Water Air Soil Pollut (2024) 235:578
https://doi.org/10.1007/s11270-024-07390-0

Air Quality Prediction System Using Machine Learning Models

Pooja Chaturvedi

Received: 20 November 2023 / Accepted: 26 July 2024 / Published online: 5 August 2024

Abstract The air quality index has a severe effect on the determination of health conditions of a city. The prediction of air quality index can aid in determining the optimum route in case of traffic and it can also aid in determining the pollutants which have severe impact on human health conditions. The paper presents an air quality prediction system using various machine learning based models. The air quality index is determined by measuring the different gases present in the atmosphere. In this paper we have considered seven such parameters as concentration levels of Particulate Matter 2.5 ($PM_{2.5}$), Particulate Matter 10 (PM_{10}), Carbon Mono oxide (CO), Nitrogen Dioxide (NO_2), Ammonia (NH_3), Sulphur Dioxide (SO_2) and Ozone (O_3) levels for the duration between the year January 2019 to October 2023 for a crowded area of Varanasi city. The various pre processing techniques have been used in the dataset for the implementation of machine learning models. The performance of the models have been compared for the prediction of the air quality. The results show that the Random Forest and Decision Tree based model achieves the maximum accuracy of approximately 100% as compared to 98%, 95% and 93% and 79% for the SVM, Multi layer Perceptron network, KNN classification and Linear Regression.

P. Chaturvedi (✉)
Institute of Technology, Nirma University, Ahmedabad, Gujarat, India
e-mail: pooja.chaturvedi@nirmauni.ac.in

Keywords Air Quality Index · Air Quality Index Category · Machine Learning Models · Air Quality Prediction System

1 Introduction

The industrial activities, population density, thermal power units, agricultural methods, automotive industries and transportation activities effect the air quality index in a region significantly (Ravindra, 2019) (Ravindra et al., 2020). In addition to the adverse effect on the atmospheric conditions, it negatively impacts the human health in terms of lung infection, breathing problems, premature death, heart failure, lung cancer and skin rashes (Manisalidis et al., 2020). The air quality index is greatly affected by the emission of greenhouse gases, meteorological factors and the gaseous pollutants such as Particulate Matter, Carbon Mono oxide, Sulphur dioxide, Ammonia, Nitrogen oxide and Ozone (Bao & Zhang, 2020)-(Balakrishnan et al., 2019). These pollutants cause other deteriorating effect in atmosphere such as global warming, climate change, decreased visibility, acid rain and the development of smog and aerosols (Malhi et al., 2021).

Out of the 15 most polluted cities in Central and South Asia in 2022, 11 are in the India. Air pollution has massive impact on the human health in India. It is second biggest risk factor and subsequently it increases the cost of air pollution by 150

billion dollars annually (https://www.greenpeace.org/static/planet4-india-stateless/2023/03/2fe33d7a-2022-world-air-quality-report.pdf. n.d.). According to the WHO air quality report published in the year 2014, Varanasi has been ranked as the third most polluted city in the world. The major contributing factor in degrading the air quality in Varanasi city is due to the vehicular exhaust followed by construction and road dust, industries and thermal power plant. This shows the severity of the air pollution in the Indian cities and its impact on human health in the coming years (https://www.ndtv.com/india-news/varanasis-air-quality-deteriorating-delhi-victim-of-negligence-report-2020928. n.d.) (https://climatetrends.in/wp-content/uploads/2019/11/Political-Leaders-Position-Action-on-Air-Quality-Indian-MPs-report-card-2014-19-April-2019-.pdf. n.d.).

The machine learning model can aid in effectively predicting and forecasting the air quality index as compared to statistical and rigid models. The availability of the historical data provides a strong aid in precisely and accuracy predicting the AQI index. The category of Air Quality and its impact on health condition for the different Air Quality Index is as shown in the Table 1.

The major contribution of the proposed work is as follows:

a. The data regarding the air quality is collected from government websites such as Central Pollution Control Board (CPCB) (Board, 2019).
b. The characteristics of the dataset is analyzed, processed and represented in a manner, so it can be used for the development of effective machine learning models for the prediction.
c. The next step is to employ the different methods such as outlier detection and removal, filling the missing value with mean value of the parameter and scaling the values of the dataset, to pre process the data.
d. The overall strategy used in the proposed approach is shown in the Fig. 1.
e. The different machine learning based models such as Linear Regression, Decision Tree, Random Forest along with hyper parameter tuning, K Nearest Neighbor and SVM with different kernels such as Linear, Polynomial and Radial Basis Forward are used for the prediction and classification of the air quality index of the Varanasi city.
f. The performance of the different approaches is evaluated and compared on the basis of different performance metrics such as Mean Absolute Error, Root Mean Square Error, Precision, Recall, Accuracy and Confusion Matrix. The comparison result show that the Random Forest achieves the highest accuracy for the considered dataset.

The organization of the paper is as follows: related work discussion in Section 2, the overview of the methodology in Section 3, results and discussion in Section 4 and Section 5 concludes the paper with conclusion and future scope.

2 Related Work

The air pollution monitoring is a crucial task in meeting the objective of mapping of air pollution levels to different cities. To predict the air quality index as a time series model two methods have proposed in Tuan-Vinh (2021): one as a hybrid model combining the ARIMA model with PCR and other as a hybrid model combining the ARIMA model with gene expression programming. In this the correlation

Table 1 Category of AQI and its health impact

AQI Range	Category	Health Effects
0–50	Good	Air quality is good
51–100	Satisfactory	Air quality is low and has ill effects on health
101–200	Moderate	Air quality is moderately polluted and can cause difficulty in breathing to the people having lung disease, asthma or heart diseases
201–300	Poor	The symptoms may be mildly aggravated in the high risk patients having heart or lung diseases
301–400	Very Poor	There may be significant aggravation of symptoms in persons having heart or lung diseases
401–500	Severe	Severe aggravation of symptoms and may be dangerous for health

