© by PSP Volume 30– No. 08/2021 pages 10057-10077 Fresenius Environmental Bulletin

PREDICTION OF AIR QUALITY INDEX BASED ON METEOROLOGICAL VARIABLES USING MACHINE LEARNING TECHNIQUES

Sinan Uguz[1], Okan Oral[2,*]

[1]Department of Computer Engineering, Isparta University of Applied Science, Isparta, Turkey
[2]Department of Mechatronics Engineering, Akdeniz University, Antalya, Turkey

ABSTRACT

It can be observed that the distribution of population worldwide is becoming increasingly concentrated in certain regions. The creation of new metropoles through a decrease in rural population leads to many disadvantageous situations. Industrialization caused by increasing population will bring about health problems in the short term and also problems resulting from global warming in the long term. The deterioration in air quality is among the most important such problem. In order to measure a complex phenomenon such as air quality and to make it easier for people to understand, each country applies certain air quality indexes in line with domestic legislation. For this study, with regard to estimations of Turkey's national air quality index (AQI) based on Environmental Protection Agency (EPA) standards, data have been collected from a station in Ankara city for the 2010-2014 period. In addition to PM_{10}, O_3, SO_2, NO_2, CO pollutant concentrations, a dataset consisting of six different sets of meteorological data has also been created. In the study, XGBoost (XGB), random forest regression (RFR) and extra trees involving tree-based ensemble learning (TBEL) algorithms have been used to estimate the air quality index. In addition to these algorithms, analyses have been carried out with the support vector regression (SVR), artificial neural networks (ANN) and k-nearest neighbors regression (k-NNR). According to the test data set results, an average R-Squared score of 0.97 was obtained with regard to the TBEL algorithms, while a score of 0.76 was obtained for the other algorithms. In addition, the TBEL algorithms showed a close prediction performance among themselves in terms of test scores, and a best R-Squared score of 0.99 has been obtained with regard to the RFR algorithm.

KEYWORDS:
Air Quality Index, Ensemble Learning, Support Vector Regression, Bagging, Boosting

INTRODUCTION

According to epidemiological research conducted in recent years, short-term changes in air pollution levels have negative effects on human health [1,2]. While air pollution has negative effects on human health in the short term, it has also become one of the causes of another important problem in the form of global warming in the long term [3]. Air pollution has important effects on human health along with other socio-economic impacts. It is known that people exposed to long-term air pollution have problems such as respiratory tract, cardiovascular, liver illnesses, eye diseases and skin infections. [4,5]. Premature deaths resulting from air pollution occur in the form of heart disease and stroke (58%), respiratory diseases (18%) and lung cancer (6%) [6].

Air pollution, which is the source of such significant health problems, has a complex structure consisting of air pollutants such as particulate matter (PM), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), tropospheric ozone (O_3), carbon monoxide (CO), rubber dust, polycyclic aromatic hydrocarbons ($PAHs$) and a range of different volatile organic compounds ($VOCs$) [1]. PM air pollutants are either liquid drops or solid particles suspended in the air. These can vary depending on air conditions, the measurement time period and place. The aerodynamic diameters of the particles display a variance of between 0.001μm and 100 μm [7]. Particles having a diameter less than 10 mm are designated as PM_{10} and those less than 2.5 mm as $PM_{2.5}$ (fine particules) [8]. Countries assign a greater importance to $PM_{2.5}$ concentration which is formed by the burning of coal, gasoline, natural gas and other organic substances. Particles having a dimension smaller than PM_{10} permeate the inside of the lungs causing more persistent health problems [9]. The health limit for $PM_{2.5}$ concentration can be expressed as $PM_{2.5} \geq 35.5\ \mu g\ m^{-3}$ averaged for 24 hours [10]. In addition, $PM_{2.5}$ is listed among the actors in terms of global warming, as it is evaluated in the same green-house gases category as O_3 [11]. Another pollutant causing air pollution is O_3. It is a pollutant which results from the release of gases of human origin such as from motor vehicles, industrial

activities and power plants, causing various health problems. O_3 is formed as a result of nitrogen oxides and hydrocarbon derivatives reacting with sunlight [12]. The health limit for the average concentration of O_3 can be expressed as $O_3 \geq 76\ ppbv$ for 8 hours [10].

In order to identify pollutant levels causing air pollution, measurement stations should become widespread. When the literature is surveyed, the majority of studies predicting pollutant levels in regions where there are no measurement station are found to use methods based on statistical procedures and machine learning to predict concentration amounts of pollutants such as $PM_{2.5}$, PM_{10} and O_3. There exist studies in which meteorological data such as temperature, humidity, atmospheric pressure, wind

TABLE 1
Studies conducted to predict AQI and Pollutants

Ref.	Predicted parameters	Techniques				Meteorological data	NSP	Location
		St	Hybrid	MLB	MLE			
Kleine Deters et al. [13]	$PM_{2.5}$			✓	✓	✓	2223	Ecuador
Sihag et al. [14]	$PM_{2.5}$			✓	✓	✓	659	India
Amanollahi, Ausati [15]	$PM_{2.5}$	✓	✓			✓	335	Iran
Feng et al. [16]	$PM_{2.5}$			✓		✓	1 year	China
Wang, Wang [7]	$PM_{2.5}$			✓			900	China
Li et al. [22]	$PM_{2.5}$	✓		✓		✓	3 years	China
Tong et al. [23]	$PM_{2.5}$			✓			6717	USA
Martínez et al. [2]	PM_{10}	✓			✓	✓		Colombia
Cujia et al. [17]	PM_{10}	✓					16 years	Colombia
Chen et al. [8]	PM_{10}	✓				✓	5 years	China
Eslami et al. [18]	O_3			✓	✓	✓	3 years	Korea
Elangasinghe et al. [21]	NO_2	✓		✓		✓	2 years	New Zealand
Kurt, Oktay [3]	SO_2, CO, PM_{10}			✓		✓	1 year	Turkey
Lei et al. [20]	NO_2, O_3, $PM_{2.5}$, PM	✓			✓	✓	5 years	China
Xi et al. [24]	AQI			✓	✓	✓	3 years	China
Zhu et al. [25]	AQI	✓	✓	✓		✓	10 years	USA
Koo et al. [26]	AQI	✓		✓			6 years	Malaysia
Arnaudo et al. [27]	AQI	✓		✓	✓	✓	3 years	Italy
Wu, Lin [28]	AQI		✓				761	China
Veljanovska, Dimoski [29]	AQI			✓			1 year	Macedonia
Nimesh et al. [30]	AQI	✓					1 year	India
Castelli et al. [31]	AQI			✓		✓	2 years	USA
Liu et al. [32]	AQI			✓	✓		9358	China
Qin et al. [33]	AQI			✓			1 year	China
Hajek, Olej [34]	AQI			✓		✓		Czech Republic

NSP: Number of samples or period, St: Statistical, MLB: Machine learning base, MLE: Machine learning ensemble

speed, wind direction, and rainfall are included among the features of the models set up for prediction purposes. In their study, Kleine Deters et al. [13] classified $PM_{2.5}$ concentration as high and low using boosted tree, linear support vector machines (L-SVM) including meteorological data. In addition to these methods, regression analysis has been carried out using ANN. As a result of the research findings, it has been observed that there exists a strong correlation between meteorological data and $PM_{2.5}$ concentration. In their study, Sihag et al. [14] tried to predict the concentration of $PM_{2.5}$. A data set consisting of points of 659 data was obtained for a three-year period. In addition to meteorological data, there were values for the pollutants SO_2, NO, O_3 and $PM_{2.5}$. The authors made use of machine learning methods such as ANN, support vector machines (SVM), Gaussian process regression, M5P, and random forest (RF). Along with other studies in which meteorological data were used as input feature in prediction methods [15,7,16], studies have also been carried out to predict PM_{10} concentration. Martínez et al. [2] used logistic regression, classification and regression trees (CART) and RF as 3 different machine learning methods to predict PM_{10} concentration levels. Cujia et al. [17] used time series forecasting-based methods to predict missing data in a 16-year dataset and to determine PM_{10} concentration levels. In the study carried out by Chen et al. [8], statistical methods were preferred to predict PM_{10} concentration by utilizing stepwise regression techniques and by removing input parameters having statistically low significance levels from the model. Studies conducted on the concentration of air pollutants other than PM can be stated to be more interested in O_3 pollutants compared with other air pollutants. Eslami et al. [18] estimated hourly O_3 concentration in their study employing extremely randomized tree (ERT) and deep neural networks, using data collected over the course of three years. Martínez-España et al. [19] utilized data collected from 5 different stations each having approximately 8,500 sample datasets by using Bagging, RF, decision tree (DT) and k-nearest neighbor (k-NN) methods, and achieved 92% accuracy. On the other hand, some studies have been directed to predict concentration of pollutants such as SO_2 and NO_2 which have less impact both on human health and on the level of global warming [20,21,3].

In Table 1, an example set of studies conducted to estimate AQI and various pollutants is provided. In the table, along with the prediction techniques used in each study, the utilization of meteorological data in the datasets is also provided. While in some of the studies the exact number of samples is given, in other studies the data collection period is provided.

Research Gaps and Motivation. Stations where pollutant and meteorological data are measured are not widely available in every country due to high installation costs. In addition, because of the large number of particles that cause air pollution, it may be difficult to access the measurement values of all these particles in existing stations. On occasions, due to limited measurement devices, incomplete or erroneous data measurements can be occur. Consequently, the available data and the prediction models created with these data are valuable. In addition, these factors constitute an important source of motivation for conducting studies dealing with air quality prediction.

AQIs, which enable a simple representation of the complex components that cause air pollution, differ depending on each country's air quality standards. Some countries publish air pollution reports according to air pollutant levels such as M_{10} ,O_3 ,SO_2 ,NO_2 ,CO. Therefore, it can be seen that studies on air quality and pollutant prediction tend to be carried out according to the specific situation found in individual countries. Thus, it can be said that motivation for an important area of research has been created for researchers from different countries.

Differences in national air quality standards have led to the creation of various AQI scales. Therefore, not only the pollutants prediction but also AQI prediction problems are among the issues that currently attract researchers. This can be seen by looking at the summary of the literature in Table 1. It is obvious that the approaches used to study AQI predictions differ from each other in a number of ways. The first of these differences is the use of meteorological data in the data sets that constitute the studies. It can be seen that in the data set of the study by Arnaudo et al. [27], besides pollutants, meteorological data such as temperature, humidity and air velocity are also included. However, meteorological data were not taken into consideration in the studies of Wu, Lin [28] and Veljanovska, Dimoski [29]. Another aspect of the studies outlined in Table 1 that causes them to differ from each other is the prediction method used. Some researchers [26,30,25] worked on statistical prediction models such as ARIMA and using multiple linear regression. Other researchers have conducted studies using prediction models based on machine learning techniques. The data for AQI predictions change over time and are of a non-linear character. They also tend to be heterogeneous, inconsistent, missing and uncertain. This character of data indicates that those models using machine learning involving uncertainty, might perform well with regard to AQI predictions [34]. Castelli et al. [31] undertook, a study including AQI estimation for the state of California, as well as the hourly-obtained pollutants O_3 ,SO_2 ,NO_2, CO, and $PM_{2.5}$. In this study, prediction models were created using only the SVR (Support Vector Regression) technique. The number of attributes in the data set was reduced by principal component analysis. While creating the prediction model using SVR, the

researchers performed hyperparameter optimization on the kernel and C parameters. As a result, they achieved an AQI prediction success of 94.1%. In this study, it can be seen that the performance of different machine learning models is not compared. Veljanovska, Dimoski [29] approached the subject as a classification problem by using the AQI categories seen in Table 2. The number of samples used in the study was limited to 365, and prediction models were developed using only ANN, k-NN and DT algorithms. The classification success obtained in this study was 92.3%, 80% and 76% for ANN, k-NN and DT algorithms respectively. The classification success remained relatively low compared to those studies reported in the literature that performed AQI prediction using regression techniques. Hajek, Olej [34] used daily data, including meteorological data, from three different stations in order to predict the AQI quality determined by the Czech National Institute of Public Health. In the study, classification accuracy scores for AQI classes were obtained by using Takagi-Sugeno fuzzy inference systems, radial basis function neural networks, ANN and SVR techniques. The results varied between 50% and 63%. Liu et al. [32] created AQI prediction models using only SVR and RFR techniques with a data set of 1,738 samples obtained from the Beijing Air Quality Dataset. It has been observed that the SVR technique gives a better result than the RFR technique.

