Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549448

ANALYSIS AND PREDICTIONS FOR CLIMATE CHANGE DATASET WITH AIR QUALITY INDEX USING DATA MINING AND MACHINE LEARNING APPROACHES

S. Ravishankar[1] and Dr. P. Rajesh[2]
[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalainagar, Tamil Nadu, India
Email: thiru.ravishankar@gmail.com

[2]Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram, (Deputed from Department of Computer and Information Science, Annamalai University, Annamalainagar) Tamil Nadu, India
Email: rajeshdatamining@gmail.com

Abstract

Nowadays, advances in improved climate change data analysis and predictions will result in significantly enhanced national economic opportunities, particularly in the agriculture and energy sectors, as well as different social benefits. Data mining algorithm and machine learning algorithms are the best way to analyze and predict various unsolved problems, particularly in climate change-related issues using data mining and machine learning algorithms like data preprocessing, classification, and decision tree approaches. This paper considers different parameters, namely $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, Xylene and AQI. This paper introduces the most critical analysis and prediction results are finding an optimum and suitable model for predicting climate change through AQI using data mining and machine learning with different accuracy parameters. Numerical illustrations were also provided to prove the proposed research.

Keywords: Climate Change, Air Quality Index, Data Mining, Machine Learning, Accuracy Methods.

1. Introduction

Natural causes include variations in the sun's energy output, volcanic eruptions, and changes in ocean circulation. Human-induced causes involve burning fossil fuels, deforestation, and agricultural practices. Climate change is often accompanied by weather changes, including temperature increases, precipitation patterns, and rising sea levels [1]. Climate change and its analysis are based on the following categories, namely, Human activities: Human activities like burning fossil fuels, deforestation, and industrial processes produce very large volumes of carbon dioxide and another type of greenhouse gases into the atmosphere, which trap heat and lead to climate change [2]. Natural events: Natural events such as volcanic eruptions, Earth orbit changes, and sun output variations can also affect the climate [3]. Oceans: Oceans absorb much heat from the atmosphere and can influence climate patterns. Large-scale ocean currents, such as the Gulf Stream, can also affect climate [4]. Natural cycles: Natural cycles, such as El Niño and La Niña, change weather patterns that affect the environment [5]. Land use: Changes in land use, such as deforestation or urbanization, can also affect climate. This is because vegetation and soil surfaces absorb and reflect sunlight differently than artificial surfaces [6].

DM techniques can be used to detect anomalies and identify extreme weather events. Data mining is the best analytical tool used in various applications, such as customer segmentation, fraud detection, market basket analysis, and predictive analytics. It can be used to uncover customer insights and trends, predict outcomes, and identify potential growth opportunities [7] and [8]. It was cleaning, transforming, and organizing the data to make better results more suitable for the specific algorithms being used. It involves various steps such as data cleaning, integration, transformation, reduction, and feature engineering. The main goal of data preprocessing is to make the data more accurate and easier to use for the machine learning model [9] and [10].

Machine learning algorithms can be supervised (labeled data) or unsupervised (unlabeled data) and can be used for various tasks such as classification, clustering, regression, and forecasting. ML algorithms can be used to process the data and make predictions on unknown data used to optimize the performance of predictive models [11]. Machine learning algorithms which are used to make future predictions and decisions based on data, and they are increasingly being used in many applications, namely self-driving cars, robotics, and medical diagnostics [12].

DM and ML can be used to analyze climate change data in order to understand the impacts of climate change better. Data mining can be used to identify relationships between climate variables such as temperature, precipitation, and sea-level rise and to uncover patterns that could inform the prediction of future climate scenarios. ML can be used to construct predictive models of future climate change trends, ultimately helping to improve our understanding of climate change [13, 14, 15].

2. Review of Literature

This author proposed the climate condition in the cities and explored willingness-to-pay (WTP) for climate change mitigation. Geographic Information System (GIS) maps cities' climate conditions, and Choice Modeling (CM) measures people's awareness for mitigating the impacts. The valuation variables are WTP, socio-economy, and alternative mitigation choice. WTP is the maximum payment in various bid choices; it is between Rp 0- to Rp 210.000-. The options are to plant trees, develop city forests, and public transportation improvement [16]. It analyses the evolution and potential future trend of research debate related to climate change impacts on the wine chain. A particular emphasis was given to sustainability evaluation in the examined literature. From a methodological point of view, sets of text analysis techniques were combined to investigate those selected scientific methods with different results [17].

The authors discussed the performances of these algorithms were compared using standard performance metrics, and the algorithm which gave the best results was used to generate classification rules for the mean weather variables. Given enough case data, the result shows that the various Data Mining techniques used to find the weather forecasting and climate change studies [18]. Bayesian model averaging (BMA) is used for model selection and ensemble projection. Posterior inclusion probability (PIP) is used as the model selection criterion. Our analysis concluded with a list of best models for maximum, minimum temperature, and precipitation, where the rank of the selected models is not the same for the listed three variables. The outputs of BMA closely followed the observed data pattern; however, it underestimated the variability. To overcome this issue, a 90% prediction interval was calculated, showing that almost all the observed data are within these intervals. The results of the Taylor diagram show that the BMA projected data are better than the individual GCMs' outputs [19].

