IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHMS FOR AIR POLLUTION PREDECTION

[1]Balu V, [2] G Lakshmi Narayani G, [3]K Bhavitha

[1]Assistant Professor, [2] Graduate Student, [3] Graduate student

Dept. of Computer Science Engineering,

SCSVMV University, Enathur, India

Abstract: As air quality is rapidly affecting people's health, the government must take required action to forecast it. The air quality index calculates the quality of the air. Other air pollutants include carbon dioxide, nitrogen dioxide, and carbon monoxide, which are emitted by the burning of natural gas, coal, and wood as well as by businesses, automobiles, and other human activity. Air pollution can result in fatal conditions like lung cancer, brain disease, and even demise. Machine learning techniques are used to calculate the air quality index. Although there have been many studies in this field, the conclusions are still not trustworthy. We have two options for completing the project. First, we can estimate the present or future air quality by knowing the precise values of SO2, Nox46, and RSPM49. Alternatively, we may use machine learning methods to compare the air quality. Air quality monitoring stations, two training and testing data sets, and the official website of the Central Pollution Control Board (CPCB) are all accessible. Decision trees, lasso regression, linear regression, K-NN regression, etc.

***Keywords*:** Air pollution, Linear Regression, Decision tree, Lasso regression, K-NN Regression.

1. INTRODUCTION

Air pollution is the primary problem in every country, developed or developing. Particularly in urban regions in emerging economies, where industrialization and an increase in the number of cars cause the release of several gaseous pollutants, health issues have been getting worse more quickly. A detrimental impact of pollution on health can range from simple adverse reactions like throat, eye, and nose irritation to more serious conditions like bronchitis, heart disease, pneumonia, lung disease, and exacerbated asthma. Research has found that just air pollution causes 50,000 to 100,000 prematurely deaths annually in the US. In contrast, there are over 3,000,000 people worldwide and 300,000 people living in the EU.

The amount of air pollution produced by private transportation has risen significantly in recent years. The main cause of vehicular emissions is carbon monoxide, which semiconductors chemical sensors can quickly identify. The pulmonary and pulmonary systems are notably affected by these pollutants' impacts on human health. These toxins disperse on surfaces such as dirt, water, as well as other a number of sensors can be used to find greenhouse gases.

The impact of urban air pollution on people's lives everywhere and at all times has drawn significant attention from people all around the world. A network of observation sites employing conventional measurement methods has been built to lessen these effects. Using the information gathered to create pollution mapping and models, it is possible to predict the condition of the environment.

2. LITERATURE SURVEY

In this paper, Agarwal S, Sharma S, Suresh R, Rahman MH, Vranckx S, Maiheu B, Blyth L, Janssen S, Gargava P, Shukla VK, Batra S (2020) Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. Sci Total Environ 735:139454 [1].

In this paper, Ali Shah SA, Aziz W, Ahmed Nadeem MS, Almaraashi M, Shim S-O, Habeebullah TM, Mateos C (2019) A novel Phase Space Reconstruction- (PSR-) based predictive algorithm to forecast atmospheric particulate matter concentration. Sci Program [2].

In this paper, Balogun A-L, Tella A (2022) Modelling and investigating the impacts of climatic variables on ozone concentration in Malaysia using correlation analysis with random forest, decision tree regression, linear regression, and support vector regression. Chemosphere 299:134250 [3].

In this paper, Chakradhar Reddy K, Nagarjuna Reddy K, Brahmaji Prasad K, Selvi Rajendran P (2021) The prediction of quality of the air using supervised learning. In: 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp 1–5 [4].

In this paper, Colchado LE, Villanueva E, Ochoa-Luna J (2021) A neural network architecture with an attention-based layer for spatial prediction of fine particulate matter. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp 1–10 [5].

2.1 PROBLEM STATEMENT

Maintaining Respectable air quality norms has come a serious challenge in civic areas with a significant attention of businesses, manufacturing, and people. As well as with population growth, further energy, electricity, and transportation are being used. We're well apprehensive that a large quantum of scrap is ditched on the ground.

The high degree of impurity in the air poses a major trouble to all feathers of life on earth, making it imperative. This study focuses on applying machine literacy styles to take over an effective analysis of all the significant workshop in this area. The position of living in smart metropolises is seriously hovered by the accumulation of poisonous feasts. It's pivotal to produce effective air quality control models that can measure global air pollution and gather information on pollutant attention when air pollution situations rise.

3. PROPOSED SYSTEM

The effectiveness of the choice four algorithm vaticinator model for the air excellent handicap is explained by the following criteria. It displays better issues in terms of categorization complexity.

