Environment Protection Engineering

Vol. 50 2024 No. 4

DOI: 10.37190/epe240401

SAMPATH HARISH KUMAR (ORCID: 0009-0007-2124-0467)[1]
THORAPADI CHANDRASEKARAN KANISH (ORCID: 0000-0002-5381-4337)[1]

MACHINE LEARNING APPROACH TO PREDICT AND COMPARE THE AIR QUALITY INDEX IN A CONFINED ENVIRONMENT

Indoor air pollution is very dangerous as people spend the majority time indoors. Cooking areas are found to be hazardous as there would be an emission of harmful pollutants. This is due to the continuous cooking process which affects people working there causing them various diseases, especially carbon monoxide poisoning. The purpose of this research is to evaluate several machine learning algorithms like support vector machine (*SVM*), *K*-nearest neighbor (KNN), logistic regression (LR), and decision tree (DT) for predicting the air quality index (*AQI*) of a Barbeque Nation Hotel kitchen's confined interior environment. This investigation was done based on real-time data that was gathered by an indoor air quality monitoring system which was placed inside the kitchen for a few weeks under various cooking conditions. Results show that DT has the highest accuracy of 98.79% followed by KNN with an accuracy of 93.01%. *SVM* has an accuracy of 80.34%, and LR has a low accuracy of 80.20%. Therefore, DT which is a classification algorithm that comes under supervised machine learning has predicted *AQI* accurately compared to others. Moreover, by segregating living from non-living particulate matter and nullifying them, airborne diseases like COVID-19 can be prevented in the future.

1. INTRODUCTION

In recent times, air pollution has evolved into a disaster that affects the whole world. Air quality has deteriorated in the majority of urban areas. There are many different and various types of contaminants, such as carbon monoxide, carbon dioxide, nitrogen dioxide, sulfur dioxide, ammonia, particulate matter (PM2.5, PM10), acrolein, asbestos, benzene, carbon disulfide, polycyclic aromatic hydrocarbons, synthetic vitreous fibers, and total petroleum hydrocarbons. Monitoring these pollutants in the air has garnered more consciousness in recent years since it has a substantial influence not only on the

[1]School of Mechanical Engineering, Vellore Institute of Technology (VIT), Vellore-632014, Tamil Nadu, India, corresponding author T. C. Kanish, email address: tckanish@vit.ac.in

health of people but also on the ecological balance [1]. Industry's impact on the environment may be gauged using various air pollutant indexes. The category for the air quality index (*AQI*), which was collected from the Central Pollution Control Board of India, the impacts on one's health are detailed in Table 1 [2]. Therefore, by inhaling the polluted air many health problems are caused especially for people working in industries where harmful gasses are released.

Table 1

Standard *AQI* ranges and their effects

Range	Category	Effects on human health
0–50	good	quality of the air is good
51–100	satisfactory	no harmful effects
101–200	moderate	discomfort in breathing for people with lungs or heart problems
201–300	poor	affects people with lung or heart diseases
301–400	very poor	exercise tolerance decreases for people with heart or lung problems
Over 400	severe	dangerous to health

Various disorders of the respiratory system and cardiovascular systems are caused by pollution from industries. There is a correlation between the concentration of pollutants like particulate matter (PM2.5), NO_2, CO and the prevalence of illnesses affecting the respiratory system among the workers and their families [3]. Heart disease risk concerning other diseases increases from 1% to 3% after chronic inhalation of PM, especially those smaller than 2.5 μm in diameter [4]. The study focuses on the increased risk of reproductive system problems caused by exposure to contaminated air [5].

Air pollution (particularly NO_2) is linked to an increased risk of illnesses affecting the reproductive system. According to estimates provided by the WHO, various pollutants present in the air are directly proportional to roughly an estimated 700 000 fatalities per year in China and India [6]. As per Dong et al. [7] if there is a 1% increase in PM2.5 concentration, then there will be a 0.05818% decrease in the annualized growth rate of GDP per capita.

The literature survey is explained clearly and very elaborately in the next subsection comprising several methods that were followed for the process of predicting and forecasting various pollutants along with their *AQI*.

The examination of air pollution is widely regarded as a significant piece of research for tracking the levels of pollution in a variety of places. To accurately anticipate, forecast, and manage the amount of pollution various machine learning techniques are increasingly being applied. Models for forecasting the levels of ozone, nitrogen dioxide, and sulfur dioxide at the ground level using three distinct machine-learning techniques are addressed [8]. The support vector machine (*SVM*), M5P model tree, and the artificial neural network (ANN) were implemented to predict air pollution which is caused by various pollutants. Findings demonstrate that peak performance in forecasting is achieved

during diverse characteristics employed in multivariate modeling using the M5P method rather than any of the other models [9]. Utilizing the real-time sensor data from Toronto's downtown to create forecasts using land use regression (LUR), ANN, and gradient boost were some of the machine learning techniques. At predetermined GPS locations, a MicroAethalometre (MicroAeth model AE51) was used to measure PM2.5, and the black carbon (BC). It has been shown that LUR performs best on small datasets, whereas ANN and the gradient boost technique improve accuracy as dataset size grows. However, the research conducted by Chen et al. [10] employs an approach in three dimensions of variation to use satellite and ground-based sensors together to compile PM2.5 data information collected in various locations like Wuhan, Xiang Yang, and Yichang (all located in China). The assimilation of data and application of the weather research and forecasting model with chemistry (WRF-Chem) are said to have improved the model's ability to forecast PM2.5 levels [11].

Ke et al. [12] developed an air quality forecasting approach that is automated based on various machine learning techniques and used to predict levels of pollution caused by six prevalent contaminants: PM2.5, PM10, SO_2, NO_2, O_3, and CO. Using the given technique, optimal model hyperparameter values may be determined automatically without human involvement. A knowledge base comprising observed meteorological information, various concentrations of pollutants, pollutant emissions, and a reanalysis data model is utilized to facilitate the incorporation of various artificial intelligence algorithm models and an ensemble model (stacked generalization) inside the system. CO, benzene, PM2.5, NO_2, O_3, and SO_2 average values were used by Wood [13] to develop the combined air quality benchmark (CLAB) for Dallas County (Texas, USA). Predictions using CLAB from a variety of ML and DL algorithms were used in the span of the 2015–2020 period. With an RMSE value of ca. 0.0450 for the 2019 data and ca. 0.0493 for the 2020 data, the support vector regression (*SVR*) model was shown to be the most accurate predictor of CLAB and ADA, respectively. Various approaches for multi-target regression have been suggested for the simultaneous prediction of several different concentrations of air pollutants. It uses a mix of a multi-target regression approach and a concept of random forest to arrive at its conclusions. While the random forest method excels at the classic single-target prediction job, multi-target regression makes use of the connection between target variables. When applied to the dataset from UCI, the innovative framework was recommended as the ensemble of regressor chains-guided feature ranking (ERCFR) managed for attaining the value 0.872 which is the R^2 score. Researchers use various techniques of deep learning to make predictions for a wider range of contaminants. Predictions of the *AQI* have been made using the long short-term memory (LSTM) [15–17].

The LSTM showed both rapid convergence and minimized cycles for training while maintaining a high level of accuracy [18]. According to the time series provided by the data from AirNet, Athiraet al. [19] deployed the recurrent neural networks (RNN), long short-term memory networks (LSTM), and gated recurrent unit networks (GRU) for PM10

predictions. As a natural progression from the LSTM model, the aggregated long short-term memory (ALSTM) model was suggested [20]. The authors of this model proposed integrated monitoring places for various sources of pollutants in a region, neighboring districts where there are various industries as well as local monitoring stations for air quality. Therefore to improve the accuracy of predictions, a predictive model was created by combining 3 LSTM models into a single one. The model consists of 3 input layers, and three subnetwork layers, with two dropout layers added between the LSTM layers to prevent overfitting. The dimension is 17, and the shape of the LSTM is a 3D array corresponding to the parameters: local, near, and chimney. This setup was utilized for prior forecasting based on data collected from sensors used for monitoring the quality of air located near industrial facilities as well as other sources of pollution. A statistical technique and a method based on deep learning were evaluated and compared by Nath et al. [21] for their ability to predict PM2.5 and PM10 contents. Based on the limited data given, the Holt –Winters statistical model outperformed deep learning approaches in terms of assessment metrics like RMSE and MSE. The accuracy of PM2.5 concentration forecasts was compared across many models based on deep learning, LSTM, bi-directional LSTM, GRU, bi-directional GRU, CNN, and a hybrid CNN-LSTM model. When compared to traditional models available on the UCI machine learning repository [22], the predictive performance of the hybrid CNN-LSTM multivariate approach is found to be superior, and it enables more precise predictions. Air pollution forecasting may also make use of the autoregressive integrated moving average (ARIMA) model [23]. It is a model for analyzing data based on a series of times to make predictions.

