JOURNAL OF OPTOELECTRONICS LASER ISSN: 1005-0086

Volume 41 Issue 8, 2022

AQI Predication Using Temporal Data Mining

Mr Shashi Bhushan[1] **and Dr. Sanjay Kumar Tiwari**[2]

[1]*Assistant Professor, Department of computer Applications, Gaya College Gaya , Bihar*

[2]*Associate Professor, PG Department Of Mathematics, Magadh University Bodh Gaya, Bihar.*

ABSTRACT

An indication of air quality standards, the Air Quality Index (AQI) is measured exploitation information about air waste that are known to cause harm to people or the environment. Many human activities release harmful chemicals, particles, or biological materials into the air, affecting both humans and their natural surroundings. One of the most pressing environmental concerns is air pollution, especially in urban and industrial areas. It emphasizes the significance of pollution forecasting and prevention. One of the least fascinating and hard challenges is air pollution prediction exploitation data mining. Data storage, stock management, and statistic generation are just some of the many tasks made easier by the proliferation of various technologies. The Air Quality Index (AQI) for India is a time-averaged measure of several air pollutants (such as SO2, NO2, rspm, spm, and so on). The major purpose of the study is to forecast near-term daily AQI using the AQI from the day before and other meteorological elements via the use of temporal data mining using a gradient descent technique and Nave Forecasting. We divided the dataset into a basic 85% and a second 15 % for use as test and train data, respectively, in order to better identify the most important seasonal swings and patterns. When it comes to estimation problems, Linear Regressions are frequently used, while Naive Forecasting and the gradient descent method are used when making predictions. Using a multivariate regression gradient descent problem, researchers can predict the air quality index using data from previous years and look forward to a single upcoming year..

Keywords: AQI, Dataset, Predication, Naïve Forecasting, Gradient Descent.

I. INTRODUCTION

Statistics, impermanent databases, impermanent shape identification, optimization, superior computation, visualization, and symmetric computation all meet in the rapidly developing field of Temporal Data Mining. Cognition revelation in impermanent information includes temporary data mining, which catalogs spatial information structures. Any method that catalogs special shape from or fits spatial data is considered a spatial information mining approach. Popular temporal data mining methods were created with an eye on mining enormous amounts of time-related data and making effective use of all available time information. [1] Particles including carbon dioxide, ozone,(Particle Matter) PM 2.5, and nitrogen oxides contribute to the worldwide problem of air pollution. Due to its small size, it is easily inhaled and absorbed into the lungs and circulatory system, where it has been linked to an increase in lung cancer. Disease, sickness, and mortality, threat to other living creatures, including food harvesting, and harm to the natural environment are only a few of the many negative outcomes associated with exposure to air pollution. That's why keeping an eye on the levels of pollution in the air is so important. Many decision-making aids aim to keep tabs on data, but few can really do so effectively. One of the most interesting and tough tasks is predicting air pollution using data mining, and here we discuss the prediction approaches used to monitor air quality in the next days and months. [2].

The AQI is a number utilized by authorities to exposit the state of the air at a certain location. Parameters related to air pollution are given weights, and those values are then converted to a individual figure or set of book of numbers using the AQI Scheme. AQI is utilized for both global and territorial air attribute control [3] in a great number of important cities. The major intention of the survey is to forecast daily AQI in the near future by using the AQI from the previous day and meteorological elements via the use of temporal data mining with a gradient descent technique and Nave Forecasting. Researchers accessed information on pollution levels in India from a government database. All of the data points in the collection can have their air quality index predicted according to a new algorithm developed by researchers. The air quality in any city in India can be predicted using our technique. India's clean air can be stored anywhere. Researchers can invert the principal pollution-causing contaminant and the place in India significantly impacted by the pollutant by projecting the air quality index. Air quality forecasts can be generated using a variety of methods, including models based on pollution source emission monitoring data, models based on meteorological data, and models based on historical statistics. Sources of pollutant emission data collection via monitoring provide significant difficulties. The focus of such an article would be on analyzing a hypothesis for predicting urban air quality based on weather and pollutant records. The same source of environmental pollution is different under various weather conditions since air quality prediction are extremely interdependent on national meteorological factors. The air quality prediction contains significant nonlinear aspects because of the influence of today's pollution levels. [4].

