Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Air Quality Index Prediction Based on Different Technique in Machine Learning in New Delhi

B. Raghavaiah[1], Lakinani Vaikunta Rao[2], Vijaya Laxmi[3], V. Ravi Kumar[4], Kollapudi Sreenivasulu[5], Amit Gupta[*6]

[1]Department of Electronics and Communication Engineering, Ramachandra College of Engineering, Eluru, Andhra Pradesh, India.

[2]Department of Chemistry, J. B. Institute of Engineering & Technology, Hyderabad, Telangana, India.

[3]Department of Computer Science and Engineering (AIML), Joginpally B.R. Engineering College Hyderabad, Telangana India.

[4]Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad, Telangana, India.

[5]Department of Chemistry, Narasaraopeta Engineering College, Narasaraopet.Andhra pradesh, India.

[6]Department of AI & ML, J. B. Institute of Engineering & Technology, Hyderabad, Telangana, India.

Email: raghavaece39@gmail.com, ivraochemistry@gmail.com, dr.ravikumar@klh.edu.in, vijumathpati13@gmail.com; sreenukollapudi@gmail.com,dramitguptacv@gmail.com*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Among the many factors contributing to the significant problem of air pollution are the combustion of fossil fuels, industrial activity, and economic growth. This study's main concern is the use of data collection methods for machine learning-based air quality forecasting. The study highlights how pollutants can affect a person's health, including cardiovascular and respiratory conditions, bronchial asthma episodes, strokes, and even death. To solve the problem, we propose to use artificial neural network and data gathering techniques. When applied to classification and regression issues, a variety of machine learning techniques, including decision trees, offer a tree-like structure of findings and alternative predictions. Until a stopping condition is satisfied, the dataset is iteratively divided using the feature that produces the most information gain or impurity reduction.By examining data from the previous year, this study provides a useful method for forecasting the air quality in New Delhi for the subsequent month. This explains how different machine learning algorithms and data acquisition are used to address difficult problems.

Keywords: Data Acquisition, Artificial Neural Network, Support Vector Machine, Linear Regression ,Air Quality Index.

INTRODUCTION

The atmosphere receives air pollutants from numerous sources, which alter the atmosphere's chemical composition and have an effect on the biotic environment. In order to survive and for living things to be healthy, it is essential to maintain good air quality. However, lung and respiratory conditions have become more prevalent. Air pollution is brought about by growing populations, industrialization, the use of fossil fuels, the creation of polluting gases (such as NOx, CO, SO2, O3, and so on) from old car exhaust.

While PM10 particles are released during building projects, landfill operations, agricultural activities, wildfires, and garbage burning, PM2.5 particles are released when gasoline, oil, petroleum, or other fuels are burned. The burning of wood, charcoal, or other fuels releases CO, but the inflammation of air in cars and other fuel-burning operations releases NOx. Sulfur dioxide (SO2) [2–3] is released into the atmosphere after the burning of fossil fuels in power stations and other industrial facilities. Ozone, a naturally occurring form of oxygen found in the planet's stratosphere, helps absorb ultraviolet radiation from the sun. All of these elements play a role in disorders such as lung disease, asthma attacks, heart attacks, miscarriages, and other illnesses. These problems can be solved with data mining and other techniques. A strong technique for looking through big data sets to find hidden patterns and connections between variables is data mining. Professionals in industries including marketing, finance, healthcare, and security can utilize information to make informed decisions. Organizations can use data mining techniques such as decision trees, clustering, and association rule learning to extract useful information from both structured and unstructured data [9–10]. Finding correlations between different types of information has become simpler thanks to big data platforms like Hadoop, which can reduce risk and save money for start-up companies. Ultimately, by exposing fresh business possibilities and insights, data mining [11–16] provides an aggressive boundary.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine learning requires data mining, which is the process of sorting through enormous data sets to identify trends and provide insights. Two examples of modern analytical methods that can be used to mine data for connections and patterns that could have otherwise gone unnoticed are artificial intelligence (AI) and natural language processing (NLP). Businesses may create models to predict customer behavior and enhance their offers using data-driven insights, empowering them to make well-informed decisions regarding their goals and future. By pinpointing areas where innovation is required or wanted, data analytics can also be utilized to bridge the gap between technical capabilities, customer desires, and business goals.

I.Decision Tree

Decision trees are a well-liked machine learning method for applications involving regression and classification. In order to construct a tree-like structure of options and their probable outcomes, it begins with a root node that symbolizes the entire dataset and branches out to several internal nodes and leaf nodes.

The edges of each internal node demonstrate the possible outcomes of a feature test, and each node within the node functions as a representation of a feature test. The path from the root node to a leaf node represents the collection of options that produce a prediction for a particular sample. The tree is the end outcome. You can locate the leaf node that has the majority class or forecast by navigating around the tree until you find it. You will then be able to predict the results for new samples. Both continuous and categorical data can be handled by decision trees, which are easy to understand. However, if the tree is allowed to grow too deeply or if there are many noisy qualities in the data, they may overfit. A "decision tree" is a graphic representation of workable answers to a problem that creates a tree-like structure according to predetermined standards. In data mining, decision trees are a collection of mathematical and computational methods that aid in the description, classification, and extrapolation of a given data set. Information is included in the following records: (x1,x2,x3,...,xn,Y) = (X,Y).
The reliant element the aim variable, Y, is what we're attempting to comprehend, categorize, or generalize. The vector x, expressed as x1, x2, x3, etc., represents the features used for that activity.

General structure: The figure 1 illustrated that decision free structure in machine learning technology. Each internal node in the decision tree above represents an attribute test, each branch represents a result, and each leaf node represents a class label.