In 2021, Mani and Volety [24] predicted the pollution levels using both LSTM and ARIMA models. Using data from the Central Pollution Control Board (CPCB) and also a simple hardware configuration that is based on IoT, LSTM proved to be superior to ARIMA in every situation. It produced lower values of root mean squared error (RMSE) and mean absolute error (MAE). The analysis and projections of Beijing's *AQI* from January 2019 to November 2021 are provided by Liu and You [25]. It averages hourly measurements of the most common contaminants that are present in the air like PM2.5, PM10, and ozone to provide everyday measurements to do various examinations and also the forecasting. Statistics of the quality of air in Beijing city were predicted by using ARIMA and the LSTM efficiently.

Based on the literature survey, the study was conducted and data were collected to predict *AQI* effectively and accurately. A clear description of the dataset that was collected with the name of each pollutant is explained in the next section.

2. DESCRIPTION OF THE DATASET

The study was conducted inside the kitchen of Barbeque Nation Hotel (where there are so many pollutants due to the continuous cooking which may affect the persons who

are inside that area for a very long period from morning to night. The major pollutants that are observed are PM2.5 and PM10. Various other factors and pollutants like temperature, humidity, pressure, and volatile organic compounds contents present in the air were also measured.

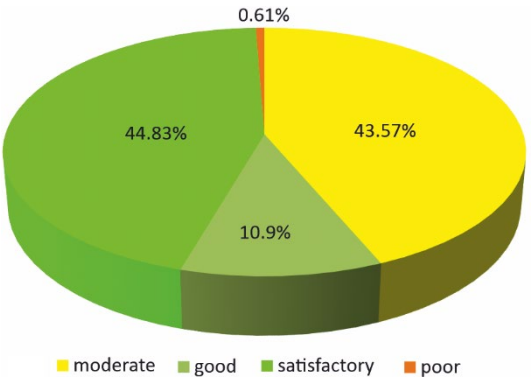

Fig. 1. Air quality index percentage from the data obtained

Figure 1 shows a pie chart for the *AQI* categories from the dataset taken. So, from the figure, it can be understood that 43.57% of the *AQI* was found to be moderate and 44.83% of the *AQI* was found to be satisfactory. Only 10.9% of the *AQI* was found to be good which does not cause any health issues when inhaled. 0.61% of the *AQI* was poor which causes various health concerns when inhaled over time.

Table 2 provides the method for calculating the PM10 index (IPM10) at varying PM10 concentrations [26]. Figure 2 shows a seaborn graph of various pollutants from the data taken and the Indoor Air Quality (IAQ).

Table 2

Equations for calculating the I_{PM10} index

Range of C_{PM10} [μg/m^3]	Equation for the I_{PM10} calculation
$C_{PM10} \leq 100$	$I_{PM10} = C_{PM10}$
$100 < C_{PM10} \leq 250$	$I_{PM10} = 100 + (C_{PM10} - 100) \times (100/150)$
$250 < C_{PM10} \leq 350$	$I_{PM10} = 200 + (C_{PM10} - 250)$
$350 < C_{PM10} \leq 430$	$I_{PM10} = 300 + (C_{PM10} - 350) \times (100/80)$
$C_{PM10} > 430$	$I_{PM10} = 400 + (C_{PM10} - 430 \times (100/80)$

The details of various instruments that were used to record the pollutant values during the study are elaborated clearly in the next subsection.

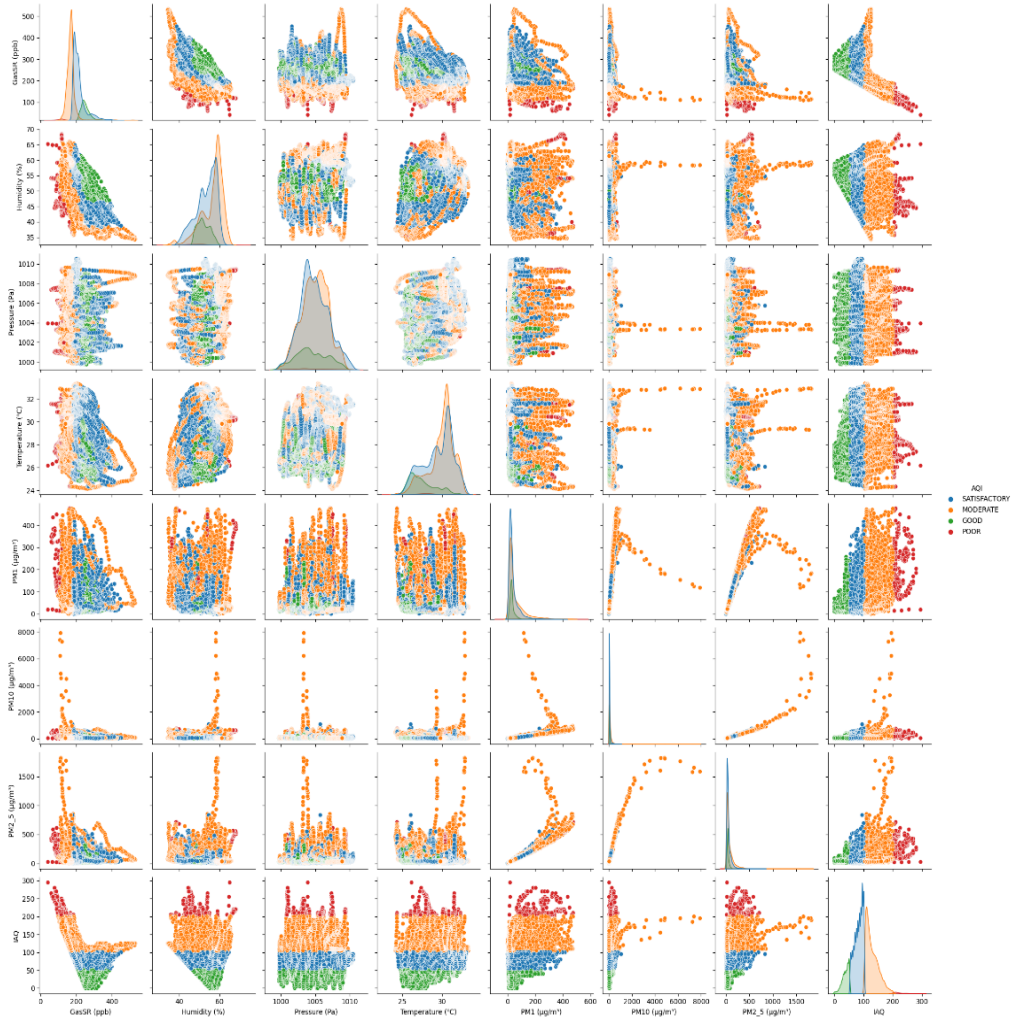

Fig. 2. Seaborn graph of various pollutants

2.1. DETAILS OF THE INSTRUMENT USED FOR MEASURING AIR POLLUTANTS:

The indoor air quality monitoring device which consists of various sensors such as MQ2, MQ3, MQ5, MQ135, SPS30, and DHT22 was used to measure various pollutant concentrations that are present in the kitchen during continuous cooking processes. Table 3 shows pollutants that can be measured by each sensor. After the data are collected, the data pre-processing should be done before training it with suitable machine learning models. The main purpose of doing this is to transform the data into a format that is suitable for building machine learning models to train them.

Table 3

Sensors and their parameters

Pollutant	Sensor
NH_3, NO*x*, alcohol, benzene, smoke, and CO_2	MQ-135
Temperature and humidity	DHT 22/11
Alcohol, gases	MQ-3
LPG, CH_4, CO, H_2	MQ-5
Combustible gas and smoke	MQ-2
PM2.5, PM10	SPS30

3. DATA PROCESSING

As a part of the data transformation process, pre-processing helps to transform the various data to a very efficient input format which would be provided for the model. Various pre-processing techniques that are used are elaborated in detail.

