FORECASTING AND PREDICTION OF AIR POLLUTANTS CONCENTRATES USING MACHINE LEARNING TECHNIQUES: THE CASE OF INDIA

Moolchand Sharma1*, Samyak Jain2, Sidhant Mittal 2, Dr. Tariq Hussain Sheikh3

Department of Computer Science & Engineering, MAIT, DELHI, INDIA1, 2
Government Degree College Poonch, Jammu and Kashmir, INDIA3

sharma.cs06@gmail.com1*, samyakjain69@yahoo.co.in2, mittal.sidhant@gmail.com2

ABSTRACT

Air quality index (AQI) is a number used by government agencies to communicate to the public how polluted the air currently. It is based on several factors like SO2, NO2, O3, RSPM/PM10, and PM2.5. Several methods were developed in the past by various researchers/environmental agencies for the determination of AQI. Still, there is no universally accepted method that exists, which is appropriate for all situations. We have developed a prediction model that is confined to standard classification or regression models. These prediction models have ignored the co-relation between sub-models in different time slots. The paper focuses on a refined model for inferring air pollutants based on historical and current meteorological datasets. Also, the model is designed to forecast AQI for the coming months, quarters or years where the emphasis is on how to improve its accuracy and performance. The algorithms are used on Air Pollution Geocodes Dataset (2016-2018), and results calculated for 196 cities of India on various classifiers. Accuracy of 94%-96% achieved from Linear Robust Regression, which increases to 97.92% after application of KNN and 97.91% after SVM and 97.47% after 5th epoch of ANN. Decision Tree Classifier has given the best accuracy of 99.7%, which increases by 0.02% on the application of the Random Forest Classifier. Forecasting achieved by Moving Average Smoothing using R-ARIMA, which offers daily values for the coming 45 days or monthly data of AQI for the next year.

KEYWORDS

Air Quality Index (AQI), Regression, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), K-Means, Decision Trees, Random Forest, Artificial Neural Networks (ANN), R-ARIMA, Particulate Matter (PM)

1. INTRODUCTION

The Adverse health impacts from air pollutants like ozone (O3), particle matter (PM), Sulphur dioxide (SO2), carbon monoxide (CO), nitrogen oxides (NOx), volatile organic compounds (VOCs), pesticides, and metals, among others, are very vulnerable to health. It is pertinent to adopt formal methods for monitoring and forecasting air quality for various areas. Air Quality Index (AQI) is such an indicator tool widely used worldwide and in India for the last 2-3 decades. There are six AQI categories, namely Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. In this paper, we will use a combination of Machine Learning (ML) algorithms with time series forecasting and also, SAP predictive analytics for predicting AQI values for 196 cities of India. Machine learning is an efficient experimental approach for both classification-regression of nonlinear systems. A broad ‘training dataset,’ which is covering most of the system’s parameters, is constructed using machine learning techniques. Generally, a random subset of the data is put aside for a completely independent validation [1]. Machine learning is the scientific study to perform a specific task using algorithms and statistical models without any explicit instructions. AQI is ongoing access to the availability of online data and low-cost computation along with the advancement of new learning algorithms in fields like healthcare, environment, and education, etc. [2]. An algorithm can provide accuracy at only a certain level, So that, Model Selection is required to get better results.
Model Selection is the process of comparing the efficiency of different classifiers and choosing the best classifier suited for model creation. In this paper, the Model selection is performed sequentially to find the AQI for various cities of India. Regression is applied to see optimal results, to the testing AQI dataset, and the resultant predicted values compared with the original dataset for deviation. We have compared the performance on five widely used machine learning classifiers, which are Support Vector Machine, K-Nearest Neighbour-Means Clustering, Decision Tree, Random Forest, and Artificial Neural Networks. And also, we have validated the result to eliminate overfitting or underfitting to create a robust and effective model.

Time series forecasting (TSF) is the use of a model to predict future values based on previously observed values. This modeling approach is used primarily when little knowledge is available on the underlying data generating process or when there is no adequate informative model that relates the prediction variable to other explanatory variables [3,4].

A statistical analysis technique, Autoregressive integrated moving average (ARIMA), is used for the development of an environmental forecasting tool with various smoothing techniques implemented for forecasting [5]. It is an essential process because there is a need for timely information about changes in the pollution level as the air pollution in cities has become so severe.

Sensitive peoples with asthma, chronic obstructive pulmonary disease (COPD), children, and older adults are more likely to experience health effects due to higher levels of AQI. This gives us the motivation to accurately forecast AQI at all levels, helpful for enhancing human health. People can then take appropriate precautions and follow tips on how to reduce air pollution.

