IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12590

PREDICTION OF AIR POLLUTION USING SUPERVISED MACHINE LEARNING TECHNIQUES

Mrs.Shakila[1], Anitha A[2], Devada Geetha Madhuri C[3], Harini S[4]

[1]Assistant Professor, Department of Computer Science and Engineering, DMI College Of Engineering, Chennai, India

[2]B.E, Department of Computer Science and Engineering, DMI College Of Engineering, Chennai, India

[3]B.E, Department of Computer Science and Engineering, DMI College Of Engineering, Chennai, India

[4]B.E, Department of Computer Science and Engineering, DMI College Of Engineering, Chennai, India

Abstract- Due to human activities, industrialization, and urbanization, air pollution has become a life-threatening factor in many countries around the world. Among air pollutants, Particulate Matter causes various illnesses such as respiratory tract and cardiovascular diseases. Hence, it is necessary to accurately predict the air pollution concentrations to prevent the citizens from the dangerous impact of air pollution beforehand. The variation of air pollution depends on a variety of factors, such as meteorology and the concentration of the other pollutants in urban areas. The aim to investigate machine learning –based techniques for the prediction of air pollution results with the best accuracy. The analysis of the dataset by supervised machine learning techniques(SMLT) captures several information like variable identification and analysis techniques like univariable analysis, bi-variate, and multivariate analysis. Compare and discuss the performance of various machine learning algorithm from the given pollution dataset with evaluation techniques.

I. INTRODUCTION

A. General

Air pollution is a mixture of solid particles and gases in the air. Car emissions, chemicals from factories, dust, pollutants and Mold spores may be suspended as particles. Ozone, a gas, is a major part of air pollution in cities. When ozone forms air pollution, also called smog. Some air pollutants are poisonous. Vehicle emissions, fuels oils and natural gas to heat homes, by-products of manufacturing and power generation, particularly coal-fuelled power plants, and fumes from chemical production are the primary sources of human-made air pollution. Inhaling them can increase the chance you'll have health problems. People with heart or lung diseases, older adults and children are at greater risk from air pollution. Air pollution isn't just outside the air inside buildings can also be polluted and affect your health.

B .Data Science

Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be major use in the formation of big business decisions.

C. Data Scientists:

Data scientists examine which questions need answering and where to find the related data. They have business acumen and analytical skill as well as the ability to mine, clean, and present data. Businesses use data scientists to source, manage, and analysis large amounts of unstructured data.

Required Skills for a Data Scientist**:**

- Programming: Python, SQL, Scala, Java, R, MATLAB.
- Machine Learning: Natural Language Processing, Classification, Clustering.
- Data Visualization: Tableau, SAS, D3.js, Python, Java, R libraries.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.12590

- Big data platforms: MongoDB, Oracle, Microsoft Azure, Cloudera.

D. Artificial Intelligence

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, and speech recognition and machine vision. The field was founded on the assumption that human intelligence "can be so precisely described that a machine can be made to simulate it". This raises philosophical arguments about the mind and the ethics of creating artificial beings endowed with human-like intelligence. In general, AI systems work by ingesting large amounts of labelled training data, analysing the data for correlations and patterns, and using these patterns to make predictions about future states. In this way, a chat box that is fed examples of text chats can learn to identify and describe objects in images by reviewing millions of examples. AI programming focuses on three cognitive skills: learning, reasoning and self-correction. This aspect of AI programming focuses on acquiring data and creating rules for how to turn the data into actionable information. The rules, which are called algorithms, provide computing devices with step-by-step instructions for how to complete a specific task.

II. LITERATURE SURVEY

Title : Forecasting Air Pollution Particulate Matter
Year **:** 2020

Air pollution are confronting the overwhelming air contamination issue. The citizens and governments have experienced and expressed the increasingly concerned regarding the impact of air pollution affecting human health and proposed sustainable development for overriding air pollution issues across the worldwide. The outcome of modern industrialization contains the liquid droplets, solid particles and gas molecules and is spreading in the atmospheric air. The heavy concentration of particulate matter of size PM10 and PM2.5 is seriously caused adverse health effect. Through the determination of particulate matter concentration in atmospheric air for the betterment of human being well in primary importance. In this paper machine learning predictive models for forecasting particulate matter concentration in atmospheric air are investigated on Taiwan Air Quality Monitoring data sets, which were obtained from 2012 to 2017. These models were compared with the existing traditional models and perform better in predictive performance. The performance of these models was evaluated with statistical measures: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Square Error (MSE), and Coefficient of Determination (R2).