Recently some researchers [27, 32, 24, 25] have been conducting studies into AQI predictions based on hybrid models and ensemble machine learning techniques. In their study, Xi et al. [24] carried out AQI estimation using SVM, RF, Gradient Boosting and DT methods employing $PM_{2.5}$, PM_{10},O_3, SO_2, NO_2, CO pollutants and meteorological data as input features. The dataset used in this study was obtained from 74 different cities throughout China, and a maximum accuracy performance of 65% was achieved. In the study, when the number of samples in the dataset increased, the increase in model performances was noted. When compared to the accuracy performance obtained by Xi et al. [24], higher performance results were obtained in our study. In addition, another aspect of our study that distinguishes it from Xi et al.'s study is that a high prediction score was obtained by using ERT as an ensemble technique. Arnaudo et al. [27] and Liu et al. [32] in their work, the only ensemble technique used was RF. Our study differs from these studies in terms of using three different ensemble techniques.

The main contributions of this paper are as follows: (1) As can be seen in some studies in Table 1, in this study not only pollutant data but also meteorological data were evaluated in the form of input features of the models established for AQI prediction purposes. (2) In addition to AQI prediction basic machine learning models and hybrid models, recently ensemble models have been used as stated in Table 1. It is observed that the use of ensemble models was more limited in studies conducted for the purposes of AQI prediction. The RF algorithm is mostly used in ensemble model studies. In this study, the predictive performance of all ensemble models (RF, ERT and XGB), and basic machine learning models such as ANN, SVR and k-NNR, were compared. (3) It has been shown that the ensemble models used in studies with regard to AQI predictions give better prediction performance compared to traditional techniques.

The paper is organized as follows. Section 2 introduces: (i) the data sources used for the experiments, and describe the process of acquisition, preprocessing and analysis; (ii) theoretical elements of the machine learning algorithms used, and the parameter optimization for each algorithm; (iii) the performance evaluation criteria used. Section 3 reports the results obtained from the learning curves. Section 4 offers a comparison with similar studies in the literature. Lastly, in Section 5, we conclude with a summary of the study, and delineate possible future works.

MATERIALS AND METHODS

The general stages of the AQI estimation in this study based on machine learning are shown in Fig 1. After the collection of pollutant concentrations and meteorological data for the time period under consideration obtained from the air quality monitoring network for AQI prediction, the data pre-processing stage with regard to the correction of deficient data was performed. Missing data analysis helps address several concerns caused by incomplete data. There was only a small amount of missing data in the data set. By taking the average of the column (feature) of the missing data, it is used instead of the missing value. Therefore there were no losses in terms of the number of data. It is important to test a model created by machine learning with new data that does not exist in the dataset, in order to determine the model's validity. To do this, the dataset was split into 80% training data and 20% test data by using the k-fold cross-validation method. In the performed experiments, the best k value was found to be 10. Consequently, in the final model, analyses were carried out by separating the dataset into 10 parts. The process of testing the model using the test set and evaluating its performance, is expressed as a generalization. The success of machine learning depends on the success of such generalization. The error obtained when the model is tested with the test set is called the generalization error. This error value reveals how good the model performs on data that has never seen before. When the data are separated very precisely, the error rate will be low, but the generalization will also be low. In such a model, the possibility of overfitting occurring is high, and this is un-solicited status. With regularization, overfitting can be prevented. Regularization is realized by changing some of the hyper

parameters of the machine learning algorithm. By means of regularization techniques, attempts are made to create the simplest possible machine learning models. For a model's regularization, tuning should be carried out on the hyper parameters specific to the algorithm used. Many models are trained with various hyper parameters using the training set. If the trained model performs more poorly when tested with the validation data, it can be of the opinion that overfitting. Following the execution of the hyper parameter tuning of the algorithms used in the study, performance evaluating metrics were applied for regression purposes. In the case of machine learning algorithms, it is possible to work with numerous features and in some machine learning algorithms, if the data exhibits the same range of values, it can provide better predictive performance [35]. In this study, feature scaling was applied for the algorithms with the exception of the TBEL models, and the samples in the data set were normalized between 0 and 1. Analyses were done by using packages written in Python and R programming languages. The scikit-learn library and some supporter libraries developed for machine learning applications in the Python language were used. Because the XGB algorithm is not included in this library, a solution was generated using the xgboost package in the R programming language. In this study, models were developed for AQI prediction purposes using ANN, SVR, and k-NNR techniques as well as TBEL techniques. The values of some parameters in the test phase of the algorithms which have an effect on prediction performance are explained in Section 2.3 and Section 2.4.

Methods Used to Assess Air Quality. In recent years, one of top issues that countries have had to deal with is air pollution. Rapid population growth in some countries and population growth leading to concentrations in urban areas have intensified the coalescence of air pollution factors resulting from industrial activity, heating and traffic. Turkey is one of the countries that have seen rapid population growth. While the urban population makes up 54.3% of the

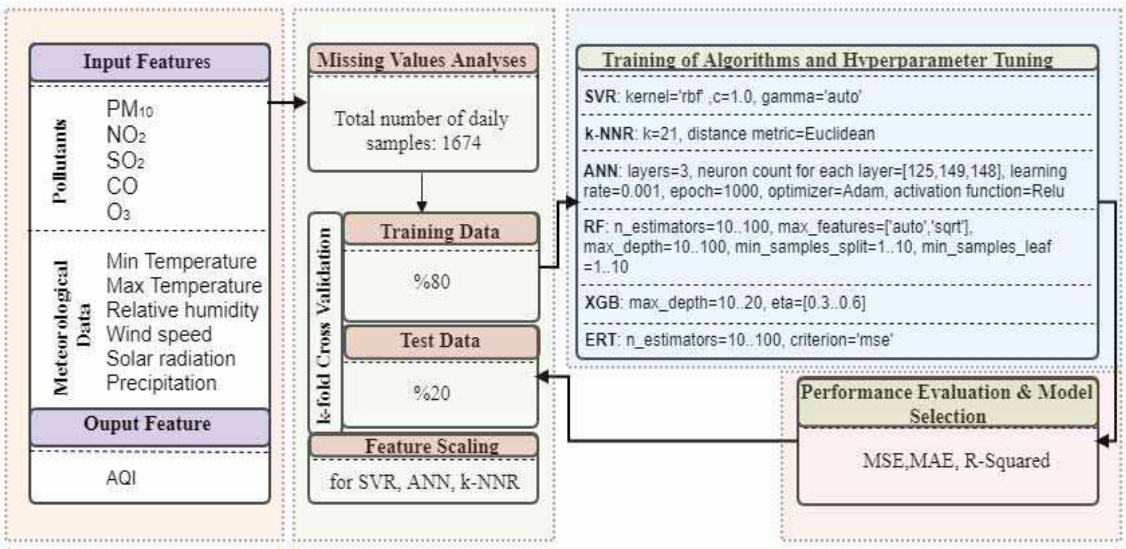


FIGURE 1
Steps of the machine learning process performed

TABLE 2
Pollutant-specific sub-indices of the AQI in Turkey [38]

AQI Categories	Index Values	SO_2(µg m^3, 1-h)	NO_2(µg m^3, 1-h)	CO(µg m^3, 8-h)	O_3(µg m^3, 8-h)	PM_{10}(µg m^3, 24-h)
Good	0-50	0-100	0-100	0-5500	0-120	0-50
Moderate	51-100	101-250	101-200	5501-10000	121-160	51-100
Unhealty for Sensitive Groups	101-150	251-500	201-500	10001-16000	161-180	101-260
Unhealty	151-200	501-850	501-1000	16001-24000	181-240	261-400
Very Unhealty	201-300	851-1100	1001-2000	24001-32000	241-700	401-520
Hazardous	301-500	>1101	>2001	>32001	>701	>521

total world population, Turkey's corresponding ratio is 74.4%, notably higher than the world average [36]. In Turkey, the metropolises especially those of Istanbul and Ankara are densely populated. AQIs have been developed for air pollution levels that have allowed them to be expressed in such a way that people can easily understand them. These are AQIs that enable the complex components that cause air pollution to be depicted in a simple way that varies in accordance with each country's specific air quality standards. Turkey has adopted air quality limit values defined by the EU as its own limit values [37]. In accordance with the decision taken in 2008, Turkey is going to apply EU regulations for all pollutant parameters from 2024 onwards, and there is going to be a gradual transition until that year. Turkey established its national air quality index by adapting the EPA AQI to its own national regulations and limit values. Statement in Eq. (1) can be used for each pollutant when computing AQI.

where p is a pollutant, the AQI_p term denotes pollutant p's individual air quality index, while C_p denotes its concentration. BP_{Lo} is the concentration's lower threshold value, namely ($\leq C$) and BP_{Hi} denotes the concentration's higher threshold value namely ($\geq C$). The AQI_{Hi} and AQI_{Lo} values are individual air quality indices corresponding to the BP_{Hi} and BP_{Lo} threshold values respectively. AQI is thus defined as the highest air quality index as stated in Eq. (2).

$$AQI = \max(AQI_1, AQI_2, .., AQI_n) \quad (2)$$

The values seen in Table 2 have been organized in accordance with Turkish regulations and in the table the AQI value intervals of each pollutants are depicted. To facilitate a better understanding, these values have been expressed in six different verbal categories. The AQI value calculated as a result of measurements obtained at any time will be equal to the value of the pollutant which has the maximum value at that time. Therefore, the concentration values belonging to the 5 pollutants seen in Table 2 should be measured one by one.