This research aims to work on air quality through emissions, pollutant concentration data, and vegetation information. The authors study the concepts to mine and collect this information spread all over the place in different formats into a knowledge base on which further data analysis and powerful Machine Learning approaches can be built to extract strong evidence helpful in making better policies around climate change [20]. Climate change forecast or detection challenging in 21 century. The Earth's surface and its temperature increased rapidly between 2000 to 2015 and the 20th century, even though greenhouse gas concentrations increased [21].

The authors explain various data mining and machine learning algorithms with accuracy for different decision tree approaches using the WEKA tool to stumble on essential parameters of the tree structure. Seven classification algorithms such as J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP), and Random Forest (RF) are used to measure the accuracy. The data mining tool WEKA (Waikato Environment for Knowledge Analysis) has been used to find experimental results of weather data set. Out of seven classification algorithms, the Random tree algorithm outperforms other algorithms by yielding an accuracy of 85.714% [22].

The authors discuss the objective of this paper is to analyze the SDGs by various independent metrics of Tamil Nadu, Kerala, and Karnataka in India and by considering three different state SDGs indexes using data mining and statistical approaches for retrieving various hidden information. Numerical illustrations are also used to prove the proposed results [23].

3. Material and Methods

3.1 Correlation coefficient

The R2 score or correlation coefficient finds the strength of the linear relationship between two different variables. It is commonly used in statistics, ranging from -1 to +1. A correlation coefficient return +1 indicates a perfect positive linear relationship between two variables. A correlation coefficient of -1 indicates a perfect negative linear relationship between two variables. A correlation coefficient return of 0 indicates no linear relationship between the two variables.

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2]-[n\sum y^2-(\sum y)^2]}} \quad \ldots (1)$$

where n means the quantity of information, Σx total of the first variable value, Σy the total of the second variable value, Σxy sum of the product of the first & second value, Σx^2 namely the sum of the squares of the first value and Σy^2 sum of the squares of the second value.

3.2 Gaussian Process

Gaussian process (GP) is a type of probability distribution over functions or a collection of random variables which can be used to describe joint distributions. It is a generalization of the multivariate normal distribution and is used to model data points that have a nonlinear relationship with each other. A mean function defines a Gaussian process μ(x) and a covariance function $\Sigma(x, x')$, which are used to determine the probability distribution over functions [24] and [25]. The formula for a Gaussian process is given by:

$$P(f(x)|x) = N(f(x)|\mu(x), \Sigma(x, x')) \qquad \dots (2)$$

where $N(f(x)|\mu(x), \Sigma(x, x'))$ is the multivariate normal distribution with mean μ(x) and covariance $\Sigma(x, x')$.

3.3 Linear Regression

Linear regression is one of the main statistical techniques used to model the relationship between dependent variables and one or more independent variables. Find predict the dependent variable's value based on the independent variables' values. Linear regression estimates actual values based on a given data set [26]. The following equation represents it:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \qquad \dots (3)$$

where y namely dependent variable, $x_1, x_2, \dots, x_n$ means independent variables, β_0 namely intercept and $\beta_1, \beta_2, \dots, \beta_n$ namely regression coefficients.

3.4 REP Tree

The reduced error pruning tree is a method used to optimize the performance of decision tree models. By pruning away branches of the tree that are not contributing to the overall accuracy, the model can be simplified, and the error rate can be reduced. The reduced error pruning tree works by starting with a fully grown decision tree and then pruning away branches of the tree that do not contribute to the model's accuracy. The pruning process is done iteratively until the best accuracy is achieved. The reduced error pruning tree is helpful in avoiding overfitting and improving the model's generalization [27].

Step 1. Start with an initial decision tree that is fully grown, meaning it includes all possible branches and leaves.

Step 2. Split the available dataset into two parts: a training set and a validation set. The training set will be used to grow the tree, while the validation set will be used to find the performance of the tree and determine whether pruning is necessary.

Step 3. Grow the decision tree using any tree-building algorithm on the training set until all the leaves are pure or no further splits are possible based on the available attributes.

Step 4. Evaluate the performance of the decision tree on the validation set using an appropriate performance metric (e.g., accuracy, error rate, F1 score). This will serve as the baseline performance for comparison.

Step 5. Start at the bottom of the decision tree and examine each internal node. For each node, temporarily remove the subtree below it, creating a pruned version of the tree.

Step 6. Evaluate the performance of the pruned tree on the validation set using the same performance metric as in step 4.

Step 7. Compare the performance of the pruned tree with the baseline performance. If the pruned tree performs better (or equivalently), replace the original subtree with the pruned subtree. Otherwise, discard the pruned tree and keep the original subtree.

Step 8. Repeat steps 5-7 for all internal nodes, moving from the bottom to top trees.

Step 9. After pruning all the internal nodes, you will be left with the final pruned decision tree.