To make a forecast, users need to come up with a number, set of figures, or scenario that represents an event in the future. Long-term and short-term preparations should be made. In contrast to a subjective prediction based on intuition, gut feeling, or supposition, a forecast is described as a prediction based on historical evidence. For instance, most definitions of regression, a method that is sometimes used in forecasting, say that its objective is to explain or predict. When the time frame of the forecast is reduced, the accuracy of the predictions improves. So, a prediction for tomorrow would be more particular than a prediction for the next period,

which is much many specific than a prognosis for the next period of time, which is more exact than a forecast for the next decade. WilliamJ. Stevenson lists a number of qualities typical of accurate forecasts.1.Accurate- Some degree of accuracy should be calculated and expressed so that the prediction can be compared to others.2.Reliable- In order to instill trust in the user, the prediction method must consistently provide accurate results.3.Timely- Certain that a given amount of time is needed to implement changes in response to the projection, it is important that the horizon for making such forecasts be sufficiently long.4.Cost Effective-It's important that the advantages gained from creating the projection are not outweighed by the costs involved. The term "air pollution index" refers to a system for translating the (weighted) values of several air pollution-related component into a individual figure or a set of numbers.
Table1: Description For various range of AQI

II. Methodology

Structure and behavior of the system are shown in a diagram called an architecture diagram, which places special emphasis on the system's architecture. It has several different pieces to the design that all work together in unison to carry out the larger program. The design is based on a menu structure that mimics database navigation and data entry. After everything has been processed, yearly statistics graphs are made, and then AQI is computed.

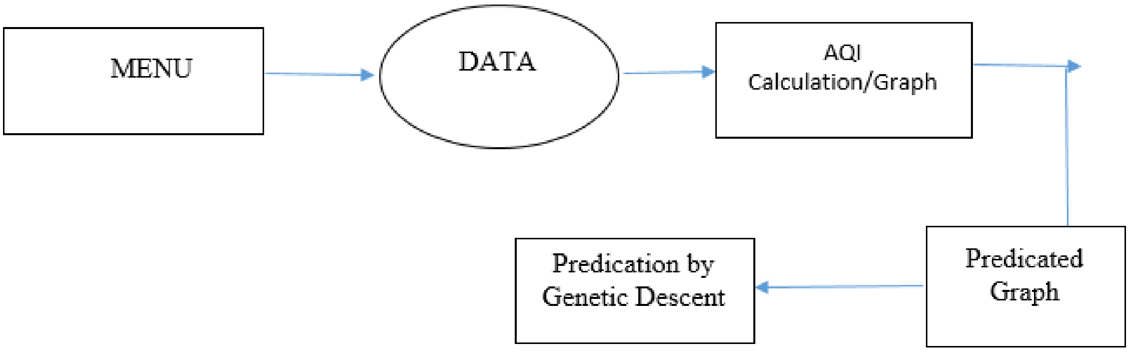


Fig1: Block Diagram

1. **Data Collection-**We gathered information from 24/7 online air monitoring stations from 1990 to 2014. Some of the air pollution concentrations studied were O3, PM2.5, and SO2. We choose the meteorological factors that would have an impact on air pollution, such as the temperature, humidity, wind speed and wind speed and direction, direction, total accumulation of precipitation, wind rainfall, visibility, dew point, pressure, wind direction, and weather.

2. **Performance Evaluation-**Researchers employ many different statistical measures, such as the correlation coefficient, the root mean squared error (RMSE), the mean absolute percent error (MAPE), and the mean absolute error (MAE), to evaluate the relative effectiveness of the various regression models (R). This page contains all the relevantequations.

