoid activation function

B. Compilation of Model

The compilation of the model has three parameters: optimizer, loss, and metrics. The optimizer controls the learning rate. The optimizer used is "Adam" which adjusts the learning rate through training. It specifies the optimization algorithm that allows the neural network to calculate the weights of the parameters from the input data and the defined loss function as shown in Figure 5.

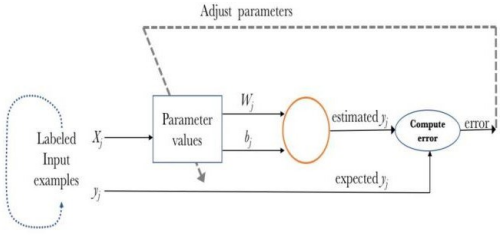


Figure 5: Compiling the model

The loss function is the function that evaluates how well the algorithm models the data set. It calculates how far the predicted values deviate from the actual values. The loss function used is "categorical_cross entropy". Metrics allow us to keep track of the loss as the model is being trained. It will monitor the learning process of the neural network. Accuracy is defined as the ratio of the number of correct predictions to a total number of predictions.

Model Training

The model can be trained to the training data available by invoking the fit()method of the model. It takes four arguments: the first two arguments are the training dataset, epochs, and verbose. Epochs are indicating the number of times the data is used in the learning process.

Model Prediction

In Keras, the classes for the new data instances can be predicted using the predict() function. It takes an argument: the testing dataset, batch size. The batch size controls the number of testing samples to work through them odes. The testing dataset issued to test the model how well it has learned.

Evaluating the model

A confusion matrix is a matrix that provides an abstract of the predictive results in a classification problem. Correct and incorrect predictions are abridged in a table with their values and broken down by each class shown in Figure 6.

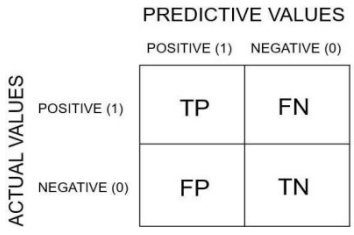


Figure 6: Confusion Matrix

True Positive (TP):predicted positive and its true. True Negative (TN): predicted negative and its true.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 14, Issue 1, 2020

False Positive (FP): predicted positive and it's false. False Negative (FN): predicted negative and its false.
Recall is defined as the ratio of the total number of correctly classified positive classes divide by the total number of positive classes.

$$Recall = \frac{TP}{TP + FN} \text{ or } \frac{\text{True Positive}}{\text{Actual Results}}$$

Precision is the ratio of the total number of accurately classified positive classes divided by the total number of predicted positive classes.

$$Precision = \frac{TP}{TP + FP} \text{ or } \frac{\text{True Positive}}{\text{Predictive Results}}$$

F1-score is defined as the Harmonic Mean of Precision and Recall. It is laborious to compare two models with distinct precision and recall values. So to make them comparable, use F-Score.

$$F\text{-}score = \frac{2 * Recall * Precision}{Recall + Precision}$$

Specificity determines the proportion of true negatives that are perfectly identified.

$$Specificity = \frac{TN}{TN + FP}$$

Metrics has a method "accuracy score" which returns an accuracy classification score.

2. Support Vector Machine
Machine learning necessitates predicting and classifying data and to do so numerous machine learning algorithms are employed according to the dataset. Support Vector Machine (SVM) is a machine learning algorithm that is used for both classification and regression problems. It can resolve both linear and non-linear problems and work well for many empirical problems. At first, an approximation of what SVMs do is to find a hyperplane between data of two classes. SVM takes the data as an input and outputs a line that separates those classes if possible as shown in Figure 7.

A. Support Vectors:
They are the data points, which are closest to the hyperplane.

B. Hyperplane:
It is a decision plane that separates between a set of objects having different class memberships.

C. Margin:
Margin is a gap between the two lines on the closest class points and is estimated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is deemed as a good margin, a smaller margin is a bad margin. The following figure flowchart shows how a Support vector machine works