Step 10. Optionally, you can perform additional pruning rounds using different training and validation sets to refine the decision tree further.

3.5 Performance Metrics

MAE stands for Mean Absolute Error. It measures the average errors in a set of predictions without considering their direction. MAE measures the average volume of the errors in a group of predictions, regardless of their approach. It is the average of absolute differences between forecast and actual values over the test sample [28].

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad \text{... (4)}$$

where Σ: summation, y_i: actual value for the i^{th} observation, x_i: calculated value for the statement, and n: total number of comments.

The RMSE (root mean square error) measures the difference between predicted and actual values and measures how well a model fits a given data set. The RMSE finds the square root of the average squared differences between predicted and actual values. The lower the RMSE, the better the model fits the data [29] and [30].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \| y(i) - \hat{y}(i) \|^2}{n}} \quad \text{... (5)}$$

where *n* is the number of elements, *y(i)* means i^{th} measurement, and *y ̂(i)* is called the prediction.

Relative Absolute Error (RAE) is a measure of accuracy that is used to compare the difference between predicted and actual values. It is calculated by taking the absolute difference between the two values and dividing it by the actual value. The result is expressed as a percentage. RAE is often used to evaluate the performance of machine learning algorithms, as it indicates how close the prediction was to the actual value [31].

$$RAE = \frac{\frac{\Sigma |y_i - \hat{y}_i|}{\Sigma |y_i - \bar{y}|}}{n} \quad \text{... (6)}$$

Where *n* is the number of elements in the observations, *y(i)* is the realized value and *y ̂(i)* is the prediction, and $\bar{y}$ means the mean values of corresponding variables.

RRSE stands for Root Relative Squared Error, a metric used to evaluate the accuracy and performance of a regression model. It measures the average relative difference between the predicted and actual values, taking into account the scale of the target variable. The RRSE calculations include the following steps.

Step 1. Obtain the predicted values from the regression model for a set of examples.

Step 2. Obtain the corresponding actual values for those examples.

Step 3. Calculate the squared difference between each predicted value and its corresponding actual value.

Step 4. Sum up all the squared differences.

Step 5. Divide the sum of squared differences by the total number of examples to get the mean squared difference.

Step 6. Take the square root of the mean squared difference to obtain the source mean squared difference (RMSE).

Step 7. Normalize the RMSE by dividing it by the range of the target variable. The content is calculated as the difference between the maximum and minimum values of the target variable.

Step 8. Multiply the normalized RMSE by 100 to obtain the RRSE.

$$RRSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \ldots (7)$$

Where n is the number of elements in the observations, $y(i)$ is the realized value, $y\hat{}(i)$ is the prediction, and $\bar{y}$ means the mean values of corresponding variables. The RRSE provides a relative measure of the model's performance, allowing comparison between regression models or datasets.

4. Numerical Illustrations

The dataset downloaded using https://www.kaggle.com/code/nareshbhat/air-quality-analysis-eda-and-classification/notebook the dataset include 615 instance with 14 parameters namely $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, Xylene and AQI. The Kaggle website is one of the world's largest data science researchers with powerful tools and resources to help you achieve your data science goals. This includes different parameters and the simple definition of the AQI-related parameters presented in the following table.

Table 1. Air Quality Index (AQI) and its definition

Sl. No.	AQI Parameters	Definition
1.	$PM_{2.5}$	PM2.5 stands for Particulate Matter 2.5.
2.	PM_{10}	PM10 refers to particulate matter 10 micrometers or smaller in diameter.
3.	NO	Nitric Oxide: In chemistry, "NO" refers to nitric oxide, which is a colorless gas with the chemical formula NO.
4.	NO_2	NO2 stands for Nitrogen Dioxide. It is a reddish-brown gas composed of nitrogen and oxygen, with the chemical formula NO2.
5.	NOx	NOx is an abbreviation used to refer to nitrogen oxides collectively. It represents a group of gases composed of nitrogen and oxygen, mainly nitric oxide (NO) and nitrogen dioxide (NO2).
6.	NH_3	NH3 stands for Ammonia. It is a compound of one nitrogen atom bonded with three hydrogen atoms, with the chemical formula NH3.
7.	CO	CO stands for Carbon Monoxide. It is a colorless, odorless, and tasteless gas composed of one carbon atom bonded with one oxygen atom, with the chemical formula CO.
8.	SO_2	SO2 stands for Sulfur Dioxide. It is a toxic gas composed of one sulfur atom bonded with two oxygen atoms, with the chemical formula SO2.
9.	O_3	O3 stands for Ozone. Ozone is a molecule composed of three oxygen atoms with the chemical formula O3. It is a colorless gas that has a distinct, pungent odor.