A training set and a testing set must first be created from the data collection. The vaticinator model is originally trained using the training dataset. The testing set will also be used to test it. an volition is to usek-fold cross confirmation. Following testing, the model's delicacy is estimated using criteria including the discovery rate, perfection, recall, F- Measure, and overall delicacy. It can be changed or edited by others who know the exact values about it. Constructing a vaticinator model Data collection should aim to have a large quantum of literal data. Undressed data and enough literal information have been collected.

3.1 SYSTEM ARCHITECTURE

It is represented in the Fig 1 below the step-by-step work done by the system in a flow type starting from Importing Dataset

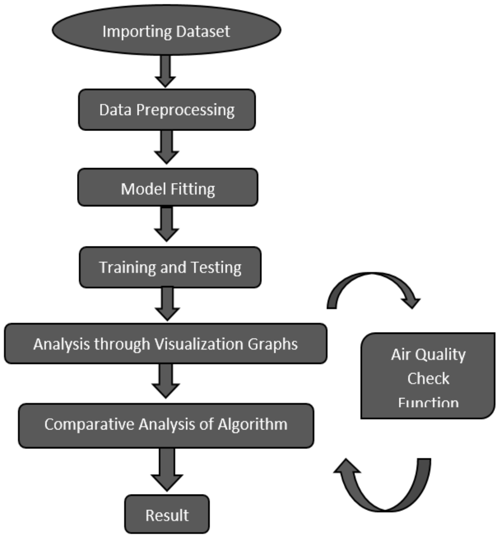

Importing Dataset
Data Preprocessing
Model Fitting
Training and Testing
Analysis through Visualization Graphs
Comparative Analysis of Algorithm
Result
Air Quality Check Function

Fig 1. System Architecture

3.2 ALGORITHMS USED

In this section, the newly discovered machine learning techniques from the literature review are described together with the experiment analysis utilizing the pre-processed data set. For the identical set of air quality data, the model is independently built using Decision Tree, Linear Regression, LASSO Regression, and KNN Regression. To obtain the AQI attribute as the experiment's output, we employed the SO2, Nox, and RSPM attributes as input.

a) Linear Regression:

In a linear model, the only dependent output variable (y) and the independent input variables (x) have a linear connection, allowing y to be evaluated or predicted from the linear combination of the independent input variables (x). The term "Linear Regression" is used to describe a linear regression where there is just one independent or input variable (x).

b) Losso Regression:

The Lasso Regression analysis technique does both variable selection and regularization in order to enhance the predictability and interpretability of the generated statistical model.

c) Decision Tree Algorithm:

In addition to fitting noisy observations, the Decision Tree method frequently does so as well. The decision tree over fits if this parameter is set too high because it learns too many intricate features from the training data as well as extensive knowledge about noise.

d) K-Nearest Neighbor Regression:

The supervised learning technique is K-Nearest Neighbor Regression. All instances that correspond to training data points are kept in an n-dimensional space via K-Nearest Neighbor. When an unknown discrete data set is retrieved, the algorithm examines the nearest k stored examples (near neighbours), returns the most suggested class, and for real-value outcomes, returns the average of the k comparable neighbours.

e) Performance metrics:

We require some metrics in order to evaluate the effectiveness of a machine learning model. We have taken MAE and RMSE into consideration as the performance indicators because our thesis is about prediction.

f) R squared (R2) :

The R square performance metric shows how well predicted and actual values line up. We may use the r2 score function of sklearn. Metrics to calculate R squared value.

$$R^2 = 1 - \frac{sumsquaredregression(SSR)}{totalsumofsquares(SST)}$$

$$R^2 = 1 - \frac{\sum(y_i - \overline{y_i})^2}{\sum(y_i - \overline{y})^2}$$

g) Root mean square error (RMSE):

The root mean square error (RMSE) measures how much the target value differs from the value predicted by the model. It is the mean square error's square root (MSE). The method of implementation is pretty comparable to MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \overline{y_j})^2}$$

Where,

yj = True value , N = Total number of data points

By contrasting performance metrics, the machine learning models are verified. The performance of the machine learning model improves with decreasing MAE, RMSE, and increasing r-squared.

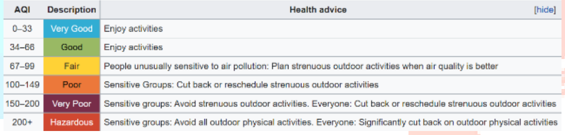
3.3 IMPLEMENTATION PROCEDURE:

To apply trial, we've used following technologies. A brief description of used technologies along with interpretation is represented below.