Selection of the features. To prepare meals, people turn to a wide range of heat sources such as gas, wood, and electricity. When used in the kitchen, all of these heaters may contribute to unhealthy levels of stale air. Toxic gases and chemicals including carbon monoxide and formaldehyde are sometimes released into the air by natural gas burners. Various pollutant indices like the particulate matters of size 2.5, 1, and 10 μm, nitrogen dioxide, sulfur dioxide, carbon monoxide, temperature, pressure, humidity, and volatile organic compounds which are the common pollutants present in the kitchen are taken into account. The dataset which is obtained from the kitchen area includes a significant amount of missing data. So these missing data are corrected by several methods. Various pollutant indices like the particulate matter of size 2.5 and 10 μm, nitrogen dioxide, ammonia, sulfur dioxide, carbon monoxide, and lead are utilized in India's air quality index computation [27].

Outlier processing. One of the most useful tools for finding outliers is the *Z-score*. *Z-scores* are used to quickly and easily quantify the dispersion of data points around the mean [28]. To assess whether or not a given data point should be labeled as an outlier, a cutoff value for the *Z-score* may be determined. The outliers have the potential to cause modeling errors which might then have an impact on the results of the model. The *Z-score* is a statistical measure that indicates how much the actual value deviates from the average value.

Handling the missing values. The dataset obtained from the kitchen area has several gaps due to missing data. Inaccurate results may be produced if all rows with invalid or missing data are deleted [29]. Multiple imputation and maximum likelihood are two of the most common approaches used when trying to fill in data gaps with predicted values. Following the data processing, the collected data are used to train appropriate machine learning models and predict output as explained in the following section.

4. MACHINE LEARNING MODELS

It is difficult to make reliable predictions using an environment quality model while attempting to simulate air quality. For the various pollutants that would be coming out from the kitchen, the modeling would be difficult due to the ever-changing nature of the environmental data. In this research, four separate models namely *K*-Nearest Neighbor (KNN), Support Vector Machine (*SVM*), Decision Tree (DT), and Logistic Regression (LR) are employed to build and perfect an *AQI* prediction model. Figure 3 explains the general steps to be followed in *AQI* prediction using the machine learning algorithm. The algorithmsused in this study are explained in the following subsections.

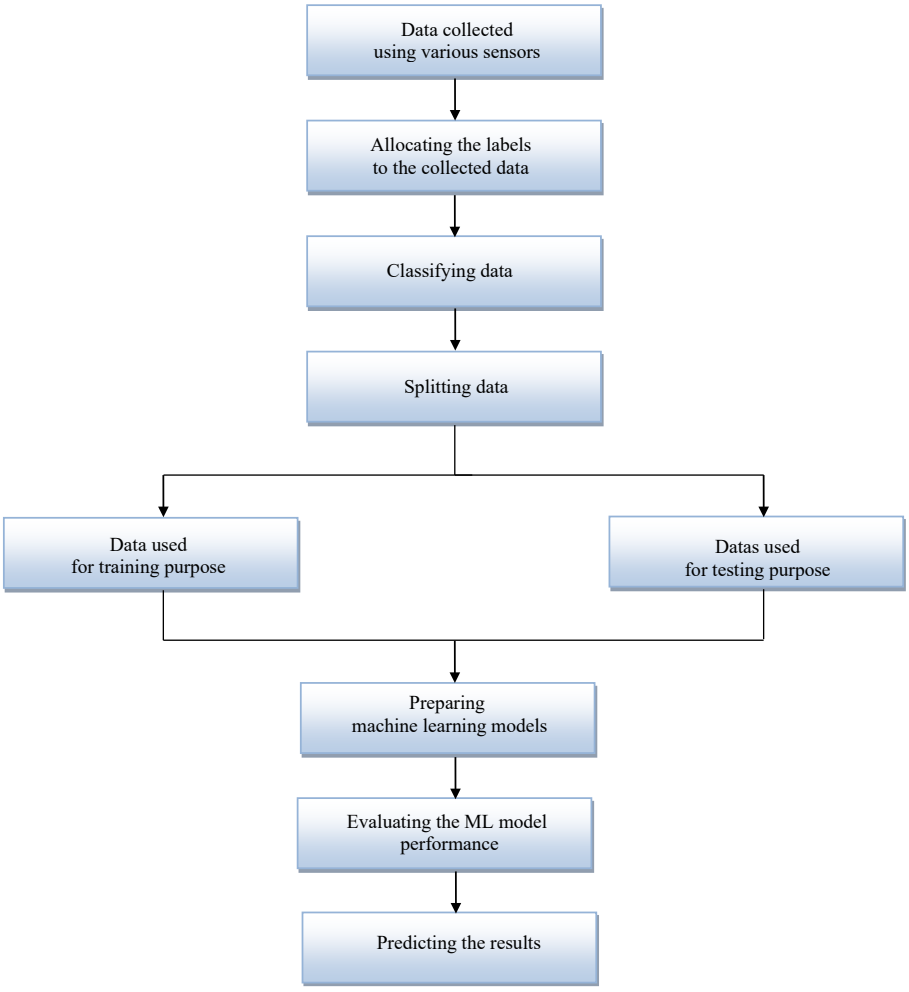

Fig. 3. Steps followed in *AQI* prediction using machine learning algorithms

4.1. SUPPORT VECTOR MACHINE

The *SVM* algorithm is very popular among various machine learning algorithms. Figure 4 shows the flowchart for the *SVM* algorithm. Since its inception, *SVM* has experienced extraordinary advancements [30].

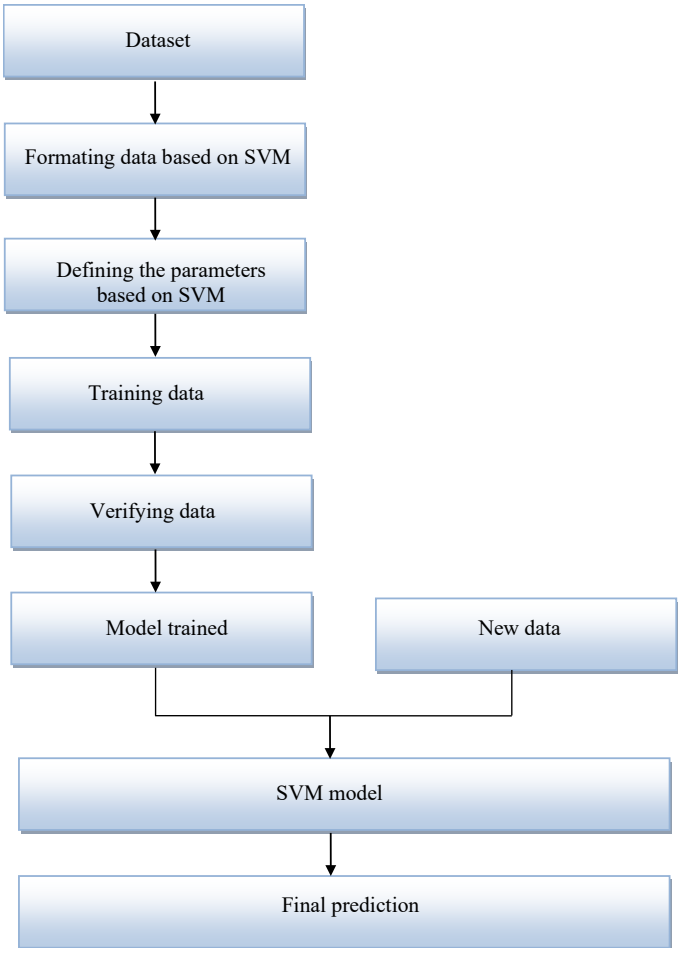

Fig. 4. *SVM* algorithm flowchart

It is a kind of supervised learning model and a classification or regression-based learning technique. Support vector regression (*SVR*) is another name for the *SVM* regression techniques. In the support vector regression, the hyperplane is the term used to describe a straight line that must be used for the data to be fitted. The radial basis function (*RBF*), the Gaussian kernel, the polynomial kernel, etc., are various examples of kernel functions that may be implemented to translate lower-dimensional data into higher-

dimensional space. The *SVM* has shown state-of-the-art outcomes in a variety of difficult situations that cannot be separated linearly [31–34]. Classification and regression issues may both be solved with the help of the algorithms available in the *SVM* and it is very effective for solving a wide variety of issues and can handle problems that are linear as well as non-linear. It is initially assumed that *SVM* locates a hyperplane connecting data from two classes. The *SVM* algorithm seeks to identify the points belonging to both classes that are situated in the vicinity of the line and therefore, these points are widely known as the support vectors. Evaluation is done to determine the length between the support vectors and the line. So, the separation that is present is known as the margin. The main aim is thus to have a margin that is as high as possible. The ideal hyperplane is one that consistently maintains a wide margin and strives to maximize this margin. Therefore, a decision boundary is generated by the *SVM* in such a manner that the gap that exists between the two classes is as large as it can be practically achievable. The goal is also to therefore choose the hyperplane which is having the highest feasible margin that can be achieved between the support vectors in the supplied dataset and to partition the given dataset in the most effective manner that is available.