10.	Benzene	Benzene is a colorless, flammable liquid with a sweet odor, and it is an organic chemical compound with the molecular formula C6H6. Benzene is a natural component of crude oil and is produced through various industrial processes.
11.	Toluene	Toluene is an aromatic hydrocarbon compound with the chemical formula C7H8. It is a colorless, volatile liquid with a sweet and spicy odor. Toluene is derived from petroleum and is commonly used as a solvent in various industries.
12.	Xylene	Xylene is an aromatic hydrocarbon compound with the molecular formula C8H10. It is a colorless liquid with a sweet, fruity odor. Xylene is derived from petroleum and is primarily composed of three isomers: ortho-xylene, meta-xylene, and para-xylene.
13.	AQI	AQI stands for Air Quality Index, a numerical scale used to measure and communicate ambient air quality based on the concentration of common air pollutants. The purpose of the AQI is to provide the public with information about the current air quality and its potential health effects.

Table 2. Climate change with AQI dataset

Date Time	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benz	Tolue	Xyle	AQI	AQI Buck
3/10/2019 21:00	18.25	50	0.7	8.48	5.1	8.78	0.44	11.18	17.62	0.9	5.55	0.65	49	Good
3/10/2019 22:00	20.5	46.75	1.73	6.38	4.78	8.77	0.51	10.97	15.2	1.05	4.75	0.45	49	Good
3/10/2019 23:00	15.5	44.5	1.27	6.1	4.28	9.12	0.39	9.48	14.22	1.3	3.65	0.3	48	Good
3/11/2019 0:00	13.25	38.5	0.42	5.8	3.33	8.55	0.34	9.18	13.72	0.82	2	0.2	47	Good
3/11/2019 3:00	6.75	39	0.5	7.22	3.92	9.07	0.36	11.57	11.8	0.9	2.23	0.2	46	Good
1/1/2019 0:00	96	165	4.15	99.45	56.3	11.52	1.77	14.67	23.53	0.1	0.35	0.1	164	Moderate
1/1/2019 1:00	100	151.5	2.75	83.95	46.9	13.25	1.93	25.15	22.55	0.1	0.2	0.1	165	Moderate

1/1/2019 2:00	112	159.25	2.75	88.6	49.38	13.62	1.38	20.3	15.7	0.13	0.25	0.1	170	Moderate
1/1/2019 19:00	67.75	129.25	5.28	92.25	53.32	17.5	0.83	22.9	30.37	0.1	0.2	0.1	187	Moderate
1/1/2019 20:00	81.75	183.25	3.65	92.1	51.95	16.6	0.88	19.07	24.47	0.1	0.23	0.13	187	Moderate
1/2/2019 20:00	118.5	253.75	16.85	175.65	107.17	10.67	1.27	68.12	14.7	0.23	0.93	0.28	246	Poor
1/2/2019 21:00	116.5	260	32.7	173.5	118.85	12.12	2.32	29.45	13.7	0.1	0.28	0.1	252	Poor
1/2/2019 22:00	132	258.25	17.48	143.95	90.75	10.08	2.35	18.8	17.18	0.13	0.25	0.15	260	Poor
1/2/2019 23:00	139.25	250.5	21.12	137.58	90.35	9.23	2.7	15.4	14.23	0.1	0.25	0.1	267	Poor
1/3/2019 0:00	176.5	264.5	27.45	155.28	104.9	9.05	2.5	12.4	10.75	0.13	0.23	0.1	276	Poor
3/7/2019 19:00	48.5	98.5	0.38	11.12	6.15	13.1	0.49	9.38	39.68	2	3.75	0.57	98	Satisfactory
3/7/2019 20:00	37.25	83.25	1.1	6.85	4.55	12.62	0.49	7.27	38.2	2.05	3.25	0.53	94	Satisfactory
3/7/2019 21:00	33	68.25	0.35	6.53	3.67	12.12	0.51	15.45	32.83	2.55	2.6	0.32	94	Satisfactory
3/7/2019 22:00	29.25	64.75	0.73	6.75	4.17	11.82	0.53	17.3	28.9	2.6	3.1	0.35	94	Satisfactory
3/7/2019 23:00	31.5	57.5	1.1	9.4	5.88	11.95	0.52	18.65	24.02	2.78	2.33	0.3	94	Satisfactory
1/3/2019 5:00	153.75	236.75	85.55	155.92	152.47	6.83	1.16	6.77	5.9	0.18	2	0.13	301	Very Poor
1/3/2019 6:00	142.5	216.25	77.65	140.17	137.67	6.52	1.28	4.38	5.9	0.15	2.95	0.2	302	Very Poor
1/3/2019 7:00	133.25	201	94.57	133.95	148.17	10.25	1.34	11	5.9	0.18	0.92	0.1	302	Very Poor

1/3/2019 8:00	142.25	216	77.4	131	132.62	8.13	0.85	10.1	6.38	0.23	0.45	0.1	303	Very Poor
1/3/2019 9:00	132.75	197.75	44.25	115.93	97.63	8.78	1.44	10.67	12.72	0.13	0.43	0.1	303	Very Poor