The key contributions that appear as a standout throughout the paper are:

- Highest Accuracy (99.79%) was obtained from the Random Forest Model.
- A comparative analysis is done on the six models by using different classifiers, as mentioned in the abstract.
- Forecasting AQI values are done via Triple Exponential Smoothing, and then Moving Average Smoothing is applied to smoothen the curve obtained for better results.
- We have built a model for some of the cities using ARIMA in R and tried to predict the daily AQI values for the coming 45 days of 2019 or monthly AQI values for the year 2019 & 2020.
- A contrast of the air pollution concentrates and AQI level between two main seasons, i.e., summer (April-May) and winter (December-February), is explained, which helps to analyze the AQI values for a specific season of a year by giving the average value for that month of the season.

The sequence of the paper is the Concept of machine learning algorithms, Time Series Forecasting, and SAP Predictive Analytics described in Section 2. The procedure followed is discussed in section 3. Dataset Collection and its Pre-Processing described in section 4. Model Results for AQI prediction produced after the execution of the algorithms presented in Section 5. The remaining of the paper deals with the conclusion and future scope followed by references.

2. LITERATURE SURVEY

Machine learning is a significant sub-field in intelligent computation. The extraction of information using computational methods is its primary objective. In environmental sciences, machine learning methods are heavily used for data processing, model emulation, climate prediction, AQI forecasting, oceanographic, and hydrological forecasting[6]. AQI prediction is grouped into three categories: (1) simple empirical approach, which predicts the values of tomorrow by the data of present-day or rely strictly on the dependence between forecasted pollutants and air pollutant variables. Second, physically-based approaches which result in biased forecasts as they are too complex to be easily represented by physically-based models. Thirdly, parametric or non-parametric statistical methods such as neural networks which outperform physically-based methods in the accuracy [7].
Forecasting time series is a need in the financial sector or other fields, economic or not. As per E. Dhamo, R is an essential tool for forecasting and especially for studying the time series models[8]. ARIMA model derived by the general alteration of an autoregressive moving average (ARIMA) model. This model type in AQI forecasting is classified as ARIMA (p, d, q), with (p, d, q) being all nonnegative integers, p denotes the autoregressive parts of the data set, d refers to integrated components of the data set, and q means moving average parts of the data set. Initially, for the AQI prediction, an appropriate ARIMA model has to be identified for the particular datasets, with the parameters having the smallest possible values so that it can analyze the data correctly and forecast more accurately[9]. Time series data for predicting the AQI has three components, i.e. (1) Trend to highlight long-term increase or decrease in the AQI values, (2) Seasonal to highlight factors like a quarter of the year, month or days of a week showing a comparison between the AQI values, (3) Cyclic to highlight irregular rise and fall in AQI value. Smoothing is done to discard the irregular roughness or noise in the data to see better patterns of AQI. Generally, Moving Average Smoothing is applied in the ARIMA time series analysis to smooth out the seasonality among the different predicted AQI values.

SAP Predictive Analytics module named Automated Analytics has been used to produce robust AQI predictive models in a short period. The main focus is to automate all the steps of the AQI predictive model workflow, without compromising the performance and thereby shielding the user from statistical complexities. SAP Automated Analytics provides a higher degree of automation. Hence, it enables us to create a robust AQI prediction without in-depth mathematical education. Expert Analytics is intended for data scientists familiar with unique statistical algorithms, their implementations, and assumptions for AQI prediction [10].

The growing rate of Globalization, Industrialization and the resultant Population growth (urbanization) have caused severe environmental concerns in India. The continuously rising environmental degradation is dangerous for human development, especially in India, where the impact of environmental pollution is more rigorous, leading to ill health, increased disabilities and mortality rate annually. The air quality is becoming essential both for the environment as well to society. ANN approach is also a novel approach for forecasting the PM2.5 pollution concentration, one of the most critical pollutant concentrations of AQI using air mass trajectory analysis and wavelet transformation [11,12,13]. Also, the AQI is affected by multi-dimensional factors, including location, time, and uncertain variables. Many researchers began to use the Big Data Analytics as well as a Machine Learning approach for handling the multi-dimensional data. This is because due to advancements in big data applications as well as in machine learning tools for managing environmental sensing data and sensor networks[14,15,16].

Air quality (AQ) forecasting is among the most common environmental forecasting applications. It is usually performed by air quality agencies or authorities responsible for the monitoring and management of the atmospheric environment in urban agglomerations. As photochemical air pollution is one of the most pronounced air-quality problems of developed countries, ozone forecasting OF is usually part of the core of every AQ forecasting system or application. The methods used for OF vary, and in general, include the following: persistence, climatology, criteria, CART, regression, neural networks, phenomenological/ intuition, and 3-D air quality models [17,18].

