IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING APPROACH FOR PREDICTING AND ANALYZING AIR QUALITY

[1]Darshan Gowda T, [2]G Sudheer Kumar, [3]Bompally Sriharsh,[4]Chandan S,[5]Mrs.Sushmitha S

[1,2,3,4]UG-Student, Department Of Computer Science And Engineering, DSATM, Bangalore, India

[5]Assistant Professor, Department Of Computer Science And Engineering, DSATM, Bangalore, India

Abstract: India has gained international attention as a result of its air pollution. The nation's rising air pollution has been one of the key concerns for both the government and the populace. We provide a comprehensive and useful approach for resolving the Air Quality Analysis in this study using machine learning techniques. This article uses information from numerous meteorological stations to calculate the Air Quality Index (AQI) for India. On the gathered data, we are applying machine learning methods including support vector machines, back propagation neural networks, decision trees, and linear regression (SVM). To assess accuracy of many models, we used the root mean square (RMS) approach.

***Index Terms* - Linear Regression (LR), Random Forest Regression (RF), Back Propagation Neural Network (BPN), Decision Tree Regression (DTR), Root Mean Square Error (RMSE), Support Vector Machine (SVM), Mean Absolute Error (MAE).**

I. INTRODUCTION

India presently ranks second in the world for pollution levels. When comparing to what would occur if pollution was lowered to satisfy the country's official standard, air pollution lowers the typical Indian's survival rate by 6.3 years. In Delhi and the surrounding areas, air pollution shortens lives by more than two decades, making some parts of India far less fortunate than the rest of the country.

India's 1.4 billion inhabitants live in areas where yearly particle pollution levels on average are higher than recommended by the WHO. In areas where the national standard for air quality in India is surpassed, 94% of people live. Fine particulate pollution has significantly increased over time. The average annual increase in particle pollution of 22% since 1998 has resulted in a 1.3-year reduction in the lifespan of the average inhabitant. There are 248 million people in northern India, which makes up 5% of the global population, who might lose more than 8 years of life expectancy if pollution levels don't go down.

The most polluted city in the world is Lucknow, the capital of the northern Indian state of Uttar Pradesh, where the pollution level is 13 times greater than the WHO recommendation. Residents of Lucknow might live 12.1 years shorter than projected if pollution persists. India's capital, Delhi, has a significant amount of air pollution. Residents' lives would be prolonged by 13 years if Delhi's pollution levels dropped to the WHO guideline level, and by 10 years if they achieved India's national norm.

The recommended feature selection and analysis technique highlight the importance of various input features to neural network predictions, and as a result, it may be able to shed light on some of the internal workings of the deep black-box models. This technique may be used for many different purposes, including the diagnosis of diseases and the detection of terrorism, in addition to the prevention and management of air pollution. In order to avoid tragic outcomes, it is difficult to act in line with a model's predictions without first verifying whether the black-box model can be trusted.

II. LITERATURE SURVEY

[1] In this research, a machine learning-based system is introduced that can use historical pollution data to predict future pollution data using a device that can accept current pollution as input. The detected data is stored for further analysis inside an Excel sheet. On the Arduino Uno platform, these sensors are used to gather information on pollutants. This study offers three prediction algorithms derived from three current areas of machine learning theory and application.

[2] The study combines PCA (known as principal component analysis) in conjunction with bp neural networks and neural networks based on genetic algorithm optimization to estimate Shanghai's (location in China) AQI (air quality index). Matlab is used for simulation and modelling. The prediction and analysis approach employs various error levels and iterations. The results show that, when compared to the PCA and bp neural network combination, that neural network optimized by the genetic algorithm may possibly minimize the prediction error of the air quality index, resulting in a prediction accuracy rate of 90.7%, considerably improving the neural network. In terms of anticipating Shanghai's air quality, the learning effectiveness performs well.

[3]. Based on pollution and weather data in New Delhi, India, this study examines how well different prediction models are able to forecast the AQI values given specific input data. Regression analysis is done on the dataset, and the findings demonstrate which meteorological variables have a greater impact on AQI levels as well as how helpful predictive models are for projecting air quality.