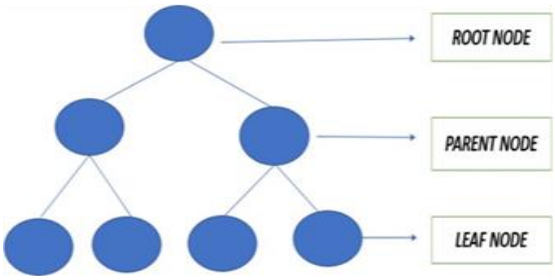


Figure 1 shows the decision tree's basic structure

II. Linear regression

A recommendation concerning the value of one variable is provided via linear regression analysis that corresponds with the appreciate of another inconstant. The term "dependent variable" refers to the variable that you are trying to forecast. To predict the value of the other variable, use the independent variable.A technique for estimating values in the actual world is linear regression, which uses continuous variables. Assuming linear regression, it is utilized in a wide range of fields, including healthcare, finance, and economics. Four assumptions need to be validated in order to do linear regression or create a relationship between one or more independent and dependent variables.

1. Variance homogeneity

2. Self-sufficiency

3. The concept of linearity

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. Normalcy

III. Support Vector Machines

Support Vector Machines (SVM) are widely employed in machine learning to solve categorization difficulties. Support Vector Regression (SVR) is a machine learning algorithm. SVM is an SL method that divides the plane into two sections by drawing a line between the two classes. The line that divides the plane into multiple pieces is known as a hyperplane. It always returns a perpendicular distance between the data point and the line of separation. It is capable of both linear and nonlinear classification. SVMs are learning tools that successfully generalize a limited set of learning patterns by using the inductive principle of structural risk minimization. Reducing both the empirical risk and the VC (Vapnik-Chervonenkis) dimension at the same time is the aim of structural risk minimization, or SRM.Based on a separable bipartition problem, Small patterns can be found in enormous data sets using a learning method called SVM. In order to forecast the classes of previously unidentified data, the method uses discriminative categorization learning by example.

2. OBJECTIVE

The Environmental Protection Agency (EPA) determines the Air Quality Index (AQI) for five major air pollutants, for which national air quality standards have been established to protect public health. The objective of the AQI is to help people understand how the local air quality affects their health by informing them about the level of air pollution and its potential impact on human health, allowing them to take appropriate protective measures.

3.LITERATURE REVIEW

A. Kashyap and colleagues' research Paper examines the problem of by pollution of the atmosphere focusing on metric PM2.5. They also used the linear regression method to do this. In their study, K. Gao et al. look at a number of machine learning and big data analytics dashboards and air characteristic prediction methods. The writers stress how crucial it is to preserve and keep an eye on the quality of the air in industrial and urban regions. The article assesses the results of published research on artificial intelligence, deep learning, and decision trees as approaches to air quality evaluation. Additionally, it discusses the challenges faced in this subject and emphasizes the necessity for additional research.

In this study, Zaini et. al. review DLNN models for prognostication air characteristic. It does not, however, compare DLNN models to other approaches; it exclusively examines DLNN models. The authors may have overlooked some important studies because they only searched two databases. The report doesn't offer a thorough analysis of the models and doesn't go into enough detail on the review process. It is therefore challenging to replicate their findings. Finally, the report makes several recommendations for further research but doesn't detail how these recommendations might be implemented.

In their study, R. Waman and D. S. Deshpande address the critical problem of classifying the health hazards related to air pollution by proposing an approach that utilizes the pollution levels. The study offers a brand-new method for categorizing health hazards by looking at air quality levels. It presents a system for categorizing air quality data and forecasting related health hazards using machine learning techniques. To examine and evaluate the hazards, the authors gather data on air characteristic from several surveillance stations and use categorization algorithms. With the use of timely information, politicians, healthcare professionals, and the general public will be better able to make judgments and take preventative action. The authors address the expandability of their approach, which one is essential for its application in the real worldwide. A more thorough analysis of the various difficulties and constraints brought on by scaling up the framework would have been useful.Huabing Ke and associates provide a general description of air pollution and apply scientific predictions in the study's first portion. They offer a solution that uses five different machine learning approaches. Due to the availability of sensor data and external sensing networks, as well as recent developments in big data applications, many academics are now adopting the big data analytics method. Through the analysis of numerous big data and machine learning techniques, this study seeks to improve air quality predictions.

This endeavor will use machine learning technology to evaluate precision of the prognosticate of fine particulate matter or PM 2.5, in environment population in Malaysia's modern city. We give a thorough analysis of the key

achievements made in the field between 2011 and 2021. Following a rigorous review, we looked at 155 papers after searching the major scientific publication databases. Physiographic dissemination, projected appreciate, prognosticator factors, accuracy measure, and machine learning technology are used to categorize the publications by M. Manuel et.al. Liu, Huixiang, et al. reported creating regression models using Support Vector Regression (SVR) and Random Forest Regression (RFR) on two publicly available datasets in order to predict the Air Quality Index (AQI) in Beijing and the greenhouse gas emissions satisfied in an Italian city.In a study by H. Kumar, artificial intelligence predictive algorithms to forecast the amount of fine particulate matter in atmospheric air were tested using data from the Taiwan Environmental Monitoring from 2012 to 2017. When it comes to making predictions, these models perform exceeding the current norms in the sector.

S. Krzysztof looked at those two methods of feature selection. One uses a linear stepwise fit approach, which is a locally favorable strategy, and the other uses a genetic algorithm, which is a worldwide strategy. Two sets of the most predictive traits are chosen on the determinant of such investigation. These sets participate in the prognostication of the contaminants PM10, SO2, NO2, and O3 in the atmosphere. Comparison is made between two prediction methods. In the first, a decision tree ensemble called a random forest (RF) is formed using the features that have been chosen.