Table 3. Mean and SD for AQI Parameters

Attributes	Mean	StdDev
PM2.5	42.499	43.961
PM10	95.163	71.082
NO	8.627	16.605
NO2	30.174	44.148
NOx	22.961	34.165
NH3	8.702	3.055
CO	0.673	0.687
SO2	11.125	6.710
O3	25.356	19.510
Benzene	1.247	3.496
Toluene	3.812	4.156
Xylene	0.417	2.059
AQI	113.418	78.887
AQI_Bucket	2.984	1.288

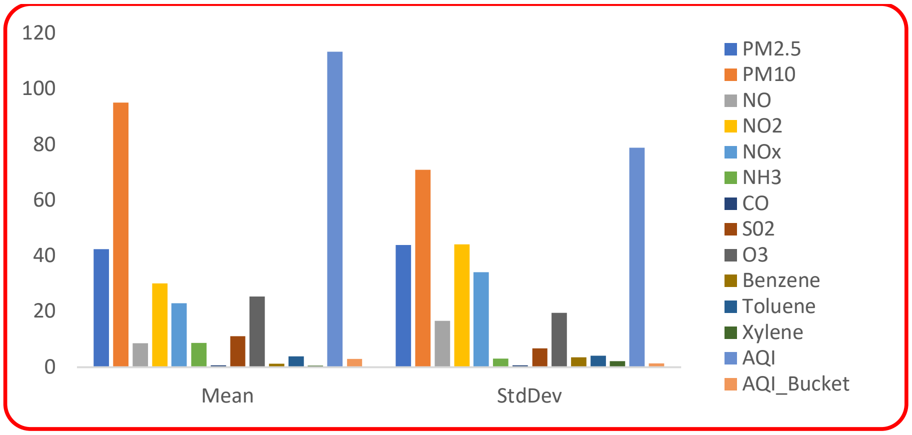

Fig. 1. Mean and Standard Deviation for AQI Parameters

Table 4. R2 Score / Correlation Coefficient for AQI Parameters

Attributes	Gaussian Process	Linear Regression	REP Tree
PM2.5	0.8876	0.9619	0.9616
PM10	0.9016	0.7711	0.9341
NO	0.9506	0.7356	0.9453
NO2	0.8415	0.9991	0.9814

NOx	0.9168	0.9996	0.9868
NH3	0.3334	0.4371	0.6221
CO	0.4271	0.2692	0.3458
S02	0.3282	0.1995	0.2397
O3	0.1104	0.0705	0.6818
Benzene	0.7025	0.6748	0.0112
Toluene	0.1090	0.1222	0.8031
Xylene	0.0837	0.2457	0.1246
AQI	0.5518	0.7360	0.9695
AQI Bucket	0.2195	0.1489	0.9951

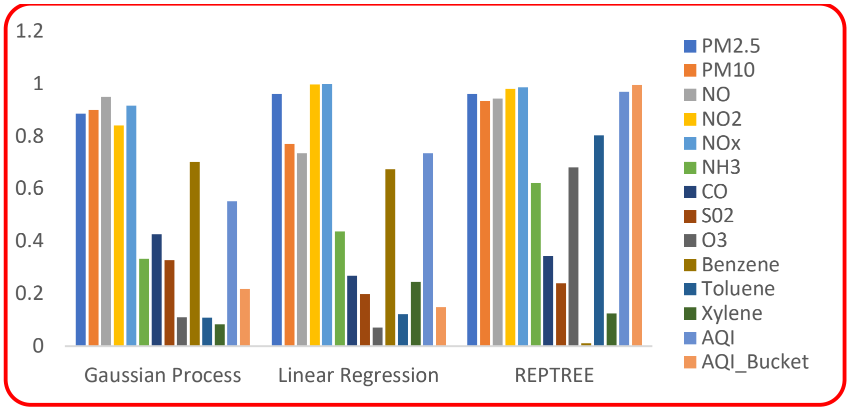

Fig 2. R2 Score / Correlation Coefficient for AQI Parameters

Table 5. Machine Learning Approaches with MAE

Attributes	Gaussian Process	Linear Regression	REP Tree
PM2.5	12.3345	7.8249	7.0893
PM10	24.9137	17.4431	16.8986
NO	3.5768	6.3435	2.464
NO2	10.2446	0.7200	3.9807
NOx	6.2088	0.3408	2.6388
NH3	2.3903	2.1974	1.7647
CO	0.2932	0.2938	0.2819
SO2	4.1726	4.3567	4.1636
O3	14.4404	14.9933	9.7156
Benzene	0.7117	0.7178	0.8262
Toluene	2.4420	2.4609	1.3229
Xylene	0.3578	0.4128	0.3317
AQI	34.4046	24.4624	9.3754
AQI_Bucket	1.1950	1.2195	0.0105