What is essential and gives this modeling effort a hybrid nature is a fact that it uses clustered datasets. Moreover, this approach improves the accuracy of existing forecasting models by using unsupervised machine learning to cluster the data vectors and trace hidden knowledge [19]. In the analysis presented in this paper, statistic estimates including relative mean errors, root mean squared errors, and the mean absolute relative error have been employed to compare performances of the models[21]. Also, When the AQI model is created, the predictive capability is monitored over time through a server-component automatically. The user is informed when the predictive ability falls below a defined threshold for model recalibration.
3. PROPOSED METHODOLOGY

We have divided the work into two main stages. The first two stages are used for model creation and forecasting of AQI. At the same time, the final step provides a front-end business outlook for the visualization and understanding of the end-users.

3.1. Model Selection

In this paper, Model Selection is performed on six widely used machine learning classifiers, which are Linear Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), K-Means Clustering, Decision Tree, Random Forest and Artificial Neural Networks (ANN). It is done to train and test on a small batch of our dataset, which includes values for the years 2016-2018. Scatter-plots, line graphs are developed for each model created using these algorithms.

Machine learning is the scientific study to perform a specific task using algorithms and statistical models without any explicit instructions. Validation will help us to determine and eliminate the possibility of over or underfitting, and testing leads to the conclusive process of determining the robustness and effectiveness of the model. The ratio for training and testing is taken as 70:30. The accuracy of each of them is compared with the help of a confusion matrix. It will be discussed in detail, along with result outputs in the next section. The procedure is shown in the flowchart (Figure 1) below.

![Figure 1: Proposed Stage I process](image)

3.2. Time Series Forecasting

This phase includes Model Creation using the dataset, which contains values for the years 2016-2018 and time-series forecasting in RStudio to help forecast or predict AQI values for the coming months/years. Here, we have built models for four main Metropolitan cities (Delhi, Mumbai, Bangalore, Kolkata) and two commercial cities (Agra and Bhilai). We have then automated these models for a total of 196 towns gathered across India and tried to predict the daily AQI values for the coming 45 days of 2019 or monthly AQI values for the year 2019 & 2020.

We have then applied a smoothing method (Moving Average Smoothing) to smoothen the curve obtained for better results. The procedure is shown in the flowchart (Figure 2) below.
Figure 2: Proposed Stage II process
4. DATASET COLLECTION, PRE-PROCESSING, AND ANALYSIS

4.1. Dataset Collection

We have collected the air pollution data from Open Government Database (OGD) Platform India (https://data.gov.in/catalog/historical-daily-ambient-air-quality-data) for the years 2016-2018. The air pollution data in this research included the concentrations of NO2, RSPM, and SO2 and PM2.5. The data collected roughly contained about 52,000 values for each year.

The collected data consisted of the pollutant levels of different states in India (Area wise). It consists of the following attributes - Station Code, Sampling Date, State, City/Town/Area, Location of Monitoring Station, Agency, Type of Location, SO2, NO2, RSPM/PM10 and PM 2.5.

Separate CSV files of each State/Union Territory (A total of 30) was downloaded, and it was recorded that the data was inconsistent, some of the values were missing for the columns PM 2.5 and RSPM. Moreover, entries such as O3(8-hour average) were missing in the CSV files. So, the pre-processing of the data has been done to make it suitable for our project.

Other Attributes (such as Latitude and Longitude) necessary for our simulation were added separately by finding coordinates of each city manually and adding them to the CSV file.

4.2. Data Pre-Processing

Pre-Processing mentions replacing or deleting the dirty and raw data (Normalization of the dataset) by identifying incomplete, or irrelevant parts of the data and then. In our model, the first step is to create the formulas to calculate the missing values in the dataset. Various methods were generated to calculate the missing data.

Attributes to which formulas could not be generated were filled by taking the mean of the values of previous and upcoming dates. Individual CSV files did not contain the data of contributory variables, which are essential to calculate AQI; hence, they were calculated using formula. To calculate the value of AQI, the officials have generated a generalized method that must consist of at least three factors contributing towards AQI, out of which one must be a PM level (PM 2.5 or RSPM). Hence with the limited attributes on our site, the formula we generated to calculate our AQI level corresponding to each row.

Attributes of half processed data sheet with their meanings are listed below in table 1. The list of formulas is listed in table 2. The process was repeated in the CSV files for all the States and Union Territories, to add all the missing values and remove redundant data to prevent under or overfitting. A total of 196 cities all over India were recorded to have complete data, and the remaining towns whose data could not be collected or generated were discarded.