S. Taneja et al. reported that data mining has been utilized to examine current patterns in Delhi's air pollution and develop future projections. Multilayer perceptron's and linear regression are the data mining methods employed. Using the aforementioned methods, we have seen that the amount of PM 10 will rise by 45.9% in the upcoming years. But because there are more two-wheelers on the road, CO and NO2 levels can be slightly higher.

According to M. Yadav, apriori-based association rule mining, a modified version of the Continuous Target Sequential Pattern Discovery (CTSPD) technique, is used to generate a set of association rules that help predict the engrossment of air pollutants. This algorithmic program only creates rules for continuous events since it takes the temporal component of the input into account. The algorithm's performance is evaluated by examining climatological data and ambient air quality that were gathered in Anand Vihar, New Delhi, between September 1, 2015, and August 31, 2016. When compared to an existing prediction system, the suggested method's forecast is proven to be more accurate than SAFAR's.

J. k. Sadhasivam el. al. reported that the open source software Prophet Algorithm is used to prognosticate the trend pollution of the air in Mumbai, Maharashtra. A machine learning system called Prophet forecasts and predicts chronological order data. It is founded on an additive model in which seasonality is matched to non-linear patterns on a weekly and annual scale. This method produces the graphical results that display the trending pattern of the contaminants in the air of Mumbai [20].

Since data mining is a promising technique for estimating PM2.5 change, Shenyang, one of the most significant industrial cities in Northeast China with significant air pollution, is selected as the example city. Weather data from 2013 to 2015, including temperature, humidity, precipitation, wind speed, and PM2.5 concentrations, are used in this projection. Three data mining methods have been established by the World Health Organization (WHO) that provide PM2.5 estimates straight from meteorological data. After analysis, it was shown that the random forest model outperformed the other two in terms of prediction accuracy. Lastly, the accuracy of the generated models is assessed.

A multilayer neural network was used by Pérez et al. to predict the hourly PM2.5 concentrations in Santiago [17]. However, the authors of the experiment employed a single indication (previous AQI data) to forecast the current AQI, ignoring the link between several pollutants (e.g., PM2.5, PM10, etc.). Corani et al. forecasted Milan's PM10 and ozone levels using feed-forward deep neural networks [18]. The recursive neural network model gave correct estimations 95% of the time. However, the 30% false positive rate highlights how poorly the neural network model replicates concentration peaks. An empirical model was developed by Fuller et al. to forecast PM10 concentrations at London background and roadside locations. [19].

Network training may now be done without having long-term parameters "explode" or "vanish" as a result of several learning updates thanks to the implementation of techniques like LSTM for RNNs (Pascanu, Mikolov, and Bengio 2013). This work uses machine learning models, such as Decision Tree (DT), Random Forest (RF), and Gradient

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Boosting Regressor (GBR), to anticipate PM 2.5 hourly scale concentrations using meteorological data and PM 2.5 concentrations from neighboring stations. The dataset was collected in China's Beijing, a research area. According to the experiments, the gradient boosting repressor model outperforms the other models proposed for hourly PM 2.5 concentration forecasting in terms of predictive precision, with an R2 value ranging from 0.9 to 0.97.

The present research illustrated how the ANN may be used to predict air quality in urban areas like Ahvaz in order to prevent harmful health effects. To evaluate the spatial-temporal profile of contaminants and air quality indicators, Tey came to the conclusion that air quality regulators may employ a computerized neural network [24]. In order to predict air pollution, machine learning (ML), and in particular deep learning (DL), models for regression issues, have recently attracted a lot of attention [25], [26].A light gradient boosting machine concept is put out in [27].

Beijing has 35 air quality monitoring facilities to handle extremely comprehensive, considerable data collected from those sites. The concentrations of PM2.5 over the upcoming 24 hours are predicted.

We used gradient boosting decision trees (GBDT), deep neural networks (DNN), and extreme gradient boosting (XGboost) in accordance with standard mean absolute, mean absolute error (MAE), and symmetric mean absolute percentage error. The results show that their model performed better than the mean absolute error (MAE).The performance of the model is also improved by the wide use of historical data [28].

This was based on the comparison of predicted AOD data with AOD data from the Aerosol Robotic Network (AERONET) and Moderate Resolution Imaging Spectroradiometer (MODIS) satellites [29].

In order to do this, this study suggests a Bayesian Optimization (BO)-based Lag-FLSTM (Lag layer-LSTM-Fully Connected network) model for multivariant air quality prediction. To test the approach, a case study is carried out in the United States. According to the findings, Lag-FLSTM has an RMSE that is at least 23.86% lower than that of other approaches [30]. Autoregression is used to predict future PM2.5 values based on historical PM2.5 data. We can reduce PM2.5 levels below the hazardous threshold by knowing what they will be in the next years, months, or weeks. This technique attempts to determine air quality and forecast PM2.5 levels using a data collection of the daily atmospheric conditions in a specific city [31].

The most hazardous environment for people and a threat to life is one with a high AQI rating. Consequently, AQI monitoring and forecasting have emerged as crucial tools for sustainable development worldwide (Rybarczyk and Zalakeviciute, 2021) [32].