[4] This study attempted to achieve relevant analysis and short-term prediction of pm2.5 concentrations in Beijing, China, using data mining from multiple sources. Based on multivariate statistical analysis, they built a correlation analysis model of PM2.5 using both physical media data and social media data. According to this study, there is a strong mathematical relationship between PM2.5(<2.5microns) concentration and average wind speed, CO, NO2, PM10(<10microns), and the number of microblogging comments per day. We further explored the correlation study using a backpropagation neural network model for big data machine learning. The BPNN approach has been found to be more effective in correlation mining. In this article, they used an autoregressive integrated moving time series model to examine PM2.5 forecasts in short-term time series.

[5] The study uses a variety of machine learning methods to forecast the AQI, which is used to reduce pollution and lessen serious health risks. The air pollution quality is indicated by the air quality index. Particulate particles, nitrous dioxide (NO2), sulphur dioxide (SO2), and (CO) carbon monoxide are the main pollutants (CO). The air quality is forecasted using earlier approaches like probability and statistics, however, these techniques are exceedingly difficult to predict. To get around problems with earlier methods, machine learning algorithms provide a superior method of estimating air pollution levels. Random forest regression, support vector regression, and linear regression are examples of different machine learning techniques. The RMSE method is used to assess the accuracy of various models.

III. METHODOLOGY

Our Proposed methodology can be divided into seven (7) steps. The first step includes the collection of the dataset. Secondly, we are pre-processing datasets using various techniques. Next, in the third step, we are applying feature extraction to the pre-processed data. In step four, we are dividing the dataset into the training and the test data. Next, we apply machine learning algorithms for training data and train our models. In step 6, we predict the output based on our testing data. Lastly, for the analysis step, we are using the Root mean square error (RMS) technique for checking the accuracy of our models.

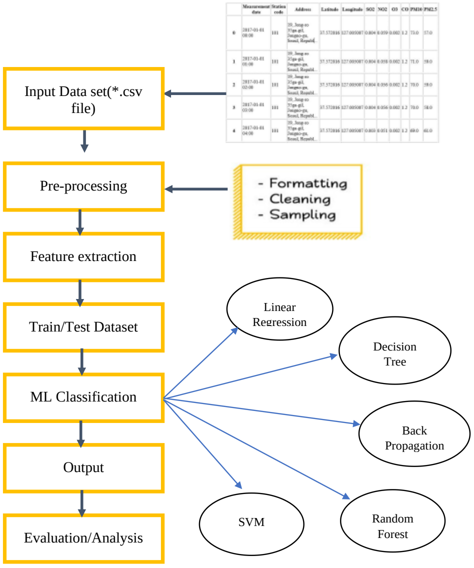
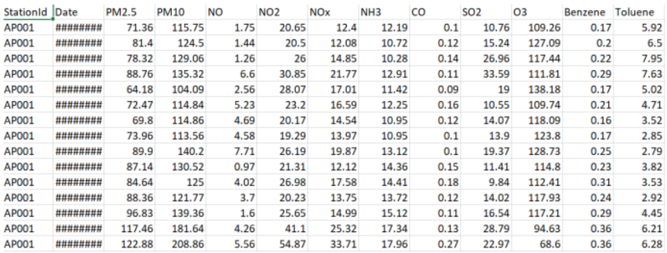


fig 1. flow chart of air quality index

i. Input Dataset:

The dataset is the combined data for Historical Daily Ambient Air Quality. The dataset contains the following features: stn_code, state, location, so2, no2, Location monitoring station, pm2_5 (<2.5 microns particle), pm10 (<10 microns particle), no, co, Benzene, Toluene.



StationId	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene
AP001	########	71.36	115.75	1.75	20.65	12.4	12.19	0.1	10.76	109.26	0.17	5.92
AP001	########	81.4	124.5	1.44	20.5	12.08	10.72	0.12	15.24	127.09	0.2	6.5
AP001	########	78.32	129.06	1.26	26	14.85	10.28	0.14	26.96	117.44	0.22	7.95
AP001	########	88.76	135.32	6.6	30.85	21.77	12.91	0.11	33.59	111.81	0.29	7.63
AP001	########	64.18	104.09	2.56	28.07	17.01	11.42	0.09	19	138.18	0.17	5.02
AP001	########	72.47	114.84	5.23	23.2	16.59	12.25	0.16	10.55	109.74	0.21	4.71
AP001	########	69.8	114.86	4.69	20.17	14.54	10.95	0.12	14.07	118.09	0.16	3.52
AP001	########	73.96	113.56	4.58	19.29	13.97	10.95	0.1	13.9	123.8	0.17	2.85
AP001	########	89.9	140.2	7.71	26.19	19.87	13.12	0.1	19.37	128.73	0.25	2.79
AP001	########	87.14	130.52	0.97	21.31	12.12	14.36	0.15	11.41	114.8	0.23	3.82
AP001	########	84.64	125	4.02	26.98	17.58	14.41	0.18	9.84	112.41	0.31	3.53
AP001	########	88.36	121.77	3.7	20.23	13.75	13.72	0.12	14.02	117.93	0.24	2.92
AP001	########	96.83	139.36	1.6	25.65	14.99	15.12	0.11	16.54	117.21	0.29	4.45
AP001	########	117.46	181.64	4.26	41.1	25.32	17.34	0.13	28.79	94.63	0.36	6.21
AP001	########	122.88	208.86	5.56	54.87	33.71	17.96	0.27	22.97	68.6	0.36	6.28