Mean absolute error (MAE)

$$MAPE = \frac{\sum_{k=1}^{n} \left| \frac{t_k - y_k}{t_k} \right|}{n} \times 100\%$$

Mean absolute percent error (MAPE)

$$MAE = \frac{\sum_{k=1}^{n} | t_k - y_k |}{n}$$

$$R = \frac{\sum_{k=1}^{n} (t_k - \bar{t})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n} (t_k - \bar{t})^2 \sum_{k=1}^{n} (y_k - \bar{y})^2}}$$

Correlation Coefficient®

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n} | t_k - y_k |^2}$$

Root Mean Square Error (RMSE)

If n is the figure of information points, *k y* is the anticipated quantity, *k t* is the discovered quantity, *t* is the ascertained information mean, and *y* is the ascertained information ordinary. The attempt content choice is assessed because it represents the exactitude of each regression model.

3. **Naive Forecasting-** The estimation technique of simply using the previous period's actuals as the basis for the forecast for the current period, without making any adjustments or attempting to identify causative elements. Its only purpose is to evaluate competing projections from more sophisticated approaches. In order to determine whether the original final forecast has been enhanced, Naive Forecast provides a comparative benchmark independent of the final prediction. Many businesses have yet to explore the potential of Nave Forecast, which is similar to a baseline forecast based on facts and information. It's a basic approach to predicting that helps spread the most basic benchmark for comparison. Organizations evaluate if the nave prediction is worse or better than the organization's own final forecast.

III. Techniquesused

1. Linearregression

Mathematically, linear regression examines the relationship between two continuous variables and makes adjustments where necessary. Linear regression uses a linear approach to characterize the relationship between a continuous dependent y variable and one or more explanatory factors denoted by the X-coordinates. One example of a descriptive inconsistent value is seen in the context of simple linear regression. When there are many variables to consider, this technique is referred to as multiple linear regression. We use linear predictor features to model the relationships, and their data can be used to infer the values of the models' black boxes.Thesemodelsareadverttoaslinearmodels.Specifiesthelastcubedrop-off linefor theplacedof ncontent points,

y =ax +b

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\operatorname{Cov}[x, y]}{\operatorname{Var}[x]} = r_{xy}\frac{s_y}{s_x},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x},$$

2. Gradientdescent Algorthim

Finding the parameters (coefficients) of a function (f) that minimizes a cost function is the goal of the optimization technique known as gradient descent. When it is not possible to determine the parameters analytically (through linear algebra, for instance), gradient descent is a useful optimization tool for finding the optimal values. The objective of the method known as gradient descent is to reduce the value of a function as much as possible.When working with massive datasets, gradient descent's execution speed might suffer. When working with a large training dataset (tens or hundreds of millions of examples), one iteration of the gradient descent technique needs a prediction for each occurrence. Stochastic gradient descent is a kind of gradient descent that is useful when dealing with a huge quantity of data.

The more broad approach known as gradient descent is utilized in linear regression. This section will teach you how to utilize the gradient descent technique to find the least value of a cost function given an arbitrary feature f. First, we'll take educated wild guesses about what 0 and 1 could be, and then we'll keep tinkering with them until the formula's predictions line up:

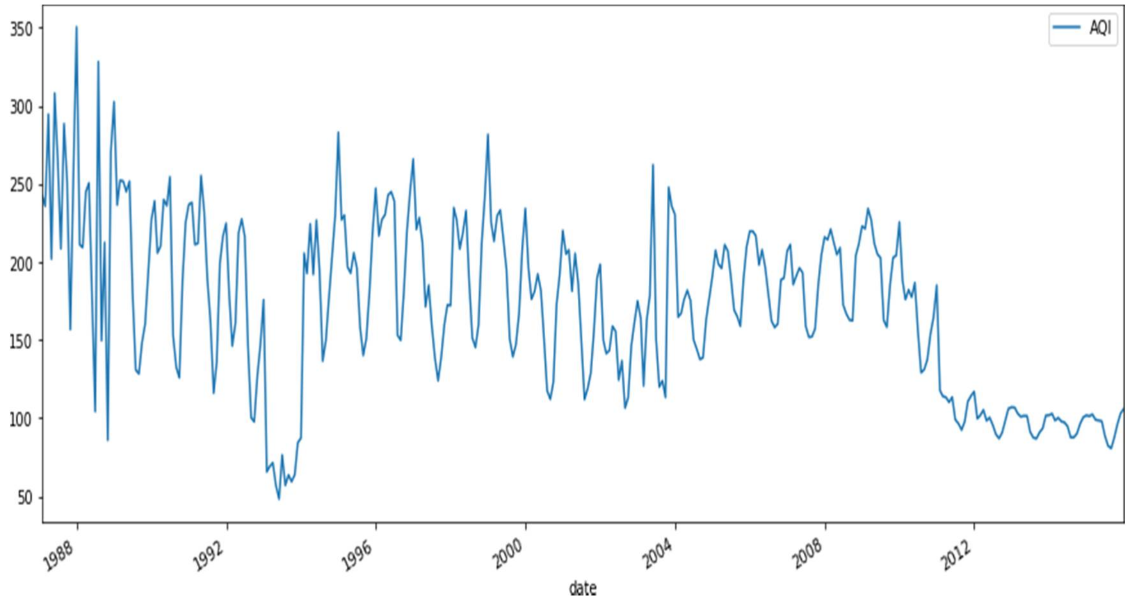
θj:=θj−α∂∂θjf(θ0,θ1)forj=0,1

Here, the acquisition charge is name *dα*, and it determines how bigger a step is required when change the constant. The acquisition rate is a optimistic figure at all period. We want to modify both j=0 andj=1at the same time, i.e. Compute the right-hand side of the higher up equalization, and then modify the constant values to the recently measured ones. Until convergence is reached, this procedure will be repeated. In this study, we utilize gradient descent to make predictions about the relationship between actual value and predicted value.

IV. Results and Discussion

Predicting the air quality index of a given area requires consulting the cpcb.nic.in website, which details the contaminating concentration of all gases and contains all the information that pollutes the cities yearly. Formulas for the AQI can be used in conjunction with a linear regression method to determine the AQI for a given year. Multiple datasets are copied into the directory, and the endless data is blanked out.