The SUB-INDEX and CHECK fields of the dataset were also removed upon normalization to reduce the size of the dataset and increase the accuracy of the model. The data cleaned was arranged alphabetically by the cities.
Table 1: List of Attributes

| Attributes            | Complete Meaning                                |
|-----------------------|------------------------------------------------|
| Stn Code              | Station Code                                    |
| Sampling Date         | The date on which the data was recorded         |
| State                 | State/Union Territory                           |
| City/Town/Village/Area| City in the corresponding state                 |
| Location of Monitoring Station | Location of AQI monitoring Station |
| Agency                | The organization responsible for collecting data|
| Type of Location      | Residential/Rural/Industrial/Others             |
| SO2,NO2,RSPM,03       | The concentration of air pollutants             |
| SI O3, SI PM10, SI SO2, SI NO2 | Sub Index values to calculate AQI |
| CHECK Values          | To keep a check whether the fields are filled or not |
| AQI                   | The calculated value of the Air Quality Index   |

To implement classification models on the dataset, we have added a new field called AQI SEGMENT, which is assigned with a pre-defined class from 1-6 (1- being the lowest and 6- being hazardous) depending upon the values in the AQI field. Figure 3 highlights the pre-defined AQI Segments.

![AQI Level Segments](image)

4.3. Data Analysis

The result set obtained after applying the model on the validation dataset is analyzed, which gives the probability, mean, standard deviations, and percent error of each of the AQI Forecasted values. Now our main aim is to link each city with its respective geocodes. It was done by manually adding Latitude and Longitude for each City to connect our AQI model with Maps to make it more readable.

Moreover, factors such as day of the week and public holidays were added as these attributes affect the AQI daily. It is then sent as a CSV input file to SAC (SAP Analytics Cloud) to create the front-end business story.
Hence the final dataset obtained after pre-processing has the following attributes:

i. Stn_Code  
ii. Sampling_Date
iii. Day_of_Week  
iv. Is_Holiday
v. Reason_for_Holiday  
vi. City/Town/Village/Area
vii. State  
viii. Location_of_Monitoring_Station
ix. Agency  
x. Type_of_Location
xi. SO2
xii. NO2
xiii. RSPM/PM10  
xiv. O3
xv. PM2.5  
xvi. AQI
xvii. AQI Segments  
xviii. Latitude and Longitude

Table 2: List of Formulas

| Variable | Formula |
|----------|---------|
| **PM 2.5** | $PM_{2.5} = \begin{cases} 
\frac{AQI \times 3}{5} & \text{if } AQI < 50 \\
30 + \frac{(AQI - 50) \times 3}{5} & \text{if } 50 \leq AQI < 100 \\
60 + \frac{(AQI - 100) \times 3}{10} & \text{if } 100 \leq AQI < 200 \\
90 + \frac{(AQI - 200) \times 3}{10} & \text{if } 200 \leq AQI < 300 \\
120 + \frac{(AQI - 300) \times 13}{10} & \text{if } 300 \leq AQI < 400 \\
250 + \frac{(AQI - 400) \times 13}{120} & \text{if } AQI \geq 400 
\end{cases}$ |
| **SIO3** | $SI_{O3} = \begin{cases} 
\frac{O3}{50} & \text{if } O3 \leq 50 \\
50 + \frac{(O3 - 50) \times 50}{50} & \text{if } 50 < O3 \leq 100 \\
100 + \frac{(O3 - 100) \times 100}{68} & \text{if } 100 < O3 \leq 168 \\
200 + \frac{(O3 - 168) \times 100}{40} & \text{if } 168 < O3 \leq 208 \\
300 + \frac{(O3 - 208) \times 100}{539} & \text{if } 208 < O3 \leq 748 \\
400 + \frac{(O3 - 400) \times 100}{539} & \text{if } O3 > 748 
\end{cases}$ |
| **SIPM10/RSPM** | $SI_{PM10} = \begin{cases} 
\frac{PM_{10}}{PM_{10}} & \text{if } PM_{10} \leq 50 \\
100 + \frac{(PM_{10} - 100) \times 100}{150} & \text{if } 50 < PM_{10} \leq 250 \\
200 + \frac{(PM_{10} - 250)}{150} & \text{if } 250 < PM_{10} \leq 350 \\
300 + \frac{(PM_{10} - 350) \times 100}{80} & \text{if } 350 < PM_{10} \leq 430 \\
400 + \frac{(PM_{10} - 430) \times 100}{80} & \text{if } PM_{10} > 430 
\end{cases}$ |
5. IMPLEMENTATION AND RESULTS

In this section, the model is implemented against the chosen machine learning classifiers and has endured time series forecasting, as discussed above. Various models have been created and checked for accuracy and percentage error.