fig 2. input dataset

ii. Data Pre processing:

Data preprocessing is the second and most important stage in creating accurate ML models is data preprocessing. The crucial phase in data mining is data pre-processing, which identifies missing important values, inconsistencies, noise, errors, and outliers.

Managing outliers: Outliers are data patterns that drastically deviate from the data's normal variance. This may result in incorrect scientific findings and unreliable predictions.

Univariate method: Having this approach, data points with extreme values of a variable are sought after. In a univariate analysis, every variable in the dataset is looked at separately. Look at the values' central tendency as well as the values' range. describes how people react to different stimuli.

Missing Values: Because of malfunctions at the measurement stations or other external factors, air quality data frequently include missing values. Missing values can be handled by some algorithms (Naive Bayes, Classification and Regression Tree, CN2, etc.), whereas they must be replaced or eliminated by others.

Interpolation methods: Interpolation is a method of constructing a fundamental function from a discrete data set such that the function traverses the given data points. This makes it easier to identify the data points between the given data points.

iii. Feature Extraction:

A feature is information that can be used to solve a computational issue specific to an application. By generating new features from existing features, feature extraction tries to reduce the amount of features in a dataset.

iv. Train/Test Dataset:

Training and training are two ways to gauge how accurate your model is. He divides the dataset into two sets; thus, train/test.

A test set and a training set. Splitting the dataset is also necessary to check for underfitting or overfitting, two highly prevalent issues that might affect models.

v. ML Classification:

Here we are mainly applying five machine learning algorithms. Each of these is briefly described below:

a) *Linear Regression*: This method of performing predictive analytics is statistical. Continous, real, or numerical variables are predicted by linear regression. To forecast air quality, one uses the R-squared value. The amount of variation in the dependent variable of a system that can be explained by the independent variables is determined by R-Squared.

b) *Decision Tree Regression*: It keeps track of an object's characteristics, builds a model based on the tree's structure to forecast future dates, and generates actionable, continuous output. DT designs have a root system, branches, and leaves that resemble a tree. For instance, several predictors (X) that offer categorical AQI on a daily or monthly basis predict AQI as response (Y).

c) *Random Forest:*It is a classifier that uses numerous decision trees for various subsets of a given dataset and averages them to increase the dataset's predictive accuracy. Random forests use predictions from each tree to forecast the ultimate result based on the predictions' majority vote rather than relying on decision trees.

d) *BP neural network algorithm*: A classical neural network is the bp neural network. The main justification for using bp neural networks to anticipate air quality is their non-linear mapping. A three-layer neural network can simulate nonlinear continuous functions with high accuracy. It is perfect for handling different pollution problems related to the Air Quality Index.

e) *Support Vector Machine (SVM)*:Support vector machines are supervised machine learning techniques that are applied to both regression and classification tasks. It is most appropriate for classification and is also referred to as a regression problem. The SVM algorithm seeks to identify an N-dimensional hyperplane that categorises the data points in a single way. The number of features determines the hyperplane's size.

vi. Output:

Using machine learning approaches, we attempt to estimate the concentrations of all pollutants, including PM2.5, CO, NO, Benzene, Toluene, NO2, and SO2. Assess the training data to predict whether the area or location is safe from pollution by determining whether the air quality is bad, moderate, or unsafe. Correlations between pollutant values and meteorological data factors were discovered by studying historical datasets. The end result is a formula that may be applied to forecast contamination levels in the future.

vii. Evaluation/ Analysis:

We attempt to assess our findings using test data in this step. Evaluation is crucial for determining how accurate our project is. Try to increase accuracy by applying several strategies, such as handling missing and outlier values, feature engineering, or utilizing numerous algorithms. Performance measures that demonstrate how effectively a model can forecast the outcomes are used to gauge how accurate the models are. We are utilizing assessment metrics such as R2 score, MAE, and RMSE to evaluate the correctness of the results.