Study Area and Dataset. Air quality is a complex phenomenon that cannot be fully represented only in terms of pollutants and meteorological data. Under the same meteorological conditions, the air quality of a mountainous region and that of a region by the sea can be different. Therefore, when the literature is examined, it should be noted that the various studies have been carried out in different regions. In order for this study, the second largest city in Turkey, Ankara where air pollution is intense, has been chosen. There are eight different air quality monitoring stations in Ankara. However, after querying the database, it was found that Keçiören station as shown in Figure 2 was found to be the station that has measurement values for all the pollutants, and has provided the greatest number of measurements throughout the year. For this reason, the dataset for this study has been formed with the concentration values of 5 major pollutants provided by the air quality monitoring station in Keçiören, which serves an urban area of approximately one million people. Regional data with regard to the five major pollutants used in the calculation of air quality index in Turkey can be monitored real-time with the help of the air quality monitoring network [38]. Using the Turkish national air quality monitoring network applications, daily data with regard to PM_{10}, O_3, SO_2, NO_2 and CO pollutants were accessed. However, meteorological data cannot be accessed through this system. For this reason, in order to access the necessary meteorological data, the Global Weather Data Project [39] was used. With this project, which has been undertaken by the National Centers for Environmental Prediction over a period of 36 years, temperature, precipitation, wind, relative humidity, and solar radiation data for any region around the world can be obtained. Besides daily recordings of pollutants in the data set PM_{10} ,O_3, SO_2, NO_2, CO , the study used meteorological data consisting of minimum and maximum temperatures, absolute humidity, rainfall, wind speed and solar radiation. The samples in both data sets consist of daily data for the period 01.01.2010 to 07.30.2014. Some days were not included in the data set due to missing data. Thus, the dataset consists of a total of 1,674 daily data inputs. The descriptive statistical information of the dataset are depicted on a yearly basis in Table 3. The features seen in Table 3 formed the input of the algorithms used and the output of the AQI values. As stated in Eq. (2), while creating the AQI values, the value of the pollutant with the highest value on that day was used.

The box plots of the input features of the pollutants and the meteorological data in the dataset are shown in Figure 3 and Figure 4 respectively. Box plots are an important type of statistical chart that provides information about the distribution of data in a data set, and in which areas the data are concentrated. When the box charts are examined, the left bar indicates the smallest value in the data set, and the right bar represents the largest value in the data set. The outliers are indicated by the circle symbol. In addition, the green triangle symbol indicates the average. In the box charts, the region between the lower quartile and the upper quartile corresponds to 50% of the data set. For this reason, the blue colored region represents half of the pollutant values. When Figure 3 is examined, the pollutant with the least outlier value see to be O_3. When Figure 4 is analyzed, it can be seen that the outliers of the meteorological features are less than the outliers of pollutants. In particular, there are no outliers in terms of maximum temperature, absolute humidity and solar radiation. The distributions of the features in Figure 4 are more symmetrical than the features in Figure 3. In Figure 4, only the distribution of precipitation is not symmetrical.

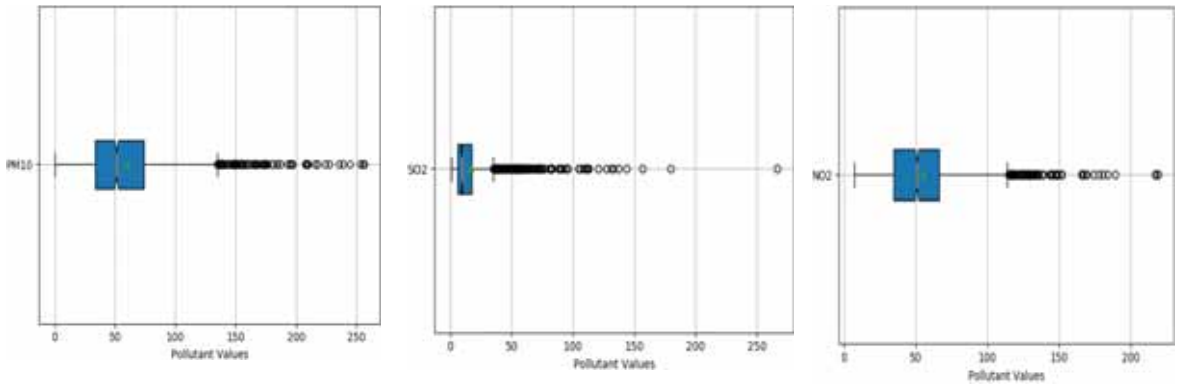
$$AQI_p = \frac{AQI_{Hi} - AQI_{Lo}}{BP_{Hi} - BP_{Lo}} \times (C_p - BP_{Lo}) + AQI_{Lo} \quad (1)$$

FIGURE 2
The location of the air monitoring station used in the study.

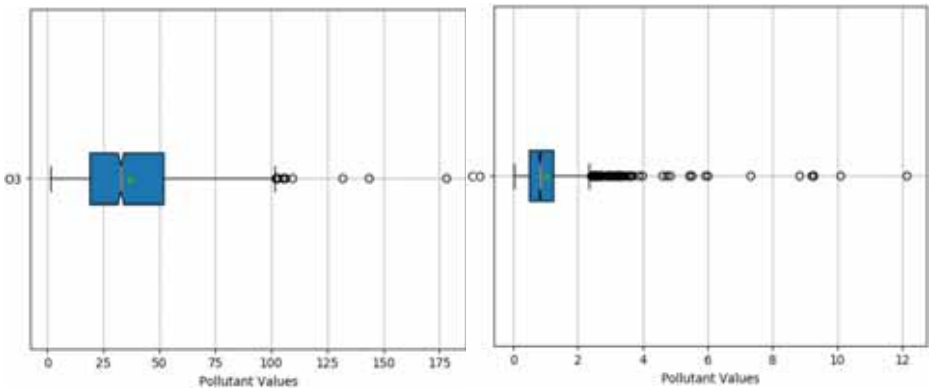
TABLE 3
The descriptive statistics of the input features used in the dataset

		Pollutants					Meteorological Variables					
Year	SP	PM_{10}	SO_2	NO_2	CO	O_3	T_{min}	T_{max}	RH	Wind	SR	Pre
2010	X_{Min}	0	2.4	28.4	0.18	9.8	-10.5	-4	0.23	0.61	1.3	0
	X_{Max}	196	34	138	3.60	94	21.6	39	0.95	7.43	32	31.2
	X_{Mean}	54	9.2	58	0.80	44	7.2	19	0.61	2.81	17.4	1.6
	X_{Std}	33	4.3	17.1	0.64	24.3	6.7	9.8	0.17	1.17	8.7	3.6
2011	X_{Min}	9.9	0.5	13.2	0.01	5.5	-17	-3.4	0.25	0.56	0.7	0
	X_{Max}	207	82.8	219.7	12.1	95	18	37.4	0.97	6.93	32.9	26.1
	X_{Mean}	55.8	11.4	56	1.42	39.6	4	15.4	0.66	2.15	17.3	2
	X_{Std}	28.5	11.7	36	1.45	19.8	7.2	10.1	0.17	0.85	8.2	4.1
2012	X_{Min}	3.2	3.6	9.6	0.25	3.11	-25.4	-6.7	0.21	0.62	0.6	0
	X_{Max}	244	266	111	3.24	178.2	21.4	40.8	0.99	6.28	32.8	37.2
	X_{Mean}	64.3	38.3	35.7	0.89	33.6	4.1	16.2	0.66	2.29	17.3	2
	X_{Std}	50.3	30.5	20.4	0.57	28.4	9.3	11.9	0.22	0.94	9.1	4.7
2013	X_{Min}	12.7	1.6	7	0.33	1.5	-18	-5.4	0.28	0.64	1.4	0
	X_{Max}	256	44.3	131	2.60	67.6	17.8	34.6	0.98	6.67	32.7	33.2
	X_{Mean}	71.3	11.2	65.2	1.09	33	4.5	16.8	0.61	2.28	17.8	1.4
	X_{Std}	41	6.9	27	0.48	16.2	7.3	9.8	0.18	0.90	8.6	3.6
2014	X_{Min}	8.3	1.9	11.8	0.14	1.3	-8.1	1.6	0.25	0.53	1.2	0
	X_{Max}	176	63.6	184.2	2.91	72.6	19.3	35	0.97	5.04	32.5	50.3
	X_{Mean}	53.7	9.9	64.6	0.68	30.3	5.2	17.4	0.62	2.22	18.7	2.1
	X_{Std}	30.6	9.7	35.6	0.52	18	6.3	8.8	0.15	0.82	7.9	5

SP=Statistical parameters, X_{Min}=Minimum value, XMax=Maximum value, X_{Mean}=Mean value, X_{Std}=Standard deviation,T_{min}= Minimum temperature (℃), T_{max}= Maximum temperature (℃), RH=Relative humidity (%), Wind=Wind speed (m/s), SR= Solar radiation (℃), Pre= Precipitation (mm)

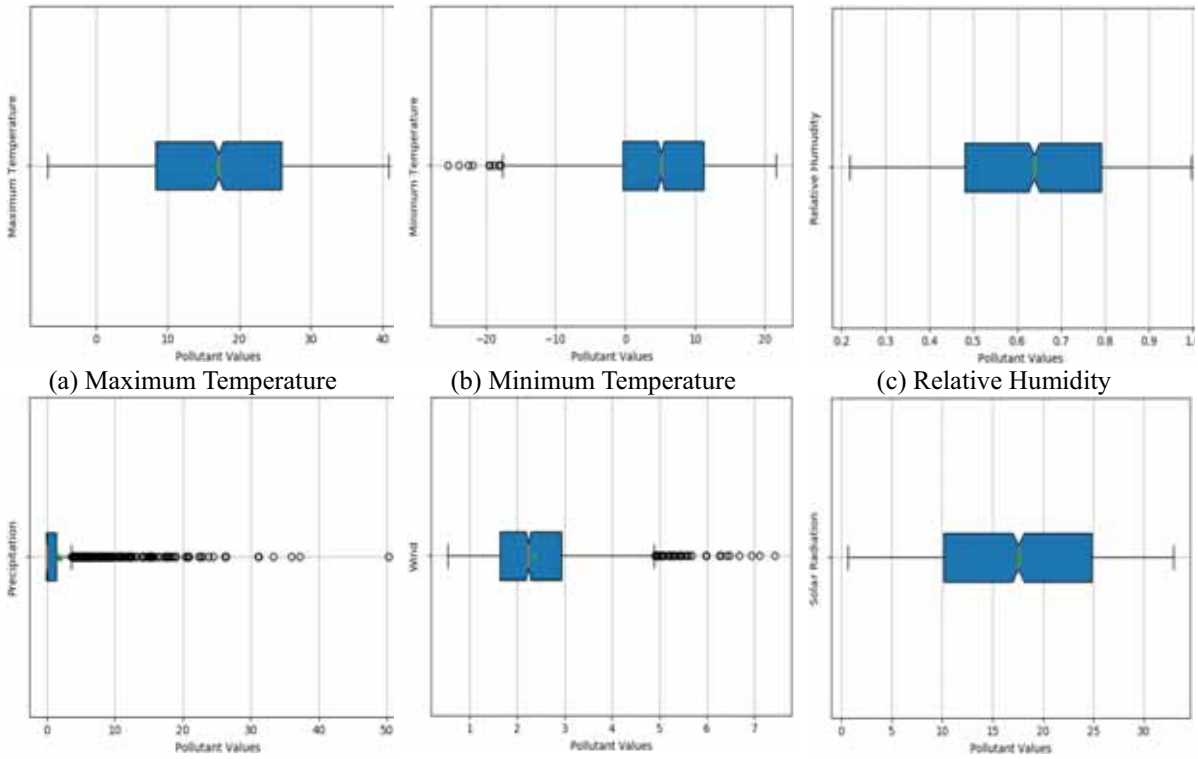


(a) PM_{10} (b) SO_2 (c) NO_2



(d) O_3 (e) CO

FIGURE 3
Representation of input features of the pollutants that are in dataset by box plots