November and August were determined to be the most polluted and cleanest months in the city, respectively. Usually tested between October and February, the air quality index (AQI) ranges from poor to severe. Higher levels were observed in November as a result of the combined effects of stubble burning in Delhi's neighboring states and pyrotechnics from Diwali celebrations [33]. The inhalable fraction of ambient particles (particulate matter with aerodynamic diameter less than 10 μm, represented by PM10) collected at three residential locations in Delhi, India between December 2008 and November 2009 revealed the presence of eight major and trace metals (Fe, Mn, Cd, Cu, Ni, Pb, Zn, and Cr)[34]. The average annual 24-hour PM10 levels at the locations varied between 166.5 and 192.3 μg m−3 (8–10 times the WHO limit). Weekday and weekend impacts on PM10 and associated metals were investigated. Principal component analysis–multiple linear regression (PCA–MLR) identified three primary sources: crustal (49–65%), vehicular (27–31%), and industrial (4–21%). As part of health risk assessment, theoretical estimates of children's blood lead levels varied from 10.2 to 13.3 μg dL-1, indicating a certain degree of lead toxicity [35].

4PROPOSED SYSTEM

Environmental awareness among the general public has grown recently. Therefore, it's critical to monitor changes and anticipate those that may be hazardous or irregular. Therefore, it could be beneficial to use a machine learning technique like a decision tree. These algorithms are then compared in order to determine the most accurate and effective method. The objective is to analyze and deliver it effectively. Decision trees are graphs that resemble trees and are used to represent decisions and their results. It is an easy-to-use approach to solving problems with

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

decision-making in artificial intelligence and machine learning.Here is an overall description of a Decision Tree's operation.

Determining the root node: Identifying the root node: The root node is the population or sample of the entire group.

Splitting the population into smaller groups; The root node is divided into two or more sub-nodes, which stand for more compact and homogeneous groupings, according to a certain criterion.

Creating decision nodes: As decision nodes, the sub-nodes reflect a question that must be answered in order to choose which branch to take. Establishing leaf nodes, the leaf nodes, or final results or decisions, are reached via branches from the decision nodes.

Choosing the best split: Several techniques, including as the chi-square test, information gain, and Gini impurity, are used to break the population up into smaller groups. The split that produces the most homogeneous groupings is chosen. That air quality index parameter was depicted in picture 2.

idea of how to categorize air quality for simpler understanding and action.

Figure 2. The illustration above presents an approximate in different level. The figure 3 illustrated that parameter using in training data set in the machine learning model.

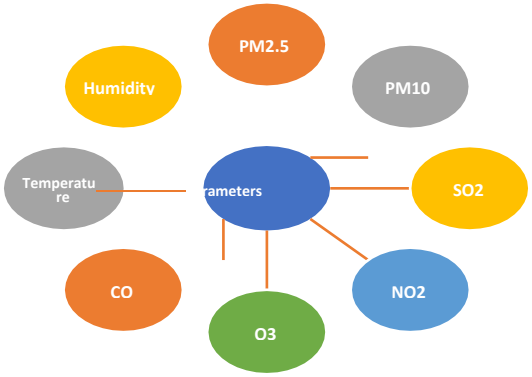


Figure 3. Parameters used for training machine

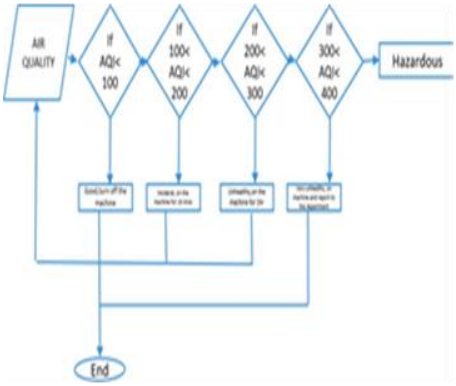


Figure 4. Delhi's air quality graphic output of air quality.

I. **Dataset used:** This model was trained using one year's worth of data from the Delhi government's official website.

I. Pseudocode

First, import libraries. Step 2: Load the data 3. Examine the dataset.

Verify any missing data in step four.

Separate data into goal and attributes in step five.

Create training and test sets from the data.

Step 6: In step 7, construct a decision tree regressor model.

Teaching the model is step eight.

Based on the test sets from Step 9, make predictions.

10. Evaluate the model.

Step 11: Value of the feature

Step 12: Tuni hyper parameter

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1.1 DECISION TREE ALGORITHM

Step 1: From a selection of training situations, pick one characteristic.

Step 2: To start with, pick a subset of the training examples.

Step 3: Utilize the property and the subset of cases to build a decision tree.

Step 4: Stop after you've accurately classified every case.

Step 5: If a classification is incorrect, the instance should be added to the original subset and a new tree should be constructed.

Step 6: Carry out steps 5 and 6 again until you have a tree that properly classifies every occurrence or a tree built from the whole training set.

II. Flowchart

The decision tree's procedure is described in [1] above. The air quality index in Delhi was depicted in Figure 4. Figure 5, which shows the air quality graphic output, provides a visual representation of Delhi's air quality. The output includes information on the concentrations of pollutants such as sulfur dioxide, carbon monoxide, nitrogen oxide, particulate matter, and ozone.