Fig2: Time Series Visualization



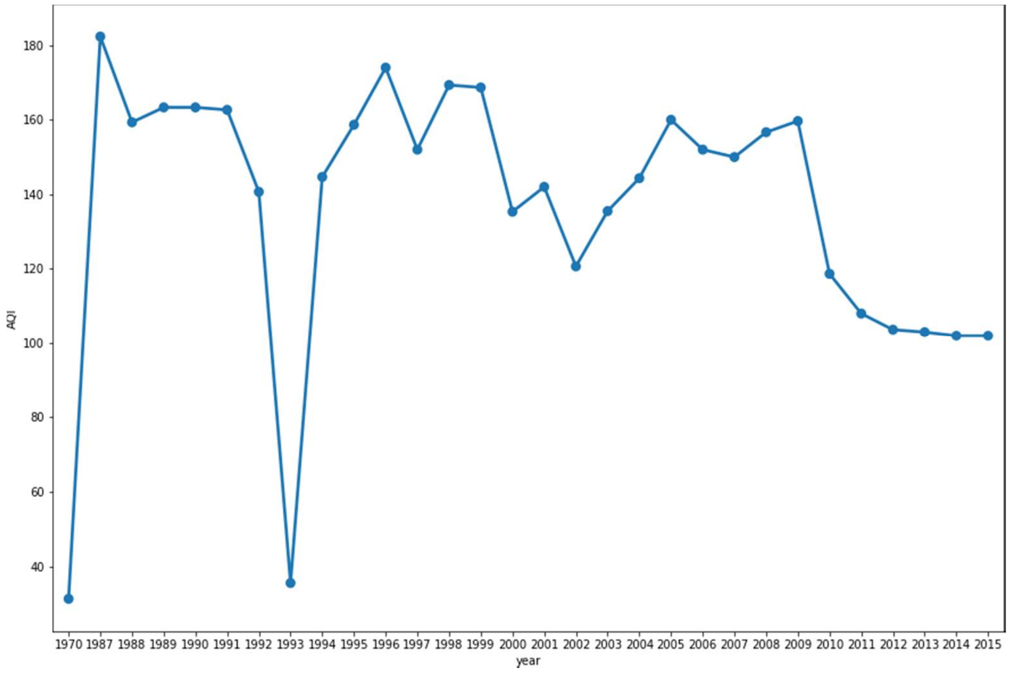


Fig3: Visualization output year wise

The average AQI for each year in India is shown here. The air quality index at a given data point is the sum of the indices for all of the pollutants in that area that are over a certain threshold. Pollutants' maximum sub index is used as the air quality index for that area. Air Quality Index Predication- To further understand the massive seasonal variances and patterns, we used the Naive Forecast technique to divide the data set in half, storing the first 85% in train datasets, and the remaining 15% in test datasets. Many businesses have yet to explore the potential of Naive Forecast, which is similar to a baseline forecast based on facts and information. In the study, we provide an annualized definition of the outcome of a naive forecasting technique. In such section, the author provide a comparison of the AQI's actual and predicted values over the course of a year (1988-2016). To visualize our data, researchers computed a moving average of each data point and displayed it. Researchers identified the ranges of the moving averages one year at a time (2008-2012), i.e., there are disparities at x minimum and x maximum prior to 2008, and at y minimum and y maximum after 2012.