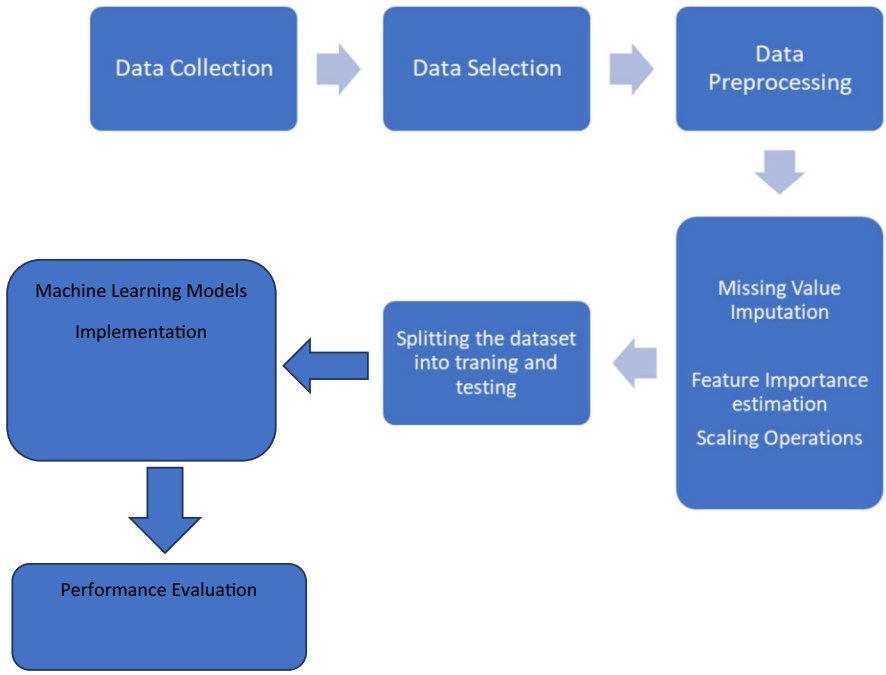


Fig. 1 Flow chart of the proposed method

between the urban nature and urban traffic were used to determine the $PM_{2.5}$ levels. The sensor and IOT based system were developed to determine the AQI values. The data collected was used to predict the air quality level for the next day using the Linear Regression Model. The performance of the proposed model is evaluated in terms of Mean Absolute error and Root Mean Square Error (Kumar et al., 2020). The five regression models as: partial least square, principal component, partial component with one out, cross validation and multiple regression were used for the prediction of air quality. The performance of the models is evaluated for the data collected from different cities (Londhe May, 2021). The daily temperature was also considered as a contributing in determination of air quality index. The deep learning model such as LSTM and other were used for the prediction of the air quality in the Dhaka city (Chowdhury et al., 2020). The improved neural network using nonlinear auto regressive neural network is developed for the precise prediction of air quality index using the weather stations and atmospheric monitoring (Zhou et al., 2018). The different pollutant levels were considered for the prediction of the air quality. The different regression models such as random forest regression, stochastic gradient descent regression and linear regression were implemented and evaluated against the various performance metrics such as MAE, MSE and R score (Srivastava et al., 2018). SVM and ANN based models were used for the prediction of the air quality for the Delhi region in Raturi and Prasad (2018). The survey was conducted to understand the effect of different pollutants on the air quality. The ANN proved to be best model for the prediction task (Mahalingam et al., 2019). The supervised learning models, SVM and ANN were used for the prediction of the air quality in the Delhi region (Sethi & Mittal, 2019). The supervised machine learning models such as Decision Tree, SVR and stacking ensemble methods proved to perform best for the air quality prediction. The emission of different hazardous gasses and their impact on the environment and human health has been investigated in the works presented in the (Wen et al., (2024), Zhang et al., (2021), Luo et al., (2024), IoT-Based Air Quality Monitoring in Hair Salons, (2023), Samuel et al., (2023), Yin et al., (2023), Shang et al., (2023), Liu et al., (2023)-(Blessy et al., 2023).

Based on the literature survey, we conclude that the open access datasets are not efficient in predicting the air quality due to the large number of incorrect and missing values. So there is need of the incorporation of machine learning models which can enhance the efficiency of air quality prediction.

3 Proposed Methodology

The motivation to select the Varanasi city is due to the fact that the air quality index has reached to 490 according to the report published in the reference (https://climatetrends.in/wp-content/uploads/2019/11/Political-Leaders-Position-Action-on-Air-Quality-Indian-MPs-report-card-2014-19-April-2019-.pdf. n.d.). The report concludes the major construction work going on in the city is the root cause of degraded air quality over the years. To address the alarming degradation in the air quality index of Varanasi city several AQI stations have been implemented at prominent places with the objective of the continuous monitoring and control of AQI. The list of AQI station in Varanasi city are as shown in the Table 2.

3.1 System Design

The proposed system consists of six stages as: i. Dataset collection, ii. Dataset selection iii. Data pre processing, iv. Splitting the dataset into training and testing dataset, v. implementation of different machine learning models and vi. Performance evaluation and comparison of different models. The flow chart of the proposed approach is as shown in the Fig. 1.

The detailed description of the task performed in each step is as follows:

1. The first step of the proposed system is the collection of data from the CPCB website. To collect the data, we have considered the station IESD Banaras Hindu University, Varanasi UPPCB and the parameter option is considered as Select All. The format of the data collection is selected as Tabular, the criteria is considered as 1 Hr. and the duration of the data collection is considered as 1st January 2019 to 22nd October, 2023. The result of data collection is stored in the.csv file which consists of 22 parameters.
2. The second step of the proposed approach is the data selection. It has been established that the AQI depends majorly on 7 parameters such as $PM_{2.5}$, PM_{10}, NO_2, CO, SO_2, NH_3, Ozone. So out of the 22 parameters we have considered 7 parameters and removed the remaining columns from the dataset. After pruning the dataset, the size of the dataset is considered as (2543,9).

Table 2 List of AQI stations in Varanasi city

S. No	Station Name
1	Ardhali Bazar
2	Bhelupur
3	IESD Banaras Hindu University
4	Maldahiya
5	Nirala Nagar

The next step of the proposed model consists of the data understanding. For better analysis of the dataset, the correlation matrix of the different parameters is determined which establishes the relationship of the different parameters on the AQI. The result of the data visualization of correlation matrix is as shown in the Fig. 2.

After the correlation analysis, the pair plot is plotted to visualize the distribution of air quality index in different classes. The pair plot for each of the parameters is as shown in the Fig. 3.

3. After the data visualization step, the various data preprocessing methods are implemented. The box plot for all the considered parameters is plotted to determine the outliers in the dataset. The outlier are the values which lie beyond the normal range of the values for the particular parameter. The visualization of the box plot is as shown in the Fig. 4.

From the box plot it is clear that the parameter $PM_{2.5}$ has highest number of outlier values as 30. The Inter quartile range is used to determine the maximum and minimum value for the parameter. The inter quartile range is determined as the difference of the 75 quartile and 25 quartile. The values which are lower than minimum value and which are greater than the maximum value is considered as the outlier. To remove the outliers the outlier values are replaced with NA values. Then these values are filled with the scaled value using standard scalar method. The scaled value for a parameter is calculated using the equation 1.