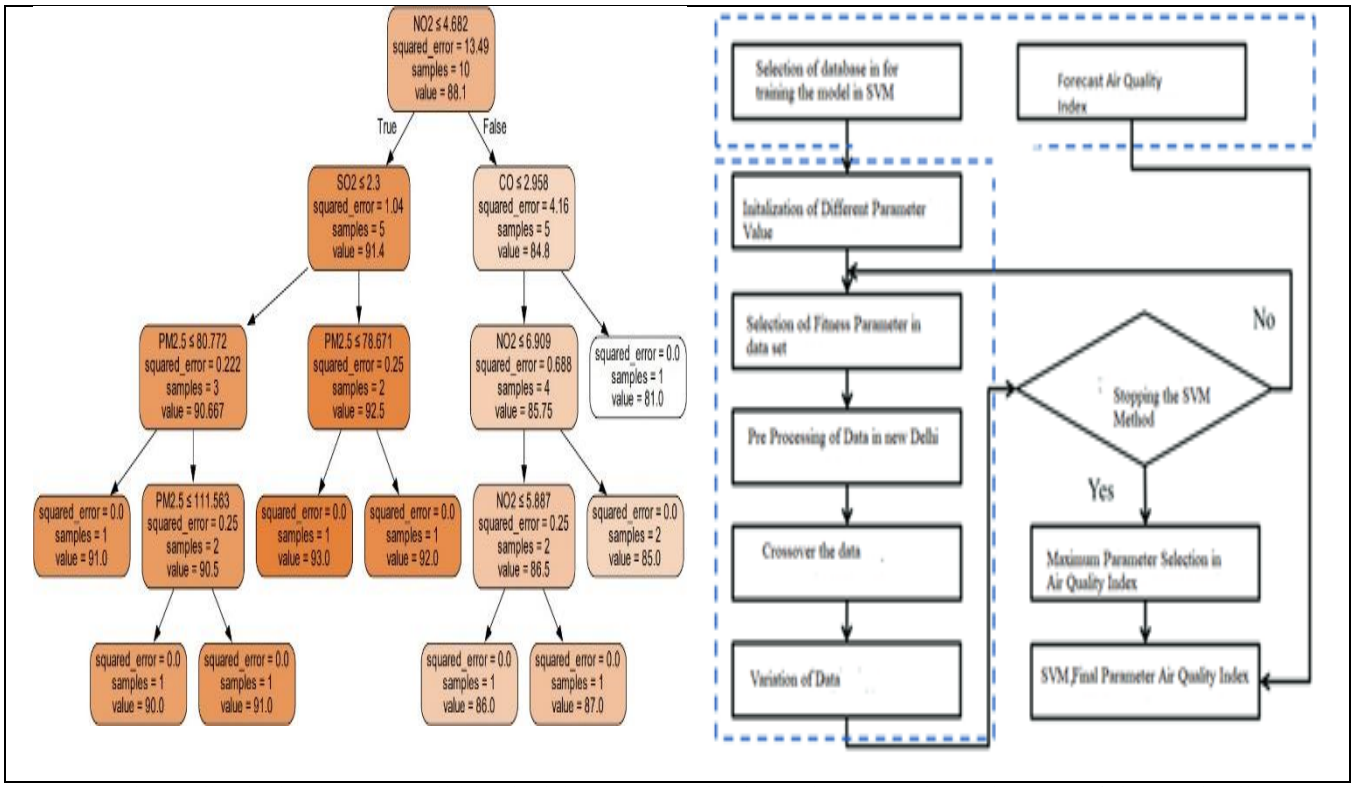


Figure 5. Air quality index in new Delhi in decision tree method

Figure 6. Flow Diagram of the methodlogy

IV. SUPPORT VECTOR MACHINE (SVM)

The support vector machine used the following stage to run the algorithm**.**Since the line equation is Y = Wx + b, SVR and LR are similar. SVR refers to this line of straight as a hyperplane. The support vectors—the data points nearest to the hyperplane on either side—are used to plot the boundary line. Within a certain range—the distance between the boundary line and the hyperplane—SVR seeks to fit the optimum line[36].

1)Stage 1: Data Collection: In this stage, we collect all data related to air pollution attributes. In smart cities, there are several sensors that detect pollutants.

2)Stage 2: Preprocessing the data involves removing noise and filling in missing values.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3)Stage 3: Select features using GA. Feature selection is the process of selecting the most pertinent inputs for the prediction model selection. This technique can be used to find and eliminate unnecessary, redundant, and non-essential traits that don't increase or decrease the prediction model's accuracy.

4)Stage 4: Multivariate Multistep Time Series Prediction Using Random Forest: In this step, we forecast air pollution using multivariate, multistep time series data by applying a random forest algorithm. A separate subset of the time-series data was used to train each tree.

5) Stage 5: Prediction in this case, our method forecasts air pollution.

V.RANDOM FOREST ALGORITHM

The inner product threshold was found using the grid search approach, with the acquired average classification accuracy serving as a reference. Accordingly, CARTs with low average classification accuracy and CART pairs with inner product values above the inner product threshold were deemed deletable. The average classification accuracies and correlations of CARTs were carefully analyzed in order to construct the random forest; those with large correlations and weak classification effects were removed, while those of higher quality were retained. The lead was shown to be consistent in several testing, and the proposed enhanced random forest performed better in terms of average classification accuracy than the five random forests used for comparison. Since the proposed upgraded random forest outperformed the five random forests in terms of G-means and out-of-bag data (OBD) scores, the lead was more obvious. Additionally, the results of three non-parametric tests show a substantial difference between the five random forests and the proposed enhanced random forest. This effectively illustrates the augmented random forest's superiority and feasibility [37].

VI.LINEAR REGREESION:

Selecting features well can lower the computational cost, lower the chance of overfitting, and increase the model's interpretability. Filtering, wrapping, and embedding are the three categories into which feature selection techniques fall [38-39].

Using statistical testing, the filtering process chooses traits. By building several models, the wrapping approach assesses the significance of features. The feature selection procedure is incorporated into the model training procedure by the embedding approach. A detailed discussion of the various problems that can affect the execution and interpretation of linear regression analysis is provided here. Through real-world examples, the reader is made aware of common interpretation problems. The limits of linear regression analysis as well as its potential applications are discussed.

In real-world applications, feature selection typically necessitates a blend of data attributes and domain expertise to identify the optimal approach.

