AG-3

by Vamshikrishna Namani

Submission date: 08-Mar-2023 05:34PM (UTC+1000)
Submission ID: 2031955603
File name: conference_paper.docx (1.06M)
Word count: 1595
Character count: 9258

Air Quality Index Analysis using Machine Learning

Maddula Venkata Jyoshna
Student
Department of Computer Science and Engineering
Narasaraopet, India
jyoshnamaddula7@gmail.com

Chandana Chala
Student
Department of Computer Science and Engineering
Narasaraopet, India
Chandanachandu2043@gmail.com

Vankayala Sri Mallika
Student
Department of Computer Science and Engineering
Narasaraopet, India
Vsrimallika2000@gmail.com

M.Venkata Rao
Asst.Professor,M.Tech
Department of Computer Science and Engineering
Narasaraopet, India
jpnecnrt@gmail.com

Abstract - Air pollution is a major issue in today's world, caused by the release of hazardous gases into the atmosphere from industries, vehicles, and other sources. To maintain good air quality, this mechanism measures various air toxins in different areas. However, the pollution level in all cities has exceeded the air quality index value set by the government, which significantly affects human health.Thankfully, machine learning (ML) research has advanced to the point where it is now able to forecast contaminants using historical data. This study describes a gadget that can measure current pollution levels and run an ML-based algorithm for estimating future pollution levels using historical pollution data.

Keywords—MachineLearning, Random Forest, Gaussain Naïve Bayes, Decision tree, Linear Regression.

I. INTRODUCTION

Air pollution monitoring is crucial in today's world, as it has a significant impact on both human health and the environment. Harmful emissions not only affect the environment, but also impact the productivity and efficiency of individuals. Thus, continuous monitoring is necessary to effectively control and mitigate air pollution . It is impossible to overlook the effects of climate change on human health., as it has been linked to numerous adverse health effects. Furthermore, air pollution can also negatively impact the environment and its delicate balance. Thus, effective monitoring of air pollution levels is essential.

Given the dangers of air pollution, it is important to monitor and control its levels to minimize its harmful effects. This can be achieved through continuous monitoring and effective measures to mitigate pollution levels. By implementing these measures, we can work towards maintaining a healthier environment and promoting human well-being. It's critical to pinpoint the cause, extent, and origin of air pollution in order to control it. The state government's environmental department typically observes pollution levels by tracking the concentration of toxic gases in various regions. The World Health Organization (WHO) also provides data on pollution levels in the country, which highlights the urgent need for air monitoring.

Monitoring air pollution has become increasingly critical due to the rising levels of pollution. Air tracking has grown to be a significant task for measuring continuous levels of air contaminants in the environment. It is essential to monitor air pollution levels regularly to take appropriate measures and control its impact on human health and the environment.

II. LITERATURE SURVEY

The public is informed of the degree of polluted air in such a specific location using the index of air quality (AQI) . The AQI has been widely used by governments and organizations worldwide as a tool for air quality management and public health protection. In recent years, many researchers have

focused on developing and improving the AQI system to make it more accurate and useful for the general public.

One study by Kaur and Bhangra (2021) reviewed the literature on AQI and its application in India. The authors found that AQI is an important tool for air quality management in India, which is facing severe air pollution problems. The study also highlighted the need for public awareness and education regarding the AQI system to improve its effectiveness.

Another study by Zhang et al. (2020) proposed a new AQI system based on machine learning algorithms. The authors used a combination of multiple linear regression and random forest algorithms to predict AQI values based on meteorological and air pollutant data.

The proposed AQI system was found to be effective in communicating air pollution levels to the public and in guiding air quality management efforts. Overall, the literature suggests that AQI is an important tool for air quality management and public health protection. Improvements in AQI systems, such as those proposed in recent studies, can help to make the AQI more accurate and effective in communicating air pollution levels to the public. It has been discovered that linear regression models are suitable for estimating or predicting pollution. While SVM-based and neural network approaches are chosen for forecasting pollution levels.

