cp_db13.docx

by A a

Submission date: 20-Apr-2024 02:14AM (UTC-0400)
Submission ID: 2329597597
File name: cp_db13.docx (1.49M)
Word count: 2301
Character count: 13805

AIR QUALITY PREDICTION USING MACHINE LEARNING

Abstract: *The biggest threat to the environment and public health in the world today is air pollution. It can be characterized as one of the gravest threats to humankind in recorded history. It harms forests, crops, animals, and other things. The quantity of hazardous gases and particulate matter in the atmosphere, the discharge of toxic gases by industry, and automobile emissions are all contributing factors to air pollution. Human health is significantly impacted by air quality. The government can protect the most vulnerable from air pollution by taking the appropriate precautions thanks to its ability to forecast air quality. By integrating advanced technologies like Machine Learning, it has the potential to revolutionize environmental monitoring and contribute significantly to safeguard public health*

.

Keywords- Random Forest Algorithm, SO2 ,NO2 ,RSPM, SPM

I. INTRODUCTION

Air quality stands as a crucial aspect of public health and environmental well-being, yet it faces significant challenges due to industrial emissions and urbanization. These challenges often lead to harmful levels of pollutants in the air, posing risks to human health and the ecosystem.. [1].

To address these challenges, we have developed an innovative system for air quality monitoring and prediction. Key attributes such as sulphur dioxide concentration, nitrogen dioxide concentration, and particulate matter levels RSPM, SPM, PM2.5 are analyzed to provide accurate forecasts of air pollution levels [2]. By offering advance predictions, our system enables authorities and individuals to take proactive measures, mitigating the adverse effects of air pollution and safeguarding public health [3].

In addition to pollution prediction, our system integrates modules for environmental monitoring and pollutant analysis. The environmental monitoring module tracks air quality trends over time, facilitating informed decision-making for policy formulation and urban planning [3]. Furthermore, the pollutant analysis module identifies sources of pollution and evaluates their impact on air quality, aiding in the development of targeted interventions for pollution control.

The system serves as a valuable tool for environmental agencies, providing them with essential information and insights to enhance air quality management and protect public health. By integrating advanced technologies like Machine Learning, this system has the potential to revolutionize air quality management and significantly contribute to environmental sustainability.

II. LITERATURE SURVEY

Numerous studies have explored the application of machine learning in predicting air quality parameters. One study aimed to forecast air pollutant concentrations in urban areas, utilizing historical datasets to predict pollutant levels and their impacts [4]. Researchers employed various machine learning algorithms to anticipate pollutant concentrations and their variations over time. Another study focused on predicting air quality indices using data mining techniques based on factors such as meteorological conditions and emission sources [5], offering modules to aid policymakers in decision-making regarding air quality management strategies.

The process of building analytical models is automated by a data analysis technique known as machine learning..The suggested model would identify the air quality index of a few Indian cities using the Random Forest technique.[3].

The Random Forest classifier leverages numerous decision trees across various dataset subsets. An algorithm based on the random forest concept is suggested for an urban sensing system to forecast the air quality in urban areas by utilizing distribution data, historical traffic and road status data, meteorological data, and historical air quality data. These information is gathered from a variety of urban sensors, including sensors that monitor the weather.

To enhance the system's functionality, visualizing pollutant levels could enable authorities to observe pollution hotspots in nearby areas, while graphical representations of predicted air quality indices could improve comprehension [3].

The objective is to assist policymakers in anticipating air quality trends and identifying optimal strategies for pollution mitigation using methods such as Random Forest, and Decision Tree algorithms [6]. Various machine learning algorithms and technologies with user-friendly interfaces, such as Logistic Regression, Naive Bayes, Decision Tree, Random Forest, AdaBoost, KNN, GNB, XGBoost, and SVM, have been utilized to predict air quality.

Collaborating with environmental agencies and other relevant stakeholders could further 11ance models and support policymakers in implementing effective air quality management measures. Developing a framework for recommending pollution control measures and interventions could streamline decision-making processes, while a user-friendly interface could enhance the system's accessibility and usability for policymakers and stakeholders involved in air quality management.

III. METHODOLOGY

Fig1: Implementation Methodology

A. Data Collection and Preprocessing:

In this study, we obtained our data from Kaggle, a popular website where data scientists share datasets. After collecting the data, which comprised over 435743 records and 13 different pieces of information [8], we uploaded it to Google Colab, an online platform for analysing data and performing machine learning tasks. This diverse dataset provides a comprehensive foundation for our air quality prediction analysis.