Title: Detection and Prediction of Air Pollution using Machine Learning Model
Author: Aditya C R , Chandana R Deshmukh , Nayana D K, Praveen Gandhi Vidyavastu
Year**:** 2018

In the populated and developing countries, governments consider the regulation of air as a major task. The meteorological and traffic factors, burning of fossil fuels, industrial parameters such as power plant emissions play significant roles in air pollution. Among all the particulate matter that determine the quality of the air, Particulate matter (PM 2.5) needs more attention. When it's level is high in the air, it causes serious issues on people's health. Hence, controlling it by constantly keeping a check on its level in the air is important. In this paper, Logistic regression is employed to detect whether a data sample is either polluted or not polluted. Autoregression is employed to predict future values of PM2.5 based on the previous PM2.5 readings. Knowledge of level of PM2.5 in nearing years, month or week, enables us to reduce its level to lesser than the harmful range. This system attempts to predict PM2.5 level and detect air quality based on a data set consisting of daily atmospheric conditions in a specific city.

Title: Air Pollution Prediction System Using Deep Learning
Author: Thongsuk Xaya souk & Hawman Lee
Year **:** 2019

One of the most influential factors on human health is air pollution, such as the concentration of PM10 and PM2.5 is a damage to a human. Despite the growing interest in air pollution in Korea, it is difficult to obtain accurate information due to the lack of air pollution measuring stations at the place where the user is located. Deep learning is a type of machine learning method has drawn a lot of academic and industrial interest. In this paper, we proposed a deep learning approach for the air pollution prediction in South Korea. We use Stacked Autoencoders model for learning and training data. The experiment results show the performance of the air pollution prediction system and model that proposed. Keywords: fine dust, PM10, PM2.5, air pollution prediction, deep learning.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12590

Title: Machine Learning-Based Prediction of Air Quality
Author: Yun-Chia Liang , Yona Maiuri , Angela Hsiang-Ling Chen , and Josue Rodolfo Cuevas Juarez 1
Year **:** 2020

Air, an essential natural resource, has been compromised in terms of quality by economic activities. Considerable research has been devoted to predicting instances of poor air quality, but most studies are limited by insufficient longitudinal data, making it difficult to account for seasonal and other factors. Several prediction models have been developed using an 11-year dataset collected by Taiwan's Environmental Protection Administration (EPA). Machine learning methods, including adaptive boosting (AdaBoost), artificial neural network (ANN), random forest, stacking ensemble, and support vector machine (SVM), produce promising results for air quality index (AQI) level predictions. A series of experiments, using datasets for three different regions to obtain the best prediction performance from the stacking ensemble, AdaBoost, and random forest, found the stacking ensemble delivers consistently superior performance for R 2 and RMSE, while AdaBoost provides best results for MAE.

Title: Air pollution prediction in smart city, deep learning approach
Author: Abdellatif Bekkar , Badr Hasina , Samira Douse and Khadija Douse
Year **:** 2021

Over the past few decades, due to human activities, industrialization, and urbanization, air pollution has become a life-threatening factor in many countries around the world. Among air pollutants, Particulate Matter with a diameter of less than 2.5µm (PM2.5) is a serious health problem. It causes various illnesses such as respiratory tract and cardiovascular diseases. Hence, it is necessary to accurately predict the PM2.5 concentrations in order to prevent the citizens from the dangerous impact of air pollution beforehand. The variation of PM2.5 depends on a variety of factors, such as meteorology and the concentration of other pollutants in urban areas. In this paper, we implemented a deep learning solution to predict the hourly forecast of PM2.5 concentration in Beijing, China, based on CNN-LSTM, with a spatial-temporal feature by combining historical data of pollutants, meteorological data, and PM2.5 concentration in the adjacent stations. We examined the difference in performances among Deep learning algorithms such as LSTM, Bi-LSTM, GRU, Bi-GRU, CNN, and a hybrid CNN-LSTM model. Experimental results indicate that our method "hybrid CNN-LSTM multivariate" enables more accurate predictions than all the listed traditional models and performs better in predictive performance

III. PROPOSED SYSTEM

The proposed method is to build a machine learning model for the classification of air pollution. The process carries from data collection where past data related to air pollution are collected. Data mining is a commonly used technique for processing enormous data. Air pollution is found before, can lower respiratory problems. Machine learning is now applied where it reduces manual effort and a better model makes error less which leads to preventing respiratory problems. The data analysis is done on the dataset, proper variable identification is done that is both the dependent variables and independent variables are found. The proper machine learning algorithms are applied to the dataset where the pattern of data is learned. After applying different algorithms a better algorithm is used for the prediction of the outcome.

