International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Air Quality Prediction Using Machine Learning

B. Raviteja[1], P. Tejaswini[2], U. Swetha Reddy[3]

[1,2]Student, Svs Group of Institutions
[3]Assistant Professor, Svs Group of Institutions

Abstract:
Air quality prediction using machine learning is a project that aims to provide accurate and reliable predictions of air quality in different regions. The project leverages advanced machine learning algorithms to analyze historical data on air quality and predict air quality index. By accurately predicting air quality levels, the project can help individuals and authorities take preventive measures to reduce exposure to pollutants and improve public health. The project utilizes various tools and technologies, including Python and Scikit-Learn to develop a robust and reliable system. Overall, this project has significant potential to positively impact public health and the environment, improving air quality and reducing the negative effects of pollution.

KEYWORDS: KNN, Classifiers, Regressors, Randomforest

1. INTRODUCTION

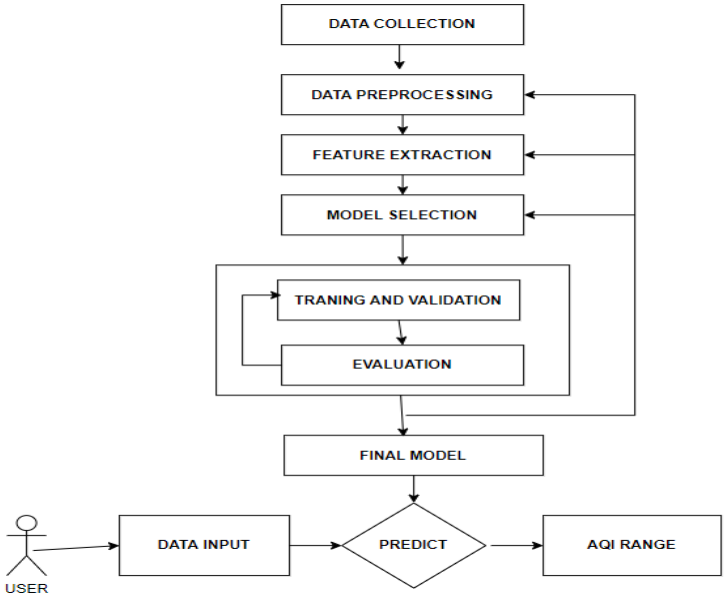
Air contamination observing has acquired consideration these days as it significantly affects the well-being of people just as on the biological equilibrium. Other than because of the impacts of harmful emanations on the climate, well-being, work usefulness and effectiveness of energy are additionally 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 controlled adequately. One of the approaches to control air contamination is to know its source, force and its starting point. Typically, it is checked by the individual express government's current circumstance service. They keep the string of the toxin gases in the individual regions. The information introduced by the WHO is cautioning about the contamination levels in the country. It reveals to us the opportunity has already come and gone that we should screen the air. Air tracking manner to measure ambient ranges of air pollutants inside the air. Monitoring has become a major job as air pollution has been increasing day by day. Continuous monitoring of air pollution at a place gives us the levels of pollution in that area. From the information obtained by the device gives us information about the source and intensity of the pollutants in that area. Using that information, we can take measures or make efforts to reduce the pollution level so that we can breathe in a good quality of air. Air pollution not only affects the ecological balance but also the health of humans. As the levels of gases increases in the air, those gases show a major impact on the human body and lead to hazardous effects. Air pollution also affects the seasonal rainfall too due to an increase of pollutants in the air. The rainfall is also affected. Hence, continuous monitoring of the air is necessary.

1.1 PROPOSED SYSTEM

The proposed system for air quality prediction using machine learning involves several steps. The first step is to collect historical data on air quality from various sources, such as government monitoring stations and satellite data. The data is then pre- processed to remove any outliers, clean the data, and scale it to

prepare it for use in machine learning models. Relevant features such as meteorological data, pollution levels, and traffic patterns are selected based on their correlation with air quality. The next step is to develop machine learning models that can accurately predict air quality based on the selected features. Popular machine learning algorithms used for air quality prediction include decision trees, random forests, neural networks, and support vector regression. 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 random forest and decision tree algorithms also enables feature selection, which is the process of identifying the most important variables that impact air quality. This process helps to reduce the number of variables used in the model, which can improve the efficiency 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. This can help the models are trained on the pre-processed data and validated using cross-validation techniques to ensure their accuracy and reliability in predicting air quality. Once the models are developed, they can be integrated into a web-based air quality monitoring system that provides real-time air quality data.