To calculate the AQI, we first computed individual pollutant indices for SO2, NO2, RSPM, and SPM using specific formulas tailored to each pollutant's concentration levels. These indices were used to derive the overall AQI for each data point. Subsequently, we categorized the AQI values into different ranges—Good, Moderate, Poor, Unhealthy, Very Unhealthy, and Hazardous—based on predefined threshold values.

This approach enabled us to classify air quality levels accurately and providing valuable insights for decision-making processes. Through our analysis, we aimed to contribute to a better understanding of air quality dynamics and facilitate efforts towards improving air quality standards and public health.

Pre-processing of Dataset:

Before we could use the data, we had to clean it up by getting rid of any missing information and unusual values. This makes the data ready for training and testing our models.

Data Visualization:

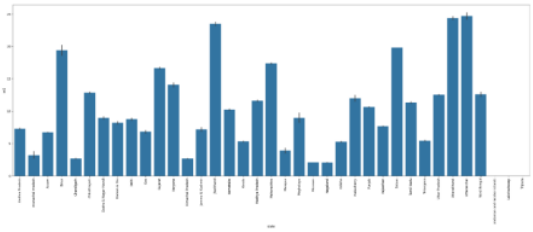

Fig 2: SO2 concentration for different states

We used graphs and charts to look at the data and see if there were any patterns. In below figure (2) we divided the dataset by states on one side and SO2 in different states.it was use analyzing average in every state.

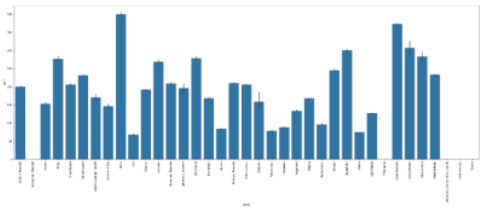

Fig 3: SPM levels in different states

We utilized graphs to visualize the dataset, specifically plotting the average SO2 levels across different states (Figure 3). This analysis helped us identify any notable patterns or variations in SO2 concentration levels among various states.

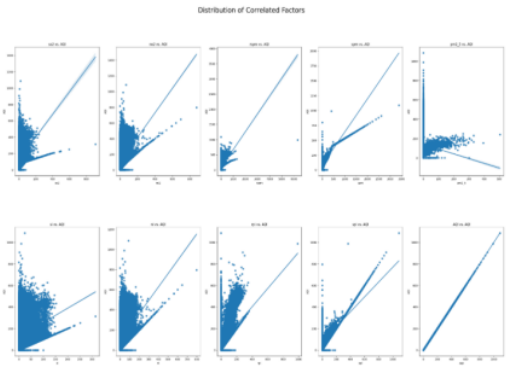

Fig 4: Distribution of Correlated Factors

We utilized regression plots to examine the distribution and correlation between important independent variables and the dependent variable, AQI (Air Quality Index). Each plot in the grid illustrates the relationship between an independent variable (e.g., SO2, NO2) and AQI, aiding in understanding their impact on air quality.

B. Model Assessment and Selection:

In To accurately measure the performance of our models, we first divided our data into training and testing sets throughout the model assessment and selection process. The first method we used was regression, and we used a number of different algorithms, including K-Nearest Neighbors, Lasso, Decision Tree, Random Forest, and XGBoost. Metrics including Mean Squared Error , R-squared , Mean Absolute Error , and Root Mean Squared Error were used to evaluate each model after it had been trained on the training set.

Decision Tree Regression and Random Forest Regression performed the best out of all of these models, obtaining remarkably low MSE and RMSE values that demonstrated strong predictive abilities.

.

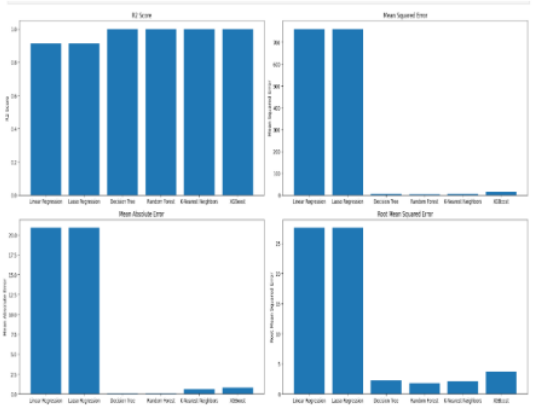

Fig 5: Analyzing the MSE, R^2 and MAE, RMSE

C. Evaluation Metrics of Classification Algorithms:

In assessing the performance of classification algorithms for predicting air quality index (AQI) ranges, several key metrics. These metrics provide a comprehensive understanding of the models' effectiveness in classifying AQI ranges correctly. Among the evaluated algorithms,

Decision Tree and Random Forest models exhibited exceptional accuracy and precision, achieving near-perfect classification of AQI ranges. Moreover, these models demonstrated high recall values, indicating their ability to correctly identify true positives while minimizing false negatives. Additionally, the F1 scores of both Decision Tree and Random Forest models were notably high, reflecting a balanced performance between precision and recall.