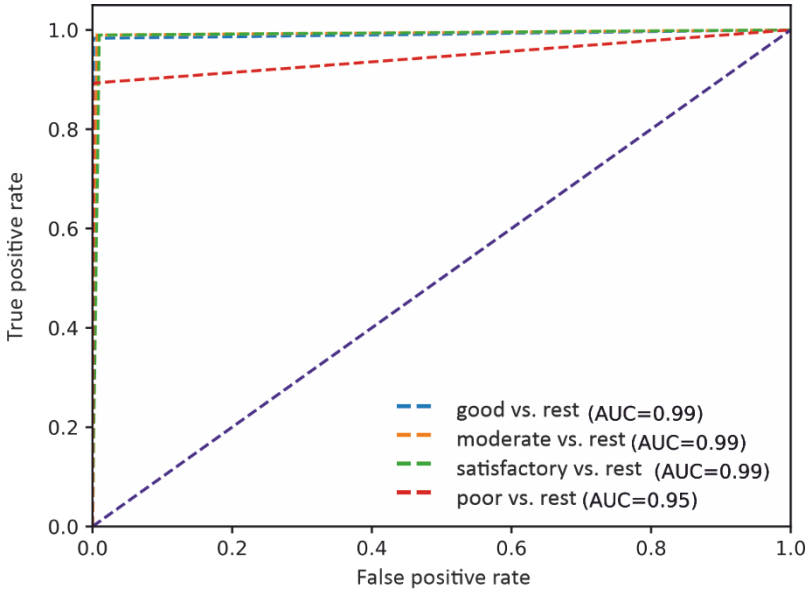

Fig. 5. Receiver operating characteristic curve for *SVM*

The *SVM* goal is to optimize the separation of hyperplanes to get the highest possible margins across all classes. The support vectors make use of a hyperplane which is an object for the data that is included inside the provided dataset. Also, the margin is defined as the furthest point from the support vector to the hyperplane [35]. Considering a dataset with the elements $(A_1, B_1, \ldots, A_n, B_n)$ if $(A_1, \ldots, A_n)$ denotes the set of variables

to be input, (B_1, ..., B_n) denotes the variables to be output and C denotes the point of intercept [35], then the equation is given as

$$SVM = \sum_{i=1}^{n} \beta_i - \frac{1}{2} \sum_{i,j=1}^{n} b_i b_j C\left(a_i, a_j\right) \beta_i \beta_j \tag{1}$$

where $i = 1, 2, 3, \ldots, n$, $C = b_i\beta_i + b_j\beta_j$.

The true positive rate and false positive rate of various categories for the *AQI* are compared using the graph for the *SVM* algorithm. Figure 5 shows the receiver operating characteristic curve for *SVM*. The steps that follow are the procedures for the *SVM* to search for the hyperplane with the greatest marginal distance:

- It produces a hyperplane that most effectively differentiates the various types.
- If there are three hyperplanes, the two hyperplanes would have the higher stratification error, and the other hyperplane separates the two classes accurately.
- Then it selects the hyperplane that is furthest from the two closest data points.

4.2. *K*-NEAREST NEIGHBOR

The KNN algorithm's relevance has grown with the processing capacity of various computers. With the acceleration of processors, computers have become more important and are now viable. The KNN method is widely used because it is simple to implement and comprehend. KNN relies heavily on distance calculations. So, it is essential to use the KNN algorithm while working with datasets with numerical values. An example of a non-parametric algorithm is the KNN algorithm. Depending upon the issue or dataset that is provided, the learning and the forecasting analysis will be performed. Unlike other classification methods, the KNN makes no assumptions about the dataset before making its prediction. The number of closest neighbor data values is denoted by the letter *K*. The KNN method makes its determination on the classification of the dataset which is provided depending on the value of *K* which implies the value of the dataset's closest neighbors. The KNN model can perform direct classification on the dataset which is used for training purposes. This also implies that predicting the new instance is done based on looking for comparable *K* neighbor examples across the entire complete training dataset and by categorizing it depending on the class with the largest number of instances. The Euclidean distance equation is used to identify similar cases [36]. In its most basic form, the KNN algorithm is composed of four stages. In the first stage, the distance between the newly added data and the previously collected data is computed. During the stage after, distances would be arranged in descending order. The class is decided in the last stage, which involves selecting the k values with the least absolute value. To calculate the Euclidean distance $E_{i,j}$ between the new instance x_i and the old instance y_j, the square root of the sum of the squared differences is taken [36].

$$E_{i,j} = \left(\sum_{k=1}^{n} \left(x_{ik} - y_{jk} \right)^2 \right)^{1/2} \tag{2}$$

The true positive rate and false positive rate of various categories depending upon the *AQI* are compared in Fig. 6 showing the receiver operating characteristic curve for the KNN algorithm.

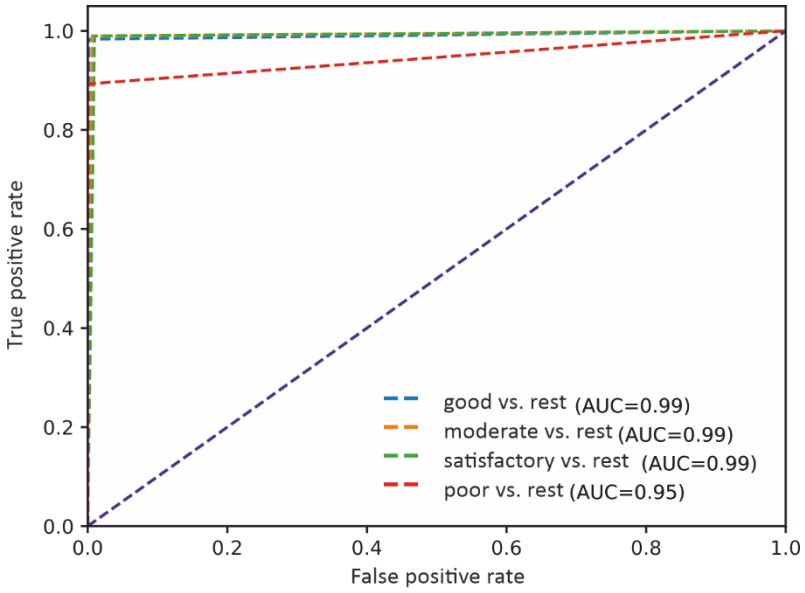

Fig. 6. Receiver operating characteristic curve for KNN

4.3. DECISION TREE ALGORITHM

The decision trees (DT) algorithm is very efficient and has many applications including those in the disciplines of machine learning, image processing, and pattern recognition. It can be used to provide assumptions based on the class names of various categories, to categorize various information which is based on training datasets and the different class labels, and to also categorize the new data that are obtained. This algorithm is one of the strong models used for the prediction of both problems like classification and regression. Decision trees are iterative models that cohesively and effectively combine several instances of a fundamental test for comparing the numerical characteristic to the threshold value [37]. It is considerably simpler to build the conceptual rules compared to the numerical weights that are present in the neural network's inter-node connections [38]. The components that make up each tree are the nodes and the branches. Each node stands for a collection of characteristics in a target classification,

and its value is defined by a set of subsets [39]. Binary trees are used to depict the DT models. It indicates that the given issue or dataset may be addressed by partitioning it into binary trees or categorizing the data in that format. Predictions are formed in the decision tree by selecting the terminal node which is present in the binary tree, given a single input variable which is denoted as x. Datasets are partitioned on the variable. The nodes at the dataset's leaves, the resulting binary tree representing the output variable are denoted by y. The tree is followed from its root node out via its branches and their sub-branches, the conclusion is drawn from the information on the tree's leaves. The dataset can be thus divided into two different parts by using a greedy algorithm [40]. The Gini index function is used by the DT classification technique to ascertain the level of impureness present in the leaf nodes of the tree to make predictions. For making accurate predictions, the classification technique in the decision tree utilizes the Gini index (G) function to rank the purity of the tree's leaf nodes:

$$G = \sum_{k=1}^{n} xk\left(1 - xk\right) \tag{3}$$

where x is the fraction of examples used for training that belongs to the k input class. The prediction is as easy as clicking a button when using a binary tree to describe the data [40]. Figures 7 and 8 show the ROC curve of the DT classifier. The true and false-positive rates of various categories for the *AQI* are compared using the graph for the DT algorithm. The area under the curve for each category is listed in Figs. 7 and 8.