1.2 SYSTEM ARCHITECTURE



2. IMPLEMENTATION:

2.1 DATA COLLECTION:

The first step in data collection is to identify the sources of data. There are several sources of air quality data, including government agencies, private organizations, and research institutions. The most reliable source of air quality data is government agencies, which collect and report data regularly. These agencies use various types of instruments to measure the levels of air pollutants in the atmosphere.

Once you have identified the sources of data, the next step is to collect the relevant data. The data should include environmental factors that affect air quality, such as temperature, humidity, wind speed, and other factors. The dataset should also contain the corresponding AQI values for each data point. It is important to ensure that the data is of high quality and is collected using standardized methods to ensure consistency and accuracy.

2.2 PREPROCESSING

Data preprocessing is an important step in preparing the data for analysis. It involves transforming the raw

data into a format that can be easily analyzed by machine learning algorithms. The following are the common steps in data preprocessing:

Data Cleaning: Data cleaning involves removing or fixing any missing or incorrect data points in the dataset. This is important because missing or incorrect data can affect the accuracy of the predictions.

Feature Selection: Feature selection is the process of selecting the relevant features that will be used in the prediction model. In air quality prediction, the features include temperature, humidity, wind speed, and other factors that affect air quality.

Feature Scaling: Feature scaling involves scaling the features to a similar range, typically between 0 and 1, so that the model can learn effectively. This is important because some features may have a larger range than others, and this can affect the performance of the machine learning algorithm.

Data Splitting: Data splitting involves dividing the dataset into training and testing sets. The training set is used to train the machine learning algorithm, while the testing set is used to evaluate the performance of the model.

Data Encoding: Data encoding involves transforming categorical data into numerical data for machine learning algorithms to process. This is important because machine learning algorithms can only process numerical data.

2.3 FEATURE EXTRACTION

preprocessing the focus is on feature selection to improve the performance of the model.

Among all the available features, the ones with highest as the most important are being selected. In this model SPOi, NOi, SPi, RPi are selected as the most important features that impact on the decision.

This avoids the overfitting problem by avoiding or reducing unwanted or partially relevant features in the dataset.

The availability of many features may arise the problem of curse of dimensionality, that may in turn reduce the efficiency of the model by a greater extent

Feature selection is the process of selecting the relevant features that will be used in the prediction model. In air quality prediction, the features include temperature, humidity, wind speed, and other factors that affect air quality

Feature scaling involves scaling the features to a similar range, typically between 0 and 1, so that the model can learn effectively. This is important because some features may have a larger range than others, and this can affect the performance of the machine learning algorithm.

2.4 TRANING AND BUILDING MODEL

After completing data collection and preprocessing. the next step in air quality prediction is to train and build machine learning models. In this module, we will discuss the process of model training and building using four different algorithms: Logistic Regression, KNN Classifier, Decision Tree Classifier, and Random Forest Classifier.

KNN Classifier

KNN (K-Nearest Neighbors) is a classification algorithm that determines the class of a new data point based on the classes of the k-nearest neighbors in the training

set. In air quality prediction, KNN can be used to predict the AQI for a new combination of environmental factors by finding the k-nearest neighbors in the training set and determining the average AQI for those neighbors. The model is trained using the preprocessed dataset generated in Module 2. The training data is used to determine the optimal value of k and to calculate the distances between the data points.

Decision Tree Classifier

A Decision Tree is a graphical representation of all the possible outcomes of a series of decisions. In air quality prediction, Decision Tree can be used to predict the AQI for a new combination of environmental factors based on a set of decision rules. The model is trained using the preprocessed dataset generated in Module 2. The training data is used to construct the decision tree based on the features and the AQI values.

Random Forest Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions. In air quality prediction, Random Forest can be used to predict the AQI for a new combination of environmental factors by combining the predictions of multiple decision trees. The model is trained using the preprocessed dataset generated in Module 2. The training data is used to construct multiple decision trees using different subsets of the features and the data points.