These evaluations underscore the importance of considering multiple metrics when assessing the performance of classification algorithms for air quality prediction. While accuracy and precision provide insights into overall classification correctness and positive prediction accuracy, recall and F1 score offer valuable information on the models' ability to correctly identify positive instances and maintain a balance between precision and recall, respectively.

By leveraging these metrics, stakeholders can make informed decisions regarding the selection and deployment of classification algorithms for air quality management and prediction initiatives.

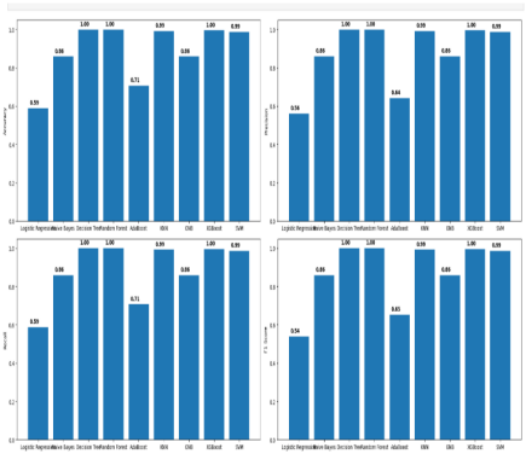

Fig 6: Accuracy's of different models

Table -1 Accuracy of different algorithms

Algorithms	Accuracy
Random Forest	0.985
Decision Tree	0.97
Naive Bayes	0.86
AdaBoost	0.71
Logistic Regression	0.59
KNN	0.95
GNB	0.86
SVM	0.97

. Random Forest emerged as the top performer with an accuracy of 0.985, demonstrating its effectiveness. Following Decision Tree algorithm with an accuracy of 0.97, indicating its suitability for the task. In contrast, the linear models all displayed similar accuracy scores which, although lower than the tree-based models, can still be valuable in certain contexts. Logistic Regression achieved an accuracy of 0.59, respectable but falling behind the ensemble methods. These results emphasize the effectiveness of ensemble methods for the task, particularly Random Forest and Decision Tree, while also highlighting the importance of considering other factors such as interpretability, computational efficiency, and scalability when choosing an algorithm for a specific application. Random Forest, with its impressive accuracy of 0.98, stands out as the top performer in his evaluation.

IV. RESULTS AND DISCUSSIONS

In this project, machine learning algorithms, particularly Random Forest, have transformed air quality prediction practices. The project focused on evaluating algorithms—KNN, Random Forest, Linear Regression, Ridge, Lasso, and Decision Tree—for predicting air quality, with Random Forest emerging as the most effective, boasting a 9[illegible] accuracy rate. This high accuracy is especially beneficial for predicting air quality in regions with severe pollution, providing valuable insights into previously unexplored pollution patterns.

Beyond air quality prediction, the project explored machine learning models' versatility in identifying key factors influencing pollution levels, recommending targeted interventions, and determining optimal pollution mitigation strategies. These applications streamline air quality management processes, empowering policymakers with tailored insights for mitigating pollution and safeguarding public health. The development of a user-friendly web application, with over 98% accuracy, demonstrates the reliability and effectiveness of machine learning in predicting air quality.

Fig 7: Prediction of Air Quality

We utilized Random Forest Algorithm to predict the air Quality by providing the input fields of SO2 , NO2, RSPM, SPM.

In this research we showcases the transformative potential of machine learning in predicting air pollution levels and identifying key factors influencing air quality, machine learning facilitates proactive measures to mitigate pollution and safeguard public health. By providing accurate predictions, it enables stakeholders to implement targeted interventions and policies aimed at improving air quality and promoting environmental sustainability.

V. CONCLUSION

This study concentrates on the prediction of the pollutants present in the air. The collected data was trained using machine learning algorithm and the predicted data has been tested. Thereafter selective algorithms are applied on the data and found the accurate ones for predicting the air quality. By predicting the air quality priorly we can take necessary steps to prevent ourselves from harmful health diseases.

If there is increased awareness about Air Quality Index India and it's health impacts depending on the various categories can help to reduce the incidence of air pollution to the most vulnerable people. Since acute exposure to acute exposure to air emissions may cause substantial harm to the health of the masses in general. Therefore, there are variables that can be taken to make people aware of the air-emission reports so that they can plan they're outdoor activities accordingly to reduce the intake of highly polluted. Every problem has a solution. And as it already created the big issue of Air pollution, it just can't be eradicated by only one's support but it needs the hands of many persons to reduce the pollutants.