(a) Maximum Temperature (b) Minimum Temperature (c) Relative Humidity

(d) Precipitation (e) Wind (f) Solar Radiation

FIGURE 4
Representation of the input features of meteorological data that are in dataset by box plots.

$$\begin{aligned} &Minimize && \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle \\ &subject\ to && y_i(\langle \boldsymbol{w}, \boldsymbol{x_i} \rangle + \theta_0) \geq 1\ , \qquad \forall i \end{aligned} \tag{3}$$

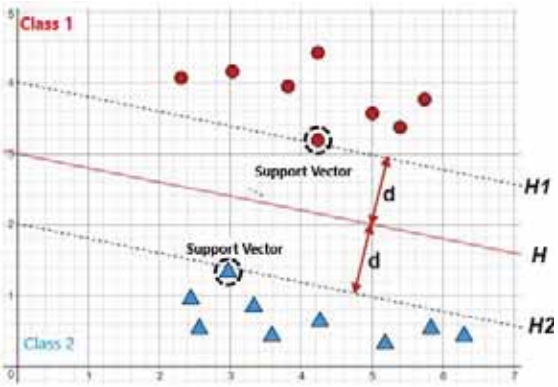


FIGURE 5
Linear separation of classes

Support Vector Regression. The support vector machines developed by Cortes, Vapnik [40] related to a machine learning technique which is widely used in regression applications where the dependent variable consists of continuous data as well as linear and nonlinear classification problems. A dataset can be classified linearly by means of planes in three-dimensional space. $\boldsymbol{x}$ and $\boldsymbol{w}$ represents the vectors in the two-dimensional plane. θ_0 and θ_1 are the weight values. The inner product of the vectors $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{w} = (\theta_1, -1)$ is shown as $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$. Also, $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \theta_0 = 0$ represents a hyperplane. In Fig. 5, the data of Class 1 and Class 2 are separated by hyperplanes $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$. Sample data on hyperplanes $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$ are called support vectors. Support vectors serve as a boundary. Two support vectors should be selected such that both the distance between the hyperplanes $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$ passing through these support vectors should be the largest, and the samples of Class 1 and Class 2 should be able to be separated from each other in the best way. This is the main goal that support vector machines want to achieve. In Figure 5, mathematical representation of $\langle \mathbf{w}, \mathbf{x} \rangle + \theta_0 = 0$ is used for H hyperplane. In this case, f$\langle \mathbf{w}, \mathbf{x} \rangle + \theta_0 - \mathrm{d} = 0$ can be used for the $\boldsymbol{H_1}$ hyperplane, and $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \theta_0 + d = 0$ can be used for the $\boldsymbol{H_2}$ hyperplane. The distance d in these expressions can be taken as 1 for ease of calculation. In this case $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \theta_0 = 1$ is obtained for the $\boldsymbol{H_1}$ hyperplane and $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \theta_0 = -1$ is obtained for the $\boldsymbol{H_2}$ hyperplane. In order to maximize the distance 2d between the hyperplanes $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$, and to separate the class 1 and class 2 data correctly, the optimization problem seen in the Eq. (3) must be solved. In Eq. (3), y_i represents class labels. Since the minimization problem in Eq. (3) has a quadratic objective function ($\frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle$) and a linear constraint, it can be solved by using quadratic programming techniques (Lagrange multipliers) [35].

In datasets that cannot be separated linearly, samples can be moved from a lower dimensional space (input space) to a higher dimensional space (attribute space) and thus their dimensions are changed. Suppose that the vectors in the input space x and z are represented as vectors a and b in higher dimensional space. If the value obtained as a result of the inner product of vectors ($\langle \boldsymbol{a}, \boldsymbol{b} \rangle$) a and b in high dimensional space can be written in terms of the inner product of x and z vectors in the input space, the function containing this inner product expression is called the kernel. Since some datasets are not separated linearly, core functions can be used to convert a linear model to a nonlinear model. If the objective function of any linear classification problem includes an inner product of the vectors $\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle$ an appropriate $K(\boldsymbol{x_i}, \boldsymbol{x_j})$ kernel function can be written instead of this inner product expression. After the publication of SVM, Smola [41] advanced an alternative loss function, which also allowed SVM to be applied to regression problems. The technique used in regression applications for datasets where the dependent variable consists of continuous data, is called support vector regression (SVR). In this study, an SVR was performed because the dependent variable consisted of continuous data.

The kernel parameter can be shown as the most important parameter used in the support vector regression model. The kernel parameter takes values as 'linear', 'poly', 'rbf', 'sigmoid' and 'precomputed' in the scikit-learn library. In this paper, the best results were obtained with the rbf kernel. Another important parameter is the C parameter. This is a penalty parameter that determines the tolerance shown to incorrectly-classified samples. The default value of 1.0 for this parameter was used. The final parameter with regard to SVM is gamma (γ), which gives the coefficient of the core functions. The default value of auto for this parameter was used.

Ensemble Methods. The main idea of the ensemble methods is to bring forth an ensemble's general prediction performance rather than the prediction performance of a single learner resulting in compensation for this single learner's prediction errors by other learners [42]. Using ensemble methods, the aim is to create models that have more powerful prediction capabilities by combining multi-machine learning methods' performances. Algorithms of ensemble methods can be used for solutions of both classification problems and regression problems. In ensemble methods, various machine learning algorithms based on techniques such as bagging and boosting have been developed. The work done by algorithms using the bagging method [43] can be sum-

marized as follows. New datasets are obtained by selecting samples from a basic dataset that is used on the condition that samples are selected randomly. New datasets that are obtained, and basic datasets are of the same size, and due to random selection, there can be repeating samples in the new datasets. In addition, the random selection of samples does not guarantee that all of the samples in the basic data will remain in the new datasets [44]. Among popular bagging algorithms, tree-based algorithms such as RF, Bagged Decision Tree and ERTs can be mentioned. Another ensemble method is boosting technique that forms the basis for algorithms such as Gradient Tree Boosting and XGBoost (XGB). In this technique, samples that are erroneously predicted by a learner and which can be expressed as being "weak", are collected in a sub dataset by assigning a greater weight to them. Then, a dataset containing erroneously predicted samples are submitted to new learners. Since each new learner focuses on previous erroneous samples, the performance of a learner that is expressed as "weak" at the beginning, turns into a "strong" prediction performance in the final model, as the summation or weighted summation of all the learners' individual predictions [45].

Random Forest Algorithm. RF is a bagging principle-based ensemble technique that operates by using numerous decision trees in the solution of both classification and regression problems [46]. In an RF algorithm, T bootstrapped sample sets are formed by selecting random samples from the original dataset that has n samples. Data belonging to the bootstrapped sample sets are divided as InBag and OOB in two sections. While Inbag corresponds to two-thirds of the original dataset, the remaining one-third is made up of OOB (Out of Bag) which contains test data reserved to evaluate performance. Later, a DT algorithm is run with features randomly selected from each bootstrapped sample test's InBag data [47].

While all features in the dataset are considered when forming split points in classical decision tress, randomly selected features are taken into consideration in the RF algorithm. A tree is grown into branches with the feature having the highest knowledge gain among the selected features until the branch number specified at the outset is reached, or no branch is left to be spanned. These steps are iterated over T epochs. In the end, a T number of decision trees are created. When a sample out of the original dataset is predicted, if there is a classification problem in question, the number of class labels with which each decision tree categorizes the sample is looked into. Prediction of the new sample is determined as the class label that the new sample is most included in. In other words, majority voting is accepted. If the problem is a regression problem, then the average of the numerical values generated by each decision tree is taken.

$\hat{y}_t$, denoting prediction of tth tree and T denoting number of trees (or bootstrapped sample sets) , $\hat{y}_{Rf}$ is the average of each tree's prediction and is expressed as stated in Eq. (4).

$$\hat{y}_{Rf} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t \tag{4}$$

In Figure 6, a flow diagram depicting the stages of bagging based running RF algorithm that solves a regression problem is seen.

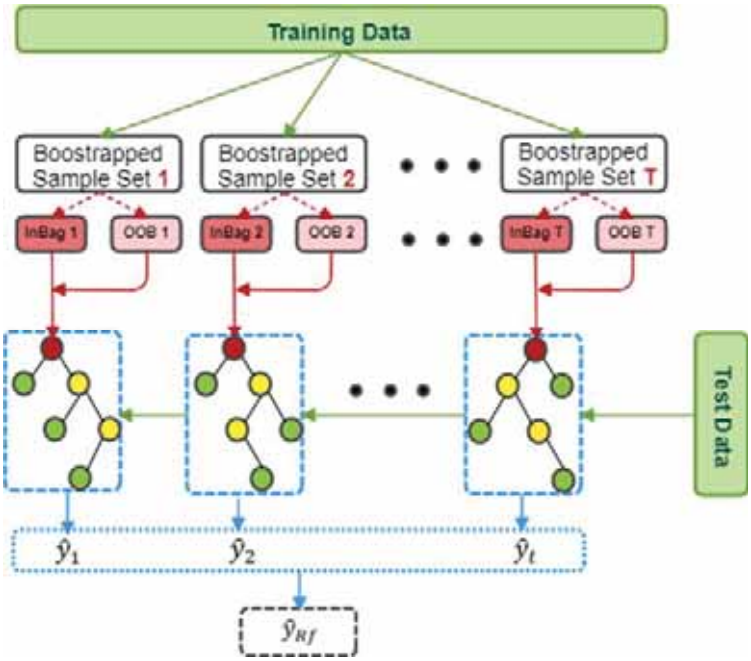


FIGURE 6
Flow diagram of RF algorithm

In the RF algorithm, random feature selection instead of selecting the best one among all features in splitting a decision tree causes bias values to increase. It may be possible to overcome this situation by using majority voting or averaging when prediction values are obtained. In this study, parameter optimization was carried out using the scikit-learn library. The most important parameter used in the experiments was the n_estimators parameter, which determines the number of trees in the forest. This parameter has been tested by taking 10 different values ranging from 10 to 100. Another parameter is max_features, which gives the maximum number of features considered when it comes to splitting a node. For this parameter, "auto" and "sqrt" values were tested. The max_depth parameter gives the maximum number of levels in each decision tree. For this parameter, experiments were made using 10 different values ranging from 10 to 100. The min_samples_split parameter returned the minimum number of data points placed on a node before the node is split. The min_samples_leaf parameter gives the minimum number of data points allowed in the leaf node. For these two parameters, parameter optimization was performed by giving values varying between 1 and 10. By using the values that each parameter can take, the model was chosen that gives the best performance among the many variations.

XGBoostAlgorithm. XGBoost (Extreme Gradient Boosting) [48] is a tree-based ensemble machine learning algorithm that uses the boosting method. XGB is based on the Gradient Boosting Decision Tree algorithm that plans to create a stronger model by reducing the errors (residuals) of the learner that is preceding itself in the gradient direction [49]. Overfitting ranks as the foremost of the problems that negatively affects prediction success for traditional decision tree learners. The regularization expression added to the objective function in the XGB algorithm plays an important role, both in terms of overcoming this problem and to in terms of resolving model complexity.

In the XGB algorithm, a model belonging to the regression tree represented by the first learner is formed by calculating the residual values which can be expressed as the difference between the target value in the dataset and the forecasted values.