- Python (Version 3.10.4) - Python is an interpreter, high level and Object-Oriented programming language. It is an open source.
- Anaconda navigator (Version 2.1.1) - It is a graphical user interface (GUI) that allows to launch applications easily and manage packages.
- Jupyter notebook (Version 4.8) - It is a web based interactive computing platform. It helps in developing, documenting, and executing code.
- Pandas (Version 1.4.2) - It is a free open source python library. It is mainly used for data analysis. It helps to perform various data manipulation operations.
- Numpy (Version 1.22.3) - It is a python library which is used for scientific computing in python. It is used to perform wide mathematical operations on data.
- Matplotlib (Version 3.5.2) - It is a Python package used to create static, animated, and interactive visualizations.
- Seaborn (Version 0.11.2) - It is a python library which is used for data visualization. It is based on matplotlib library. It helps to make statistical graphics using python.
- Scikit-learn (sklearn) (Version 1.0.2) - It is a python library which is used for machine learning. It is largely written in python.

4. RESULTS:

Researchers will compare which machine learning algorithm is providing the most accurate value of AQI after predicting the AQI using several machine learning algorithms. The figures in the table below indicate whether the air quality at that place is very good, good, fair, etc.



AQI	Description	Health advice [hide]
0–33	Very Good	Enjoy activities
34–66	Good	Enjoy activities
67–99	Fair	People unusually sensitive to air pollution: Plan strenuous outdoor activities when air quality is better
100–149	Poor	Sensitive Groups: Cut back or reschedule strenuous outdoor activities
150–200	Very Poor	Sensitive groups: Avoid strenuous outdoor activities. Everyone: Cut back or reschedule strenuous outdoor activities
200+	Hazardous	Sensitive groups: Avoid all outdoor physical activities. Everyone: Significantly cut back on outdoor physical activities

Fig.4.1 Comparison Table

Graphs :

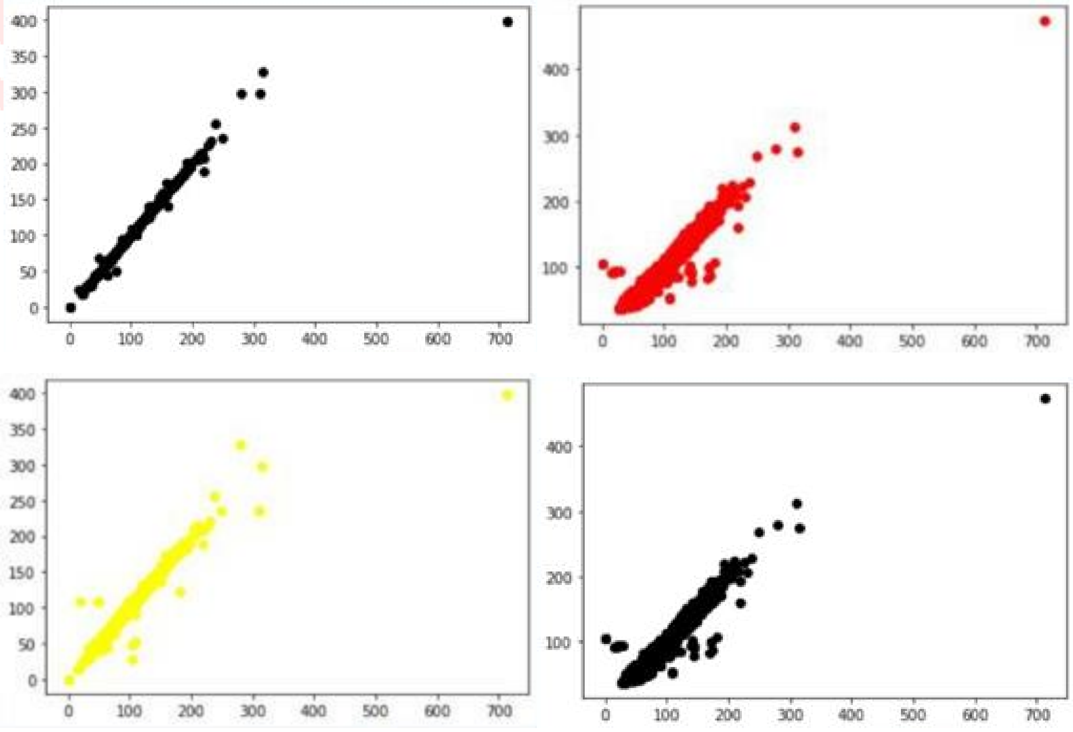
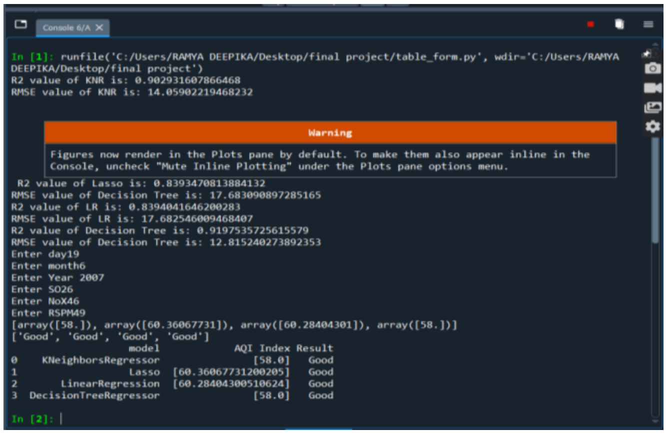


