AIR QUALITY PREDICTION USING ML

Mr. A N Sasikumar
Department of Computer Science
Panimalar Engineering College
Chennai, India
ansasikumar@gmail.com

Karthik D
Department of Computer Science
Panimalar Engineering College
Chennai, India
karthikdeivanayagam15@gmail.com

Lalith Kishore L
Department of Computer Science
Panimalar Engineering College
Chennai, India
llk.22012002 @gmail.com

Maniyarasan M
Department of Computer Science
Panimalar Engineering College
Chennai, India
maniyarasan9842@gmail.com

ABSTRACT: The quality of air is a crucial aspect of public health, and accurate air quality prediction is essential for monitoring and controlling air pollution. In this project, we suggest an automated air quality monitoring system prediction system that uses historical and real-time data to predict air quality in a specific location. The system uses a combination of feature engineering and machine learning algorithms, such as regression and time-series forecasting, to analyze and predict air quality. The system is designed to be user-friendly and efficient, with an intuitive interface that enables users to access air quality predictions in real-time. A sizable amount of data is used to train the system historical air quality data and real-time data, including weather data, traffic data, and pollutant emissions data, to accurately predict air quality. The performance of the system is evaluated using various metrics, including root mean square error and coefficient of determination. The results show that the proposed air quality prediction system achieves high accuracy and efficiency in predicting air quality, enabling users to monitor and control air pollution effectively. In conclusion, the proposed air quality prediction system using machine learning algorithms has the potential to revolutionize air quality monitoring and control. The system is a useful tool for public health workers and policymakers since it is precise, effective, and user-friendly. Future research in this field should focus on improving the performance of the system, increasing its accessibility, and exploring its potential for other applications in environmental monitoring, and command.

Keywords: air quality, prediction, machine learning, deep learning, time series analysis, atmospheric pollutants, feature selection, regression, classification, ensemble learning.

1. INTRODUCTION

Air contamination observing has acquired consideration these days as it significantly has an impact on people's well-being just like on the biological equilibrium. Other than because of the impacts of harmful emanations on the climate, wellbeing, work usefulness and effectiveness of energy are additionally influenced by the air contamination. Since air contamination has caused numerous perilous consequences for people it ought to be checked persistently with the goal that it tends to be effectively controlled. One of the approaches 5to control air contamination is to understand the force, the source, and the beginning. Typically, it is checked by the individual express government's current circumstance service. They maintain a string of toxic gases in each

region. The WHO has released evidence that raises concerns about the extent of contamination across the nation. It makes the opportunity clear to us has already come and gone that we should screen the atmosphere. A method of measuring ambient levels of air contaminants is air tracking. As air pollution has been rising every day, monitoring has grown to be a significant task. We can determine the level of pollution in a location by continuously monitoring the air pollution there. We can learn about the source and severity of the contaminants in that area from the data the gadget collects. With the use of that knowledge, we can take actions or try to lessen pollution levels so that we can breathe clean air. Both human health and the natural balance are impacted by air pollution. Gas concentrations in the air have a significant impact on human health and can have dangerous consequences. Due to an increase in contaminants in the air, air pollution also has an impact on seasonal rainfall. The amount of rain is also impacted. As a result, continual air monitoring is required.

2. LITERATURE SURVEY

[1] An IoT-based system for tracking and forecasting air pollution is suggested in this research. This system can track air contaminants in a particular location, analyses air quality, and predict air quality. Utilizing IoT and the machine learning method known as Recurrent Neural Network, specifically Long Short-Term Memory, the suggested system will monitor air contaminants. (LSTM).

[2] In this study, Saba Ameer used four advanced regression techniques to forecast pollution and presented a comparative analysis to identify the most effective model for quickly and reliably predicting air quality. Using numerous datasets and Apache Spark experiments, the researchers calculated pollution levels. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were employed as evaluation metrics for comparing these regression models.

[3] Yi-Ting Tsai suggests a strategy in this research for forecasting PM2.5 concentrations that combines LSTM and RNN (Recurrent Neural Network). (Long Short-Term Memory). The researchers construct a neural network and execute RNN with LSTM through Tensorflow using Keras, a Python-based high-level neural networks API. The network's test data for predicting is from 2017, and the training data was acquired from Taiwan's EPA (Environmental Protection Administration) and integrated into 20-dimensions data from 2012 to 2016.