Model output $\hat{y}_i$ can be expressed as in Eq. (5), with m fetures, x_i inputs, K denoting the number of trees, and F corresponding to each independent tree in a probable F tree space (known also as CART). In machine learning applications, the final objective is to set up a model that generates the most accurate $\hat{y}_i$ value. At the same time, the sum of the residuals is expected to be minimized, in other words, the model fits to the dataset to the greatest possible extend. The sum of the errors can be expressed with functions such as mean squared error (MSE) or logistic loss. These functions are also referred to as objective functions and machine learning learners want to minimize the objective function.

The objective function that we try to minimize in the XGB algorithm is seen in Eq. (6). The expression $\hat{y}_i^{(t)}$ yields the prediction value of the ith sample of the algorithm at the tth iteration. The expression l represents a differentiable convex loss function like MSE.

The statement $\Omega(f_t)$ seen in Eq. (7) is expressed as regularization term, and is added to the objective function to avoid overfitting. While T in regularization term denotes the leaf number in the tree, γ and λ are parameters that enable the degree of regularization to be changed. The statement w_j represents the weight values in each leaf.

When MSE is used as a loss function in the objective function stated in Eq. (6), the function includes first and second order terms. In this case, the objective function stated in Eq. (8) is obtained when the Taylor series expansion of the loss function is extended up to the 2nd degree.

where

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right) \tag{9}$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right) \tag{10}$$

The final expression in Eq. (11) is obtained after removing the constant term at the end of the second order Taylor expansion of the objective function.

In this study, experiments were carried out using the xgboost package of the R programming language. In the experiments, the maximum depth of the tree was determined by using values between 10 and 20. The eta parameter used to prevent overfitting was determined by using values between 0.3 and 0.6. Using the values that each parameter can take, the model has been chosen that gives the best performance among the variations.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F, \qquad x_i \in \mathbb{R}^m \tag{5}$$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\ \left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{6}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\ \sum_{j=1}^{T} w_j^2 \tag{7}$$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[l\ \left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (8)$$

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \ \Omega(f_t) \quad (11)$$

Extremely Randomized Tree. The Extremely Randomized Tree (also named as Extra trees) [50], exhibits a better performance with respect to training time by diverging from the RF algorithm in some aspects. The ERT algorithm can be used for both classification and regression purposes, just like an RF. The ERT algorithm, just like the RF algorithm, makes random feature selection from feature space but one of the fundamental differences lies in the selection of threshold values. In the selection of the threshold value, the best one is chosen from the set of threshold values randomly created for each feature. This aspect makes the ERT algorithm an ensemble learning method consisting of more randomized trees than is found in the RF algorithm [45]. In the RF algorithm, bootstrapped sample sets are created by randomly selecting samples from the original dataset. In the case of ERT, independent training of each tree with all the samples of the original dataset is involved [51]. The fact that the ERT contains more randomness compared to the RF makes it possible to create models with smaller variance. This situation makes the ERT an algorithm that is more resistant to overfitting formation than the RF. When creating a tree in the ERT algorithm, observations do not resample, and best split is not used. After selecting a random subset predictor for each split, a small number of randomly-chosen split points are generated for each of the selected predictors. The best split is then selected from these few splitting points. As a result, in the ERT algorithm, the forest contains more variables and fewer related trees than trees in the RF. All these features mean that the ERT offers a better prediction performance compared to other TBEL models. In this study, experiments were conducted using the scikit-learn library. The experiments were conducted by changing the value of the n_estimators parameter to between 20 and 100, which determines the number of trees in the forest. The criterion parameter used for measuring the quality of a split is determined as "mse". Parameter optimization has been performed for max_depth, min_samples_split and min_samples_leaf parameters using the same values as the RF.

Artificial Neural Networks. A simple ANN architecture called a single layer neural network contains only the input and output layers. In multi-layer ANN architectures, in addition to these two layers, there are hidden layers that significantly increase learning success. In this study, a prediction model has been created by using multi-layer ANN architecture. Relevant analyses were made using the Keras library. In the training involved in creating the best multi-layer ANN model, parameters that affect the ANN performance experiments were used. Accordingly, the ANN architecture consisted of three hidden layers. Each hidden layer consisted of between 100 and 150 artificial neurons. In the experiments, the number of neurons was changed by a loop in each training iteration. Thus, the number of neurons in the hidden layers of the best performing model was found to be 125,149 and 148, respectively. In the training sessions, in order to minimize the error, the weights in the network are updated. For the solver parameter used in this step, “adam” and “sgd” values were used. In the experiments, it was seen that the performance of the models was better when using “adam”. “Relu” was preferred as the activation function. In the phase of updating the weights in the neural network, the initial value of the learning rate parameter was determined as 0.001. Choosing a much smaller value significantly affected the training time of the network. Choosing a larger one may cause local or global minimum error values not to be accessed. In the multi-layer ANN architecture, the training involved 1,000 iterations.

k-Nearest Neighbors Regression. In every machine learning algorithm, parameter optimization is an important criterion for improving performance. In the k-NNR algorithm seen in Table 4, the k parameter can be expressed as an important parameter in this way. The k-NNR algorithm is basically based on the principle of calculating the distances of the samples in the data set using distance measurement methods such as euclidean, manhattan and minkowski. As a result of the calculation, depending on the k parameter, samples that are close to each other are included in the same class. In this study, in the experiments performed using the scikit-learn library, training was carried out by giving values between 5 and 30 to the k parameter. It was seen that the loss value was lower when the k parameter was between 21 and 30. For this reason, models with k parameters in this range have been found to be the best performing models obtained with k-NNR. For another parameter in the form of the distance metric, Euclidean distance is preferred.

Performance Evaluation. The AQI values to be predicted in the conducted study consisted of continuous numerical data. In machine learning, performance evaluating metrics such as R-Squared (Determination Coefficient) along with error criteria such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) show to what extent samples from a dataset can be represented by a line or curve. This

can also be used if the feature of the dataset to be predicted is comprised of continuous numerical data [52]. The mathematical notation of MSE, MAE and R-Squared criteria are given in Eqs. (12), (13) and (14) respectively.

In the equations, statement $\hat{y}_i$ expresses the i th sample's prediction value, while the y_i value expresses accurate values corresponding to an $n_{samples}$ number of dataset samples.

Findings. In this section, MSE, MAE and R-Squared results are expressed by presenting the learning curves of each algorithm. The learning curve determines the cross-validated training and test scores for different training set sizes. While creating the curves, the k value for k-fold cross validation was determined as 10. While the MSE and MAE values are expected to decrease with the increase in the training set size of the curves, the R-Squared value is aimed to approach 1. The MSE and MAE results of the RF algorithm are seen at Fig. 7. The graphs contain not only training score data but also test score results. This is because the test score results should be examined to determine whether or not the model is overfitting. From this perspective it can be seen that the MSE and MAE values of the RF model created using training data, are 0.0010 and 0.0015, respectively. The test score values were found to be 0.011 and 0.012, respectively. Therefore, it can be said that the test score results are close to the training score results. It can be seen that both loss value graphs start at a high loss value and trend downwards as the training set size increases and, after a certain phase, they consistently produce results displaying around the same loss value. In particular, this situation can be seen more clearly after a training set size of 500.

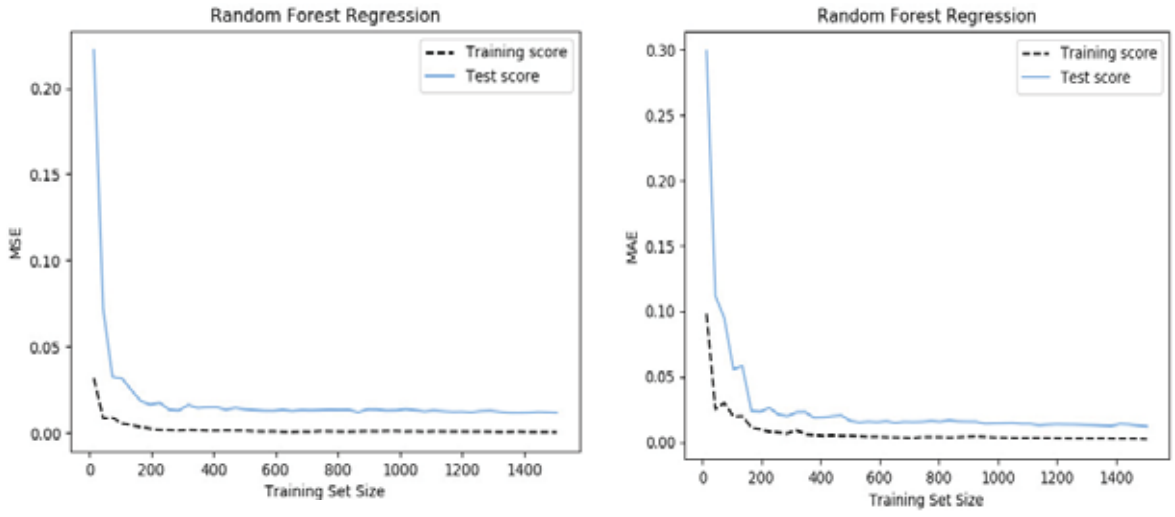
$$MSE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \tag{12}$$

$$MAE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \tag{13}$$

$$R - squared = (y,\hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{samples}-1} (y_i - \bar{y})^2}, \ \bar{y} = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} y_i \tag{14}$$

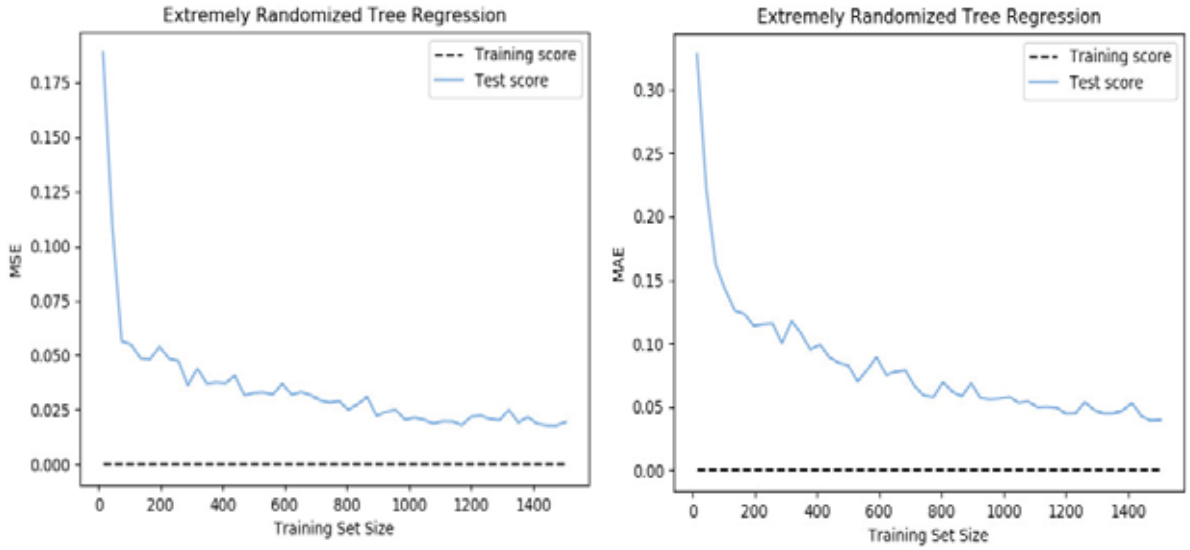
TABLE 4
The algorithm of k-NNR

Algorithm: k-Nearest Neighbors Regression		
Input	:	Attribute-value representation of instances, x_i
Output	:	Real-valued target, y_i
Step 1	:	Compute distance $D(x, x_i)$ to every training instance x_i
Step 2	:	Select k closest instances $x_{i1} \dots x_{ik}$ and their labels $y_{i1} \dots y_{ik}$
Step 3	:	Output the mean of $y_{i1} \dots y_{ik}$: $\hat{y} = f(x) = \frac{1}{k}\sum_{j=1}^{k} y_{i_j}$