III. EXISTING SYSTEM

A gap in the literature was identified as previous papers only focused on predicting PM2.5 levels using ML algorithms. However, this project aims to predict levels of all pollutants including CO, O3, NO2, SO2, PM2.5, and PM10 by utilizing meteorological data for improved prediction accuracy. These systems typically assign an AQI value on a scale from 0 to 500, where lower air quality is indicated by greater readings.

IV. METHODOLOGY

In extremely polluted locations, the central environmental control board has built a number of pollution monitoring stations. This monitoring station data are used for research and further evaluation.

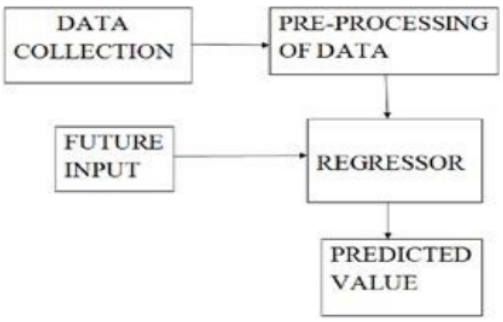

Fig. 1: Proposed system work flow

Data Collection

The Data set is collected from the Kaggle website: http://www.kaggle.com/code/sharmamanali/air-quality

The dataset contains data related to the air gases values to predict the air quality. The dataset size is 4,80,534 rows and 11 columns namely so2,no2,spm,rspm.

stn_code	sampling_	state	location	agency	type	so2	no2	rspm	spm	location_c	pm2_5
	150	February -	Andhra Pra	Hyderabac	NA	Residentia	4.8	17.4	NA	NA	NA	NA
	151	February -	Andhra Pra	Hyderabac	NA	Industrial /	3.1	7	NA	NA	NA	NA
	152	February -	Andhra Pra	Hyderabac	NA	Residentia	6.2	28.5	NA	NA	NA	NA
	150	March - M	Andhra Pra	Hyderabac	NA	Residentia	6.3	14.7	NA	NA	NA	NA
	151	March - M	Andhra Pra	Hyderabac	NA	Industrial /	4.7	7.5	NA	NA	NA	NA
	152	March - M	Andhra Pra	Hyderabac	NA	Residentia	6.4	25.7	NA	NA	NA	NA
	150	April - M0	Andhra Pra	Hyderabac	NA	Residentia	5.4	17.1	NA	NA	NA	NA
	151	April - M0	Andhra Pra	Hyderabac	NA	Industrial /	4.7	8.7	NA	NA	NA	NA
	152	April - M0	Andhra Pra	Hyderabac	NA	Residentia	4.2	23	NA	NA	NA	NA
	151	May - M0	Andhra Pra	Hyderabac	NA	Industrial /	4	8.9	NA	NA	NA	NA
	152	May - M0	Andhra Pra	Hyderabac	NA	Residentia	3.6	18.6	NA	NA	NA	NA
	150	June - M0	Andhra Pra	Hyderabac	NA	Residentia	3.9	14.1	NA		133	NA
	151	June - M0	Andhra Pra	Hyderabac	NA	Industrial /	5.6	11.8	NA		82	NA
	152	June - M0	Andhra Pra	Hyderabac	NA	Residentia	3.3	19.3	NA		111	NA
	150	July - M07	Andhra Pra	Hyderabac	NA	Residentia	3.9	8.2	NA		118	NA
	152	July - M07	Andhra Pra	Hyderabac	NA	Residentia	3.5	12.1	NA		135	NA
	151	July - M07	Andhra Pra	Hyderabac	NA	Industrial /	7.9	10.2	NA		80	NA
	150	August - M	Andhra Pra	Hyderabac	NA	Residentia	4	9.9	NA		179	NA
	151	August - M	Andhra Pra	Hyderabac	NA	Industrial /	12.4	11.5	NA		58	NA
	152	August - M	Andhra Pra	Hyderabac	NA	Residentia	4	12.3	NA		99	NA
	150	September	Andhra Pra	Hyderabac	NA	Residentia	6.3	11.5	NA		270	NA
	151	September	Andhra Pra	Hyderabac	NA	Industrial /	44.8	13.7	NA		97	NA
	152	September	Andhra Pra	Hyderabac	NA	Residentia	8.1	17.8	NA		167	NA
	150	October -	Andhra Pra	Hyderabac	NA	Residentia	7.7	11.3	NA		145	NA
	151	October -	Andhra Pra	Hyderabac	NA	Industrial /	20.6	13.6	NA		75	NA
	152	October -	Andhra Pra	Hyderabac	NA	Residentia	20.4	27.5	NA		212	NA
	150	November	Andhra Pra	Hyderabac	NA	Residentia	13.9	7.2	NA		93	NA