Fig. 2 Correlation matrix for the considered parameters

	Pmone	Pmtwo	NO2	NH3	CO	SO2	Ozone	aqi	AQI_Range
Pmone	1.000000	0.846583	0.231563	0.457906	0.713730	-0.200817	-0.193565	0.101634	0.278054
Pmtwo	0.846583	1.000000	0.320992	0.439554	0.639927	-0.041122	-0.093215	-0.028990	0.205092
NO2	0.231563	0.320992	1.000000	0.654432	0.402425	0.269758	-0.107814	-0.024491	0.016576
NH3	0.457906	0.439554	0.654432	1.000000	0.500126	-0.063577	-0.181787	-0.015546	0.073089
CO	0.713730	0.639927	0.402425	0.500126	1.000000	-0.222376	-0.324432	0.024103	0.166674
SO2	-0.200817	-0.041122	0.269758	-0.063577	-0.222376	1.000000	0.540306	0.286223	0.274734
Ozone	-0.193565	-0.093215	-0.107814	-0.181787	-0.324432	0.540306	1.000000	0.560152	0.570048
aqi	0.101634	-0.028990	-0.024491	-0.015546	0.024103	0.286223	0.560152	1.000000	0.837866
AQI_Range	0.278054	0.205092	0.016576	0.073089	0.166674	0.274734	0.570048	0.837866	1.000000

$$y = \frac{(x - \mu)}{s} \tag{1}$$

where y is scaled value of the parameter x, μ is the mean of the parameter and s is the standard deviation of the parameter.

Similarly, the missing/NULL values are identified and filled with the mean value for that parameter.

In this step, the irrelevant columns as From Date and To Date are also removed as it has no contribution in the air quality index determination. Hence after the pre processing the dataset size is (2543,7).

After pre processing the heat map is plotted to determine if there is any missing value in the dataset as shown in the Fig. 5.

It can be inferred from the heat map that there are no missing values in the dataset.

After pre processing, the next step is calculating the air quality index and air quality index category using the standard range allowed for each parameter. The two functions are defined for this functionality. The index value of each parameter is determined using the equation 2.

$$Ip\frac{I_{hi} - I_{lo}}{Bphi - Bplo}(Cp - Bplo)^{*}I_{lo} \tag{2}$$

where Ip is the index for Pollutant P, Cp is the rounded concentration of Pollutant P, $Bphi$ is the break point which is greater than or equal to Cp, $Bplo$ is the break point which is lesser than or equal to Cp, I_{hi} is the AQI corresponding to $Bphi$ and I_{lo} is the AQI corresponding to $Bplo$.

The air quality index and air quality index category are calculated using the parameter index value as shown in the Table 3.

After appending AQI category the size of the dataset is (2543,8).

4. In the next step, the considered dataset is divided into training and testing dataset. The test size is considered as 30%, hence the size of testing dataset is (751,8) and the size of training dataset is (1752,8).
5. The different machine learning models is implemented for the prediction of the air quality.
6. The performance evaluation of machine learning model is done on the basis of different performance metrics such as accuracy, confusion matrix, R2 score, precision, recall, Mean Absolute Error and Mean Square Error. The comparative evaluation of the models is also done.

3.2 Implementation Details

To predict the air quality for the considered, machine learning models as Linear Regression, Decision Tree, Random Forest, Neural Network, K Nearest Neighbor and SVM with different kernel functions is utilized. All the models have been implemented in python programming language on a Windows based operating system. In this section, a brief description of the various machine learning models used to forecast the air quality is provided.