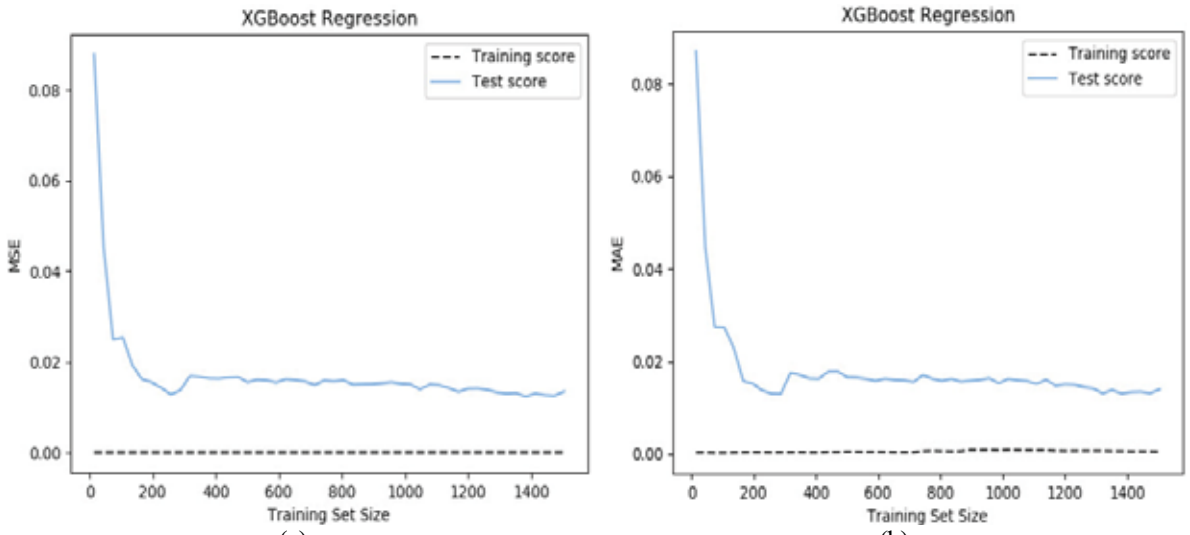
(a) (b)

FIGURE 7
The loss values for Random Forest Regression (a) MSE (b) MAE



(a) (b)

FIGURE 8
The loss values for Extremely Randomized Tree Regression (a) MSE (b) MAE



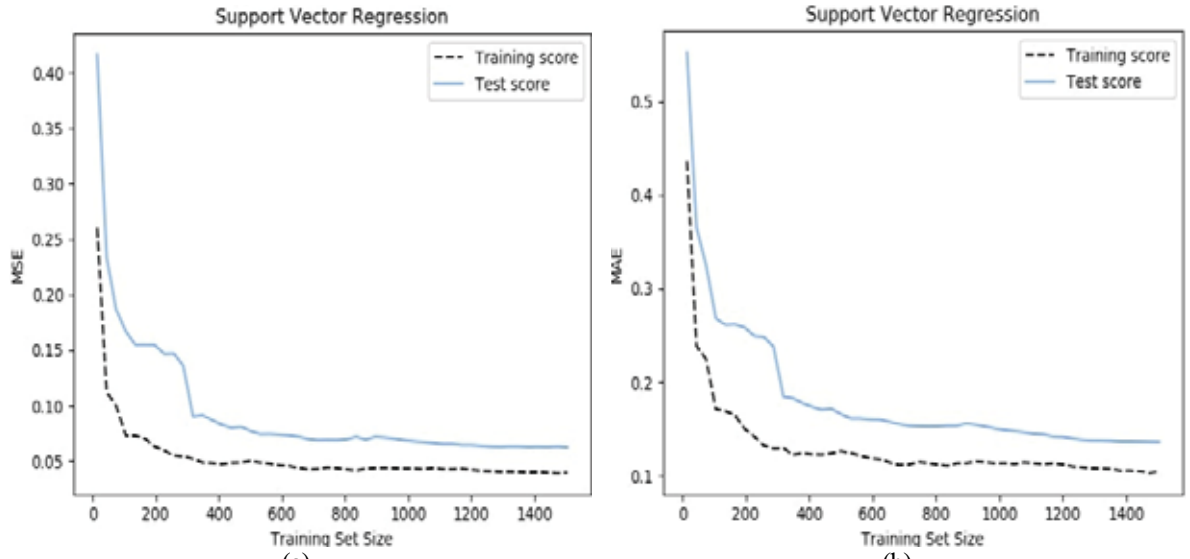
(a) (b)

FIGURE 9
The loss values for XGB Regression (a) MSE (b) MAE

The MSE and MAE results of the ERT algorithm are seen in Figure 8. It can be seen that the MSE and MAE values of the ERT model created using training data are 0.0018 and 0.0012, respectively. The test score values obtained were 0.019 and 0.040, respectively. When the ERT graphics are examined, it can be seen that the test score values are more unstable than those of the previous RF algorithm. In terms of training set size value up to 1000, this unstable situation is clearly visible. If there were more samples in the data set, it is predicted by looking at the graph that the loss value would continue to trend downwards.

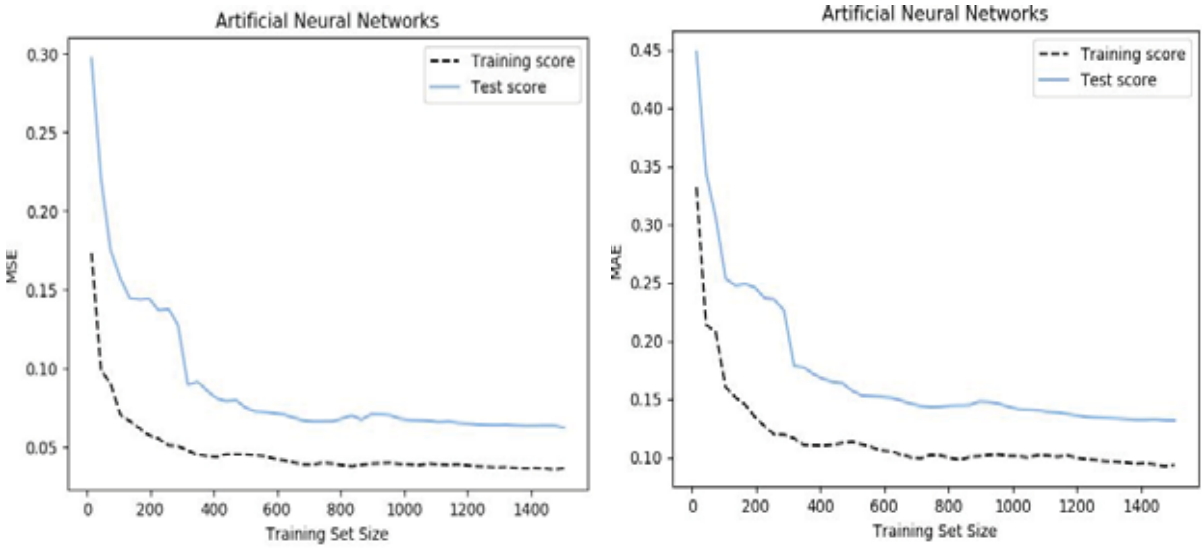
The MSE and MAE results belonging the XGB algorithm are seen in Figure 9. It can be seen that the MSE and MAE values of the XGB model created by using training data are 0.0003 and 0.0010, respectively. The test score values obtained were 0.013 and 0.014, respectively. When the graphics are examined, it can be seen that a more stable model is formed compared to the ERT algorithm. This is clearly seen after training set size values of 400.

In this study, for AQI prediction besides TBEL algorithms, some machine learning algorithms are also used. The first of these SVR models in the form of MSE and MAE results can be seen in Figure 10. It can be seen that the MSE and MAE values of the SVR model created using training data are 0.042 and 0.105, respectively. The test score values obtained were 0.062 and 0.138, respectively. These values show that the SVR model exhibits a lower prediction performance than the TBEL models. When the stability of the graphics of the model is examined, it can be concluded that the TBEL models are more stable than the ERT model.



(a) (b)

FIGURE 10
The loss values for Support Vector Regression (a) MSE (b) MAE

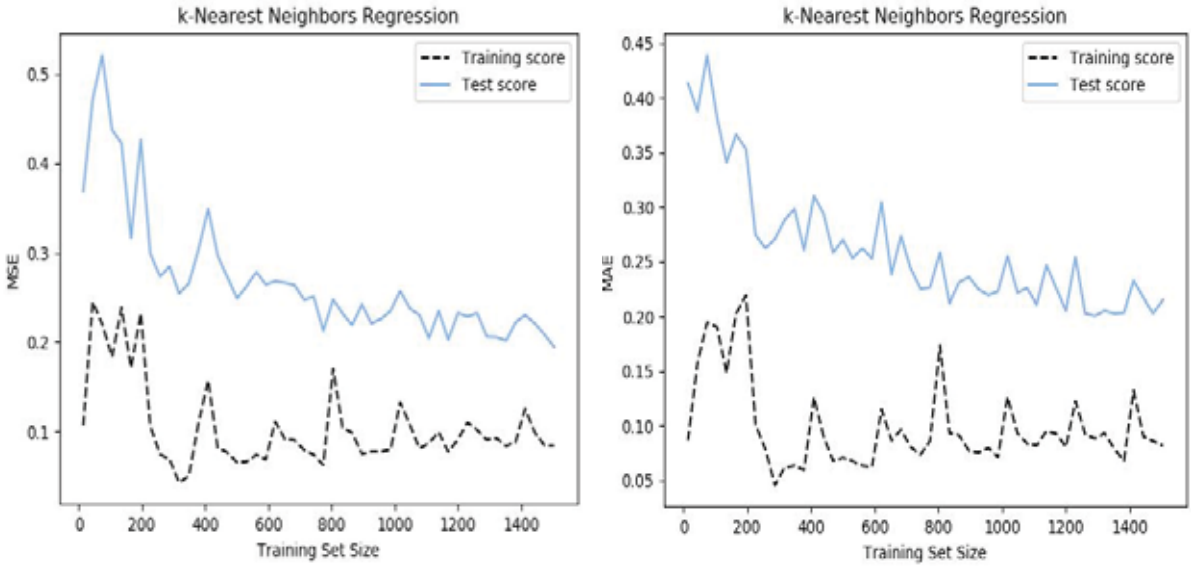


(a) (b)

FIGURE 11
The loss values for Artificial Neural Networks (a) MSE (b) MAE

Another machine learning model used in this study was ANN. The MSE and MAE results are seen in Figure 11. It can be seen that the MSE and MAE values of the ANN model created by using training data are 0.041 and 0.097, respectively. The test score values obtained were 0.060 and 0.013, respectively. With training set size values greater than 900, the model appears to have a stable shape. The ANN model is has the greatest number of parameters among the machine learning models. In this respect, the changes in these parameters with regard to the ANN model can provide lower loss values. The parameters used in this study are expressed in Section 2.4.4. Increasing the number of samples may also have an effect on ANN performance. This can be seen by looking at the graphics in that the decline in error scores in the graphics continues, even if by only a little.