5.1. Machine Learning Classifiers

Post the training to analyze the training model, and we passed feature sets upon the model that has not been trained yet. From these feature sets, the model generated a set of predictions. These predicted labels were then compared to the correct labels to help us compute the testing accuracy of the model. In this paper, the following six classifiers are used.

a) Regression

Linear regression is a linear model of the relationship between one or more independent variables and a dependent variable. When there is only one explanatory variable, the case represents a Simple Linear Regression Model. This type of model helps in choosing a line where the summation of the distance between the actual and predicted value of the data is minimal.
b) Classification

Classification is a technique that uses one or more independent variables as the basis to determine the class of the dependent variable. Following are the types of classifiers used in our project:

i.) Support Vector Machine (SVM) Classifier

A Support Vector Machine (SVM) is used to classify the dataset by separating hyperplane. The algorithm outputs an optimal hyperplane that can uniquely categorize new examples. This technique is used to perform Linear Regression in both simple planes and higher dimensional space as well. This algorithm learns from obvious cases and looks for extreme circumstances and treat them as support vectors.

ii.) K-Nearest Neighbours (KNN) Classifier

It is a type of supervised learning. This algorithm classifies new cases based on distance functions. It compares the distance of the unique data point to its nearest K neighbors and assigns it to the class of the majority ones. Choosing the optimal value for K is best done by first inspecting the data. Distance in our model is calculated using Minkowski distance (Generalization of both Euclidean distance and Manhattan distance).

iii.) Decision Tree Classifier

It is a type of supervised learning. The decision tree uses the tree representation to solve the problems. They can be used to solve both regression and classification problems. Each edge is used to ask a question to the dataset, and each edge provides the sub-category to the dataset. Leaf nodes created such that they cannot further be classified and solely represent a Unique category.

iv.) Random Forest Classifier

Random forest builds multiple decision trees for the same problem and then merges them or selects the best out of most accurate models. It is an ensemble model and also results in a better model with a wide diversity. Random decision forests overcome the disadvantage of Decision Tree Models, i.e., Overfitting.

c) Artificial Neural Networks (ANN)

An artificial neuron network (ANN) is a non-linear computational model. It can be used for both the Classification of data and predicting discrete values. Each node in a neural network is known as a perceptron. Each node is connected to the nodes in the next layer of a neural network by a synapse having a weight associated with it. When a node is fired, it multiplies all specific inputs of a node to their corresponding weight and adds them. The summation of the result is then passed through an activation function, which then gives a relevant output result. The result is then compared with the actual value (expected). The error is calculated used to backpropagate through the network and adjust the weights accordingly to minimize the error.

5.2. Time Series Forecasting Classifier

a) ARIMA

ARIMA is the abbreviation for Autoregressive Integrated Moving Average. ARIMA models utilize historical information to make predictions and hence are a popular and flexible class of forecasting a model. Since this type of model is used as a foundation for more complex models, it is called a basic forecasting technique. It is also known as the Box-Jenkins approach.
b) Moving Average Smoothing

Moving average smoothing is a time series forecasting technique. It is useful for feature engineering and data preparation. Calculating a moving average involves creating a new series where the values are the average of original time series raw observations. To calculate the average data in the new series window defined by the window width is slid along the time series. It is referred to as the “moving” part. The window width establishes several raw observations to calculate the moving average value. It provides a specification of the window size required by the moving average.

5.3. RESULTS AND DISCUSSION

In this paper, the model is tested upon the six machine learning classifiers chosen and then validated to check the accuracy and robustness. The input variables are denoted as independent variables, while the output variable is indicated as the dependent variable. In the following figure 6, the confusion matrix, along with accuracy for the classifiers, are shown from (i) to (v). The input parameters and other specifications are emphasized in Table 5. The dataset is divided into two parts, i.e., 80:20 ratios, where 80% of data is used to train the model, and the remaining 20% is used for testing and validation of the model.

The outcome of the regression model is a predicted array. It is compared with that of the actual values (refer table 3). Since it is a Multilinear model, hence it is not possible to plot the results at once. Thus two main plots are plotted individually, i.e., PM 2.5 vs. AQI & RSPM vs. AQI (refer figure 4(i), 4(ii). 5-D visualization of data set for Classification models is represented in figure 4(iii) where the x-axis represents SO2, and the y-axis represents NO2, z-axis represents RSPM, size. Color intensity represents O3, and the color represents AQI SEGMENTS.