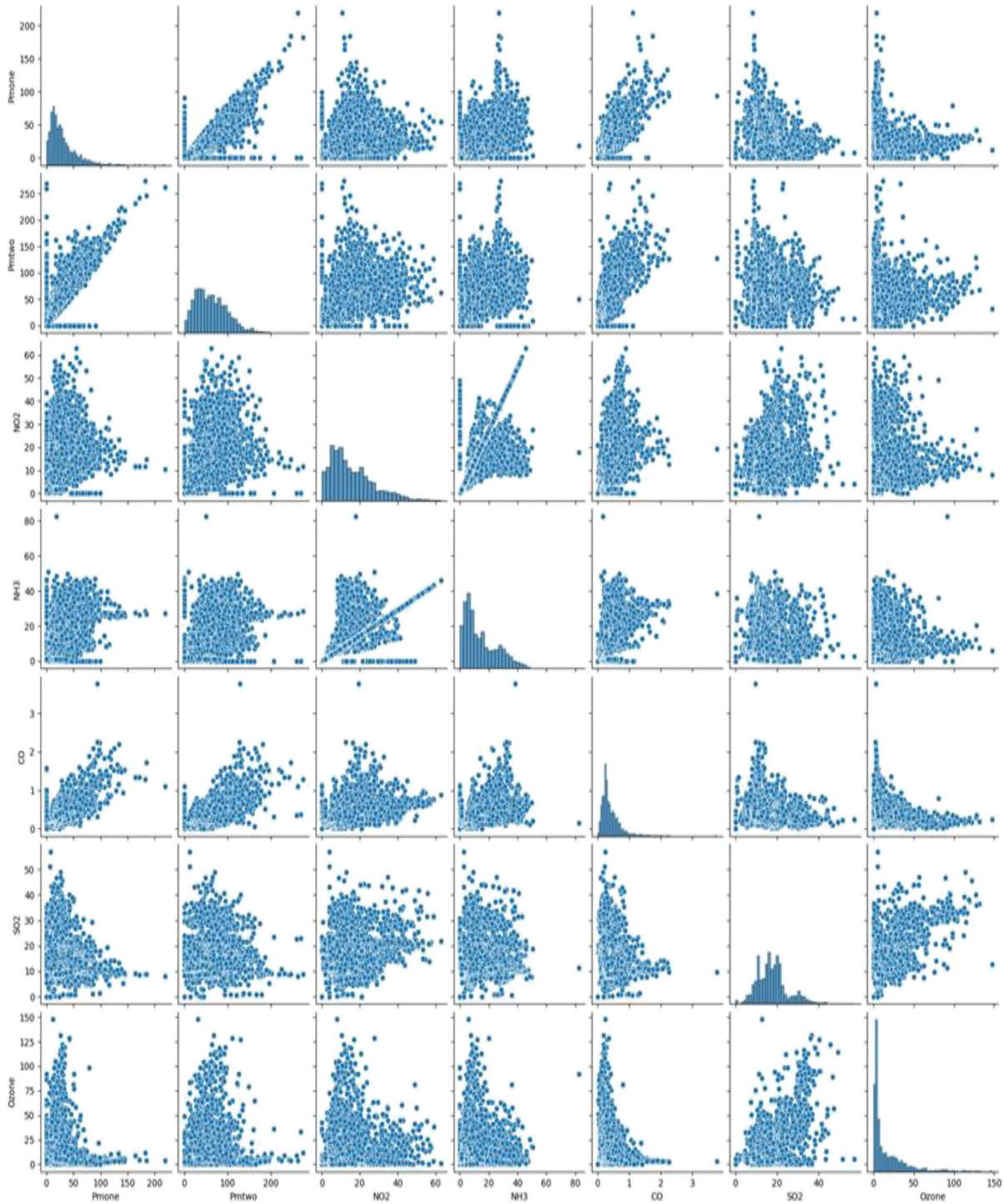


Fig. 3 Pair plot for the different parameters

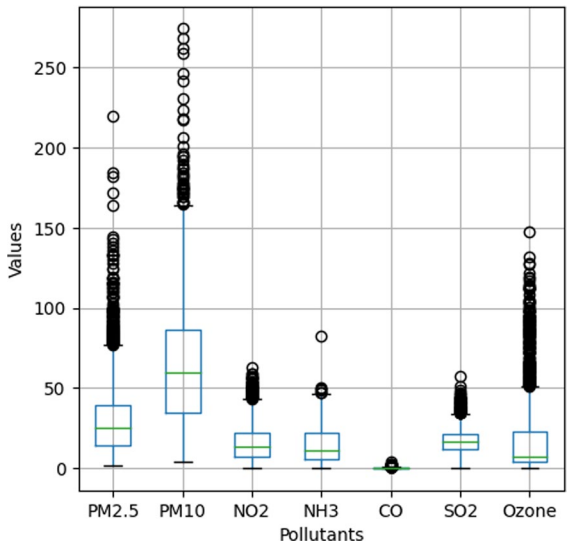


Fig. 4 Box plot

3.3 Machine Learning Models

Extra tree regressor stands for extremely randomized trees. It is an ensemble based supervised machine learning method that aggregates the results from the de-correlated decision trees to improve the efficiency and performance (Geurts et al., 2006). Extra tree regression models are useful, if the accuracy is more important as compared to the construction of the generalized model. It is different from the random forest-based method in that it does not utilize the concept of bootstrap, instead it works using the randomized split. The extra tree regressor method can also be used for the determination of importance of features in the dataset.

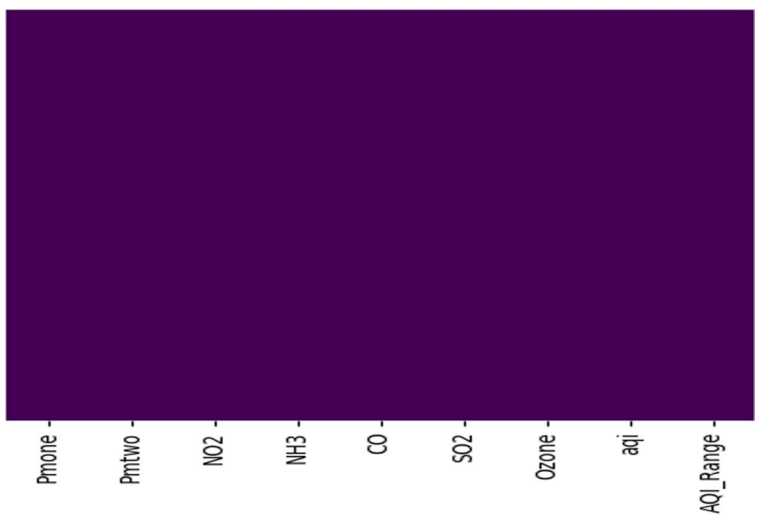


Fig. 5 Heat map of the different parameters

Table 3 AQI calculation using the different parameter concentration levels

AQI Category	AQI	Concentration Ranges						
		PM_{10}	$PM_{2.5}$	NO_2	O_3	CO	SO_2	NH_3
Good	0–50	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200
Satisfactory	51–100	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400
Moderately Polluted	101–200	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800
Poor	201–300	251–350	91–120	181–280	169–208	10–17	381–800	801–1200
Very Poor	301–400	351–430	121–250	281–400	209–748	17–34	801–1600	1201–1800
Severe	401–300	430 +	250 +	400 +	748 +	34 +	1600 +	1800 +

3.3.1 Linear Regression Model

The linear regression model is a supervised machine learning algorithm, which tries to find the best fitting line for the exploratory variables using the relationship with the dependent variables (Rogers & Girolami, 2016). The algorithm tries to find a line which is best fitted for the data under consideration using the concept of residuals. The residual value represents the distance between the exploratory variable and the actual value. The process of finding the best fitting line is done in several iterations. The Eq. 3 is the equation of the regression line:

$$a = \gamma_0 + \gamma_1\beta_1 + \gamma_2\beta_2 + \quad (3)$$

where α is the dependent variable and β is the independent variable. γ represents the coefficient of the regression model.

3.3.2 Decision Tree Model

Decision tree is a supervised prediction and classification method based on Boolean conditions. The decision tree represents a collection of nodes and links (Quinlan, 2014). The nodes in the tree are different parameters and link is the connection between the parameters. The best attribute is determined as a root node and then depending on the conditions the course of actions is determined. Since a single attribute cannot determine the class labels accurately so different parameters such as Information Gain, Gini Index are considered. These parameters are termed as impurity. The Eq. 4 is used to calculate the Gini impurity:

$$Gini(Dataset) = 1 - \sum_{i=1}^{c} P_i \quad (4)$$

where Gini (Dataset) represents the Gini impurity for the dataset, c represents the number of classes and P_i represents the probability of the instance belonging to the class i.

3.3.3 Random Forest

The random forest is a supervised machine learning algorithm which predicts the class label by considering the majority vote for a particular class (Liaw & Wiener, 2002). The class label is determined using the collection of forest of decision trees. This algorithm was designed for removing the drawback of the decision tree algorithm of overfitting the dataset. The random forest uses bagging for the aggregation of the different trees.

3.3.4 Neural Network Model

The neural network-based machine learning model consists of number of inputs, hidden and output layers. The input values are provided from input layer, the hidden layers are responsible for the processing of the input data using the activation function. The output of each layer is determined using the activation function. The neural network has been found to perform well for a specific task, as it provides the facility of the parallel processing.

3.3.5 KNN Model

KNN stands for K Nearest Neighbor. It is used for prediction tasks and it works by comparing the similarity value of a new data point by the values in the dataset. The different similarity metrics are used for the determination of the class label such as Manhattan distance, Minkowski distance and Cosine Similarity. The k nearest neighbors are evaluated and on the basis of the majority votes, the class of a new data point is determined. By varying the value of k, the model can be fine-tuned to achieve the optimal accuracy.

3.3.6 Support Vector Machine Model

The support vector machine is a supervised machine learning algorithm which tries to determine a hyperplane in a N dimensional space that distinctly classifies a data point. To determine the unique hyperplane the SVM algorithm tries to maximize the margin i. e. the distance between the data points of the two class should be maximum. The SVM algorithm can be applied for the linearly separable as well as non-linearly separable classes. For the non-linearly classes, the maximal margin hyperplane is determined using the kernel function. The objective of kernel function is to transform the input dataset into higher dimensional data points. There are three kinds of kernel functions possible in context to SVM i.e., linear, polynomial and radial basis kernel function (Dun et al., 2020).