Among the models created in this study, the lowest prediction performance was obtained with regard to k-NNR. The MSE and MAE results of k-NNR are shown in Figure 12. It can be seen that the MSE and MAE values of the k-NNR model created using training data are 0.095 and 0.10, respectively. The test score values obtained were 0.19 and 0.21, respectively. When the graphics are examined, it can be seen that the model is quite unstable. It can also be seen that the difference between the training score and the test score values is greater than that of other models. It is thought that increasing the number of samples might also have an effect on k-NNR performance. However, the instability of the model left k-NNR behind the other algorithms in terms of performance.



(a) (b)

FIGURE 12
The loss values for k-Nearest Neighbors Regression (a) MSE (b) MAE

TABLE 5
The performance in terms of training score results obtained from all machine learning models

Training Score						
	XGB	RF	ERT	SVR	ANN	k-NNR
R Squared	0.99	0.98	0.99	0.92	0.92	0.80
MSE	0.0003	0.0010	0.0018	0.042	0.041	0.095
MAE	0.0010	0.0015	0.0012	0.105	0.097	0.10

TABLE 6
The performance in terms of test score results obtained from all machine learning models

Test Score						
	XGB	RF	ERT	SVR	ANN	k-NNR
R Squared	0.97	0.99	0.96	0.84	0.83	0.61
MSE	0.013	0.011	0.019	0.062	0.060	0.19
MAE	0.014	0.012	0.040	0.138	0.0131	0.21

In Figure 13, the graphics with regard to the R-Squared results are shown with regard to the machine learning models used in this study. The R-Squared value of the models that provided good predictive performance is expected to approach 1. Training score values of R-Squared values of the TBEL models consisting of XGB, RF and ERT seem to be 0.99, 0.98 and 0.99 respectively. The test score values of the same models were 0.97, 0.99 and 0.96, respectively. These results show that the TBEL models exhibit an average performance of 0.97 R-Squared. It can be seen that the training score values of the R-Squared values of the other machine learning models used in this study (SVR, ANN and k-NNR) are 0.92, 0.92 and 0.80, respectively. The test score values of these same models were found to be 0.84, 0.83 and 0.61, respectively. According to these results, it can be concluded that with an average of 0.97, the TBEL models provide a higher prediction performance.

Training score results of all models belonging to the findings obtained in the study are shown in Table 5, and the test score results are shown in Table 6.

In Table 7 shows the real AQI values of different time periods in the data set and the prediction results produced by the models for these values.

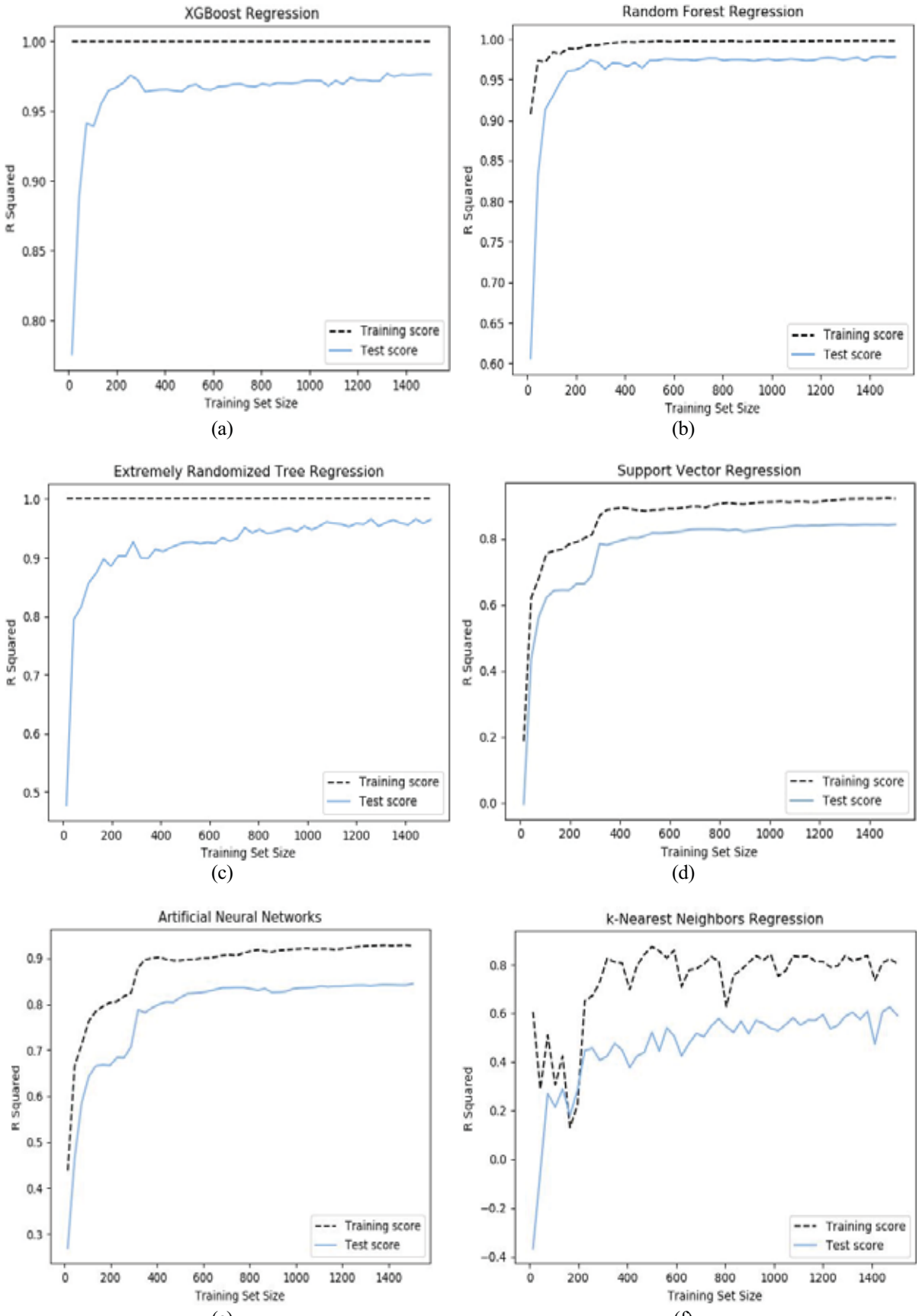


FIGURE 13
The R-Squared values for Machine Learning models
(a) XGBoost Regression (b) Random Forest Regression (c) Extremely Randomized Tree Regression (d) Support Vector Regression (e) Artificial Neural Networks (f) k-Nearest Neighbors Regression

TABLE 7
Actual and forecast results for AQI

Date	Actual	Forecasted by XGB	Forecasted by RF	Forecasted by ERT	Forecasted by SVR	Forecasted by ANN	Forecasted by k-NNR
03/01/2010	73.21	67.12	70.28	68.33	61.49	62.22	42.95
08/01/2010	120.88	126.84	124.61	129.12	141.92	144.32	169.21
12/03/2010	105.31	96.88	108.36	110.42	85.93	83.14	60.02
22/04/2010	27.58	26.68	26.92	30.02	32.36	23.48	17.41
05/11/2010	195.78	187.63	189.26	201.32	162.7	159.3	280.21
07/02/2011	207.53	219.04	195.36	196.12	175.96	241.05	155.25
27/02/2011	43.13	39.15	48.39	40.09	34.47	51.63	70.23
30/03/2011	84.16	81.03	85.33	88.09	98.30	100.58	51.69
10/07/2011	12.10	11.23	11.95	13.20	14.02	10.04	7.43
03/05/2012	266.81	259.01	263.31	278.03	221.06	223.69	164.8
15/092012	238.72	248.50	236.22	226.76	201.6	277.68	327.56
04/02/2013	116.14	120.35	114.29	121.52	135.21	136.52	73.16
18/11/2013	71.34	66.23	69.98	68.48	59.90	58.87	99.58
06/12/2013	235.53	226.13	232.17	224.36	197.02	194.28	144.02
20/01/2014	127.85	131.05	126.42	132.69	146.32	107.20	178.23
03/03/2014	103.87	107.82	102.02	109.36	86.39	84..59	62.31

DISCUSSION

Because TBEL algorithms and three different machine learning algorithms are used in this study for AQI prediction, a comparison can be made with the relevant studies noted in Table 1. Unlike in our study, Liu et al. [32] found that the prediction model created by SVR obtained a better prediction score than the TBEL-based RF algorithm. In terms of AQI prediction, they obtained an R-Squared score of 0.97 with the SVR algorithm. This score is higher than the SVR prediction performance in our study. However, the 0.84 R-Squared score that they obtained with RF with regard to in the prediction performance of nitrogen oxide pollutants remained lower than our prediction score. Xi et al. [24] developed AQI prediction models using TBEL algorithms in their study. However, the R-Squared score with the highest prediction score of 0.67 in their experiments lagged far behind that of our study. In their study, Arnaudo et al. [27] created prediction models using a TBEL-based RF algorithm. In their work, they performed their training by selecting the n_estimators and max_depth parameters as 100 and 8, respectively. In our study, we determined the parameters to be used by experimenting on values ranging from 10 to 100.

In their study, Kleine et al. [13] obtained similar results, and TBEL methods were observed to produce 4% better results compared with L-SVM. A similar assessment can also be made with regard to the work of Martínez-España et al. [19]. In this study, trainings were performed using logistic regression and RF algorithms, and the best accuracy scores were obtained with the RF algorithm. Combining the performances of multiple machine learning and ensemble methods, the general prediction performance of an ensemble rather than a single learner was observed to produce superior results. The use of ensemble methods, in terms of their lower prediction variance and decreasing the probability of the model overfitting, is recommended for prospective researchers planning to conduct similar studies. Naturally, considering that each data set may have characteristics specific to itself, experimenting using on different prediction models is deemed important in order to determine the optimum process. A number of research groups worldwide are conducting studies to forecast air pollution leading to significant health problems and global warming at an early phase. Collecting adequate and large quantities of data as part of the first phase of these research projects is one of the most challenging problems. In our study, the data belonging to pollutants, and the meteorological data were obtained from different sources. Since our data sources contained data only for the period up to 2014, no data set could be formed with regard to subsequent data. Therefore, sharing the data sets used by the conducted studies is quite important. We also consider it important to provide contributions to new researchers by openly sharing our data sets. As a step towards this goal, the data set used in this work is being made available [53].