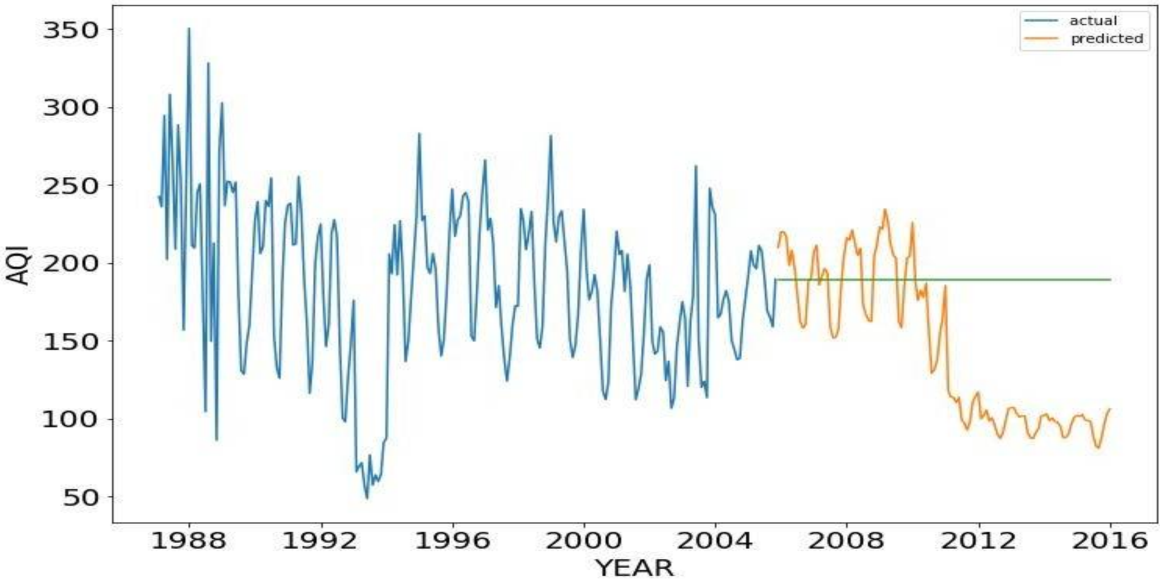


Fig4: Naïve Forecast

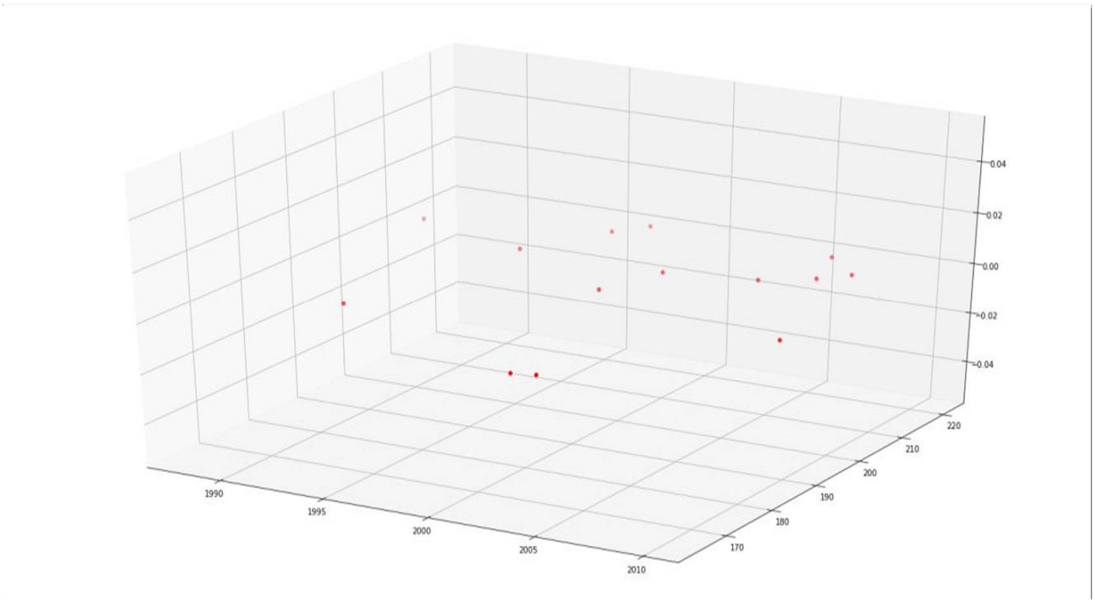


Fig5: Temporal Data Mining Output

It predicated the temporal data mining in 3-D axis form in which they show result in year wise.

DOI: 10050086.2022.08.68

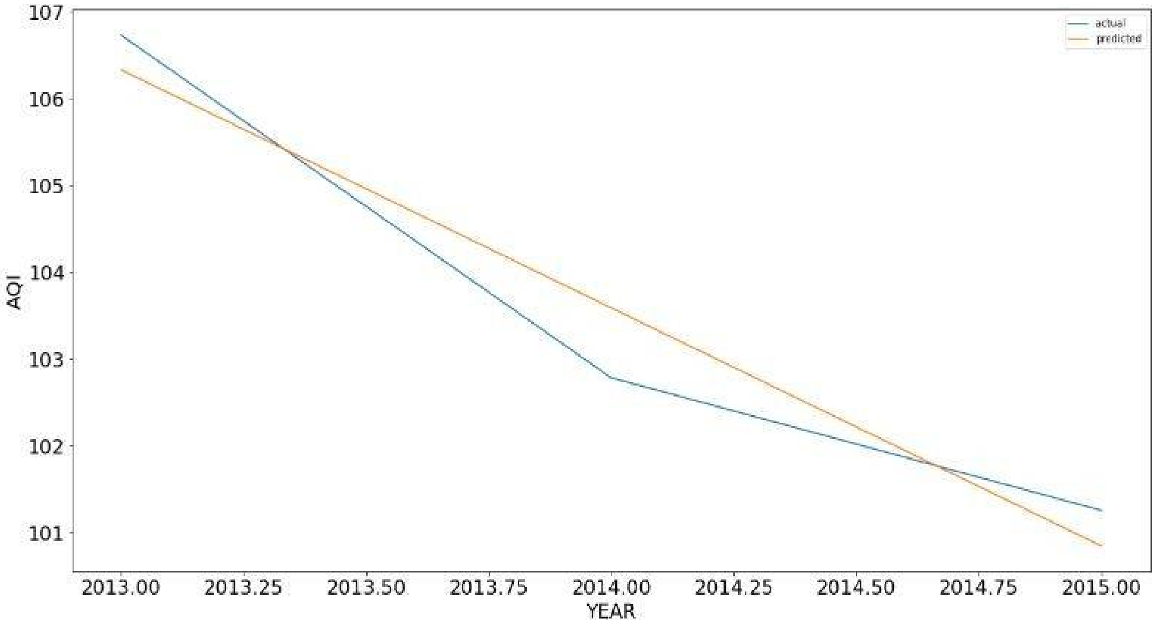


Fig6: Actual Vs. Predicated in Gradient Descent (2013-2015)

Gradient Descent employed iterations to compare historical data with future forecasts. Each year, a new set of results is applied. Since air contains innumerable elements that can be made or performed daily, air pollution is the most common difficulty individuals face.