3.3.7 Xgboost

Xgboost stands for Extreme Gradient Boosting. It is a supervised machine learning algorithm used for classification and regression tasks. This method is suitable for the large and complex datasets. The model starts with the predicting the initial value of the independent variable. The residual is computed as the difference of the predicted and actual value. The xgboost method works on decreasing the residual value to the minimum and it continues until a terminating condition is arrived. The terminating condition can be either the maximum number of iterations or the threshold value for the residual value (Tianqi & Carlos, 2016).

3.4 Performance Metrics

The performance of the considered machine learning model is evaluated in terms of following performance metrics.

3.4.1 Mean Absolute Error

Mean absolute error represents the error in the pairwise observations (Sammut, 2010). The larger MAE value indicates the larger error in the model. The formula to calculate the MAE is as shown in the Eq. 5:

$$MAE = \frac{\sum_{I=1}^{n}(y\prime - y)}{n} \tag{5}$$

where, n represents the number of observations, y' is the actual value and y is the predicted value.

3.4.2 Mean Square Error

The performance metric used to measure how well the regression line fits to the data values is known as MSE (Nevitt & Hancock, 2000). It can be considered as the mean deviation of the residuals. The Eq. 6 is used to calculate the MSE is as follows:

$$MSE = \frac{\sum_{t=1}^{n}(y\prime - y)^2}{T} \tag{6}$$

Table 4 Example of a confusion matrix

Actual/Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

3.4.3 Confusion Matrix

It is a graphical representation used to determine the performance of any machine learning model as shown in the Table 4.

It consists of four terms which are defined as follow:

a. *True Positive (TP)*- This metric represents the number of positive data points classified correctly.
b. *True Negative (TN)*- This metric represents the number of negative data points classified correctly.
c. *False Positive (FP)*- This metric represents the number of positive data points classified incorrectly.
d. *False Negative (FN)*- This metric represents the number of negative data points classified incorrectly.

Accuracy

Accuracy represents the ratio of number of instances correctly classified to the total number of observations. The formula to calculate the accuracy of a machine learning model is as follows:

$$Accuracy = TP/TP + NP \tag{7}$$

where TP represents the number of positive instances correctly classified. $TP+NP$ represents the total number of observations.

Precision

Precision represents the number of instances which are positively labeled and are correctly classified.

$$Precision = TP/TP + FP \tag{8}$$

Recall

Recall is the performance metrics which represent the efficiency of the model in predicting the positive outcomes.

$$Recall = TP/TP + FN \quad (9)$$

F1 Score

The harmonic mean of Precision and Recall is defined as the F1 score of the model. It is defined as:

$$F1Score = \left(2^{*}Precision^{*}Recall\right)/(Precesion + Recall) \quad (10)$$

4 Results and Discussion

The implementation results are discussed in two subsections as: i. performance of the difference machine learning models and ii. Comparison of considered models.

4.1 Performance of the Different Machine Learning Models

4.1.1 Extratree Feature Importance

Extra tree regressor is used to determine the importance of different features on the air quality index. The results of the feature importance are as shown in the Fig. 6. It can be inferred from the figure that the Ozone has the highest impact on air quality index.

4.1.2 Linear Regression

The result of linear regression model for the prediction of air quality is as shown in the Table 5.

Table 5 Performance evaluation of Linear Regression Model

Performance Metrics	Value
Training dataset score	0.74
Testing dataset score	0.77
Mean score	0.72
R2 Score	0.77
MAE	0.16
MSE	0.04

The distribution of training dataset and testing dataset prediction across different AQI ranges is as shown in the Fig. 7 and Fig. 8 respectively.

The scatter plot in Fig. 9 shows the distribution of the test data set and prediction values across the different AQI ranges.

4.1.3 Random Forest

The random forest is applied over the considered dataset. The accuracy of the model is best among all the models considered as 99%. To further improve the performance of the proposed model, randomized search cross validation is applied for the hyper parameter tuning. The depth values are considered randomly in the range of 1 to 20. The number of estimators is used in the range of 50 to 500. The 5 folds of cross validation is considered and the number of iterations in each fold is considered as 5. The model is then trained using the best parameters. After the cross validation, the best hyper parameters are found