4.4. LOGISTIC REGRESSION ALGORITHM

The logistic regression technique is notable for being both straightforward and effective as a predictive tool. To develop an accurate prediction model of the issue, the LR applies the sigmoid function. It creates a model of the dataset and maps each of the variables to a value that ranges from 0 to 1. The predictive analysis is carried out using the LR method, and it is predicated on the connection between the binary dependent variable and the other one or more independent variables taken from the provided dataset. To make an accurate prediction of Y, the input values (X_1, X_2, …, X_n) are linearly mixed with the coefficient values [41]. Using X_1 and X_2 as predictors of the final outcome Y, the independent variables X' and X'' the logistic regression equation can be given as

$$Y = \frac{1}{2}\left(\frac{\mathrm{e}^{mX_1+c}}{1+\mathrm{e}^{mX_1+c}} + \frac{\mathrm{e}^{mX_2+c}}{1+\mathrm{e}^{mX_2+c}}\right) \tag{4}$$

where c is the intercept and m denotes the coefficient between X_1 and X_2 values. Each input value (X_1, X_2) may be used to train the coefficient m using the data from the training set.

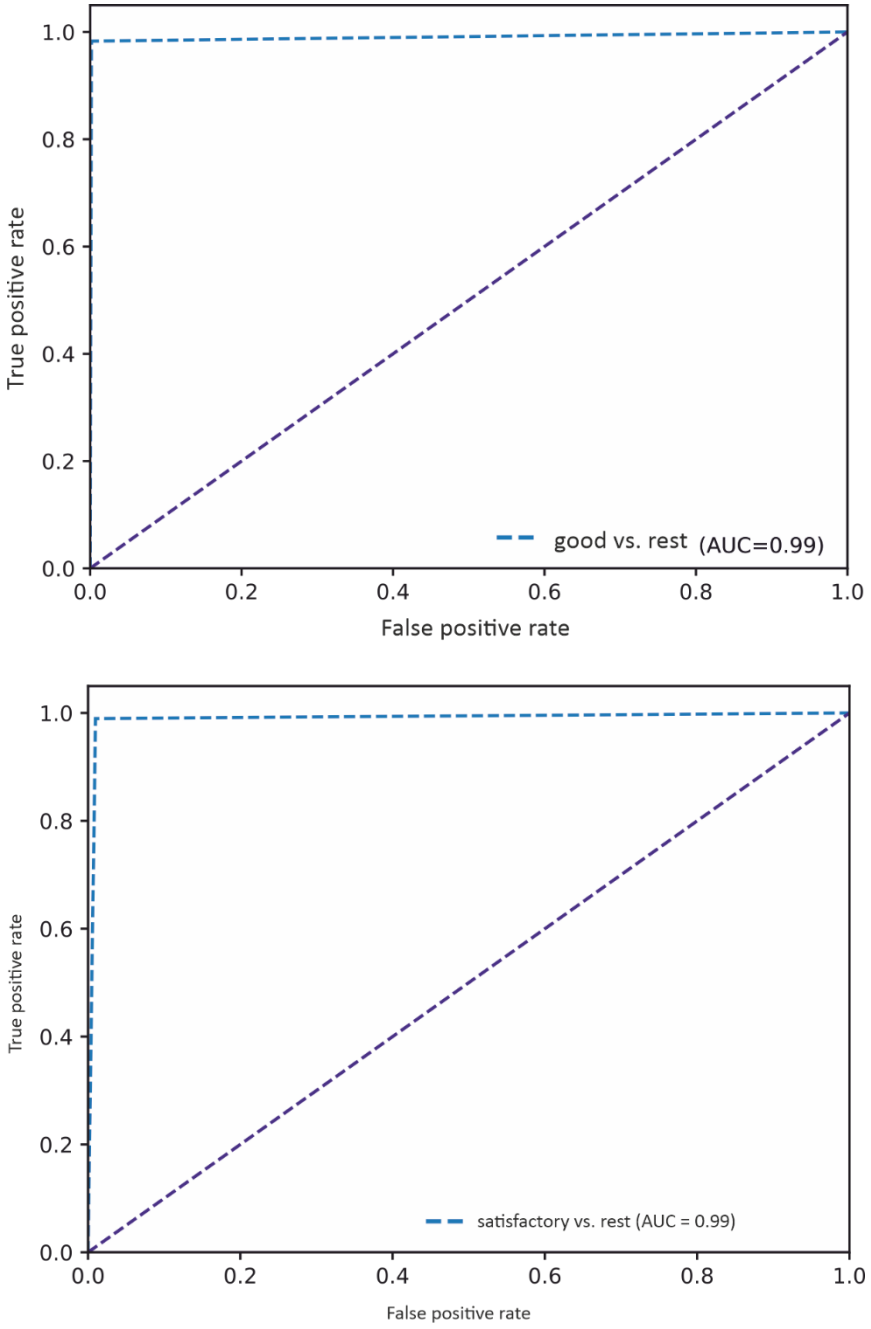

Fig. 7. Receiver operating characteristic curves of the decision tree classifier: good vs. rest (upper), satisfactory vs. res (lower)