V. CONCULSION

Containing levels of air pollution has risen rapidly in importance and is now among the top responsibilities of governments. It is crucial for people to learn about the pollution levels in their area and to take action to reduce it. This finding demonstrates the efficacy of using unsupervised learning techniques (logistic regression and gradient descent) to determine the air quality index and forecast RSPM concentrations. Using the time series data from each area, the model can forecast the AQI and relay that information to the appropriate authorities. Furthermore, it is a advanced learning model that can pinpoint the exact location that requires tending. Both regular individuals and meteorologists will benefit from the adopted model's ability to identify and anticipate pollution levels, allowing for more informed and timely responses to the problem. It would aid people in establishing a news hub for less-frequented communities.

REFERENCES

[1] Lin, W., Orgun, M.A. and Williams, G.J., 2002, December. An Overview Of TemporalDataMining. In*AusDM*(pp. 83-90).

[2] Azid,A.,Juahir,H.,Toriman,M.E.,Kamarudin,M.K.A.,Saudi,A.S.M.,Hasnam,C.N.C.,Aziz,N.A.A.,Azaman,F.,Latif,M.T. ,Zainuddin,S.F.M.andOsman,M.R., 2014.Prediction of the level of air pollution using principal component analysis and artificialneuralnetworktechniques:AcasestudyinMalaysia.*Water,Air,&SoilPollution*,*225*(8),p.2063.

[3] Ma,Y.,Richards,M.,Ghanem,M.,Guo,Y.andHassard,J.,2008.Airpollutionmonitoringandminingbasedonsensorgridin London.*Sensors*,*8*(6),pp.3601-3623.

[4] Huang,M.,Zhang,T.,Wang,J.andZhu,L.,2015,September.Anewairqualityforecasting model using data mining and artificial neural network. In *2015 6th IEEEInternational Conference on Software Engineering and Service Science (ICSESS)* (pp.259-262).IEEE.

[5] Dixian Zhu, Changjie Cai, Tianbao Yang and Xun Zhou: A Machine Learning ApproachforAirQualityPrediction: ModelRegularizationandOptimization.Bigdataandcognitivecomputing.

[6] Carbajal-Hernández,JoséJuan"Assessmentandpredictionofairqualityusingfuzzylogicandautoregressivemodels."Atmospheric Environment60(2012):37-50.

[7] Kumar,AnikenderandP.Goyal,“ForcastingofdailyairqualityindexinDelhi”,Scienceof th TotalEnvironment409, no. 24(2011):55175523.

[8] Wang Z et al , “ A nested air quality prediction modelling system for urban and regionalscales:Applicationforhighhigh-ozoneepisodeinTaiwan“Water,AirandSoilPollution130.1-4(2001):391-396.

[9] RussoAnaFrankRaischelandPedroG.Lind,“Airqualitypredictionusingoptimal neuralnetworks withstochasticvariables”, AtmosphericEnvironment 79(2013): 822-830.

[10] SivacoumarR,et al,“Airpollutionmodellingforanindustrial complexandmodelperformanceevaluation“ ,Environmental Pollution 111.3 (2001)471-477.

[11] Singh Kunwar P., Shikha Gupta and Premanjali Rai, “Identifying pollution sources andpredictionurbanairquality usingensemblelearningmethods”,Atmosphericenvironment80(2013):426-437.

[12] Peng,H.,2015.*Airqualitypredictionbymachinelearningmethods*(Doctoraldissertation,Universityof British Columbia).

[13] M. Caselli & L. Trizio & G. de Gennaro & P. Ielpo. "A Simple Feedforward NeuralNetwork for the PM10 Forecasting: Comparison with a Radial Basis Function NetworkandaMultivariateLinearRegressionModel." WaterAirSoilPollut(2009) 201:365–377.

[14] S.Bordignon,C.GaetanandF.Lisi,"Nonlinearmodelsforground-levelozoneforecasting." StatisticalMethodsand Applications, 11, 227-246, (2002).

[15] EdwinDiday.Symbolicdataanalysis:amathematicalframeworkandtoolfordatamining.InAdvances in Data Science and Classification, pages409–416.Springer,1998.