![Figure 4(i): RSPM/PM10 vs AQI Plot](image1)

![Figure 4(ii): PM 2.5 vs AQI Plot](image2)
The outcome of the Regression Model (Predicted vs. Actual Values)

| Iterations | Predicted Values | Actual / Validated Values |
|------------|------------------|--------------------------|
| 0          | 108.78           | 102.67                   |
| 1          | 204.302          | 197.5                    |
| 3          | 252.931          | 250                      |
| 3          | 102.234          | 100                      |
| 4          | 73.2078          | 73                       |
| 5          | 106              | 108.5                    |
| 6          | 86.46            | 86                       |
| 7          | 29.79            | 33.5                     |
| 8          | 64.53            | 65                       |
| 9          | 82.65            | 82.13                    |
| 10         | 53.57            | 52                       |
| 11         | 50.30            | 50                       |
| 12         | 120.11           | 115                      |
| 13         | 367.29           | 336.6                    |
| 14         | 54.25            | 56                       |
| 15         | 56.02            | 57.38                    |
| 16         | 50.56            | 51                       |
| 17         | 240.32           | 280.19                   |
| 18         | 183.35           | 180.67                   |
| 19         | 68.06            | 67.33                    |
| 20         | 99.06            | 95                       |

Table 3: Output Predicted (Left) vs. Validated Values (Right)

The below table 4 represents the individual contributions of the various factors that affect the output variable AQI. The higher the contribution of the variable, the more is its influencing factor for calculation of AQI, then classified into clusters based on the AQI Range, as shown in figure 4(iv).
TABLE 4: Variable Contribution w.r.t AQI Calculation

| VARIABLE | CONTRIBUTION (Scale of 0-1) |
|----------|-----------------------------|
| RSPM_PM10 | 0.5557                      |
| PM_2.5   | 0.4065                      |
| O₃       | 0.0184                      |
| NO₂      | 0.0149                      |
| SO₂      | 0.0044                      |

Figure 4(iv): Clustering Scatter Plot

Here, Accuracy is calculated from the confusion matrix based on a pre-defined formula, which is shown below in figure 5.
Figure 5: Accuracy from Confusion Matrix

MAPE (Mean Absolute Percent Error) value is calculated to validate forecast models. The formula for calculating MAPE is shown below as equation 1. It is an independent unit of measurement which gives the error. As a result, the MAPE value should be as low as possible for an efficient model.

\[
MAPE = \frac{\sum |A-F|}{N} \times 100
\]  

(1)

Where A: actual; F: forecast; N: no of observations
Table 5: Input Parameters for Classifiers

| Model                       | Algorithm Input/ Accuracy Parameter | Output/Accuracy Result |
|-----------------------------|-------------------------------------|------------------------|
| Multiple linear regression  | Input variables: SO2, NO2, O3, PM 2.5 & RSPM | Output variable: AQI Multilinear model |
| Artificial Neural Network (ANN) | Confusion Matrix | Accuracy:  
- After Epoch 1 – 92.57  
- After Epoch 2 – 96.66  
- After Epoch 3 – 97.08  
- After Epoch 4 – 97.37  
- After Epoch 5 – 97.47 |
| K-Nearest Neighbour (KNN)   | K (Neighbors) = 5 Minkowski distance Confusion Matrix | Accuracy = 97.92% |
| Support Vector Machine (SVM)| Confusion Matrix | Accuracy = 97.91% |
| Decision Tree               | Confusion Matrix | Accuracy = 99.77% |
| Random Forest               | Confusion Matrix | Accuracy = 99.79% |
| K-Means Clustering          | Input variables: SO2, NO2, O3, PM 2.5, RSPM AQI Segments | AQI Segments 1-6(1- lowest and 6- hazardous) (Refer figure 4(v)) |

There may be other methods available for determining the Air Quality Index, but at the moment, the methods used in the paper are the best suited. A comparative analysis is done between all the models obtained from the different techniques, and the best accuracy one is chosen.

In the following figure 6, the confusion matrix for the classifiers is shown from (i) to (v). The number of true and false predictions are represented with count values, broken down by each class.

![Figure 6(i): Confusion Matrix using SVM](image)

![Figure 6(ii): Confusion Matrix using KNN](image)
In this paper, the algorithm used for forecasting AQI values is Triple Exponential Smoothing. We have then applied a smoothing method (Moving Average Smoothing) to smoothen the curve obtained for better results.

The result set is obtained after applying the model on the validation dataset is analyzed (in this case represented by red lines), which gives the probability, mean, standard deviations, and percent error of each of the AQI Forecasted values. The forecast charts for some cities are displayed below in figure 7(i-v), respectively.
Figure 7: Forecast Plot AQI vs. Time for Various Cities
Hyperparameters and performance comparison of the model in the prediction of AQI via Triple Exponential Smoothing is listed below table 6 and table 7.