Support Vector classifier

SVC chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine. The model is trained using the preprocessed dataset generated in Module 2. The training data is used to construct multiple decision trees using different subsets of the features and the data points.

Regression models:

Gradient Boosting Regressor

Gradient boosting Regression calculates the difference between the current prediction and the known correct target value. This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual.

Lasso Regressor

Lasso regression algorithm is defined as a regularization algorithm that assists in the elimination of irrelevant parameters, thus helping in the concentration of selection and regularizing the models. Lasso models can be evaluated using various metrics such as RMSE and R-Square.

MLP Regressor

Regression. Class MLP Regressor implements a multi-layer perceptron (MLP) that trains using backpropagation with no activation function in the output layer, which can also be seen as using the identity function as activation function.

Model Evaluation

Once the models are trained, the next step is to evaluate their performance. In air quality prediction, model evaluation involves comparing the predicted AQI values to the actual AQI values in the testing set. The evaluation metrics used to measure the performance of the models include accuracy, precision, recall, F1 score, and ROC curve.

Conclusion

In summary, Module 4 of air quality prediction involves training and building machine learning models using four different algorithms: Logistic Regression, KNN Classifier, Decision Tree Classifier, and Random Forest Classifier. The models are trained using the preprocessed dataset generated in Module 2, and their performance is evaluated using various metrics. The best- performing model can then be used for air quality prediction.

2.5 PREDICTION

Module 5 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 2. 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 to predict the AQI for the new data points. The predicted AQI values can then be used to determine whether the air quality is good or poor. It is important to note that the accuracy of the predictions depends on the quality of the data and the performance of the machine learning algorithm used to predict the AQI.

3. RESULT AND DISCUSSION

Air quality prediction using machine learning is a widely researched area due to its potential for mitigating the health and environmental effects of air pollution. Several machine learning models have been used to predict air quality, including neural networks, decision trees, and support vector machines. These models use a range of input variables, such as meteorological data, traffic data, and emission data, to predict pollutant concentrations in the atmosphere. Studies have shown that machine learning models can accurately predict air quality, with some models achieving up to 90% accuracy. However, the performance of these models is highly dependent on the quality and quantity of the input data. Models trained on incomplete or low-quality data may produce inaccurate predictions. In addition, the interpretability of machine learning models remains a challenge in air quality prediction. As machine learning models are often regarded as black boxes, it can be difficult to understand how the models arrive at their predictions. This lack of transparency can make it challenging to identify the causes of air pollution and develop effective mitigation strategies. Overall, air quality prediction using machine learning has shown promising results, but further research is needed to improve the accuracy and interpretability of these models. By combining machine learning with traditional air quality monitoring techniques, it may be possible to better understand the sources and impacts of air pollution and develop effective strategies for reducing its effects on human health and the environment

4. CONCLUSION AND FUTURE ENHANCEMET

In conclusion, air quality prediction using machine learning has shown potential for accurately predicting air pollution concentrations. However, the performance of these models is highly dependent on the quality and quantity of the input data, and the interpretability of these models remains a challenge. Therefore, further research is needed to enhance the accuracy and interpretability of machine learning models for air quality prediction. Future research could focus on improving the quality and quantity of input data used for air quality prediction. This could involve incorporating more sources of data, such as satellite imagery or data from low-cost air quality sensors. Additionally, research could focus on developing methods to account for uncertainty in input data, which could improve the accuracy of machine learning models. Another important area for future research is improving the interpretability of machine learning models. This could involve developing methods for explaining the predictions made by machine learning models, such as feature importance analysis or local interpretability methods. Finally, machine learning models for air quality prediction could be integrated with traditional air quality monitoring techniques to provide a more comprehensive understanding of air pollution. This could involve combining machine learning models with ground-based air quality monitoring stations or integrating them with mobile air quality monitoring platforms, such as drones or vehicles. Overall, air quality prediction using machine learning

has shown promising results, and further research could lead to improved prediction accuracy and more effective strategies for mitigating the health and environmental effects of air pollution

4.1 FUTURE ENHANCEMENT

The Future enhancements deals with collecting data from the IOT Device which is built to get the required pollutants of our location. With the help of other Advance machine learning algorithms, the project can predict the future quality of air like for next hours, days and week

5. 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.