Prediction of Air quality index data using machine Learning

Saif Ali Khan - X22123296

March 2023

1 Introduction

1.1 Motivation

Globally, air pollution is a serious problem for both the environment and human health. . AQI, or Air Quality Index, a useful tool for determining air quality and informing the public of related health dangers. In order to reduce exposure to polluted air and lower the risk of respiratory and cardiovascular disorders, accurate forecasting of AQI levels is essential.

Machine learning algo is been increasingly used to predict AQI levels due to their ability to handle the complexity of air flow and provide real-time predictions. there are many algoritham which can use historical data and real-time air quality monitoring data to generate accurate and reliable predictions. Incorporating additional features such as meteorological data, traffic patterns, and land-use Data can also be improving prediction accuracy.

The use of machine learning in predicting AQI has several advantages. It allows for real-time prediction, which is crucial in alerting the public about potential health hazards. It is a useful technique for forecasting air quality because it can manage the non-linear interactions between many components that affect AQI. Additionally, machine learning can aid governments and environmental organizations in making informed decisions related to air pollution control policies.

This project's purpose is to develop an accurate and efficient predictive model for AQI using advanced machine learning techniques such as classification methods, and hybrid models. This model will be trained on historical AQI data and real-time air quality monitoring data to improve prediction accuracy and efficiency. By using these advanced methods, we aim to handle the volatility of AQI more effectively and provide real-time information to individuals and communities to enable them to take necessary precautions to protect themselves from air pollution.

The development of a reliable and accurate predictive model for AQI using machine learning techniques is of paramount importance in reducing the health hazards associated with air pollution. The results of this project will provide a more comprehensive understanding of air pollution patterns and trends, enabling us to take informed decisions to mitigate its harmful effects. Additionally, this project will help reduce exposure to air pollution and improve the health and well-being of individuals.

2 Main Section

2.1 Research Question

- The ultimate goal is to improve the air quality index and protect public health, how can machine learning algorithms be utilized to forecast the air quality index (AQI) and which features have the most important impact on AQI prediction accuracy?

2.2 literature review

Air pollution is a major environmental concern that affects public health and well-being globally. The AQI is a widely used measure to represent air quality in terms of the concentration of air pollutants. AQI is computed by measuring the concentration of several pollutants. With the rise of smart cities and the increasing availability of air quality data, There is growing interest in the use of machine learning (ML) algorithms to predict AQI. Several papers have been published on AQI prediction using ML algorithms. For instance, Saba Ameer et al. [4]air,conducted a comparative analysis of various ML techniques for predicting AQI in smart cities. Pooja Bhalgat et al. [7] used ML algorithms to predict AQI levels in India. Raquel Espinosa et al. [9] proposed a time series forecasting-based multi-criteria methodology for AQI prediction. Lan Gao et al. [13] explored various ML algorithms for AQI prediction. In addition, Gaganjot Kaur Kang et al. [2] reviewed big data and ML approaches for AQI prediction. K Kumar and BP Pande [6] conducted a case study on air pollution prediction with ML in Indian cities. Yun-Chia Liang et al. [3] used ML-based prediction models for AQI prediction. Huixiang Liu et al. [1] used ML algorithms to predict AQI and air pollutant concentrations. Tanisha Madan et al. [8] provided a review of AQI prediction using ML algorithms. Manuel M endez et al. [11] surveyed ML algorithms for air quality forecasting. Animesh Tiwari et al. [5] used deep learning models to predict AQI in Delhi, India. Anıl Utku and Umit Can [12] conducted a comparative analysis of various ML algorithms for AQI prediction. mendez manuel mereyo .[10] proposed a predictive data feature exploration-based approach for AQI prediction. Overall, the research on AQI prediction using ML algorithms is diverse, and a variety of techniques and models have been explored. The application of these models could have significant implications for public health and environmental policies.