[4] Now that machine learning technology has advanced, it is possible to forecast pollution using previous data. Venkat Rao Pasupuleti introduces a gadget in this work that uses historical pollutants and current pollutants to execute a machine learning algorithm to forecast future pollutant data. For further analysis, the sensed data is stored in an Excel file. On the Arduino Uno platform, these sensors are used to gather information on pollutants.

[5] Shengdong Du suggests a unique deep learning model that makes use of a hybrid deep learning architecture to learn the spatial-temporal correlation characteristics and interdependence of multivariate air quality-related time series data in order to forecast air quality (mainly PM2.5). The foundation modules of the model include one-dimensional convolutional neural networks

(1D-CNNs) and bi-directional long short-term memory networks since multivariate air quality time series data is nonlinear and dynamic. (Bi-LSTM). The latter learns geographic-temporal dependencies, whereas the former extracts local trend and spatial correlation features. The researchers then create a cooperatively hybrid deep learning framework based on one-dimensional CNNs and Bi-LSTM for shared representation features learning of multivariate air quality-related time series data.

[6] In this paper, Ke Gu suggests a heuristic recurrent air quality predictor (RAQP) to estimate air quality. The RAQP calculates air pollutant concentrations (APCs), such as fine particulate matter, using important meteorological and pollution-related data. (PM2.5). The RAQP approach repeatedly applies the 1-h prediction model, which learns recent records of meteorology and pollution-related parameters to forecast air quality one hour in the future, and then estimates air quality several hours later. Numerous tests demonstrate that the RAQP predictor outperforms nonrecurrent approaches and pertinent state-of-the-art techniques for predicting air quality.

[7] This study suggested an attention-based air quality predictor (AAQP) to better safeguard individuals from air pollution, using Beijing as an example. The AAQP is a seq2seq model that predicts future air quality indices using historical air quality data as well as weather data. The AAQP with n-step recurrent prediction surpassed the related techniques, according to the experimental data, because the training time and error accumulation were both greatly decreased when compared to the original seq2seq attention model.

[8] The Long Short-Term Memory prediction method is improved and strengthened by this model. (LSTM). Data from the IoT node and information from the national environmental protection department are combined in the experiment. The first step was to choose 96 hours of data from four cities to use as an experimental sample. The experimental results and the genuine value are quite similar. The researchers then selected daily smog data from 2014/1/1 to 2018/1/1 as a train and test dataset. For 74 cities, smog information is provided. 70% of the total data were utilized for training, while the remaining data were used for testing. The outcomes of this study demonstrate how much more accurately this model can anticipate.

[9] In order to increase prediction accuracy and time performance with a lot of data, an enhanced decision tree method is proposed in this study. The feature attribute value and the weighting of the information gain are two ways that the model is improved based on an existing approach. Both computing complexity and accuracy are improved. The experimental findings demonstrate that in terms of accuracy and computing complexity, the modified model performs better than the conventional approaches. Additionally, it handles classification and prediction with vast amounts of air quality data more effectively. Additionally, it has the ability to anticipate future data with accuracy.

[10] The Convolutional Long Short-Term Memory (ConvLSTM) model, a combination of Convolutional Neural Networks and Long Short-Term Memory that automatically manipulates both spatial and temporal data features, is suggested by Van-Duc Le in this research. The researchers demonstrate how to interpolate and estimate air quality for the

entire city at once using the ConvLSTM model by converting air pollution data into image sequences. Additionally, they show that their method outperforms earlier studies in this field and is relevant to challenges involving spatiotemporal air pollution.

3. PROPOSED SYSTEM

The proposed system for air quality prediction using random forest and decision tree algorithms has several advantages over existing systems. These algorithms are capable of handling both continuous and categorical variables, which is crucial for air quality prediction where variables such as pollutant levels and weather conditions can be interdependent. Moreover, decision trees can model non-linear relationships, interactions, and dependencies between variables. Another advantage of using random forest and decision tree algorithms is their ability to handle missing data effectively. In air quality prediction, missing data is a common problem, and traditional methods such as Naive Bayes can struggle to handle this issue. Random forest and decision tree algorithms can handle missing data by imputing missing values, reducing the impact of missing data on the accuracy of predictions. The use of these algorithms also enables feature selection, decreasing the number of variables in the model, which increases effectiveness and accuracy of air quality predictions. Finally, random forest and decision tree algorithms can provide insights into the factors that impact air quality by visualizing the decision tree structure. These advantages make them an ideal choice for air quality prediction, and they have the potential to revolutionize air quality monitoring and control.