The primary goal of linear regression techniques is to determine the best parameter estimation so that the discrepancy between the actual data and the model's predicted value is as little as feasible.The following steps usually make up the algorithmic flow:

1)**Selecting Features**: Selecting features involves deciding which independent variables to include in the model. Correlation analysis, domain expertise, or feature selection algorithms can all be used for this.

2)**Model Fitting**: Estimate a linear model's parameters using least squares or other optimization methods.

3)**Model Evaluation**: Assess the model's performance using test sets or cross-validation.

4)**Model Diagnosis**: Verify that the model complies with the fundamental presumptions of linear regression and make any necessary adjustments for potential incompatibilities.

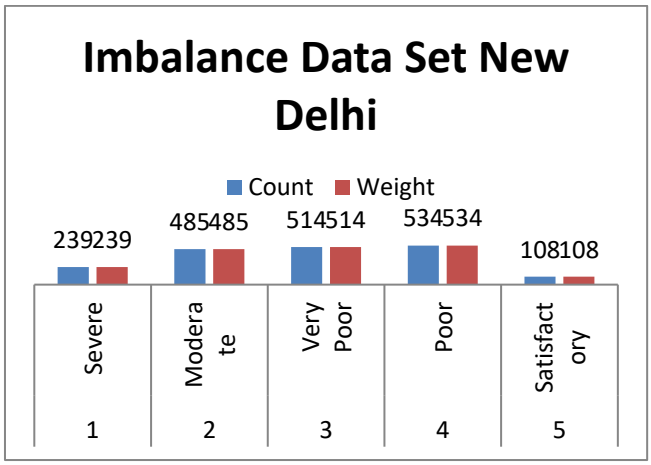
Application and prediction: Make predictions and apply them to actual issues using well-fitting models.

regression and adjust for possible model incompatibilities.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Prediction and application: Use well-fitted models to make predictions and apply them to realproblems.The figure 8 indicate that imbalanced data set new delhi air quality index different parameter in count value and weight parameter.The figure 9 illlustrated that balanced data set checking the air quality index.



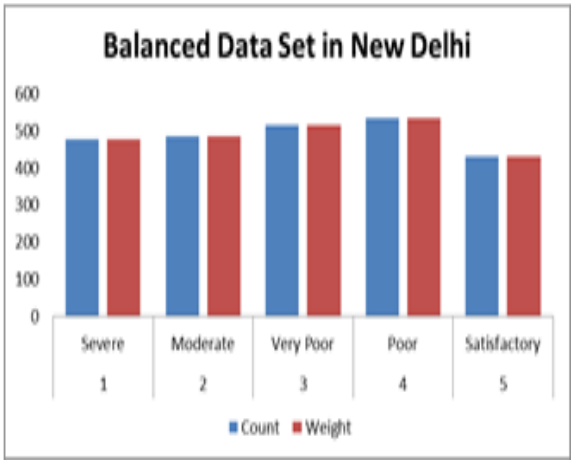
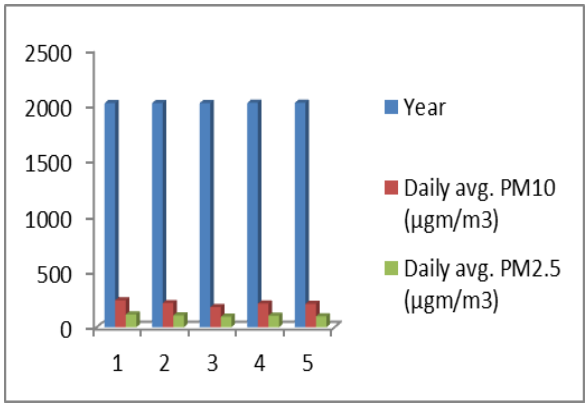


Figure 8. Imbalanced Data Set New Delhi Air Quality Index Figure 9. Balanced Data Set New Delhi Air Quality Index

The figure 10 illustrated that daily average PM10 and daily average value in the last five years. The figure 11 indicated that air quality index different value of PM10 and PM 2.5 value in the last five years.



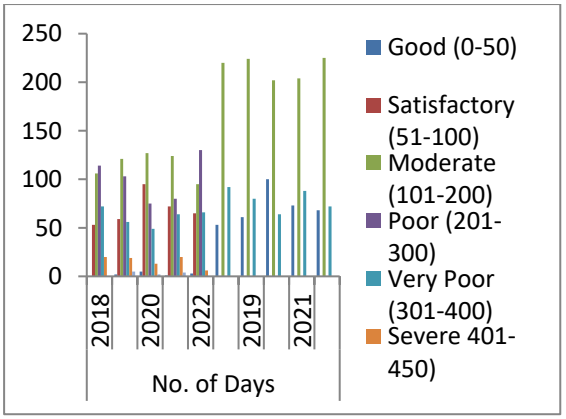


Figure 10: Daily Average PM10 and Daily Average Value and PM 2.5

Figure 11. Air Quality Index Different Value of PM10

The figure 12 illustrated that different types of the error value in support vector machine. The figure 13 percentage accuracy in three machine learning method such as support vector machine, random forest regression and linear regression method.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