Table 6: Triple Exponential Smoothing Summary

| Hyperparameters                  | Values used for AQI Calculation |
|----------------------------------|---------------------------------|
| Period                           | 12                              |
| Start Year                       | 2016                            |
| Start Period                     | 1                               |
| Alpha (coefficient for level smoothing) | 0.3                          |
| Beta (coefficient for trend smoothing) | 0.1                          |
| Gamma (coefficient for seasonality smoothing) | 0.1                          |

Table 7: Performance Comparison (Calculation of various errors)

| Performance Comparison Factors   |   |
|----------------------------------|---|
| $R^2$                            | 0.80115 |
| MSE                              | 5.9031  |
| RMSE                             | 24296.439 |
| MAPE                             | 2.70348 |
| $\chi^2$ (Goodness of Fit)      | 0.73187 |
| f-value                          | 104.5791 |

The monthly forecast and trend for the years 2016-2019 are shown below in figure 8. The plot helps to analyze the AQI values of a specific month of a year to AQI values of some other month of the same/different year.
A contrast of the air pollution concentrates and AQI level between two main seasons, i.e., summer (April-May) and winter (December-February), is shown in figure 9 (i-ii). The plot helps to analyze the AQI values for a specific season of a year by giving the average value for that month of the season.

The monthly values of the years 2016-2018, along with the forecasted values of the year 2019 is shown below in table 8 to get a better statistical understanding.
Table 8: Month-wise AQI values of INDIA (2016-2019)

| MONTHS   | 2016       | 2017       | 2018       | 2019(Predicted) |
|----------|------------|------------|------------|-----------------|
| January  | 629946.5   | 647517.7   | 650359.8   | 704175.1        |
| February | 581855.2   | 567419.2   | 586880.8   | 649925.7        |
| March    | 615295.8   | 596684.6   | 660046.9   | 694598.7        |
| April    | 58881.1    | 581932.5   | 621471.2   | 668288          |
| May      | 616647.4   | 566926.2   | 625523.2   | 698571.4        |
| June     | 546489.9   | 530034.3   | 562361.4   | 627026.3        |
| July     | 544177.2   | 480418.5   | 571190.2   | 625566.6        |
| August   | 524659.1   | 467540.9   | 565444.4   | 609424          |
| September| 530454.5   | 459507     | 561505.7   | 618515.1        |
| October  | 559391.6   | 541059.7   | 640822.5   | 665876.6        |
| November | 581508.2   | 571293.2   | 639584.8   | 692914.1        |
| December | 631947.8   | 602523     | 687226.4   | 754914.9        |

In the linear relationship between AQI vs. PM2.5 & AQI vs. RSPM, it can be seen that the predicted values (red) vs. the actual costs (green) are quite close to each other, which leads to an inference that the linear regression model provides high accuracy. But since our model is a multivariate model, linear regression is not a feasible solution to plot a graph of all the contributory variables vs. the output variable. Hence, we use other machine learning classifiers for model creation, as shown in figure 10, along with their accuracy.

![Classifier vs Accuracy(%) relation](image)

Figure 10: Accuracy of Classifiers
The results from the phase of model creation are considered to be the final measure of accuracy for the model. Validation is done, which will help us to determine and eliminate the possibility of over or underfitting. On the other hand, testing leads to the conclusive process of determining the robustness and effectiveness of the model. Testing and debugging are followed by forecasting AQI values. The same process will be followed to make plots for four metropolitan cities (Delhi, Bangalore, Mumbai, Kolkata) and Bhilai, which is then automated to get the AQI results for all the 196 cities for the upcoming months.

6. CONCLUSION AND FUTURE SCOPE

This research paper probed the relationship between AQI and supplementary factors affecting AQI. Our final goal is to develop unique and more efficient models to predict AQI with excellent efficiency in minimum computational time. Many current air quality forecasting methods use only linear techniques which would miss nonlinear relationship in the data. After implementation, the result manifested that Artificial Neural Networks can express the non-linear relationship between the contributory variables and AQI even though the accuracy to predict was less favorable. With the Decision Tree Classifier, the model has given the best accuracy of 99.7%, which increases by 0.02% on the application of the Random Forest Classifier, as shown in figure 10. The Time Series analysis is widely used to make predictions, and the ARIMA model has become one of the most popular and fastest-growing methods in the predictions and forecasting research. This study has demonstrated the potential of using nonlinear machine learning methods to improve air quality forecasts. In the future, aggregation of model selection and time series analysis can be considered for better efficiency of the model and to provide results with higher accuracy in less computational time.

FUNDING

There is no funding from any government or private bodies

COMPLIANCE WITH ETHICAL STANDARDS

1) Conflict of Interest: The author(s) declare(s) that there is no conflict of interest.