4. MODULES

MODULE 1: DATA COLLECTION AND PREPROCESSING

Air quality prediction involves forecasting air pollutant levels based on environmental elements include humidity, temperature, and wind speed. The first step is to collect a relevant dataset from sources such as government agencies and private organizations. The dataset should include environmental factors and corresponding AQI values. Data preprocessing is then performed to transform them into a format that machine learning algorithms can understand. This includes data cleaning, feature selection, feature scaling, data splitting, and data encoding. Data cleaning involves removing or fixing missing or incorrect data points, while feature selection identifies relevant features. Feature scaling scales the features to a similar range for effective machine learning. Data splitting divides the dataset into training and testing sets, while data encoding transforms categorical data into numerical data. The reliability and correctness of the data of the preprocessing steps are critical to the performance of the machine learning algorithm used for air quality prediction.

MODULE 2: MODEL TRAINING

Module 2 of air quality prediction involves training and building machine learning models using four different algorithms: KNN Classifier, Decision Tree Classifier, Random Forest Classifier, and Logistic Regression. Logistic Regression is used to predict whether a given combination of environmental factors will result in good or poor air quality, while KNN is used to predict the AQI for a new combination of environmental factors according to the k-nearest neighbors. Decision Tree utilized to predict the AQI for a new combination of

environmental factors based on a set of decision rules, and Random Forest combines the predictions of multiple decision trees to predict the AQI. Once the models are trained, their performance is employing a variety of criteria, including as accuracy, precision, recall, and F1 score, and ROC curve. The most effective model can then be applied for air quality prediction in Module 3.

MODULE 3: MODEL EVALUATION AND CREATING WEB APP

Module 3 is the final step in the air quality prediction process, which involves using the best-performing model to predict the AQI for new data points. The process of making predictions begins with collecting new data on the environmental factors that affect air quality. Once the new data has been collected, it is preprocessed using the same preprocessing steps used in Module 1. This includes data cleaning, feature selection, feature scaling, data splitting, and data encoding to ensure that the new data is suitable for machine learning analysis.

After the new data has been preprocessed, the best-performing model can be used the AQI for the fresh data points to be predicted. After that, one can utilize the anticipated AQI values to determine whether the air quality is good or poor. It is significant to note that the performance of the machine learning algorithm used to predict the AQI and the quality of the data both influence how accurate the forecasts are.

The best model must be integrated into a web application or another platform in order to be used for real-time air quality prediction. The model can be employed to forecast AQI for new data points in real-time, allowing users to make informed decisions about their activities and health. The real-time air quality prediction can be particularly useful for people who suffer from respiratory problems or other health issues that are exacerbated by poor air quality.

In order to keep the model accurate and useful, it is crucial to regularly review and update it. This can involve collecting new data and retraining the model periodically to incorporate the latest environmental factors and AQI values. Additionally, it is important to monitor the model's performance and make required adjustments to increase its correctness and effectiveness.

In conclusion, Module 3 is a critical step in the air quality prediction process as it involves using the best-performing model to make accurate predictions about air quality for new data points. By integrating the model into a web application or other platform, the predictions can be made in real-time, allowing users to make informed decisions about their activities and health. To confirm the model's efficacy and correctness, it is important to continuously monitor and update it with the latest data and performance metrics.

5. RESULT

Logistic Regression performed well in predicting binary outcomes, but its accuracy decreased when predicting AQI values with a larger range.

KNN Classifier performed well in predicting AQI values for locations with similar environmental conditions to those in the training set, but its performance decreased

when predicting AQI values for locations with different environmental conditions.

Decision Tree Classifier performed well in predicting AQI values when the relationships between the features and AQI were simple and linear.

Random Forest Classifier performed well in predicting AQI values when the relationships between the features and AQI were complex and non-linear. It also showed a high degree of robustness to noisy or missing data.

Overall, the results suggest that the best method for predicting air quality is Random Forest Classifier., due to its ability to handle complex relationships between variables and its robustness to missing or noisy data. However, the choice of algorithm should depend on the specific application and the characteristics of the dataset being used.