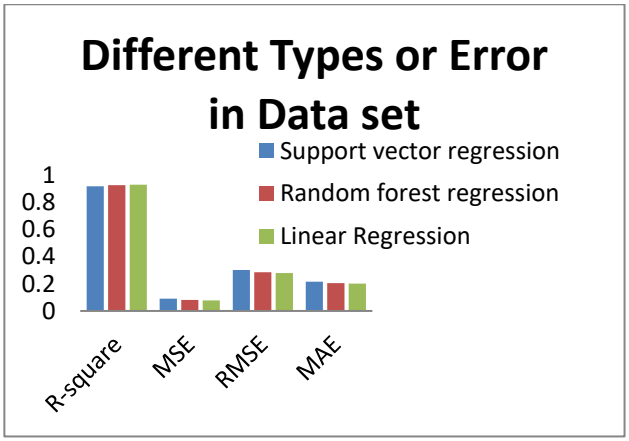


Figure 12. different types of the error value

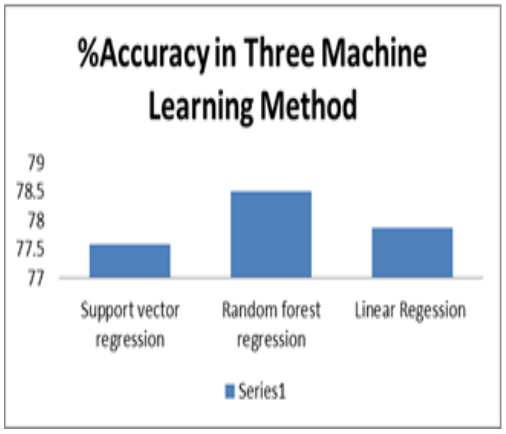


Figure 13. percentage accuracy in three machine learning method

The figure 14 indicate that air quality index in new delhi in different time interval. The figure 15 comparison of the dataset size with and without decision tree algorithm different parameter of air quality index value.

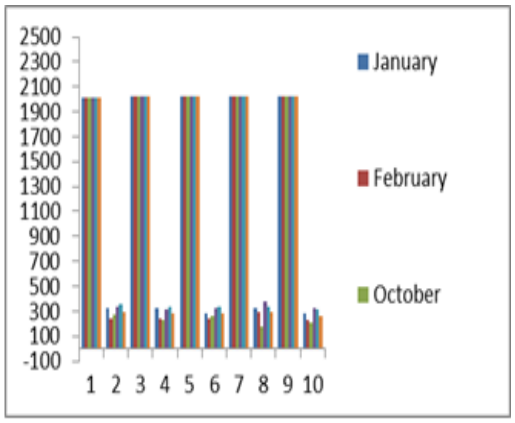


Figure 14. air quality index in new delhi in different time interval

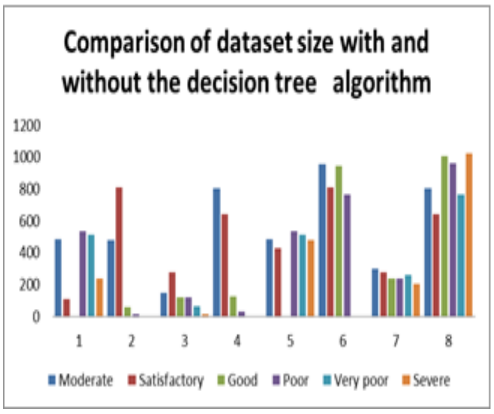


Figure 15. Comparison of the dataset size with and without decision tree algorithm

VII.CONCLUSION

We also go over how decision trees may help businesses make data-driven discoveries that allow them to develop models for forecasting user behavior and even improve their services by leveraging state-of-the-art analytical tools like AI and NLP. This study highlights the need of using artificial intelligence and data mining to predict air quality as well as the potential for using these technologies to combat air pollution globally. We collected data for a whole year and focused on Bengaluru's expected air quality, which showed a steady rise in NO2 pollutants and forecasted the pollutant level for the upcoming month**.**

REFERENCES

[1] Kashyap, Abhishek, and Soumyalatha Naveen. "A Comparative Study on Prediction of PM2. 5 Level Using Optimization Techniques." 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). IEEE, 2021.
[2] Kang, Gaganjot Kaur, et al. "Air quality prediction: Big data and machine learning approaches." Int. J. Environ. Sci. Dev 9.1 (2018): 8-16.
[3] Zaini, Nur'atiah, et al. "A systematic literature review of deep learning neural network for time series air quality forecasting." Environmental Science and Pollution Research (2022): 1-33.
[4] Gore, Ranjana Waman, and Deepa S. Deshpande. "An approach for classification of health risks based on air quality levels." 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM). IEEE, 2017.