2) Research involving human participants and animals: This article does not contain any studies with human participants or animals performed by any of the authors.

3) Informed consent: This article does not contain any studies with human participants performed by any of the authors. Therefore, obtaining informed consent does not apply.

REFERENCES

[1] Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1), pp. 3–10. doi:10.1016/j.gsf.2015.07.003

[2] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp. 255–260. doi:10.1126/science.aaa8415

[3] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, pp. 159–175. doi:10.1016/s0925-2312(01)00702-0

[4] McLeod, A. I., Yu, H., & Mahdi, E. (2012). Time Series Analysis with R. Handbook of Statistics, pp. 661–712. doi:10.1016/b978-0-444-53858-1.00023-5

[5] M. Schain, “Machine Learning Algorithms and Robustness,” vol. 7, no. 1, p. 237, 2015, Retrieved from https://www.tau.ac.il/.

[6] H. Peng, “Air Quality Prediction by Machine Learning Methods,” no. October 2015, doi:10.14288/1.0166787.

[7] P. Hajek and V. Olej, “Predicting common air quality index – The case of Czech microregions,” Aerosol Air Qual.
[8] E. D. Gjika and L. Puka, “Using the R-package to forecast time series: ARIMA models and Application Using the R-package to forecast time series: ARIMA models and Application,” no. ISTI 2010, March 2015, Retrieved from https://www.researchgate.net/.

[9] F. R. A. Evans, “The Birmingham Lunacy Prosecution,” Lancet, vol. 146, no. 3765, pp. 1066–1067, 1895, Retrieved from https://www.thelancet.com/.

[10] A. Forster, “How does Automated Analytics do it? The magic behind creating predictive models automatically,” SAP Blogs, pp. 1–17, 2015, Retrieved from https://blogs.sap.com/.

[11] Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, and Jingjie Wang, “Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory-based geographic model and wavelet transformation,” Institute of Remote Sensing and GIS, Peking University, Beijing 100871, China, 2015, doi:10.1016/j.atmosenv.2015.02.030.

[12] Niharika, Venkatadri M and Padma S. Rao, “A survey on Air Quality Forecasting Techniques,” (IJCST) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, 103-107, Retrieved from http://www.ijcst.org/.

[13] Shivangi Nigam, B.P.S. Rao, N. Kumar, and V. A. Mhaisalkar, “Air Quality Index – A Comparative Study for Assessing the Status of Air Quality,” Research J. Engineering and Technology, Raipur, Vol. 6, Iss. 2, (Apr-Jun 2015): 267-274, doi:10.5958/2321-581X.2015.00041.0.

[14] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie, “Air Quality Prediction: Big Data and Machine Learning Approaches,” International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018, doi: 10.18178/ijesd.2018.9.1.1066.

[15] Dixian Zhu, Changjie Cai, Tianbao Yang, and Xun Zhou, “A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization,” Big Data and Cognitive Computing, vol. 2, no. 1, p. 5, Feb. 2018, doi:10.3390/bdccc2010005.

[16] Konstantina Veljanovska and Angel Dimoski, “Air Quality Index Prediction Using Simple Machine Learning Algorithms,” Volume 7, Issue 1, January - February 2018, Retrieved from https://www.ijettes.org/.

[17] Slini, T., Karatzas, K., & Moussiopoulos, N. (2002). Statistical analysis of environmental data as the basis of forecasting: an air quality application. Science of The Total Environment, 288(3), pp. 227–237. doi:10.1016/s0048-9697(01)00991-3.

[18] Jiangshe Zhang and Weifu Ding, “Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong,” International Journal of Environmental Research and Public Health, vol. 14, no. 2, p. 114, 24 January 2017, doi: 10.3390/ijerph14020114.

[19] Bougoudis, I., Demertzis, K. & Iliadis, L. HISYCOL, a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens. Neural Comput & Applic, Volume 27, pp. 1191–1206 (2016). doi:10.1007/s00521-015-1927-7

[20] Fei, L.L.; Chan, L.Y.; Bi, X.H.; Guo, H.; Liu, Y.L.; Lin, Q.H.; Wang, X.M.; Peng, P.A.; Sheng, G.Y. Effect of cloud-to-ground lighting and meteorological conditions on surface NOx and O3 in Hong Kong. Atmos. Res. 2016, 182, pp. 132–141.

[21] Yeganeh, B, Motlagh, M.S.P, Rashidi, Y, Kamalan, “H. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model.”, Atmos. Environ, vol. 55, 357–365, August 2012, doi: 10.1016/j.atmosenv.2012.02.092.