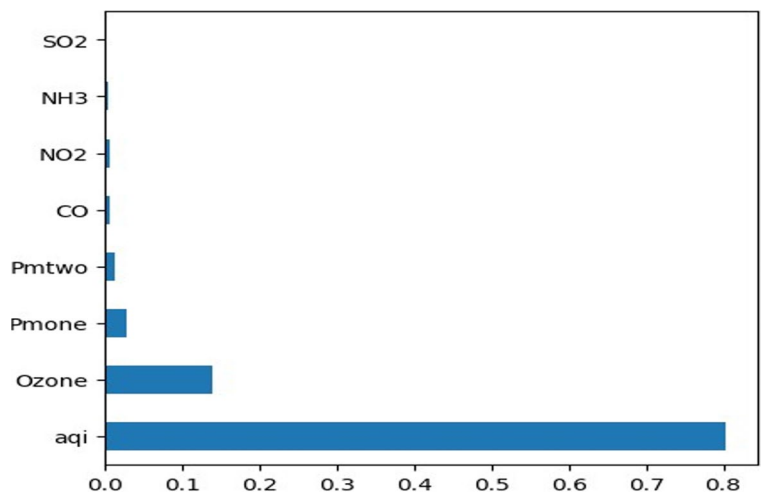


Fig. 6 Feature importance plot using Extra tree Regressor

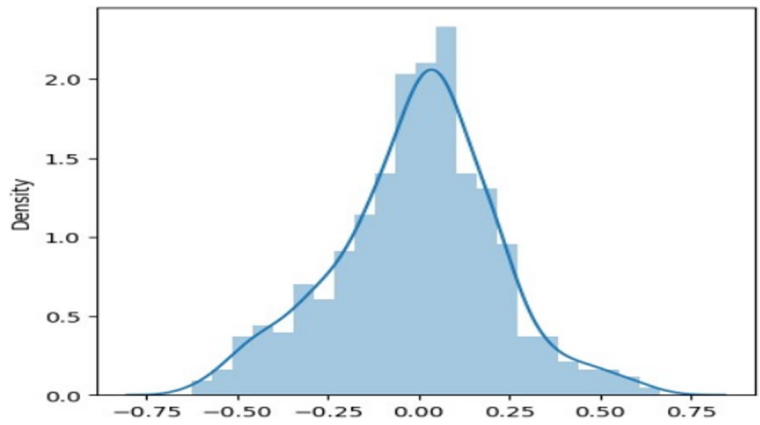


Fig. 7 Distribution plot for testing dataset

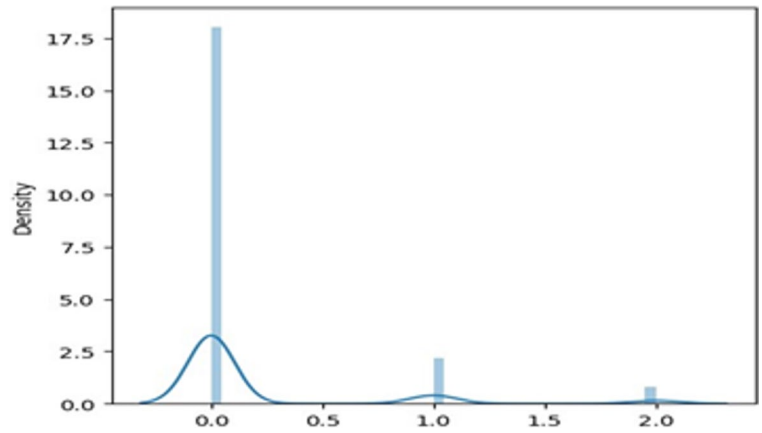


Fig. 8 Distribution plot for the training dataset

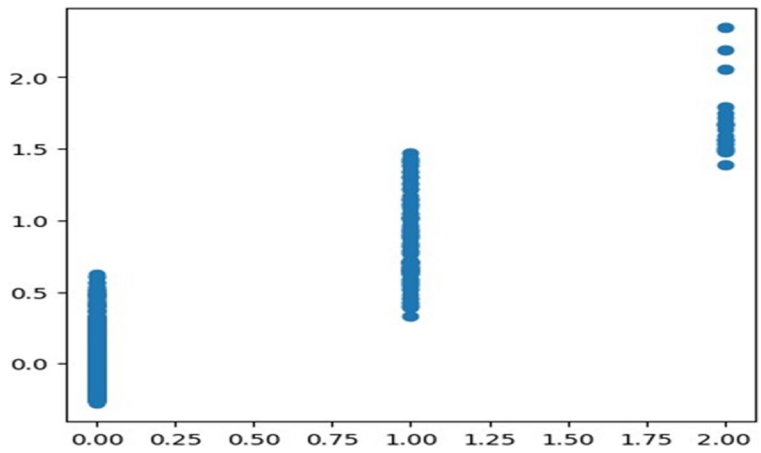


Fig. 9 Scatter plot for the testing dataset and prediction

as max depth is 14 and number of estimators is 286. The confusion matrix for the best model is as shown in the Fig. 10.

The feature importance is further determined using the best random forest model is as shown in the

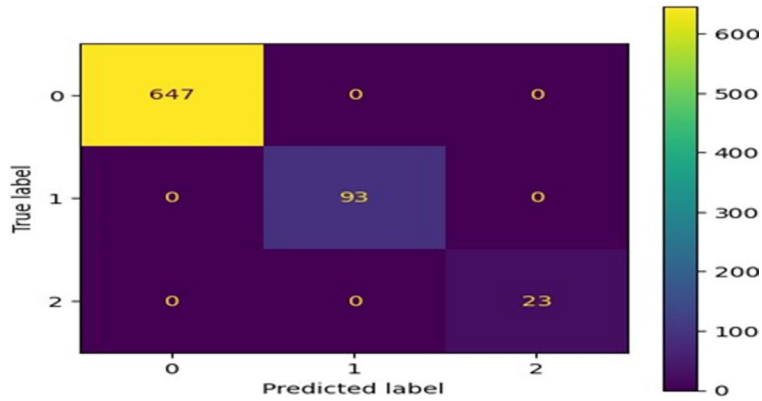


Fig. 10 Confusion matrix for the Best RBF model obtained after hyper parameter tuning

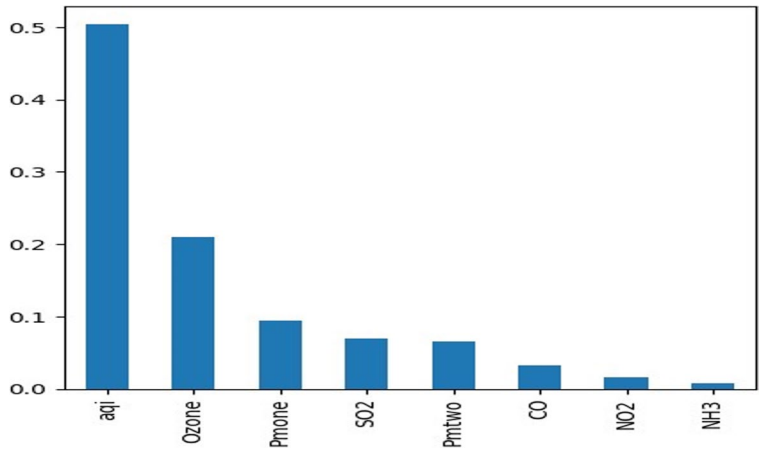


Fig. 11 Feature importance plot using best Random Forest Model

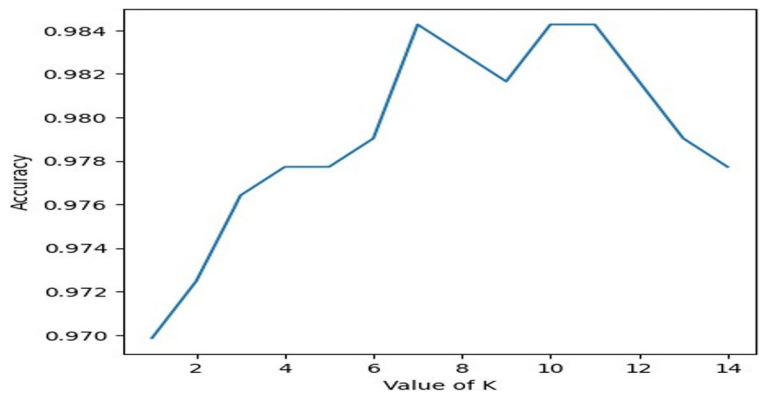


Fig. 12 Effect of varying the value of k on KNN accuracy

Fig. 13 Performance of KNN Model

```
Confusion Matrix:
[[645   2   0]
 [ 11  79   3]
 [  0   1  22]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       647
           1       0.96      0.85      0.90        93
           2       0.88      0.96      0.92        23

    accuracy                           0.98       763
   macro avg       0.94      0.93      0.94       763
weighted avg       0.98      0.98      0.98       763
```

Fig. 14 Confusion matrix for Neural Network Model

```
Confusion Matrix
[[642   5   0]
 [  6  85   2]
 [  0   0  23]]
```

Fig. 11.The results show that the ozone is the most prominent feature for the air quality prediction.