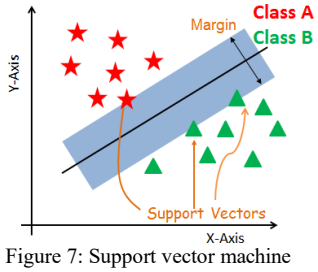
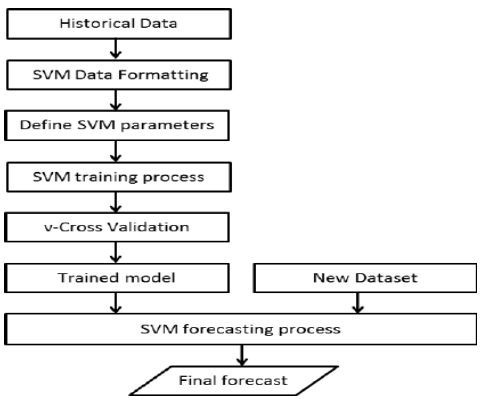


Figure 7: Support vector machine



INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 14, Issue 1, 2020

Figure8: Flowchart for Support Vector Algorithm

According to the SVM algorithm, the points closest to the line from both the classes are determined. These points are called support vectors. The distance between the line and the support vectors is evaluated. This distance is called the margin. The aim is to maximize the margin. The optimal hyper plane is a hyper plane that always has a large and maximized margin. Thus, a decision boundary is made by SVM in such a way that the separation between the two classes is as wide as possible. The intention is to select a hyper plane with the maximum attainable margin between support vectors in the given dataset and to separate the given dataset in the best possible way.
Support Vector Machine forages for the maximum marginal hyper plane in the subsequent steps: Generate hyper planes that set the classes apart in the best way.
A. The below figure9 shows three hyper plans black, blue, and orange. The blue and orange hyper planes have higher stratification errors, but the black hyper plane is separating the two classes accurately.
B. Choose the proper hyper plane with the maximum separation from the either nearest data points as manifested in the right-hand side figure.

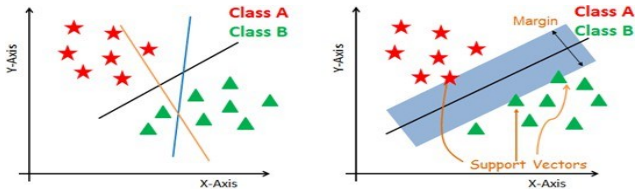


Figure 9: Support vector machine classification

SVM Kernels:
The SVM algorithm is accomplished in operation using a kernel. The input data space is transformed to the required form. The strategy used by SVM for this transformation is called the kernel trick. Now, the kernel takes a low-dimensional data input space and transfigures it into a higher-dimensional space. A better and an accurate classifier is built with the help of the kernel trick.

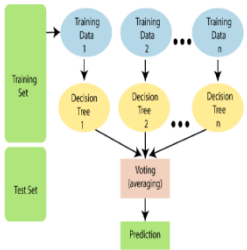
Linear Kernel:
A linear kernel can be used as a normal dot product for any two given perceptions. The sum of the multiplication of each pair of input data values is the product between two vectors. The linear kernel is being used in the current project.

Generating Model:
A. Firstly, a support vector machine model is created. Then the SVM module is imported and a support vector lassifier object is created by passing the argument kernel as the linear kernel in SVC() function.
B. Then, the model is fit on the train set using fit() and the prediction is performed on the test set using predict(). A classification report is used to estimate the peculiarity of prediction from a classification algorithm.

3. Random Forest
Random forest is another supervised learning algorithm that is used for both classifications as well as regression. Random Forest Algorithm constructs decision trees on the available data samples and then gets the prediction from each of them and finally designates the best solution by means of voting.

Algorithm for Random forest:
A. Let, the number of training cases is 'N', and the number of variables in the classifier is 'M'.
B. The number 'm' of input variables is used to ascertain the decision a node of the tree, and 'm' should be substantially less than 'M'.
C. Select a training set for this tree by choosing N times with substitution from all N available training cases.
D. Use the remainder of the cases to estimate the error of the tree by predicting their classes.
E. For each node of the tree ,'m'variables are randomly chosen based on the decision taken at the node. The best split is determined based on these m variables in the training set.
F. Each tree is fully grown and not pruned.

Defining the model:
Technically, the ensemble method is better than a single decision tree because it decreases the over-fitting by averaging the result. It is based on the divide-and-conquer approach. The assemblage of decision tree classifiers is known as a Forest. For each attribute, the attribute selection indicators such as information gain, gain ratio, and Gene index issued to generate individual decision trees.



INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 14, Issue 1, 2020

Each tree depends on an independent random sample. In a classification problem, each tree votes, and therefore the hottest class is chosen because of the outcome. In case of regression, the average of all the outputs is considered as final result as shown in Fig. 10.