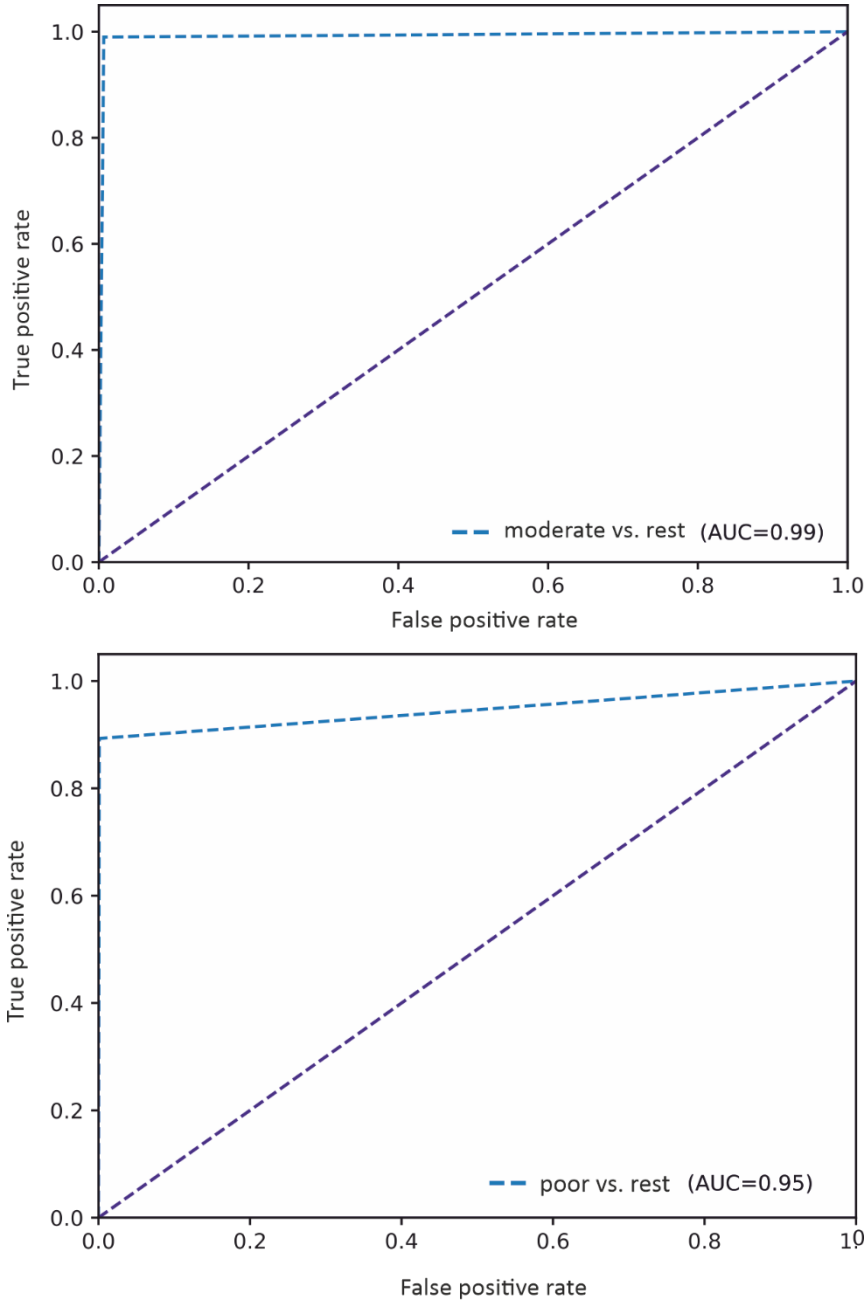

Fig. 8. Receiver operating characteristic curves of the decision tree classifier: moderate vs. rest (upper), poor vs. rest (lower)

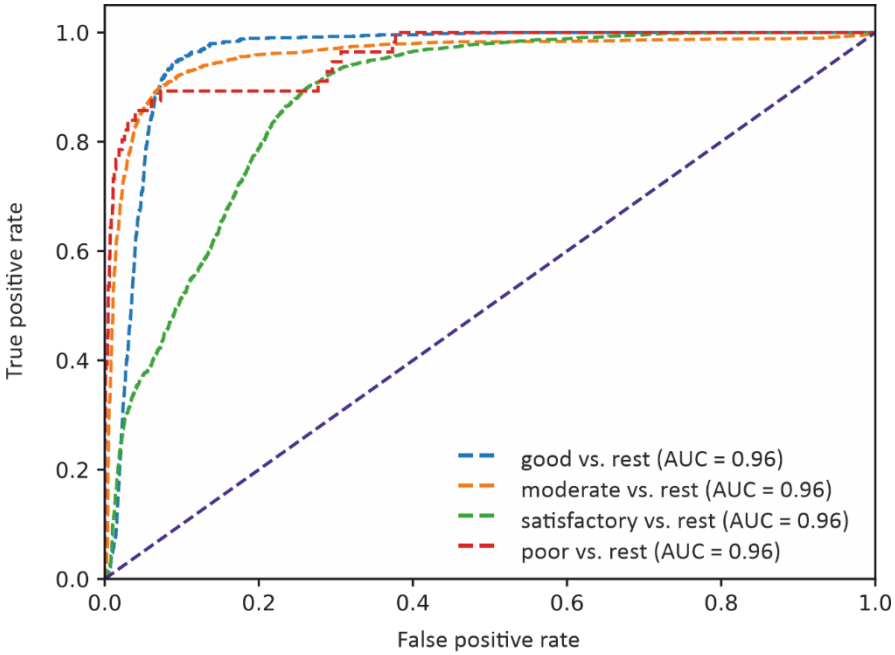

Fig. 9. Receiver operating characteristic curve of the logistic regression

Figure 9 shows the ROC curve of LR. The true positive rate and the false positive rate of various categories for the *AQI* are compared using the graph for the LR model.

Certain metrics are utilized to evaluate outcomes after the models have been trained by using the collected data. Those parameters are explained clearly in the next section.

5. EVALUATING PARAMETERS

Mean absolute error (*MAE*) refers to the measurement errors occurring between observations that are paired [26]. The error is calculated by subtracting the anticipated value $\hat{Y}_i$ from the actual value Y_i, where n is the total number of observations

$$MAE = \frac{\sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|}{n} \tag{5}$$

Root-mean-square error (*RMSE*), called also the residual standard deviation, refers to fitting the regression line to the data points [26, 42]. In a regression with variables observed across T periods, the *RMSE*

$$RMSE = \left(\frac{\sum_{t=1}^{T} \left(\widehat{Y}_t - Y_t \right)^2}{T} \right)^{1/2} \tag{6}$$

of predicted values at time t of the dependent variable Y_t is calculated.

R^2 *score* is a statistical measure of the degree to which an independent variable predicts another variable of interest. The R^2 score, which measures the coefficient of determination, is more precise and illuminating than the ambiguous t-tests and z-scores. The Coefficient of determination (sometimes abbreviated as R^2 or r^2) is a statistical measure of how well one independent variable can explain the observed value of another dependent variable. The equation for R^2 is given as

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{m} \left(Y_i - \widehat{Y}_i \right)^2}{\sum_{i=1}^{m} \left(Y_i - \overline{Y} \right)^2} \tag{7}$$

where SSR is the sum squared regression, and SST is the total sum of squares. In this equation, m represents the total number of observatories, $\widehat{Y}_i$ the projected ith value, Y_i the actual ith value, and $\overline{Y}$ the mean of actual values [26, 43].

Accuracy. The prediction algorithm's accuracy (A) is measured by comparing the number of accurate classification predictions to the actual classification of the dataset. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the four different ways of outputs of any given forecasted model. The model's precision may be determined using the equation [44, 36]

$$A = \frac{TP + TN}{TP + TN + FN + FP} \tag{8}$$

Cohen's kappa score measures the reliability of a forecasted model. The model's projected outcome is correlated to the actual outcome. As a statistical value, it may range from 0 to 1. Consistency is likely to be high for values close to 1 [44, 45].

$$K = \frac{\left(TP + \frac{TN}{N} \right) - \frac{(TP + FN)(TP + FP)(TN + FN)}{N^2}}{1 - \frac{(TP + FN)(TP + FP)(TN + FN)}{N^2}} \tag{9}$$

Confusion matrix. The effectiveness of a prediction model may be seen in its entirety by inspecting its confusion matrix. Prediction results are shown as a matrix including data on the proportion of accurate, wrong, right predictions and as well as associated error rates [44, 45].

Receiver operating characteristic – area under curve (ROC-AUC). The graphical representation of the effectiveness of the predictive model is provided by the ROC-AUC curve. The ROC curve depicts the connection that exists between recall and accuracy for a given set of threshold values. The point at which the threshold is reached is when the model produces positive predictions. The ROC-AUC curve is produced by maintaining the false positive rate along the *x*-axis and the genuine positive rate along the *y*-axis. Its value might be anything between 0 and 1 [44, 45].

After the models are trained by the above-mentioned algorithms and evaluated, the results obtained from the study are explained in the following section.

6. RESULTS AND DISCUSSION

This section comprises the major findings from data obtained inside the kitchen. It clearly explains the maximum, minimum, and average values of temperature, humidity, volatile organic compounds, and particulate matter (PM1, PM2.5, PM10). The percentage of the *AQI* that was good, moderate, satisfactory, and poor from the entire data obtained is also clearly explained. Accuracy rates that are obtained as a result of training the data are explained for each algorithm. The algorithm that predicted the *AQI* most accurately and efficiently is explained. The evaluation procedure of the efficient algorithm is clarified followed by the future scope of the research that can be carried out.

- The minimum temperature recorded was 25 °C and the maximum was around 35 °C. The average temperature of the kitchen area during cooking was found to be around 30 °C.
- The humidity in the kitchen was found to be a minimum of 33% and a maximum of 71%. The average humidity calculated was 55%.
- The minimum concentrations of volatile organic compounds were 65 ppb and the maximum 475 ppb. The average VOC concentration observed was 100 ppb.
- The concentration of PM1 ranged from 0 to 453 μg/m³. The average concentration of PM1 was 26 μg/m³.
- The concentration of PM2.5 ranged from 19 μg/m³ to 568 μg/m³. The average concentration of PM2.5 was 43 μg/m³.
- The concentration of PM10 ranged from 23 μg/m³ to 558 μg/m³. The average concentration of PM10 was 65 μg/m³.

Therefore, from the entire data collected during the study, nearly 43.57% of the *AQI* was found to be moderate, and 44.83% – satisfactory. Only 10.9% was found to be good

which does not cause any health issues when inhaled, and 0.61% *AQI* was poor which causes various health concerns when inhaled over time.