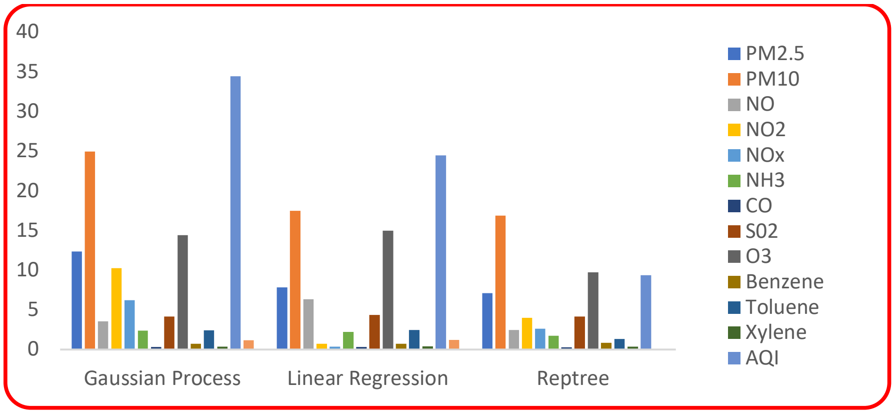

Fig. 3. Machine Learning Approaches with MAE

Table 6. Machine Learning Approaches with RMSE

Attributes	**Gaussian Process**	**Linear Regression**	**REP Tree**
PM2.5	20.7848	12.0172	12.1004
PM10	31.6369	52.7685	25.3599
NO	5.5438	11.2509	5.4247
NO2	26.1090	1.8492	8.4983
NOx	14.1880	1.0106	5.5368
NH3	2.8957	2.7522	2.4048
CO	0.6225	0.7655	0.6497
SO2	6.3338	6.6042	6.6147
O3	38.1824	60.3370	14.3395
Benzene	3.0530	2.9982	3.5114
Toluene	11.1076	15.0303	2.4857
Xylene	2.0649	1.9979	2.0410
AQI	88.1906	62.7834	19.3592
AQI_Bucket	1.2725	1.3671	0.1279

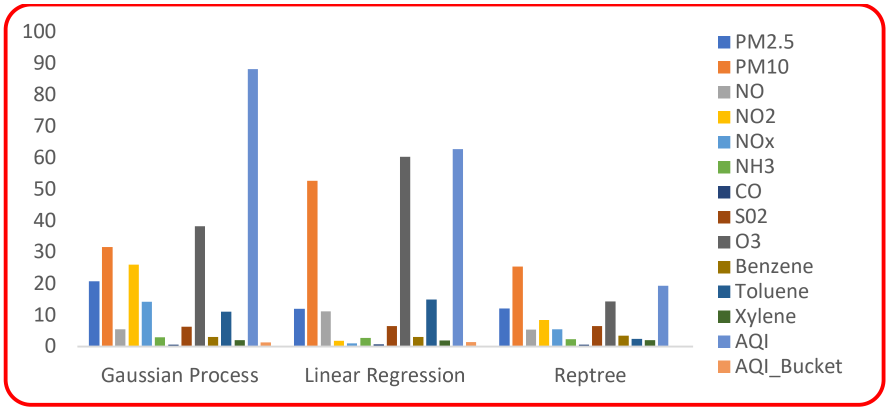

Fig. 4. Machine Learning Approaches with RMSE

Table 7. Machine Learning Approaches with RAE

Attributes	Gaussian Process	Linear Regression	REP Tree
PM2.5	39.4605	25.0333	22.6799
PM10	46.6872	32.6876	31.6672
NO	37.1058	65.8068	25.5615
NO2	32.5712	2.2891	12.6560
NOx	26.1579	1.4360	11.1175
NH3	94.6089	86.9717	69.8477
CO	78.4552	78.6200	75.4499
S02	96.8798	101.1556	96.6713
O3	92.8539	96.4089	62.4728
Benzene	80.9180	81.6070	93.9338
Toluene	99.8831	100.6581	54.1105
Xylene	96.1187	110.9111	89.1081
AQI	59.0836	42.0096	16.1005
AQI_Bucket	100.5781	102.6394	0.8835

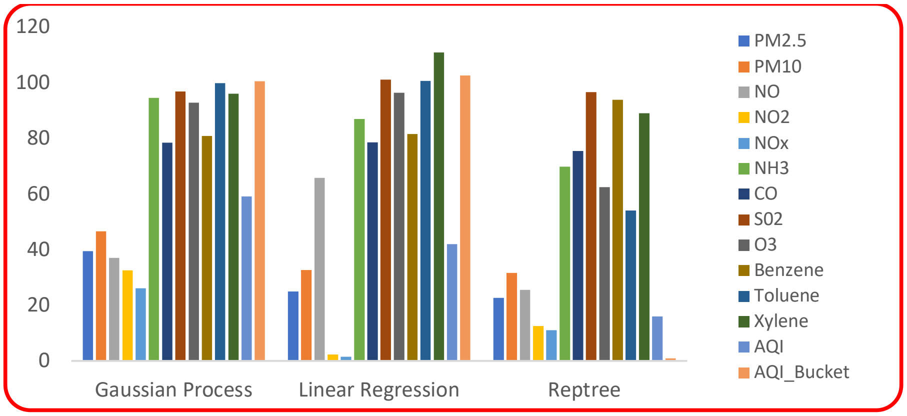

Fig. 5. Machine Learning Approaches with RAE

Table 8. Machine Learning Approaches with RRSE%

Attributes	Gaussian Process	Linear Regression	REP Tree
PM2.5	47.2318	27.3081	27.4972
PM10	44.4348	74.1148	35.6186
NO	33.3641	67.7110	32.6475
NO2	59.0954	4.1854	19.2351
NOx	41.4973	2.9559	16.1942
NH3	94.7356	90.0432	78.6764
CO	90.5012	111.2954	94.4672
S02	94.3711	98.3996	98.5559
O3	195.4406	308.8408	73.3983