Figure10: Random Forest Classifier

Building a Classifier:

The goal of the ensemble method is to mix the predictions of several base estimators built with a given learning algorithm to enhance robustness over one estimator. The sklearn. Ensemble module includes two averaging algorithms supported randomized decision trees; the Random Forest algorithm method. This means a various set of classifiers is made by introducing randomness within the classifier construction. The prediction of the ensemble is given because of the averaged prediction of the individual classifiers.In Random Forest Classifier each tree within the ensemble is made from a sample drawn with replacement from the training set.

IV. COMPARATIVE ANALYSIS

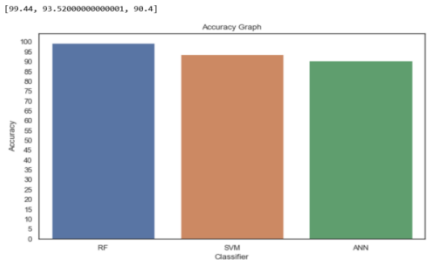


Figure 11: Comparative analysis form a chine learning algorithms

The above figure shows the performance evaluation of Random forest, SVM, and Artificial neural networks by computing classification reports, confusion matrix, an accuracy score. The most efficient algorithm among the three algorithms for air quality prediction is Random Forest. The Random Forest is less prone to over-fitting because it combines the predictions of many decision trees into a single model.

V. CONCLUSION

The result demonstrates that the proposed methods are effective and reliable for use. The accuracy score of the Artificial Neural Network, Support Vector Machine, and Random Forest-based model is 90.4%, 93.5%, and 99.4% respectively. The most efficient algorithm among the three algorithms for air quality prediction is Random Forest Algorithm. Air quality prediction may be a worthwhile investment on multiple levels-individual, communities, national and global. The accurate prediction helps people plan, decreasing the effects of harmful air pollutants on health and the cost associated and creating a cleaner and healthier environment.

REFERENCES

[1] www.wikipedia.com
[2] Gaganjot Kaur, Jerry Zeyu, Shengqiang Lu, “Air Quality Prediction: Big data and Machine Learning Approach”, Index of Community Socio-Educational Advantage, pp. 150-158, May2017.
[3] Rajeev Tiwari, Shuchi Upadhyay, Parv Singhal, “Air Pollution Level Prediction System”, International Journal of Innovative Technology and Exploring Engineering, vol.8, pp. 201-207, April2019.
[4] https://expertsystem.com/machine-learning-definition/
[5] Ziyue Guan, Richard O. Sinnott “Prediction of Air Pollution through Machine Learning Approaches on the Cloud”, Institute of Electrical and Electronics Engineers International Conference, Zurich Switzerland, pp. 51-60, December 2018.
[6] Ibrahim KOK, Mehmet Ulvi, Suat Ozdemir, “A Deep Learning Model for Air Quality in Smart Cities”, Institute of Electrical and Electronics Engineers International Conference, Boston USA, pp.1983-1990, December 2017.
[7] Chavi Srivastava, Amit Singh, Shymali Singh, “Estimation of AirPollution in Delhi using Machine Learning Techniques”, Institute of Electrical and Electronics Engineers International Conference, NoidaIndia,pp.304-309,September2018
[8] Nadjet Djebbri, Mounria Rouainia, “Artificial Neural Networks Based Air Pollution Monitoring in Industrial Sites”, Institute of Electrical and Electronics Engineers International Conference ,Antalya Turkey, August 2017.
[9] Ping Wei Soh, Jia Wei Chang, Jen Wei Huang, “Adpative Deep Learning-Based Air Quality Prediction Model using the most Relevant Spatial-Temporal Relations”, Institute of Electrical and Electronics Engineers, vol. 6, pp. 38186-38199, June 2018.
[10] Sankar Ganesh, Sri Harsha Modali, Soumith Reddy, “Forecasting Air Quality Index using Regression Models”, Institute of Electrical and Electronics Engineers International Conference, TirunelveliIndia, pp. 248-254, May 2017.
[11] Adven Masih, “Application of Random Forest Algorithm to Predict the Atmospheric Concentration of NO2”, Institute of Electrica land Electronics Engineers International Conference, Yekaterinburg Russia, April2019
[12] https://towardsdatascience.com
[13] C. Ma, Real time Mobile Pollution Detection Monitoring for Improved Route Planning, Master Dissertation, University of Melbourne, 2016. 38.