A very effective and efficient model for the prediction of the air quality index is presented from the data collected in the cooking area of the hotel making use of an indoor air quality monitoring system. Many pre-processing techniques including the elimination of outliers, standardization of data, selection of features, and management of missing values are used to provide a more accurate representation of the data. Within the scope of this study, a total of four different machine learning models which are namely *SVM*, KNN, LR, and DT were used to build an accurate prediction model.

Out of these, the DT algorithm performed well by predicting the *AQI* efficiently and categorized it more effectively than the other algorithms used. When the pollutant values were given, the algorithm efficiently predicted the *AQI* value and categorized it as "Moderate". This implies that the air quality inside the area is acceptable but some pollutants present can be harmful to a group of individuals who are sensitive to air pollution.

Table 4

Predicted *AQI* with the decision tree algorithm

VOC [ppb]	Humidity [%]	Pressure [Pa]	Temperature [°C]	PM1 [µg/m³]	PM10 [µg/m³]	PM2.5 [µg/m³]	*AQI*	Category
233.9	51.38	1005.44	25.39	48	73	58	50	good
199.49	58.91	1005.68	27.81	21	33	30	65	satisfactory
268.57	51.72	1005.78	25.01	18	26	25	20	good
117.85	42.54	1003.28	28.05	287	380	363	250	poor
130	64.31	1009.46	30.39	22	46	32	102	moderate
107.95	49.59	1002.64	28.25	21	40	32	90	satisfactory
98.81	51.84	1003.4	28.03	130	175	165	215	poor
174.41	64.44	1007.96	30.5	21	34	30	130	moderate

This model was tested by using the additional raw data obtained from the kitchen area apart from the data used for testing and training purposes. When various pollutant values were given, the model predicted the *AQI* value and its category accurately. Table 4 shows the predicted *AQI* and its categorization using the raw data obtained additionally to validate the model.

Table 5

Accuracy rate of classifiers

No.	Classifier	Accuracy [%]
1	support vector machine (*SVM*)	80.34
2	*K*-nearest neighbor (KNN)	93.01
3	decision tree (DT)	98.79
4	logistic regression (LR)	80.20

The accuracy rates of various classifiers used are given in Table 5 and they are compared with each other in Fig. 10. The DT model performed well with an accuracy rate of 98.7% followed by the KNN algorithm with an accuracy rate of 93.01%. The *SVM* and the LR Algorithms performed with an accuracy rate of 80.3% and 80.2%, respectively.

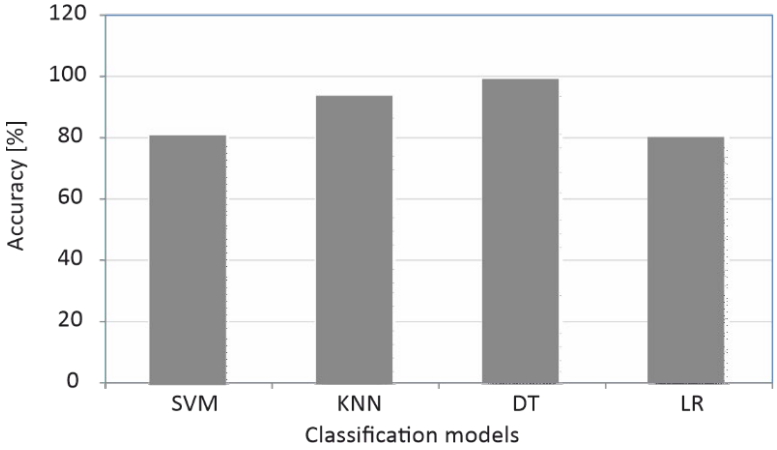

Fig. 10. Accuracy rates for different classifiers

Therefore, by predicting the *AQI* efficiently in advance, the concerned authorities can take some necessary actions to regulate air pollution. Some innovative approaches can also be implemented to improve the *AQI* inside the cooking area thereby creating a safer environment for the chefs and other people who are working there for a very long time in a day. Future work can be done to find some new methods for segregating the living from the non-living particulate matters and nullifying them so that various airborne diseases like COVID-19 that are caused by those living particulate matters can be prevented in the future.

7. CONCLUSION

Common sources of air pollution include the combustion of fossil fuels like natural gas, coal, wood, various industrial processes, and transportation. Since the kitchen of the star hotels which is indulging in a continuous cooking process was also observed as one of the areas where these harmful air pollutants are emitted on a large scale, this research helps in identifying a suitable model for predicting the *AQI*. These harmful air pollutants that are emitted in the kitchen affect the humans working there full time when they inhale it continuously by causing various lung diseases, etc. The quality of air can be measured through the *AQI*. Due to the advancement in technology, several machine

learning algorithms help in determining the *AQI* efficiently. Various star hotels are taking many initiatives to minimize the emission of these harmful pollutants while cooking. This concept of predicting the quality of air is a valuable investment in various and different areas that can be used widely at global and national levels. Therefore, by predicting the quality of the air accurately there would be a significant decrease in the number of individuals who fall sick due to the effect of various harmful air pollutants which are present in the air. On the other hand, it also helps in creating an environment that is clean and healthy to live.

REFERENCES

[1] ESPINOSA R., JIMÉNEZ F., PALMA J., *Multi-objective evolutionary spatio-temporal forecasting of air pollution*, Fut. Gen. Comp. Syst., 2022, 136, 15–33. DOI: 10.1016/j.future.2022.05.020.

[2] MALTARE N.N., VAHORA S., *Air quality index prediction using machine learning for Ahmedabad City*, Dig. Chem. Eng., 2023, 7, 100093.

[3] LEONG W.C., KELANI R.O., AHMAD Z., *Prediction of air pollution index (API) using support vector machine (SVM)*, J. Environ. Chem. Eng., 2020, 8 (3), 103208. DOI: 10.1016/j.jece.2019.103208.

[4] EJAZ A., MALIK A.O., LOPEZ-CANDALES A., *Ambient air pollution and cardiovascular disease: learned from the COVID-19 pandemic*, Postgrad. Med., 2021, 133 (6), 587, 588. DOI: 10.1080/00325481.2021.1921504.

[5] LI Y., DU J., LIN S., HE H., JIA R., LIU W., *Air pollution increased risk of reproductive system diseases: a 5-year outcome analysis of different pollutants in different seasons, ages, and genders*, Environ. Sci. Poll. Res., 2022, 29 (5), 7312–7321. DOI: 10.1007/s11356-021-16238-7.

[6] HE G., PAN Y., TANAKA T., *The short-term impacts of COVID-19 lockdown on urban air pollution in China*, Nat. Sust., 2020, 3 (12), 1005–1011. DOI: 10.1038/s41893-020-0581-y.

[7] DONG D., XU B., SHEN N., HE Q., *The adverse impact of air pollution on China's economic growth*, Sust., 2021, 13 (16), 9056. DOI: 10.3390/su13169056.

[8] SHABAN K.B., KADRI A., REZK E., *Urban air pollution monitoring system with forecasting models*, IEEE Sens. J., 2016, 16 (8), 2598–2606. DOI: 10.1109/JSEN.2016.2514378.

[9] WANG A., XU J., TU R., SALEH M., HATZOPOULOU M., *Potential of machine learning for prediction of traffic-related air pollution*, Trans. Res. D Trans. Environ., 2020, 88, 102599. DOI: 10.1016/j.trd.2020.102599.

[10] CHEN L., MAO F., HONG J., ZANG L., CHEN J., ZHANG Y., GAN Y., GONG W., XU H., *Improving PM2.5 predictions during COVID-19 lockdown by assimilating multi-source observations and adjusting emissions*, Environ. Poll., 2022, 297, 118783. DOI: 10.1016/j.envpol.2021.118783.

[11] GRELL G.A., PECKHAM S.E., SCHMITZ R., MCKEEN S.A., FROST G., SKAMAROCK W.C., EDER B., *Fully coupled online chemistry within the WRF model*, Atm. Environ., 2005, 39 (37), 6957–6975. DOI: 10.1016/j.atmosenv.2005.04.027.

[12] KE H., GONG S., HE J., ZHANG L., CUI B., WANG Y., MO J., ZHOU Y., ZHANG H., *Development and application of an automated air quality forecasting system based on machine learning*, Sci. Total Environ., 2022, 806, 151204. DOI: 10.1016/j.scitotenv.2021.151204.