2.3 Data Source

Below 3 data-sets will be used in the paper

2.3.1 U.S. Pollution Data

: The dataset available on Kaggle provides a comprehensive collection of air pollution data in the US, covering a significant period of time. The information included in the dataset can be useful for researchers and analysts interested in investigating air pollution levels, trends, and patterns across different locations and time periods. The dataset is sourced from the United States Environmental Protection Agency (EPA) and contains measurements from monitoring stations located across the country, which were taken at different frequencies, depending on the pollutant and location. The dataset also includes information on the geographic coordinates of the monitoring stations, as well as other relevant details such as the type of monitoring equipment used. Overall, this dataset has the potential to support a wide range of research and analysis related to air pollution in the United States.(source: kaggle, no.of rows:100000+, no of columns: 28)

2.3.2 Air quality in northern Taiwan

The air quality dataset available on Kaggle provides a detailed collection of air quality measurements from Northern Taiwan, covering several years of data. The dataset includes information on various air pollutants and meteorological data, which can help researchers to understand the relationship between atmospheric conditions and air quality. The air pollutants measured in the dataset include nitrogen dioxide, sulfur dioxide etc. The data was collected from multiple monitoring stations located across Northern Taiwan, and the frequency of measurements varies depending on the pollutant and station. In addition to air quality data. Overall, this dataset has the potential to support a wide range of research and analysis related to air quality in Northern Taiwan, and can help to identify patterns and trends in air quality levels over time.

2.3.3 India Air Quality Data

The air quality dataset available on Kaggle provides a comprehensive collection of air quality measurements from various cities across India, covering a period of several years. which are known to have significant impacts on human health and the environment. The measurements were taken at different monitoring stations located in different parts of the country, and the frequency of measurements varies depending on the pollutant and station. In addition to air quality data, the dataset also includes information on the geographic coordinates of the monitoring stations. Overall, this dataset has the potential to support a wide range of research and analysis related to air quality in India, and can help to identify patterns and trends in air quality levels over time and across different regions of the country (source: kaggle, no.of rows:43000, no of columns: 13)

2.4 Machine Learning Methods

- Linear Regression: Linear regression is a supervised learning method for modeling the relationship between one or more independent variables and a dependent variable.. In the context of air quality index prediction, linear regression can be used to model the relationship between air pollutants and meteorological parameters. The rationale for choosing linear regression is its simplicity and interpretability, making it a good starting point for modeling the air quality index.

- K-Nearest Neighbors (KNN): It works by determining the k -nearest neighbors of a data point using some similarity measure, and then predicting the class or value using the majority vote or average of the neighbors. In the context of air quality index prediction, KNN can be used to identify similar air quality conditions and predict the air quality index on the historical data. The rationale for choosing KNN is its simplicity and ability to handle noisy data.

- Decision Trees: Decision trees are a non-parametric, and interpretable machine learning algorithm used for both classification and regression tasks. They work by recursively splitting data into the most informative features until a stopping criterion is met. In the context of air quality index prediction, decision trees can be used to identify the most important meteorological parameters and air pollutants that affect the air quality index. The rationale for choosing decision trees is their simplicity, interpretability, and ability to handle noisy data.

- Random Forest: used for both classification and regression tasks. At training time, it builds Multiple decision trees are used, and The output class in a classification task is determined by identifying the mode of the predicted classes, while in a regression task, it is determined by calculating the mean prediction of the individual samples. In the context of air quality index prediction, random forest can be used to model the complex nonlinear relationships between air pollutants and meteorological parameters. The rationale for choosing The ability of random forest to handle high-dimensional data

- Gradient Boosting Machines (GBM): used for classification and regression tasks. It works by building multiple weak models (e.g., decision trees) in a sequential manner and combining them to make the final prediction. In the context of air quality index prediction, GBM can be used to capture complex nonlinear relationships between air pollutants and meteorological parameters. The ability of GBM to handle high-dimensional data capture interactions between variables, and provide accurate predictions.