REFERENCES

[1] J. Kotcher, et al, "Fossil fuels are harming our brains: identifying key messages about the health effects of air pollution from fossil fuels," BMC public health vol. 19, no.1, p. 1079, 2019.

[2] Khedo K.K., et al, A Wireless Sensor Network Air Pollution Monitoring System. Int. J. Wirel. Mob. Netw. 2010; 2:31–45,2019.

[3] K. Mahesh Babu, et al , "Air Quality Prediction based on Supervised Machine Learning Methods," International Journal of Innovative Technology and Exploring Engineering (IITEE) ISSN: 2278-3075, Volume-8, July 2019.

[4] Srinidhi jha, et al ," Air quality modelling using long short-term memory(LSTM) over NCT-Delhi, India," Springer Nature B.V, 2019.

[5] Alireza Rahimpour, et al , "Air quality data series estimation based on machine learning approaches for urban environments," Springer Nature B.V, 2020.

[6] Qjang Zhang, et al ,"A deep learning and image-based model for air quality estimation," Elsevier, 2020.

[7] Shrddha Sagar, et al, "Air Quality Prediction using Machine Learning Algorithms ICACCCN", 2020.

[8] Karlapudi Saikiran, et al, "Prediction of Air Quality Index Using Supervised Machine Learning Algorithms" , International Conference on Advance of Computing Communication and Embedded System(ACCESS), 2021.

[9] B.T.G.s.Kumara, "Machine Learning approach for predicting Air Quality Index", International Conference on Decision Aid Science and Application (DASA), 2021.

[10] Chenchen Li , "Research on Air Quality Prediction Based on Machine Learning", International Conference on ICHCI, 2021.

[11] Gopalakrishnan V Hyperlocal air quality prediction using machine learning towards data science, 2021.

[12] Liang Y, Maimury Y, Chen AH, Josue RCJ Machine learning-based prediction of air quality. Appl Sci 10(9151):1–17, 2020.

[13] Rybarczyk Y, Zalakeviciute R Assessing the COVID-19 impact on air quality: a machine learning approach. Geophys Res Lett, 2021.

[14] Madan T, Sagar S, Virmani D Air quality prediction using machine learning algorithms–a review. In: 2nd international conference on advances in computing, communication control and networking (ICACCCN) pp 140–145, 2020.

[15] Mahalingam U, Elangovan K, Dobhal H, Valliappa C, Shrestha S, Kedam G. 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, 2019.

cp_db13.docx

ORIGINALITY REPORT

9%	5%	5%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	Submitted to Universita' La Sapienza Student Paper	1%
3	Submitted to University of Cincinnati Student Paper	1%
4	repozitorij.uni-lj.si Internet Source	1%
5	www.sciepub.com Internet Source	1%
6	Wallace Duarte de Holanda, Lenardo Chaves e Silva, Álvaro Alvares de Carvalho César Sobrinho. "Machine learning models for predicting hospitalization and mortality risks of COVID-19 patients", Expert Systems with Applications, 2024 Publication	1%
7	courses.cs.ut.ee Internet Source	1%

8	Kambhampati Teja, Ruhul Amin Mozumder, Nirban Laskar. "Forecasting the impact of meteorological parameters on air pollutants in Andhra Pradesh using machine learning techniques", Environmental Quality Management, 2023 Publication	<1 %
9	japh.tums.ac.ir Internet Source	<1 %
10	link.springer.com Internet Source	<1 %
11	Lorraine Craig, Jeffrey R. Brook, Quentin Chiotti, Bart Croes et al. "Air Pollution and Public Health: A Guidance Document for Risk Managers", Journal of Toxicology and Environmental Health, Part A, 2008 Publication	<1 %
12	Binzhe Zhang, Min Duan, Yufan Sun, Yatong Lyu, Yali Hou, Tao Tan. "Air Quality Index Prediction in Six Major Chinese Urban Agglomerations: A Comparative Study of Single Machine Learning Model, Ensemble Model, and Hybrid Model", Atmosphere, 2023 Publication	<1 %
13	Reema Gupta, Priti Singla. "Chapter 23 Predictive Analysis of Air Pollutants Using	<1 %

Machine Learning", Springer Science and Business Media LLC, 2023
Publication

14 Thomas M. T. Lei, Shirley W. I. Siu, Joana Monjardino, Luisa Mendes, Francisco Ferreira. "Using Machine Learning Methods to Forecast Air Quality: A Case Study in Macao", Atmosphere, 2022 <1%
Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		