[13] WOOD D.A., *Local integrated air quality predictions from meteorology (2015 to 2020) with machine and deep learning assisted by data mining*, Sust. Anal. Model., 2022, 2, 100002. DOI: 10.1016/j.samod.2021.100002.

[14] MASMOUDI S., ELGHAZEL H., TAIEB D., YAZAR O., KALLEL A., *A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection*, Sci. Total Environ., 2020, 715, 136991. DOI: 10.1016/j.scitotenv.2020.136991.

[15] LI X., PENG L., YAO X., CUI S., HU Y., YOU C., CHI T., *Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation*, Environ. Poll., 2017, 231, 997–1004. DOI: 10.1016/j.envpol.2017.08.114.

[16] ZHAO J., DENG F., CAI Y., CHEN J., *Long short-term memory. Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction*, Chemosphere, 2019, 220, 486–492. DOI: 10.1016/j.chemosphere.2018.12.128.

[17] LIU B., GUO X., LAI M., WANG Q., *Air pollutant concentration forecasting using long short-term memory based on wavelet transform and information gain. A case study of Beijing*, Comp. Int. Neurosci., 2020, 834699. DOI: 10.1155/2020/8834699.

[18] NAVARES R., AZNARTE J.L., *Predicting air quality with deep learning LSTM. Towards comprehensive models*, Ecol. Inform., 2020, 55, 101019. DOI: 10.1016/j.ecoinf.2019.101019.

[19] ATHIRA V., GEETHA P., VINAYAKUMAR R., SOMAN K.P., *Deepairnet: Applying recurrent networks for air quality prediction*, Procedia Comp. Sci., 2018, 132, 1394–1403. DOI: 10.1016/j.procs.2018.05.068.

[20] CHANG Y.S., CHIAO H.T., ABIMANNAN S., HUANG Y.P., TSAI Y.T., LIN K.M., *An LSTM-based aggregated model for air pollution forecasting*, Atm. Poll. Res., 2020, 11 (8), 1451–1463. DOI: 10.1016/j.apr.2020.05.015.

[21] NATH P., SAHA P., MIDDYA A.I., ROY S., *Long-term time-series pollution forecast using statistical and deep learning methods*, Neural Comp. Appl., 2021, 1–20. DOI: 10.1007/s00521-021-05901-2.

[22] BEKKAR A., HSSINA B., DOUZI S., DOUZI K., *Air-pollution prediction in smart city, deep learning approach*, J. Big Data., 2021, 8 (1), 1–21. DOI: 10.1186/s40537-021-00548-1.

[23] GOPU P., PANDA R.R., NAGWANI N.K., *Time series analysis using ARIMA model for air pollution prediction in Hyderabad city of India*, Soft Comp. Sign. Proc.: Proc. 3rd ICSCSP 2020, Springer, Singapore 2021, 1, 47–56. DOI: 10.1007/978-981-33-6912-2_5.

[24] MANI G., VOLETY R., *A comparative analysis of LSTM and ARIMA for enhanced real-time air pollutant levels forecasting using sensor fusion with ground station data*, Cogen. Eng., 2021, 8 (1), 1936886. DOI: 10.1080/23311916.2021.1936886.

[25] LIU T., YOU S., *Analysis and forecast of Beijing's air quality index based on ARIMA model and neural network model*, Atmosphere, 2022, 13 (4), 512. DOI: 10.3390/atmos13040512.

[26] MALTARE N.N., VAHORA S., *Air quality index prediction using machine learning for Ahmedabad city*, Digit. Chem. Eng., 2023, 7, 100093. DOI: 10.1016/j.dche.2023.100093.

[27] SHAH D.P., PATEL P., *A comparison between national air quality index, India and composite air quality index for Ahmedabad, India*, Environ. Chall., 2021, 5, 100356. DOI: 10.1016/j.envc.2021.100356.

[28] CURTIS A.E., SMITH T.A., ZIGANSHIN B.A., ELEFTERIADES J.A., *The mystery of the Z-score*, Aorta, 2016, 4 (4), 124–130. DOI: 10.12945/j.aorta.2016.16.014.

[29] EMMANUEL T., MAUPONG T., MPOELENG D., SEMONG T., MPHAGO B., TABONA O., *A survey on missing data in machine learning*, J. Big Data, 2021, 8 (1), 1–37. DOI: 10.1186/s40537-021-00516-9.

[30] CHERVONENKIS A.Y., *Early history of support vector machines, empirical inference. Festschrift in Honor of Vladimir N. Vapnik*, 2013, 13–20. DOI: 10.1007/978-3-642-41136-6_3.

[31] LEONG W.C., KELANI R.O., AHMAD Z., *Prediction of air pollution index (API) using support vector machine (SVM)*, J. Environ. Chem. Eng., 2020, 8 (3), 103208. DOI: 10.1016/j.jece.2019.103208.

[32] KE H., GONG S., HE J., ZHANG L., CUI B., WANG Y., MO J., ZHOU Y., ZHANG H., *Development and application of an automated air quality forecasting system based on machine learning*, Sci. Total Environ., 2022, 806, 151204. DOI: 10.1016/j.scitotenv.2021.151204.

[33] ESPINOSA R., JIMÉNEZ F., PALMA J., *Multi-objective evolutionary spatio-temporal forecasting of air pollution*, Futur. Gen. Comp. Syst., 2022, 136, 15–33. DOI: 10.1016/j.future.2022.05.020.

[34] BALOGUN A.L., TELLA A., *Modeling and investigating the impacts of climatic variables on ozone concentration in Malaysia using correlation analysis with random forest, decision tree regression, linear regression, and support vector regression*, Chemosphere, 2022, 299, 134250. DOI: 10.1016/j.chemosphere.2022.134250.

[35] SCHÖLKOPF B., SMOLA A.J., BACH F., *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, 2022.

[36] ALTAY O., ULAS M., *Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children*, 6th International Symposium on Digital Forensic and Security (ISDFS), IEEE, 2018, 1–4. DOI: 10.1109/ISDFS.2018.8355354.

[37] DAMANIK I.S., WINDARTO A.P., WANTO A., ANDANI S.R., SAPUTRAW., *Decision tree optimization in C4.5 algorithms using genetic algorithm*, J. Phys., Conf. Ser., IOP Publ., 2019, 1255 (1), 012012. DOI: 10.1088/1742-6596/1255/1/012012.

[38] ROKACH M., ROKACH L., MAIMON O., *Top-down induction of decision trees classifiers-a survey*, IEEE Trans. Syst., Man, Cyber., Part C (Appl. Rev.), 2005, 35 (4), 476–487. DOI: 10.1109/TSMCC. 2004.843247.

[39] SWAIN P.H., HAUSKA H., *The decision tree classifier: Design and potential*, IEEE Trans. Geosci. Electron., 1977, 15 (3), 142–147. DOI: 10.1109/TGE.1977.6498972.

[40] ELSON J., TAILOR A., BANERJEE S., SALIM R., HILLABY K., JURKOVIC D., *Expectant management of tubal ectopic pregnancy: prediction of successful outcome using decision tree analysis*, Ultras. Obst. Gyn., 2004, 23 (6), 552–556. DOI: 10.1002/uog.1061.

[41] CHIANG W.Y.K., ZHANG D., ZHOU L., *Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression*, Dec. Supp. Syst., 2006, 41 (2), 514–531. DOI: 10.1016/j.dss.2004.08.016.

[42] NEVITT J., HANCOCK G.R., *Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling*, J. Exp. Educ., 2000, 68 (3), 251–268. DOI: 10.1080 /00220970009600095.

[43] CHICCO D., WARRENS M.J., JURMAN G., *The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation*, Peer J. Comp. Sci., 2021, 7, 623. DOI: 10.7717/PEERJ-CS.623.

[44] PRASANNAVENKATESAN T., JACOB I.J., RUBY A.U., VAMSIDHAR Y., *Prediction of COVID-19 possibilities using KNN classification algorithm*, Research Square, 2020. DOI: 10.21203/rs.3.rs-70985/v2.

[45] WU H., YANG S., HUANG Z., HE J., WANG X., *Type 2 diabetes mellitus prediction model based on data mining*, Inf. Med. Unl., 2018, 10, 100–107. DOI: 10.1016/j.imu.2017.12.006.