A. Architecture Diagram:

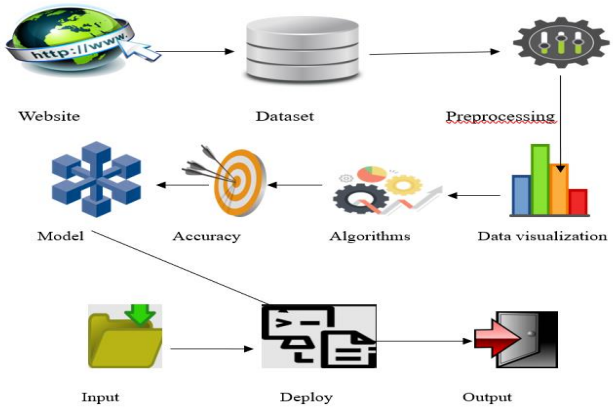


Fig.1. Architecture Diagram

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12590

B. Use Case Diagram:

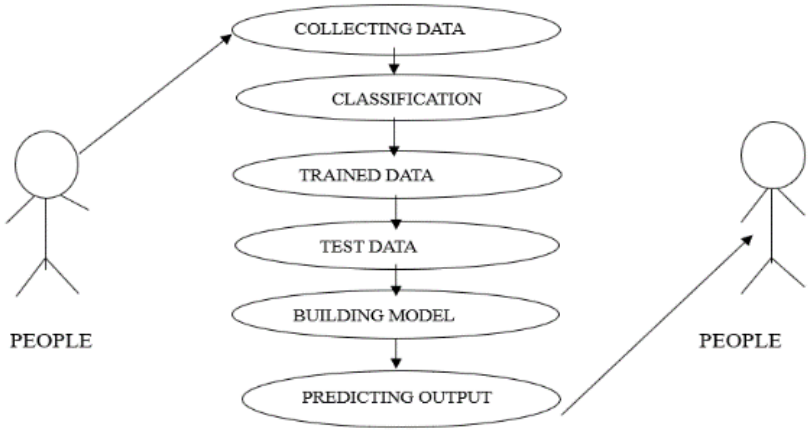


Fig.2. Use Case Diagram

IV. MODULE DESCRIPTION

A. Data Preprocessing:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

B .Data Visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

C. Algorithm Implementation:

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

D. Deployment.

After giving the input data among four algorithms random forest classifier makes best choice for deploying the output during deployment the pollution of air could be identified.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

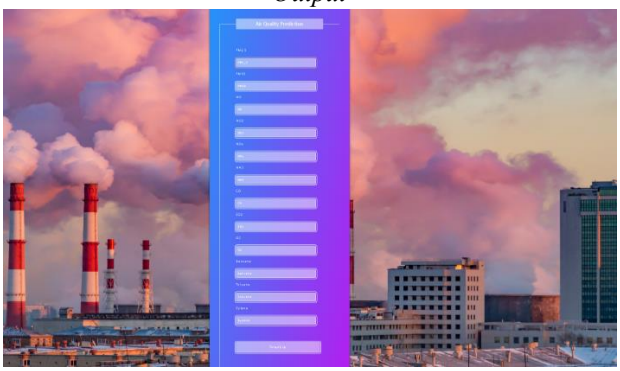
International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12590

V. PROJECT OUTCOMES

Output



VI. RESULT AND DISCUSSION

The Decision Tree Algorithm gave the best results among all the algorithms, with an overall accuracy of 99.8. The prediction model precision findings, helped in evaluating and contrasting current work on air quality assessment which is based upon Big Data Analytics and Machine Learning. Air Pollution is one of the major factors in our healthcare domain. There are lot of patients who are actively present in the world. It is difficult to find the Air Pollution. So this project can easily find out the Air Pollution.

VII. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set of higher accuracy score algorithm will be find out. The founded one is used in the application which can help to find the Air Pollution.

.

REFERENCES

1.M. A. Zaidan et al., “Intelligent calibration and virtual sensing for integrated low-cost air quality sensors,” IEEE Sensors J., vol. 20, no. 22,pp. 13638–13652, Nov. 2020.
2.P. Ferrer-Cid, J. M. Barcelo-Ordinas, and J. Garcia-Vidal, "Graph signal reconstruction techniques for IoT air pollution monitoring platforms, "2022, arXiv:2201.00378.
3. T.-B. Ottosen and P. Kumar, "Outlier detection and gap filling methodologies for low-cost air quality measurements, "Environmental Sci., Processes Impacts, vol. 21, no. 4, pp. 701–713, 2019.
4.N. H. Motlagh et al., "Toward massive scale air quality monitoring, IEEE Commun. Mag., vol. 58, no. 2, pp. 54–59, Feb. 2020.
5. Acharjya, D. P. "A survey on big data analytics: challenges, open research issues, and tools.\", (2019)
6. Yue, G., Gu, K., and Qiao, J. "Effective and efficient photo-based pm2.5 concentration estimation." (2019) IEEE Transactions on Instrumentation and Measurement, PP,1–10.
7. Wang, Wei, Fei Yue Mao, Lin Du, Zeng Xin Pan, Wei Gong, and Shanghai Fang. "Deriving hourly PM2. 5 concentrations from Himawari-8 aids
8. Pan, Qingyue. "Application of XG Boost algorithm in hourly PM2. 5 concentration prediction." In IOP Conference Series: Earth and Environmental Science, IOP Publishing vol. 113, no. 1 (6), p. 012127.
9. Sun, Xiaotong, Wei Xu, and Hongyun Jiang. "Spatial-temporal prediction of air quality based on recurrent neural networks." In Proceedings of the 52nd Hawaii International Conference on System Sciences 9.
10. Sharma, Nidhi, Shweta Taneja, Vaishali Sagar, and Arshita Bhatt. ”Forecasting air pollution load in Delhi using data analysis tools.” Proceeding.