Benzene	87.2724	85.7062	100.3738
Toluene	267.0872	361.4091	59.7696
Xylene	100.2411	96.9912	99.0839
AQI	111.6774	79.5038	24.5149
AQI_Bucket	98.7222	106.0561	9.9195

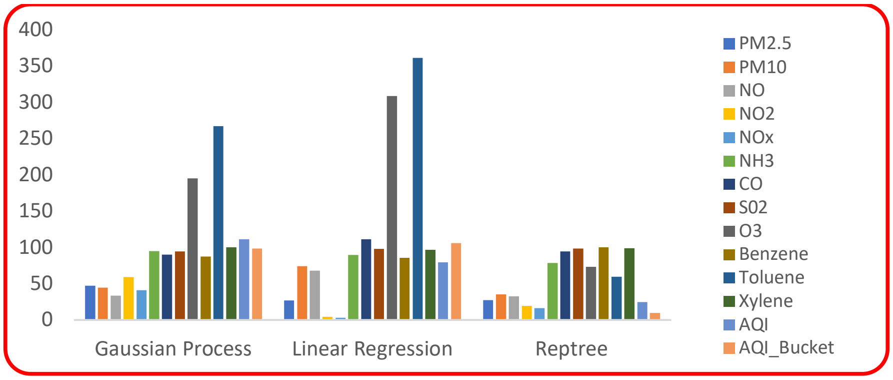

Fig. 6. Machine Learning Approaches with RRSE

5. Result and Discussion

Table 1 explains 14 parameters and their simple definition, and Table 2 indicates the dataset. Comparison of average test scores and standard deviations presents different interpretations. Namely, the xylene is the minimum mean score, and the air quality index returns the maximum mean value. The SD is used to find the deviations based on different groups. In this case, CO (Carbon Monoxide) returns minimum SD values, and AQI returns the maximum SD score. The related results are shown in Table 3 and Figure 1. The graph in Figure 1 further illustrates the trends in the average test scores over time for both all the parameters (groups).

Based on the dataset, it is evident that three different machine learning approaches are used to find the hidden patterns and also which is the best or influencing parameter to decide the critical parameter. They are based on Table 4 and Figure 2 using Equation 1, which is used to find the R2 score or correlation coefficient by comparing 13 parameters. Numerical illustrations suggest that there may be a significant difference from one parameter to another. In this case, the Gaussian process (equation 2) returns a strong positive correlation of 0.9506 when using NO.

In another case, using linear regression (equation 3), No_2 and NO_x return a strong positive correlation of 0.999. REP Tree algorithm (steps 1 to 10) is one of the decision tree approaches which is used to solve the problems. In this case, the AQI Bucket, NO_2, and NO_x return strong positive correlations, likewise 0.9951, 0.9814, and 0.9868. The related results and discussion are shown in Table 4 and Figure 2.

Further analysis of the data revealed a gradual improvement in test scores over time. The MAE is used to find the model error using Equation 4. In this case, three different machine learning algorithms are to be used, namely the Gaussian process, linear regression, and REP Tree return minimum error compared to other parameters while using CO and Xylene. These findings raise questions about potential factors that may have contributed to the variations in using other parameters in accuracy performance between the different groups. The related numerical illustrations are shown in Table 5 and Figure 3.

The RMSE (root mean square error) is a measure of the difference between predicted and actual values using Equation 5. Linear regression (equation 3) is used to find the prediction values for using dependent and independent variables. In this case, the model returns nearly 1, which means strong positive correlations for using NO_2 and NO_x parameters. Linear regression and its prediction are also good with maximum accuracy for using RMSE (equation 5) except for three parameters, namely PM10 (52.7685%), O3 (60.3370%), and AQI (62.7834%). The related numerical illustration is shown in Table 6 and Figure 4.

Relative Absolute Error (RAE) is a measure of accuracy using equation 6 that is used to compare the difference between a predicted value and an actual value. One of the decision tree approaches, REP Tree, solves the problem using tree approaches with a reduced error pruning approach. Using different parameters, the final outcome class indicates whether the AQI index returns good, moderate, poor, satisfactory, or very poor. In this case, the outcome variable AQI Bucket return minimum error compare to the Gaussian process. Similarly, the linear regression returns the relative absolute error using AQI Bucket as NO2 (2.28) and NOx (1.43). Similar numerical illustrations are shown in Table 7 and Figure 5.