[5] Ke, Huabing, et al. "Development and application of an automated air quality forecasting system based on machine learning." Science of The Total Environment 806 (2022): 151204.
[6] Shaziayani, Wan Nur, et al. "Classification Prediction of PM10 Concentration Using a Tree-Based Machine Learning Approach." Atmosphere 13.4 (2022): 538.
[7] Liang, Yun-Chia, et al. "Machine learning-based prediction of air quality." applied sciences 10.24 (2020): 9151.
[8] Méndez Manuel, Mercedes G. Merayo, and Manuel Núñez. "Machine learning algorithms to forecast air quality: a survey." Artificial Intelligence Review (2023): 1-36.
[9] Kumar, K., and B. P. Pande. "Air pollution prediction with machine learning: A case study of Indian cities." International Journal of Environmental Science and Technology (2022): 1-16.
[10] Liu, Huixiang, et al. "Air quality index and air pollutant concentration prediction based on machine learning algorithms." Applied Sciences 9.19 (2019): 4069.
[11] Harishkumar, K. S., K. M. Yogesh, and Ibrahim Gad. "Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models." Procedia Computer Science 171 (2020): 2057-2066.
[12] Siwek, Krzysztof, and Stanisław Osowski. "Data mining methods for prediction of air pollution." International Journal of Applied Mathematics and Computer Science 26.2 (2016): 467- 478.
[13] Shweta Taneja; Nidhi Sharma; Kettun Oberoi; Yash Navoria "Predicting trends in air pollution in Delhi using data mining." 2016 1st India international conference on information processing (IICIP). IEEE, 2016.
[14] Yadav, Mansi, Suruchi Jain, and K. R. Seeja. "Prediction of air quality using time series data mining." International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2. Springer Singapore, 2019.
[15] Jayakumar Sadhasivam, V Muthukumaran, J Thimmia Raja, V Vinothkumar, R Deepa and V Nivedita , "Applying data mining technique to predict trends in air pollution in Mumbai." Journal of Physics: Conference Series. Vol. 1964. No. 4. IOP Publishing, 2021.
[16] Zhao, Chang, and Guojun Song. "Application of data mining to the analysis of meteorological data for air quality prediction: A case study in Shenyang." IOP conference series: earth and environmental science. Vol. 81. No. 1. IOP Publishing, 2017.
[17] Pérez, P.; Trier, A.; Reyes, J. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. 2000, 34, 1189–1196.
[18] Corani, G. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model. 2005, 185, 513–529.
[19] Fuller, G.W.; Carslaw, D.C.; Lodge, H.W. An empirical approach for the prediction of daily mean PM10 concentrations. Atmos. Environ. 2002, 36, 1431–144.
[20] Chavi Srivastava; Shyamli Singh; Amit Prakash Singh,"Estimation of Air Pollution in Delhi Using Machine Learning Techniques"International Conference on Computing, Power and Communication Technologies (GUCON), ISBN:978-1-5386-4491-1.
21) Abdellatif Bekkar; Badr Hssina; Samira Douzi; Khadija DouzAir Quality Forecasting using decision trees algorithms.
22) Liu, B.; Yan, S.; Li, J.; Li, Y. Forecasting PM2.5 concentration using spatio-temporal extreme learning machine.In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 950–953.
23) H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, "Air pollution prediction by using an artifcial neural network model," Clean Technologies and Environmental Policy, vol. 21, no. 6, pp. 1341–1352, 2019.
[24] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," Atmospheric Environment, vol. 80, pp. 426–437, 2013.
[25] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, "Comparative analysis of machine learning techniques for predicting air quality in smart cities," IEEE Access, vol. 7, pp. 128 325– 128 338, 2019.
[26] Q. Chen, W. Wang, F. Wu, S. De, R. Wang, B. Zhang, and X. Huang, "A survey on an emerging area: Deep learning for smart city data," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 3,no. 5, pp. 392–410, 2019.
[27] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang,and L. Huang, "A predictive data feature exploration-based air quality prediction approach," IEEE Access, vol. 7, pp. 30 732–30 743, 2019.
[28] Doreswamy, K. S. Harishkumar, Y. Km, and I. Gad, "Forecasting air pollution particulate matter (pm2.5) using machine learning regression models," vol. 171. Elsevier B.V., 2020, pp. 2057–2066.
[29] Y. Bao et al.Assessing the impact of Chinese FY-3/MERSI AOD data assimilation on air quality forecasts: sand dust events in northeast China Atmos. Environ.(2019)

[30J Ma, Y. Ding, J. C. Cheng, F. Jiang, V. J. Gan, and Z. Xu, "A lagflstm deep learning network based on bayesian optimization for multisequential-variant pm2.5 prediction," Sustainable Cities and Society,vol. 60, 9 2020.
[31] C. R. Aditya, C. R. Deshmukh, N. D K, P. Gandhi, and V. astu, "Detection and prediction of air pollution using machine learning models," International Journal of Engineering Trends and Technology, vol. 59, no. 4, pp. 204–207, 2018.
[32] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters," Journal of Electrical and Computer Engineering, vol. 2017, Article ID 5106045, 14 pages, 2017.
[33] A. Garg, N.C. Gupta "The great smog month and spatial and monthly variation in air quality in ambient air in Delhi, India",J. Heal. Pollut., 10 (2020), 10.5696/2156-9614-10.27.200910.
[34] S.K. Guttikunda, R. Goel, P. Pant,"Nature of air pollution, emission sources, and management in the Indian cities"Atmos. Environ., 95 (2014), pp. 501-510, 10.1016/J.ATMOSENV.2014.07.006.
[35] P.S. Khillare, S. Sarkar "Airborne inhalable metals in residential areas of Delhi, India: distribution, source apportionment and health risks"Atmos. Pollut. Res., 3 (2012), pp. 46-54, 10.5094/APR.2012.004.
[36] Amit Gupta,Dr. Shashi Kant Dargar and Dr. Abha Dargar," House Prices Prediction UsingMachine Learning Regression Mode", DOI: 10.1109/ICMNWC56175.2022.10031728, ElectronicISBN:978-1-6654- 9111- 2, Print on Demand(PoD) ISBN:978-1-6654-9112-9.https://ieeexplore.ieee.org/document/100317.
[37] Amit Gupta, Shaik Meeravali Arun Singh Chouhan, K.Y.Srinivas", Activity Based Learning System Educational Institutions for Measuring Students Preference using Machine Learning Technique" DOI: 10.1109/ICONSTEM56934.2023.10142321. Electronic ISBN:979-8-3503- 4779-1,ISBN:979-8-3503-4780-7. https://ieeexplore.ieee.org/document/10142321.
[38] Amit Gupta, Meeravali Shaikh, G. Jithender Reddy, Arun Singh Chouhan," Detection of Cardiac Failure using the Convolutional neural network (ConvNets CNN's) Technique", DOI: 10.1109/AISP57993.2023.10134963, Electronic ISBN:979-8-3503-2074-9, ISBN:979-8-3503- 2075- 6. https://ieeexplore.ieee.org/document/10134963.