4.1.4 KNN Model

The KNN based model is used for the prediction of air quality. The accuracy of the model is found as 98% for the value of k as 7. The effect of the accuracy on varying the value of k is as shown in the Fig. 12.

The other performance metrics related to the KNN model is as shown in the Fig. 13.

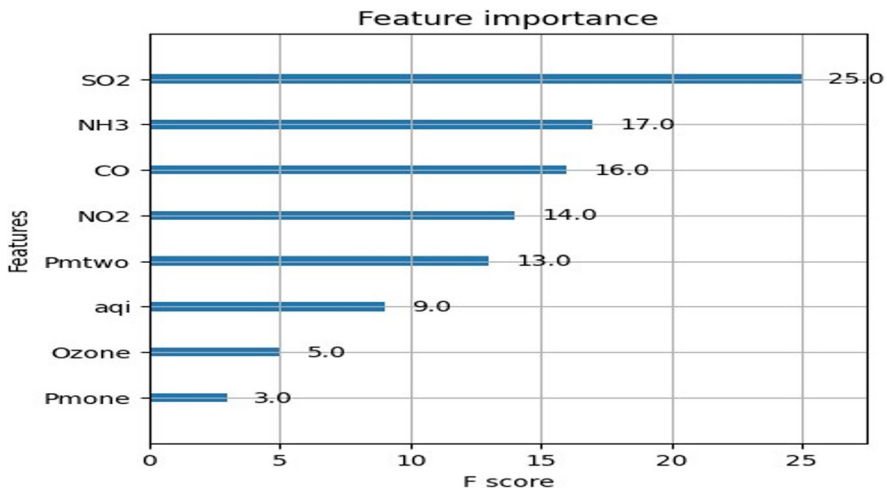
4.1.5 Confusion Matrix MLP

In the section, we have implemented the neural network model for the prediction of the air quality. The accuracy of the model is found as 98.5%. The confusion matrix for the model is as shown in the Figs. 14 and 15.

Table 6 Performance of different kernel functions in SVM

Model	Accuracy	Precision	Recall	F1 Score
SVM Linear	98.55	98.54	98.55	98.54
SVM Polynomial	98.55	98.56	98.55	98.52
SVM RBF	98.29	98.28	98.29	98.26

Fig. 15 Feature importance plot using Xgboost Model



	train-rmse-mean	train-rmse-std	test-rmse-mean	test-rmse-std
0	0.454776	0.006017	0.455263	0.011878
1	0.413543	0.005301	0.414276	0.010763
2	0.406203	0.005278	0.407915	0.011698
3	0.396672	0.004859	0.399134	0.011457
4	0.361322	0.004271	0.364057	0.010483

Fig. 16 Performance of the Xgboost Model for Training and Testing dataset

4.1.6 Support Vector Machine Model

Support vector machine-based model is implemented for the classification of the considered dataset in the different air quality ranges. The different kernel functions are used for the evaluation of the performance of the model. The performance results in terms of precision, recall and F1 score for the different kernel functions are as shown in the Table 6.

The results show that the polynomial and linear SVM models have the highest accuracy as compared to RBF SVM.

4.1.7 Xgboost Feature Importance

In this section, we have described the feature importance for the different parameters using the Xgboost based method. The result show that SO_2 and NH_3 has highest impact on the prediction of air quality.

The results of mean score for the training and testing dataset using the Xgboost based model is as shown in the Fig. 16.

It can be inferred from the table that the error value is significantly low for both the training and testing dataset.

4.2 Comparison of considered models

In this section we have compared the performance of the different models in terms of their accuracy for the prediction of air quality as shown in the Fig. 17. The results show that the random forest and decision tree have the highest accuracy as 100% and linear regression has the lowest accuracy as 79%.

5 Conclusion and Future Scope

The paper presented a machine learning based model for the effective prediction and classification of air quality index for the Varanasi city collected from CPCB website. The various data pre processing techniques are employed to improve the data representation such as outlier detection, missing value

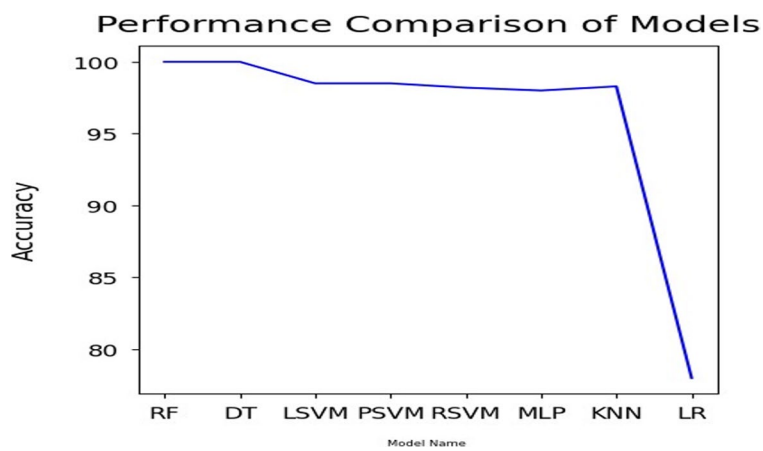


Fig. 17 Performance comparison of Different Models

imputation and scaling the data. In the proposed ExtraTree Regressor method is employed to determine the importance of the different features. The results show that the concentration level of Ozone has significant impact on air quality index. Six different machine learning models such as Linear Regression, Decision Tree, Random Forest, K Nearest Neighbor, Neural Network and SVM with different kernels along with the hyper parameter tuning has been implemented for the determination of the most efficient machine learning model. The results show that the random forest and decision tree models have highest accuracy in the prediction of the air quality whereas SVM with RBF kernel is most efficient for the classification task on the basis of several performance metrics such as accuracy, precision and recall.

The proposed models can be implemented for the city or state level air quality prediction. The model can be also be investigated for the real time air quality monitoring and prediction.

Funding There is no funding received for the work.

Data Availability The data will be made available as per the suitable request.

Declarations

Ethical Approval Yes.

Consent to Participate Yes.

Consent to Publish Yes.

Competing Interests The author declares that there is no competing interests.

References

Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., Stanaway, J. D., Beig, G., Joshi, T. K., Aggarwal, A. N., Sabde, Y., Sadhu, H., Frostad, J., Causey, K., Godwin, W., Shukla, D. K., Kumar, G. A., Varghese, C. M., Muraleedharan, P., … Dandona, L. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017. *Lancet Planet, Health, 3*, e26–e39.