Root Relative Squared Error provides a relative measure of the models with performance, and the result is expressed as a percentage using equation 7. In this case, the REP Tree with AQI Bucket returns a minimum error. The linear regression returns the absolute relative mistake using AQI Bucket is NO2, and NOx also returns the minimum error rate. Similar numerical illustrations are shown in Table 8 and Figure 6.

6. Conclusion and further research

It is essential to consider the limitations of this study. The sample size of each group was relatively small, which could impact the generalizability of the results. Additionally, other variables could influence climate change and also AQI performance. The findings presented in this study contribute to our understanding that NO_2 and NO_x are the best parameters for deciding climate change with AQI. The reduced error pruning tree is a method used to optimize the performance of decision tree models.

The REP Tree is the best solution to climate change analysis with an air quality index. In this case, the proposed results acquire maximum accuracy. Future studies can build upon these findings to develop targeted interventions to improve accuracy and include other climate change-related parameters to find the accuracy using different machine learning and decision tree approaches.

Reference

1. United Nations Framework Convention on Climate Change. (2021). What is climate change? Retrieved June 8, 2021, from https://unfccc.int/topics/what-is-climate-change
2. McMichael, A.J., et al., Climate Change and Human Health: Risks and Responses. Lancet, 2006. 367(9513): p. 859-869.
3. IPCC (Intergovernmental Panel on Climate Change), Climate Change 2013: The Physical Science Basis. 2014.
4. NOAA (National Oceanic and Atmospheric Administration), Ocean and Climate, https://www.noaa.gov/education/resource-collections/ocean-climate.

5. IPCC (Intergovernmental Panel on Climate Change), Climate Change: The Science of Climate Change. 2014.
6. EPA (Environmental Protection Agency), Effects of Land Use on Climate. https://www.epa.gov/climate-indicators/effects-land-use-climate.
7. Han, J. and M. Kamber, Data Mining: Concepts and Techniques. 3rd ed. 2011, Morgan Kaufmann.
8. Shmueli, G., et al., Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. 2017, Wiley.
9. Shah, A. (2020). What is Data Preprocessing? Medium. https://towardsdatascience.com/what-is-data-preprocessing-c2f6de17d25e)
10. Raschka, S. (2015). Python Machine Learning. Packt Publishing Ltd
11. "What is Machine Learning? - Towards Data Science." Towards Data Science, Towards Data Science, 17 Mar. 2020, towardsdatascience.com/what-is-machine-learning-8ba1c45fb18b.
12. Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.
13. Kumar, P., & Goyal, P. (2018). Machine learning algorithms for climate change analysis: A survey. Environmental Modelling & Software, 104, 4-17
14. Liu, A., Paruchuri, P., & Liu, Y. (2019). Data mining techniques for climate change analysis. Journal of Environmental Informatics, 33(1), 1-8.
15. Sutanto, D., & Mukhopadhyay, S. (2020). Climate change analytics: A comprehensive review of machine learning applications. Renewable and Sustainable Energy Reviews, 133, 110406.
16. Gravitiani, E. and Antriyandari, E., 2016. Willingness to pay for climate change mitigation: application on big cities in Central Java, Indonesia. Procedia-Social and Behavioral Sciences, 227, pp.417-423.
17. Sacchelli, S., Fabbrizzi, S. and Menghini, S., 2016. Climate change, wine, and sustainability: a quantitative discourse analysis of the international scientific literature. Agriculture and agricultural science procedia, 8, pp.167-175.
18. Jeslet, D.S. and Jeevanandham, S., 2015. Climate Change Analysis using Data Mining Techniques. Int. J. Adv. Res. Sci. Eng, 8354(4), pp.46-53.
19. Khan, F. and Pilz, J., 2018. Statistical Methodology for Evaluating Process-Based Climate Models. In Climate Change and Global Warming. IntechOpen.
20. Babu Saheer, L., Shahawy, M. and Zarrin, J., 2020, June. Mining and analysis of air quality data to aid climate change. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 232-243). Springer, Cham.
21. Lean, J.L., 2018. Observation-based detection and attribution of 21st-century climate change. Wiley Interdisciplinary Reviews: Climate Change, 9(2), p.e511.
22. Rajesh, P. and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using the WEKA tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243.
23. Rajesh, P. and Kumar, B.S., 2020. Comparative studies on Sustainable Development Goals (SDG) in India using Data Mining approach. J. Sci, 14(2), pp.91-93.

24. Rasmussen, C. E. (2006). Gaussian Processes for Machine Learning. MIT Press.
25. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
26. "Linear Regression." Investopedia, Investopedia, 16 Mar. 2020, www.investopedia.com/terms/l/linearregression.asp.
27. Kohavi, R., & Sahami, M. (1996). Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286).
28. Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/
29. S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, "Root mean square error (RMSE): A comprehensive review," International Journal of Applied Mathematics and Statistics, vol. 59, no. 1, pp. 42–49, 2019.
30. J. C. Nash, "The use of the root mean square error (RMSE) in climatological studies of space-time fields," Journal of Climate, vol. 5, no. 2, pp. 565–572, 1992.
31. Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566