6. PERFORMANCE ANALYSIS

In general, the performance analysis of air quality prediction using different machine learning algorithms showed that all four algorithms can be effective in predicting air quality. The performance of each algorithm can be evaluated using various metrics, including accuracy, precision, recall, F1 score, and ROC curve. Logistic Regression showed good performance in predicting binary outcomes, but it may not be as effective in handling complex relationships between variables. KNN Classifier showed good performance in identifying the nearest neighbors and predicting AQI, but the performance may suffer when the number of features is large. Decision Tree Classifier showed good performance in modeling non-linear relationships, interactions, and dependencies between variables, but the performance may suffer when the tree becomes too complex. Random Forest Classifier showed good performance in handling missing data, reducing the number of variables used in the model, and combining multiple decision trees to improve predictions.

7. CONCLUSION AND FUTURE SCOPE

Predicting the state of the air we breathe is important for several reasons, including public health, and precise forecasting is necessary for keeping air pollution under control. While conventional techniques like as Naive Bayes have seen extensive use, they are not without shortcomings that might compromise their performance in the real world. In order to better forecast air quality, methods for machine learning, such as decision trees and random forests can handle complicated interactions between variables, efficiently deal with missing data, enable feature selection, and reveal the elements that affect air quality. The proposed machine learning-based air quality prediction system has the potential to significantly improve current approaches to gauging and managing air quality. This system is a great resource for public health professionals and policymakers since it is reliable, effective, and simple to use. Better public health outcomes may be achieved if policymakers and public health professionals are able to reliably anticipate air quality in order to implement more effective policies and interventions to enhance air quality. The capacity of machine learning algorithms to swiftly and effectively assess big datasets is a major benefit when used to air quality prediction systems. These algorithms may describe intricate

dependencies amongst variables and provide precise predictions in real time because of their ability to work with both continuous and categorical data. Machine learning methods are useful in air quality prediction systems because of their capacity to deal with missing data. Imputing missing values is a frequent technique used by machine learning algorithms, which may help mitigate the impact of lacking information on the precision of air quality forecasts.

Feature selection, or the determination of the most influential factors influencing air quality, is another use of machine learning algorithms. This method aids in minimizing the model's reliance on a large number of input variables, which ultimately leads to more precise and effective air quality forecasts. In conclusion, there is a tremendous possibility to enhance the efficacy of air quality monitoring and management via using machine learning algorithms to improve air quality prediction systems. The suggested system, built on top of the random forest and decision tree algorithms, is an effective, efficient, and user-friendly resource for public health administrators and policymakers. These algorithms' ability to reliably forecast future air quality has the potential to enhance health and save lives in heavily polluted urban areas. Using machine learning methods to improve air quality prediction systems has the potential to significantly enhance public health outcomes as a result of recent developments in both technology and data science.

8. REFERENCES

[1] Temesegan Walelign Ayele, Rutvik Mehta, "Air pollution monitoring and prediction using IoT", Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018

[2] Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif Ul Islam, Muhammad Nabeel Asghar, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities", IEEE Access (Volume: 7), 2019

[3] Yi-Ting Tsai, Yu-Ren Zeng, Yue-Shan Chang, "Air Pollution Forecasting Using RNN with LSTM", IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 2018

[4] Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth, Hari Kiran Reddy, "Air Quality Prediction of Data Log by Machine Learning", 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020

[5] Shengdong Du, Tianrui Li, Yan Yang, Shi-Jinn Horng, "Deep Air Quality Forecasting Using Hybrid Deep Learning Framework", Transactions on Knowledge and Data Engineering (Volume: 33, Issue: 6), 2021

[6] Ke Gu, Junfei Qiao, Weisi Lin, "Recurrent Air Quality Predictor Based on Meteorology- and Pollution-Related Factors", IEEE Transactions on Industrial Informatics (Volume: 14, Issue: 9), 2018

[7] Bo Liu, Shuo Yan, Jianqiang Li, Guangzhi Qu, Yong Li, Jianlei Lang, Rentao Gu, "A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction", IEEE Access (Volume: 7), 2019

[8] Baowei Wang, Weiwen Kong, Hui Guan, Neal N. Xiong, “Air Quality Forecasting Based on Gated Recurrent Long Short-Term Memory Model in Internet of Things”, IEEE Access (Volume: 7), 2019

[9] Yuanni Wang, Tao Kong, “Air Quality Predictive Modeling Based on an Improved Decision Tree in a Weather-Smart Grid”, IEEE Access (Volume: 7), 2019

[10] Van-Duc Le, Tien-Cuong Bui, Sang-Kyun Cha, “Spatiotemporal Deep Learning Model for Citywide Air Pollution Interpolation and Prediction”, IEEE International Conference on Big Data and Smart Computing (BigComp), 2020