Fig.4.2 Algorithm Graphs

The model has Date, Month, Year, SO2, RMSE, and NOX values, and the result will be compared using the employed algorithms Decision tree, lasso regression, linear regression, and knn regression The output is shown in the image below.



```
In [1]: runfile('C:/Users/RAMYA DEEPIKA/Desktop/final project/table_form.py', wdir='C:/Users/RAMYA
DEEPIKA/Desktop/final project')
R2 value of KNR is: 0.902931607866468
RMSE value of KNR is: 14.05902219468232

        Figures now render in the Plots pane by default. To make them also appear inline in the
        Console, uncheck "Mute Inline Plotting" under the Plots pane options menu.
 R2 value of Lasso is: 0.8393470813884132
RMSE value of Decision Tree is: 17.683090897285165
R2 value of LR is: 0.8394041646200283
RMSE value of LR is: 17.682546009468407
R2 value of Decision Tree is: 0.9197535725615579
RMSE value of Decision Tree is: 12.815240273892353
Enter day19
Enter month6
Enter Year 2007
Enter SO26
Enter NoX46
Enter RSPM49
[array([58.]), array([60.36067731]), array([60.28404301]), array([58.])]
['Good', 'Good', 'Good', 'Good']
                   model            AQI Index Result
0    KNeighborsRegressor                 [58.0]   Good
1                  Lasso  [60.36067731200205]   Good
2       LinearRegression  [60.28404300510624]   Good
3  DecisionTreeRegressor                 [58.0]   Good

In [2]: 
```

Fig.4.3 Final output of comparison of machine learning algorithms.

5. CONCLUSION

The unpredictable and volatile nature of contaminants in space and time, as well as the dynamic environment, make it difficult to anticipate air quality. The quality of the air must be regularly monitored and analyzed, especially in developing nations, due to the serious effects that air pollution has on people, animals, plants, historical sites, the climate, and the environment. But scientists haven't given much thought to the AQI prediction for India. The current study examines air pollution data spanning six years from 23 Indian cities. By filling in NAN values, resolving outliers, and normalizing data values, the dataset is first prepared and cleaned.

Then, for further investigation, pollutants that have an effect on AQI are filtered using a correlation-based feature selection technique and logarithms are used to convert the skewed characteristics. Using exploratory data analysis approaches, latent patterns in the dataset are found in different places. Prior to being rigorously examined, incomplete record analysis, and model creation, the data was cleansed and processed.

The decision tree approach procedure shows outstanding accuracy on the public test set when measured against classification records. Depending on their ability to take action, this application may help India's meteorological division anticipate the future of air quality and its reputation. In this experiment, three machine learning algorithms and a deep learning algorithm were used to compare air pollution.

REFERENCES:

[1] S. Simu et al., "Air Pollution Prediction using Machine Learning," 2020 IEEE Bombay Section Signature Conference (IBSSC) and ,2020,pp.231-236,doi:10.1109/IBSSC51096.2020.9332184.

[2] Bekkar, A., Hssina, B., Douzi, S. et al. Air-pollution prediction in smart city and deep learning approach. J Big Data 8, 161 (2021). https://doi.org/10.1186/s40537-021-00548-1

[3] Shakir Muhammad Abdullah, 2021, To Predict Air Pollution using Machine Learning and Arima Model, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH AND TECHNOLOGY (IJERT) Volume 10, Issue 11 (November 2021)

[4] Sethi, Jasleen & Mittal, Mamta. (2021). Prediction of Air Quality Index Using Hybrid Machine Learning Algorithm 10.1007/978-981-15-5421-6_44.

[5] https://www.ijitee.org/wp-content/uploads/papers/v8i9S4/I11320789S419.pdf

[6] https://www.diva-portal.org/smash/get/diva2:1681590/FULLTEXT02

[7] https://www.nature.com/articles/s41598-021-00804-7#Sec1

[8] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00548-1#Sec25

[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9107909/

[10] https://www.researchgate.net/publication/334058882_Comparative_Analysis_of_Machine_Learning_Techniques_for_