Fig. 2: Dataset diagram

A. *Data Cleaning and Feature Extraction*

Before giving the data to any machine learning algorithms the data must be cleaned. In data cleaning process null values and outliers are removed. In the collected dataset no null values are present. After data cleaning ,features are extracted from the data set. In dataset we have 6 different features which affects our model output and we have to check the correlation among the features.

	stn_code	so2	no2	rspm	spm	pm2_5	Soi	Noi	Rpi	SPMi	AQI
stn_code	1.000000	-0.074017	-0.250173	-0.065792	-0.670591	NaN	-0.075818	-0.252804	NaN	-0.695725	-0.678123
so2	-0.074017	1.000000	0.429088	-0.007617	0.175886	NaN	0.992479	0.430772	NaN	0.165794	0.192352
no2	-0.250173	0.429088	1.000000	0.187069	0.351881	NaN	0.441215	0.999327	NaN	0.349513	0.412102
rspm	-0.065792	-0.007617	0.187069	1.000000	0.235284	NaN	-0.006148	0.189495	NaN	0.214176	0.250063
spm	-0.670591	0.175886	0.351881	0.235284	1.000000	NaN	0.180506	0.355006	NaN	0.992247	0.988764
pm2_5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Soi	-0.075818	0.992479	0.441215	-0.006148	0.180506	NaN	1.000000	0.443056	NaN	0.170239	0.196961
Noi	-0.252804	0.430772	0.999327	0.189495	0.355006	NaN	0.443056	1.000000	NaN	0.352611	0.415356
Rpi	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
SPMi	-0.695725	0.165794	0.349513	0.214176	0.992247	NaN	0.170239	0.352611	NaN	1.000000	0.993589
AQI	-0.678123	0.192352	0.412102	0.250063	0.988764	NaN	0.196961	0.415356	NaN	0.993589	1.000000

Fig. 3:Correlation

Correlation helpful to remove the features which shows negligible affect on the model output. Fig 3 shows the correlation between the features of the dataset.

B. Model Architecture

In this step dataset is splitted in 75% and 25% for training and testing. During training, machine learning algorithms finds the relation between the input and output features. By using this relation the model able to predict the outputs to the new input values. In this we used the following four machine learning models:

1. Random Forest: A machine learning approach called random forest is employed for both classification and regression applications. It is a member of the family of evolutionary algorithms, which combine a number of weak learners to produce a powerful learner. In a random forest, multiple decision trees are built on randomly selected subsets of the data and features. Each tree is expanded to its maximum depth throughout the training process, and the algorithm chooses the optimum split from a random group of features at each node.

2. Decision Tree: Decision trees are highly interpretable, easy to understand, and can handle both categorical and numerical data. They can also handle high-dimensional data and are highly scalable. The risk of overfitting with decision trees increases with tree depth and noisy data, respectively. Techniques such as pruning, regularization, and ensemble methods such as random forests can help to overcome these limitations.

3. Linear Regression: A statistical technique used to evaluate the connection between two parameters is linear regression,where one variable is dependent on the other. It involves finding a linear equation that best describes the relationship between the variables. The premise underlying the linear regression technique is that there exists a linear connection among the variables.