2.5 Evaluation Methods

- To evaluate performance of machine learning algorithms for air quality index prediction, various evaluation methods can be used. Some commonly used evaluation methods include:
- Mean Absolute Error (MAE): the average absolute disparity between the predicted and actual values of the air quality index is measured by MAE. A lower MAE indicates better predictability.
- Root Mean Square Error (RMSE): difference between the predicted and actual values of the air quality index. It penalizes larger errors more heavily than MAE and is commonly used when larger errors are more problematic.
- Coefficient of Determination (R-squared or R2): R2 quantifies the proportion of the variance in the air quality index that can be explained by the five machine learning algorithms that were chosen.
- cross-validation : Cross-validation is a resampling technique that involves dividing the data into multiple folds, each consisting of a training and testing set. The training set is used to train the machine learning algorithm, and the testing set is used to evaluate it. This process is repeated for each fold, allowing the algorithm's performance to be assessed across different subsets of the data. By examining the algorithm's performance on multiple testing sets, cross-validation can help evaluate its generalizability and guard against overfitting to the training data.
- Receiver Operating Characteristic (ROC) curve: also known as sensitivity, against the false positive rate, which is the complement of specificity, for various threshold values used in a binary classification model. It can be useful in assessing the trade-off between true positives and false positives in binary classification situations.
- Precision-Recall curve: a visualization of recall (sensitivity) and precision (positive predictive value) at different threshold values. It is frequently used for datasets with imbalances and can aid in assessing the trade-off between recall and precision.

2.6 Bibliography

References

[1] Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif Ul Islam, and Muhammad Nabeel Asghar. Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, 7:128325–128338, 2019.

[2] Pooja Bhalgat, Sachin Bhoite, and Sejal Pitare. Air quality prediction using machine learning algorithms. *International Journal of Computer Applications Technology and Research*, 8(9):367–370, 2019.

[3] Lan Gao, Changjie Cai, and Xiao-Ming Hu. Air quality prediction using machine learning. *Machine Learning in Chemical Safety and Health: Fundamentals with Applications*, pages 267–288, 2022.

[4] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie. Air quality prediction: Big data and machine learning approaches. *Int. J. Environ. Sci. Dev*, 9(1):8–16, 2018.

[5] Ibrahim Kok, Metehan Guzel, and Suat Ozdemir. Recent trends in air quality prediction: An artificial intelligence perspective. In *Intelligent Environmental Data Monitoring for Pollution Management*, pages 195–221. Elsevier, 2021.

[6] K Kumar and BP Pande. Air pollution prediction with machine learning: A case study of indian cities. *International Journal of Environmental Science and Technology*, pages 1–16, 2022.

[7] Yun-Chia Liang, Yona Maimury, Angela Hsiang-Ling Chen, and Josue Rodolfo Cuevas Juarez. Machine learning-based prediction of air quality. *applied sciences*, 10(24):9151, 2020.

[8] Huixiang Liu, Qing Li, Dongbing Yu, and Yu Gu. Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Applied Sciences*, 9(19):4069, 2019.

[9] Tanisha Madan, Shrddha Sagar, and Deepali Virmani. Air quality prediction using machine learning algorithms–a review. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 140–145. IEEE, 2020.

[10] Manuel Méndez, Mercedes G Merayo, and Manuel Núñez. Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, pages 1–36, 2023.

[11] Animesh Tiwari, Rishabh Gupta, and Rohitash Chandra. Delhi air quality prediction using lstm deep learning models with a focus on covid-19 lockdown. *arXiv preprint arXiv:2102.10551*, 2021.

[12] Anıl Utku and Umit Can. Machine learning-based a comparative analysis for air quality prediction. In *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2022.

[13] Ying Zhang, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao, Rongrong Zhang, Qingqing Wang, and Linyan Huang. A predictive data feature exploration-based air quality prediction approach. *IEEE Access*, 7:30732–30743, 2019.