Bao, R., & Zhang, A. (2020). Does lockdown reduce air pollution? Evidence from 44 cities in northern China. *Science of the Total Environment, 731*, 13905.

Blessy, A., John Paul, J., Gautam, S., et al. (2023). IoT-Based Air Quality Monitoring in Hair Salons: Screening of Hazardous Air Pollutants Based on Personal Exposure and Health Risk Assessment. *Water, Air, and Soil Pollution, 234*, 336. https://doi.org/10.1007/s11270-023-06350-4

Board, C.P.C., 2019. National air quality monitoring programme (NAMP). URL: https://cpcb.nic.in. Accessed 30 Mar 2023

Chowdhury, A. S., Uddin, M. S., Tanjim, M. R., Noor, F., & Rahman, R. M. (2020). Application of data mining techniques on air pollution of Dhaka city. In 2020 IEEE 10th International Conference on Intelligent Systems (IS). IEEE, 562–567

Dun, M., Xu, Z., Chen, Y., & Wu, L. (2020). Short-term air quality prediction based on fractional grey linear regression and support vector machine. *Mathematical Problems in Engineering, 2020*, 1–13. https://doi.org/10.1155/2020/8914501

Geurts, P., Ernst, D., & Wehenkel, L. (2006). *Extremely Randomized Trees. Machine Learning, 63*, 3–42.

https://climatetrends.in/wp-content/uploads/2019/11/Political-Leaders-Position-Action-on-Air-Quality-Indian-MPs-report-card-2014-19-April-2019-.pdf.

https://www.greenpeace.org/static/planet4-india-stateless/2023/03/2fe33d7a-2022-world-air-quality-report.pdf

https://www.ndtv.com/india-news/varanasis-air-quality-deteriorating-delhi-victim-of-negligence-report-2020928.

IoT-Based Air Quality Monitoring in Hair Salons. (2023). Screening of Hazardous Air Pollutants Based on Personal Exposure and Health Risk Assessment. *Water Air and Soil Pollution, 234*, 336. https://doi.org/10.1007/s11270-023-06350-4

Kumar, R., Kumar, P., & Kumar, Y. (2020). Time series data prediction using IoT and machine learning technique. *Procedia Computer Science, 167*(2020), 373–381.

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R News, 2*(3), 18–22.

Liu, Z., Feng, J., & Uden, L. (2023). Technology opportunity analysis using hierarchical semantic networks and dual link prediction. *Technovation, 128*, 102872. https://doi.org/10.1016/j.technovation.2023.102872

Londhe, M. (2021). Data mining and machine learning approach for air quality index prediction. *International Journal of Engineering and Applied Physics, 1*(2), 136–153.

Luo, J., Zhuo, W., Liu, S., & Xu, B. (2024). The Optimization of Carbon Emission Prediction in Low Carbon Energy Economy Under Big Data. *IEEE Access, 12*, 14690–14702. https://doi.org/10.1109/ACCESS.2024.3351468

Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019). A machine learning model for air quality prediction for smart cities. In 2019 International conference on wireless communications signal processing and networking (WiSPNET). IEEE, 452–457

Malhi, G. S., Kaur, M., & Kaushik, P. (2021). Impact of climate change on agriculture and its mitigation strategies: A review, 2021 *Sustain. Times, 13*, 131.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts

of air pollution: A review. *Frontiers in Public Health, 8*, 505570.

Nevitt, J., & Hancock, G. R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *The Journal of Experimental Education, 68*(3), 251–268. https://doi.org/10.1080/00220970009600095

Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

Raturi, R., & Prasad, J. R. (2018). Recognition of future air quality index using artifcial neural network. *International Research Journal of Engineering and Technology (IRJET), 5*, 2395–56.

Ravindra, K. (2019). Emission of black carbon from rural households kitchens and assessment of lifetime excess cancer risk in villages of North India. *Environment International, 122*, 201–212.

Ravindra, K., Singh, T., Pandey, V., & Mor, S. (2020). Air pollution trend in Chandigarh city situated in Indo-Gangetic Plains: Understanding seasonality and impact of mitigation strategies. *Science of the Total Environment, 729*, 138717.

Rogers, S., & Girolami, M. (2016). A first course in machine learning. Chapman and Hall/CRC.

Sammut, Claude Webb, G.I., 2010. Mean squared error. In: Sammut, Claude and Webb, G.I. (Ed.), *Encyclopedia of Machine Learning*. Springer US, Boston, MA, p. 653. https://doi.org/10.1007/978-0-387-30164-8_528.

Samuel, C., et al. (2023). Exposure and health: A progress update by evaluation and scientometric analysis. *Stoch Environ Res Risk Assess, 37*, 453–465. https://doi.org/10.1007/s00477-022-02313-z

Sethi, J. K., & Mittal, M. (2019). Ambient air quality estimation using supervised learning techniques. *EAI Endorsed Transactions on Scalable Information Systems*, *6* (22), e8–e8

Shang, K., Xu, L., Liu, X., Yin, Z., Liu, Z., Li, X.,... Zheng, W. (2023). Study of Urban Heat Island Effect in Hangzhou Metropolitan Area Based on SW-TES Algorithm and Image Dichotomous Model. SAGE Open, 13(4). https://doi.org/10.1177/21582440231208851

Srivastava, C., Singh, S., & Singh, A. P. (2018). Estimation of air pollution in Delhi using machine learning techniques. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 304–309

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, 785–794

La, T. V., Dao, M. S., Tejima, K., Kiran, R. U., & Zettsu, K. (2021). Improving the awareness of sustainable smart cities by analyzing lifelog images and IoT air pollution data. In 2021 IEEE International Conference on Big Data (Big Data). IEEE, 3589–3594

Wen, Z., Shang, Y., Lyu, L., Tao, H., Liu, G., Fang, C.,... Song, K. (2024). Re-estimating China's lake CO2 flux considering spatiotemporal variability.*Environmental Science and Ecotechnology*, 19, 100337 https://doi.org/10.1016/j.ese.2023.100337

Yin, Z., Liu, Z., Liu, X., Zheng, W., & Yin, L. (2023). Urban heat islands and their effects on thermal comfort in the US: New York and New Jersey. *Ecological Indicators, 154*, 110765. https://doi.org/10.1016/j.ecolind.2023.110765

Zhang, S., Bai, X., Zhao, C., Tan, Q., Luo, G., & Wang, J.,... Xi, H. (2021). Global CO2 Consumption by Silicate Rock Chemical Weathering: Its Past and Future. *Earth's Future, 9*(5), e1938E–e2020E. https://doi.org/10.1029/2020EF001938

Zhou, Y., De, S., Ewa, G., Perera, C., & Moessner, K. (2018). Data-driven air quality characterization for urban environments: A case study. *IEEE Access*, 6, 77996–78006