V. RESULT AND ANALYSIS

After created all the four models by using the training data , we test the models by using the test data and calculates the accuracies of the models. The following table shows the accuracies of different models:

Type of Models	Accuracy
Linear Regression	98.91
Decision Tree	99.95
Naïve Base	95.00
Random Forest	99.96

Fig.4: Accuracy Table

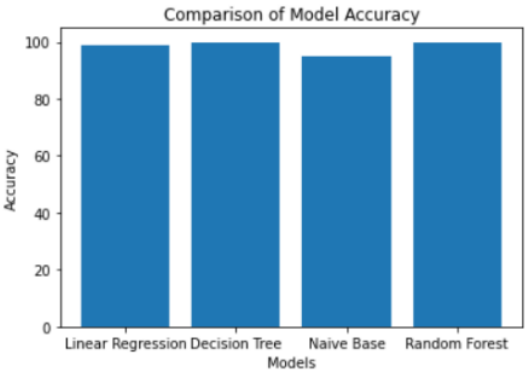

Fig.5: Accuracy Bar Chat

VI. DEPLOYMENT

The final model is deployed into an application by using Flask module of python by creating the user interface which can be accessible easily by the people. The application simply takes the input and provides output.

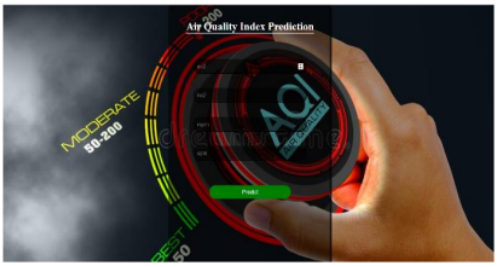

Fig.6 Input Screen

Fig .7 Output Screen

VII. CONCLUSION

By using unique formulas like necessarily imply, Decision Tree, and Random Forest, we forecast the air quality list. We concluded from the results that its Random Forest approach provides a superior expectation of the list of air quality.

This project has highlighted the benefits of Machine Learning in knowing its value of air quality index value, and how it can help improve its efficiency and effectiveness . By automating the recommendation process, farmers can save time and effort, while also reducing the risk of making costly mistakes.

VIII. REFERENCES

[1] https://en.wikipedia.org/wiki/Air_quality_index

[2]. Kennedy Okokpujie, Etinosa Noma-Osaghae, Odusami Modupe, Samuel John, and Oluga Oluwatosin, "A SMART AIR POLLUTION MONITORING SYSTEM," International Journal of Civil Engineering and Technology (IJCIET), vol. 9, no. 9, pp. 799–809, Sep. 2018.

[3]. Kostandina Veljanovska and Angel Dimoski, "Air Quality Index Prediction Using Simple Machine Learning Algorithms," International Journal of Emerging Trends & Technology in Computer Science, vol. 7, no. 1, 2018.

[4]. D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," Big Data and Cognitive Computing, vol. 2, no. 1, p. 5, Mar. 2018.

[5]. A. Masih, "Machine learning algorithms in air quality modeling," Global Journal of Environmental Science and Management, vol. 5, no. 4, pp. 515–534, 2019.

[6]. https://archive.ics.uci.edu/ml/datasets/Air+quality

[7]. Rokach,Lior,Maimom,O.(2008).Data mining with decision trees : theory and applications.World Scientific Pub Co Inc .ISBN 978-9812771711.

[8].BreimanL(2001)."RandomForests". MachineLearning.45(1) :32.doi:10.1023/A:1010933404324

Ritik Sharma, Gaurav Shilimkar, Shivam Pisal, " Air Quality Prediction by Machine Learning" ,International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN:2395-6011,Online ISSN:2395-602X,Volume 8,Issue 3,pp.486-492, May-June-2021.Available at doi:https://doi.org/10.32628/IJSRST218396

Journal URL:https://ijsrst.com/IJSRST218396

AG-3

ORIGINALITY REPORT

6%	3%	3%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth, Hari Kiran Reddy. "Air Quality Prediction Of Data Log By Machine Learning", 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020 Publication	2%
2	"Classification of Student Performance Dataset using Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering, 2019 Publication	1%
3	Submitted to Liverpool John Moores University Student Paper	1%
4	ijsrst.com Internet Source	1%
5	ijireeice.com Internet Source	1%
6	neuroquantology.com Internet Source	1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		