R2 Score: R2 Score: On a scale from 0 to 1, the R2 score (called "R-Squared Score") indicates how effectively our model is performing all of its predictions. The R2 number, on the other hand, may be used to assess the model's correctness in terms of distances or residuals.

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

Mean Absolute Error (MAE): This is simply the sum of all distances and residuals divided by the total number of points in the dataset. Our model predicted accurate average distances.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Root Mean Square Error (RMSE): RMSE is defined as the square root of all distances divided by the total number of points. With a few differences, RMSE serves a similar purpose as MAE (that is, use it to assess how close predictions are to the actual numbers on average).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

IV. DATA VISUALIZATION

We utilize the Python package matplotlib to visualize the dataset. There are several plot types available in matplotlib. Charts help you understand patterns, trends, and relationships. They are typically tools for inferring quantitative data. The various visualizations in our project are:

- ***Plotting highest and lowest ranking states:*** For this result, we are using a bar plot, taking each pollutant on Y-axis versus Indian states on the X-axis.
- ***Plotting the highest ever recorded levels***: Here we are plotting the bar graph taking the highest value of each pollutant in Y-axis and states in the X-axis.
- ***Plotting yearly trends:*** Here we are plotting yearly trends of each pollutant of each state of India. We are using the point-plot graph available in matplotlib. Taking Y-axis with each pollutant and X-axis with the year.
- ***Plotting a heatmap for a particular indicator:*** Here we plot a heatmap for the yearly median average for a given indicator. Taking Y-axis with the states and X-axis with the year. We are using the heatmap technique available in matplotlib.
- ***Plotting pollutant averages by locations/state:*** Here we are plotting the bar graph of a pollutant average for a given indicator by locations in a given state.

V. CONCLUSION

It has been demonstrated that ML-based AQI prediction models are more dependable and consistent. Data collecting was made simple and accurate by modern technology and sensors. Only Machine learning algorithms can effectively analyze the voluminous environmental data needed to make accurate and trustworthy forecasts. In this study, we employ a variety of regression approaches to identify the most suitable and effective algorithm for predicting air quality. Additionally, we attempt to determine each model's correctness using a variety of assessment criteria, such as the R2 score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

REFERENCES

[1] Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan , Srikanth, Hari Kiran Reddy,in year 2020 “Air Quality Prediction Of Data Log By Machine Learning”.

[2] Ruijun Yang ,Xueqi Hu, Lijun He, “Prediction of Shanghai air quality index based on BP neural network optimized by genetic algorithm”(2020)

[3] Soubhik Mahanta, T. Ramakrishnudu, Rajat Raj, Jha Niraj Tailor, : Urban Air Quality Prediction Using Regression Analysis (2019)

[4] 4. X.Y. Ni, H. Huang, W.P. Du : Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data.(2016)

[5] Karlapudi Saikiran, Gottapu Lithesh, Birru Srinivas, Prediction of Air Quality Index Using Supervised Machine Learning Algorithms (2021)

[6] Elias Kalapanidas and Nikolaos Avouris, Applying Machine Learning Techniques in Air Quality Prediction.

[7] Ioannis N. Athanasiadis, Air quality assessment using Fuzzy Lattice Reasoning (FLR).

[8] Giorgio Corani, Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning

[9] Duong, Dat Q., Quang M. Le, Tan-Loc Nguyen-Tai, Dong Bo, Dat Nguyen, Minh-Son Dao, and Binh T. Nguyen. "Multi-source Machine Learning for AQI Estimation." In 2020 IEEE International Conference on Big Data (Big Data), pp. 4567-4576. IEEE, 2020.

[10] Amado, Timothy M., and Jennifer C. Dela Cruz. "Development of machine learningbased predictive models for air quality monitoring and characterization." In TENCON 2018-2018 IEEE Region 10 Conference, pp. 0668-0672. IEEE, 2018.

[11] Mahalingam, Usha, Kirthiga Elangovan, Himanshu Dobhal, Chocko Valliappa, Sindhu Shrestha, and Giriprasad Kedam. "A machine learning model for air quality prediction for smart cities." In 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), pp. 452-457. IEEE, 2019.