CONCLUSIONS

Air pollution is one of the major problems caused by a rapidly increasing world population. Air pollution is a common problem facing all humanity in terms of both its negative effects on human health and its environmental damage in the long term. In the literature, it can be seen that studies regarding predictions of AQI and pollutants causing air pollution, show some differences. It can be said that the

studies differ from each other in terms of factors such as the prediction methods used, the use of meteorological data in the data set, and the regions in which the data were obtained. In terms of prediction methods used, there is a tendency is to use statistical methods, basic machine learning methods, ensemble methods and hybrid methods. There are limited number can be said of studies using ensemble methods and hybrid methods, especially in studies related to AQI prediction. Differences in national air quality standards cause researchers in different countries to make predictions based on their own standards. In this study, tree-based ensemble learning and some basic methods associated with machine learning algorithms using AQI prediction were realized, based on Turkey's national air quality standards.

The dataset used in this study were obtained from an air quality monitoring station that is located in Ankara, the second largest city in Turkey. In this data set, not only did we include measurement values with regard to five different pollutants (PM_{10}, O_3, SO_2, NO_2, CO) for the period 2010 to 2014, but also meteorological data consisting of minimum and maximum temperature, absolute humidity, rainfall, wind speed and solar radiation for that same period.

In the study, XGB, RF and extra trees were used in terms of tree-based ensemble learning algorithms for AQI prediction. In addition, analyses were performed with SVR, ANN and k-NNR techniques. Performance evaluation criteria for regression problems, MSE, MAE and R-Squared were used, and the performance of all algorithms in terms of AQI predictions were compared. It has been observed that the TBEL algorithms performed better than other algorithms. According to the test data set results, while an average R-Squared score of 97% was obtained with regard to the TBEL algorithms, the score of the other algorithms was found to be 76%. In addition, the tree-based ensemble algorithms displayed a similar prediction performance in terms of test scores. The best R-Squared score of 0.99 was obtained by the RFR algorithm. In terms of the other machine learning algorithms used, a close prediction performance was obtained in terms of the SVR and ANN algorithms.

With regard to air quality prediction, not only the data based on measurements such as pollutant concentrations and meteorological data, but also factors such as population, socio-economic structure etc. of the region where the measurement occurred also played an important part. Consequently, it can be see that air quality prediction is a complex process with regard to which a large number of factors need to be taken into account. Owing to the variability of all the factors in this field, specific studies of the situation in every country is valuable. We also conducted a study that produced high AQI prediction scores for Ankara, which is one of the foremost regions of Turkey in terms of high levels of air pollution. We are planning to focus future studies on comparing these results with the help of statistical methods. In addition to the prediction of air quality and the existence of pollutants at an early stage, the consequences of air pollution levels on human health and its related costs are also among the topics we plan to research.

REFERENCES

[1] Chen, H., Copes, R. (2012) P-300: Review of the air quality health index and the air quality index. Epidemiology. 23(5S), 704.

[2] Martínez, N. M., Montes, L. M., Mura, I., Franco, J. F. (2018) Machine Learning Techniques for PM 10 Levels Forecast in Bogotá. In: 2018 ICAI Workshops (ICAIW), Bogota D.C., Colombia, IEEE, 1-7.

[3] Kurt, A., Oktay, A. B. (2010) Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Systems with Applications. 37(12), 7986-7992.

[4] Gautam, D. B., Bolia, N. (2020) Air pollution: impact and interventions. Air Quality, Atmosphere & Health. 13, 209-223.

[5] Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., Asghar, M. N. (2019) Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access. 7, 128325-128338.

[6] WHO. (2020) Ambient (outdoor) air quality and health. http://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health, (Accessed 02.2.2020)

[7] Wang, X., Wang, B. (2019) Research on prediction of environmental aerosol and PM2.5 based on artificial neural network. Neural Computing and Applications. 31(12), 8217-8227.

[8] Chen, Y., Shi, R., Shu, S., Gao, W. (2013) Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. Atmospheric Environment. 74, 346-359.

[9] Gu, K., Qiao, J., Lin, W. (2018) Recurrent air quality predictor based on meteorology and pollution related factors. IEEE Transactions on Industrial Informatics. 14(9), 3946-3955.

[10] Ryan, W. F. (2016) The air quality forecast rote: Recent changes and future challenges. Journal of the Air & Waste Management Association. 66(6), 576-596.

[11] Gu, K., Zhou, Y., Sun, H., Zhao, L., Liu, S. (2019) Prediction of air quality in Shenzhen based on neural network algorithm. Neural Computing and Applications. 32, 1879-1892.

[12] Freeman, B. S., Taylor, G., Gharabaghi, B., Thé, J. (2018) Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association. 68(8), 866-886.

[13] Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., Rybarczyk, Y. (2017) Modeling PM 2.5 urban pollution using machine learning and selected meteorological parameters. Journal of Electrical and Computer Engineering, 1-14.
[14] Sihag, P., Kumar, V., Afghan, F. R., Pandhiani, S. M., Keshavarzi, A. (2019) Predictive modeling of PM2.5 using soft computing techniques: case study-Faridabad, Haryana, India. Air Quality, Atmosphere & Health 12 (12), 1511-1520.
[15] Amanollahi, J., Ausati, S. (2020) PM2.5 concentration forecasting using ANFIS, EEMD-GRNN, MLP, and MLR models: a case study of Tehran, Iran. Air Quality, Atmosphere & Health. (13), 161-171.
[16] Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J. (2015) Artificial neural networks forecasting of PM 2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmospheric Environment. 107, 118-128.
[17] Cujia, A., Agudelo-Castañeda, D., Pacheco-Bustos, C., Teixeira, E. C. (2019) Forecast of PM10 time-series data: A study case in Caribbean cities. Atmospheric Pollution Research. 10(6), 2053-2062.
[18] Eslami, E., Salman, A. K., Choi, Y., Sayeed, A., Lops, Y. (2020) A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks. Neural Computing and Applications. 32, 7563-7579.
[19] Martínez-España, R., Bueno-Crespo, A., Timon-Perez, I. M., Soto, J., Muñoz, A., Cecilia, J. M. (2018) Air-Pollution prediction in smart cities through machine learning methods: A case of study in Murcia, Spain. Journal of Universal Computer Science. 24(3), 261-276.
[20] Lei, M. T., Monjardino, J., Mendes, L., Gonçalves, D., Ferreira, F. (2019) Macao air quality forecast using statistical methods. Air Quality, Atmosphere & Health. 12(9), 1049-1057.
[21] Elangasinghe, M. A., Singhal, N., Dirks, K. N., Salmond, J. A. (2014) Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmospheric Pollution Research. 5(4), 696-708.
[22] Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T. (2017) Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental Pollution. 231 (1), 997-1004.
[23] Tong, W., Li, L., Zhou, X., Hamilton, A., Zhang, K. (2019) Deep learning PM 2.5 concentrations with bidirectional LSTM RNN. Air Quality, Atmosphere & Health. 12(4), 411-423.
[24] Xi, X., Wei, Z., Xiaoguang, R., Yijie, W., Xinxin, B., Wenjun, Y., Jin, D. A. (2015) comprehensive evaluation of air pollution prediction improvement by a machine learning method. In: 10th IEEE International Conference on Service Operations and Logistics, and Informatics, Tunisia. IEEE. 176-181.
[25] Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., Che, J. (2017) Daily air quality index forecasting with hybrid models: A case in China. Environmental Pollution. 231, 1232-1244.
[26] Koo, J. W., Wong, S. W., Selvachandran, G., Long, H. V., Son, L. H. (2020) Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models. Air Quality, Atmosphere & Health. 13(1), 77-88.
[27] Arnaudo, E., Farasin, A., Rossi, C. (2020) A comparative analysis for air quality estimation from traffic and meteorological data. Applied Sciences. 10(4587), 1-20.
[28] Wu, Q., Lin, H. (2019) A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Science of the Total Environment. 683, 808-821.
[29] Veljanovska, K., Dimoski, A. (2018) Air quality index prediction using simple machine learning algorithms. International Journal of Emerging Trends & Technology in Computer Science. 7(1), 025-030.
[30] Nimesh, R., Arora, S., Mahajan, K. K., Gill, A. N. (2014) Predicting air quality using ARIMA, ARFIMA and HW smoothing. Model Assisted Statistics and Applications. 9(2), 137-149.
[31] Castelli, M., Clemente, F. M., Popovič, A., Silva, S., Vanneschi, L. (2020) A machine learning approach to predict air quality in California. Complexity, 1-11.
[32] Liu, H., Li, Q., Yu, D., Gu, Y. (2019) Air quality index and air pollutant concentration prediction based on machine learning algorithms. Applied Sciences. 9(19), 4069.
[33] Qin, Z., Cen, C., Guo, X. (2019) Prediction of air quality based on KNN-LSTM. Journal of Physics: Conference Series. 1237(4), 042030.
[34] Hajek, P., Olej, V. (2015) Predicting common air quality index-the case of czech microregions. Aerosol and Air Quality Research. 15(2), 544-555.
[35] Uguz, S. (2020) Theoretical aspects of machine learning and an artificial intelligence school with Python applications, Nobel academic publishing, Turkey.
[36] Eurostat. (2020) The EU in the world population. https://ec.europa.eu/eurostat/statistics-explained/index.php. (Accessed 01.29.2020).
[37] Büke, T., Köne, A. Ç. (2016) Assessing air quality in Turkey: A proposed, air quality index. Sustainability. 8(1), 73.

[38] Turkey national air quality monitoring network. (2020, http://www.havaizleme.gov.tr, (Accessed 03.06.2020).
[39] Global weather data for SWAT, (2020). https://globalweather.tamu.edu, (Accessed 02.01.2020).
[40] Cortes, C., Vapnik, V. (1995) Support-vector networks. Machine learning. 20(3), 273-297.
[41] Smola, A. (1996) Regression Estimation with support vector Learning machines. Master, Technical University of Munich, Germany, 1-78.
[42] Sagi, O., Rokach, L. (2018) Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 8(4), 1-18.
[43] Breiman, L. (1996) Bagging predictors. Machine learning. 24(2), 123-140.
[44] Harrington, P. (2012) Machine learning in action. Manning Publications, Shelter Island, 1-384.
[45] Julian, D. (2016) Designing machine learning systems with Python. Packt Publishing Ltd, Birmingham, 1-209.
[46] Breiman, L. (2001) Random forests. Machine learning. 45(1), 5-32.
[47] Rodriguez-Galiano, V., Chica-Olmo, M., Chica-Rivas, M. (2014) Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. International Journal of Geographical Information Science. 28(7), 1336-1354.
[48] Chen, T., Guestrin, C. (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco California USA, 785-794.
[49] Friedman, J.H. (2001) Greedy function approximation: A gradient boosting machine. The Annals of statistics. 29(5), 1189-1232.
[50] Geurts, P., Ernst, D., Wehenkel, L. (2006) Extremely randomized trees. Machine Learning. 63(1), 3-42.
[51] Pinto, A., Pereira, S., Rasteiro, D., Silva, C.A. (2018) Hierarchical brain tumour segmentation using extremely randomized trees. Pattern Recognition. 82, 105-117.
[52] Zheng, A. (2015) Evaluating machine learning models: A beginner's guide to key concepts and pitfalls. CA: O'Reilly Media, Sebastopol, 1-59.
[53] Github. (2020) Github. https://github.com/sinanuguz/airquality. (Accessed 05.07.2020).

Received: **26.03.2021**
Accepted: **18.04.2021**

CORRESPONDING AUTHOR

Okan Oral
Department of Mechatronics Engineering,
Akdeniz University,
Antalya – Turkey

e-mail: okan@akdeniz.edu.tr