

HIGHER ACCURACY OF AIR QUALITY INDEX PREDICTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the
award of Bachelor of Technology degree in
Information Technology

By

S.Gokul krishna
(Reg.No.39120034)



DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119

MAY - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE

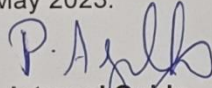
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

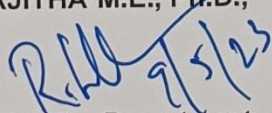
DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **S Gokulkrishna (Reg. No. 39120034)** who carried out the ProjectPhase-2 entitled “ **HIGHER ACCURACY OF AIR QUALITY INDEX PREDICTION USING MACHINE LEARNING**” under my supervision from January 2023 to May 2023.


Internal Guide

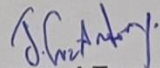
Dr P. AJITHA M.E., Ph.D.,


Head of the Department

Dr. R. SUBHASHINI M.E., Ph.D.,

Submitted for Viva voce Examination held on 10/05/2023


Internal Examiner


External Examiner

DECLARATION

I S Gokulkrishna (Reg. No 39120054) here by declare that the Project Phase-2 Report entitled "HIGHER ACCURACY OF AIR QUALITY INDEX PREDICTION USING MACHINE LERANING" done by me under the guidance of Dr. P. AJITHA M.E., Ph.D., is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Information Technology.

DATE: 10/05/2023

PLACE: Chennai

S. Gokulkrishna

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr. R. SUBHASHINI M.E., Ph.D.**, Head of the Department, Department of Information Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. P. AJITHA M.E., Ph.D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-2 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

ABSTRACT

The aim of this project is to use a heterogeneous ensemble of differential evolution with random forest technique for pollution prediction in Indian Cities. This method is different from existing work like mathematical model for the prediction of air quality of the region. We will use Random Forest algorithm for the prediction of air pollution in the urban region of Indian cities. A random forest algorithm is an ensemble method for regression and classification. We will determine the presence of air pollutants like carbon mono-oxide, carbon dioxide, nitrogen, sulphur dioxide, ozone, pm2.5, pm10. We will also use many more data for our purpose like meteorology data, point of interest data, traffic and road data for determining air quality of large metro cities. This paper will work on an algorithm called Random Forest; the random forest algorithm will be used as classification. The main aspect for determining the air pollution of any region, we will need to know about the real-time information of air quality from monitoring stations like PM2.5, PM10, NO2 as these are main air pollutants which causes more damages to human beings and our environment. If we talk about air quality, then particulate matter is main cause affecting the health of humans and management of city. It also affects government policies. However, in big cities there are very few monitoring stations of air quality. So, in our paper we will talk about Random Forest Technique to predict and Measure Pollution This algorithm is also used for data training and prediction.

TABLE OF CONTENTS

Chapter No.	TITLE	Page No.
	ABSTRACT	v
	LIST OF FIGURES	viii
1	INTRODUCTION	
	1.1 Overview	11
	1.2 Problem Statement	11
2	LITERATURE SURVEY	
	2.1 Inferences from Literature Survey	15
	2.2 Open problems in Existing System	15
3	REQUIREMENTS ANALYSIS	
	3.1 Feasibility Studies/Risk Analysis of the Project	16
	3.2 Software Requirements Specification Document	16
4	DESCRIPTION OF THE PROPOSED SYSTEM	
	4.1 Project Management Plan	20
	4.2 System Architecture	21
	4.3 Selected Methodology	21
	4.4 Description of Software for Implementation	26
	4.5 Transition/Software to operation plan	26
5	DESIGN & DEVELOPMENT	
	5.1 Development and Deployment setup	27
	5.2 Description of software for implementation	27
	5.3 Decision Tree Classifier	28
	5.4 Logistic Regression	28
	5.5 Random Forest	28
	5.6 Random Forest Regression	30

6	RESULT & DISCUSSION	
	6.1 Result	31
	6.2 Discussion	32
7	CONCLUSION & FUTURE WORK	
	7.1 Conclusion	33
	7.2 Future Work	34
	7.3 Research Issues	34
	REFERENCES	35
	APPENDIX	
	A. SOURCE CODE	37
	B. SCREEN SHOT	43
	C. RESEARCH PAPER	44

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	Page No.
4.2.1	System Architecture	21
4.3.1	Proposed Model	25
5.1.1	Random Forest Decision Tree Chart	29

CHAPTER 1

INTRODUCTION

This project forecast the air quality of India by using machine learning to predict the air quality index of a given area. Air quality index of India is a standard measure used to indicate the pollutant (so₂, no₂, rspm, spm. etc.) levels over a period. We developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient decent boosted multivariable regression problem. we improve the efficiency of the model by applying cost Estimation for our predictive Problem.

This model will be capable for successfully predicting the air quality index of a total county or any state or any bounded region provided with the historical data of pollutant concentration. This model by implementing the proposed parameter reducing formulations, we achieved better performance using the standard regression models. Air pollution monitoring has gained attention these days as it has a major impact on the health of humans as well as on the ecological balance. Besides because of toxic emissions on the environment, health, work productivity and efficiency of energy are also affected by the air pollution. Since air pollution has caused many hazardous effects on humans it should be monitored continuously so that it can be controlled effectively. One of the ways to control air pollution is to know its source, intensity and its origin.

Usually, it is monitored by the respective state government's environment ministry. They keep the cord of the pollutant gases in the respective areas. The data presented by the WHO is warning about the pollution's levels in the country. It tells us it's high time that we should monitor 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 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 These gases are cannot be seen or noticed which are produced from burning of fossil fuels, wood burnings, industrial boilers and from the explosion of volcano. They may cause the affects in humans and are the main reason for causing cancer, birth defects and breathing-related problems.

Air Quality Index- Nowadays pollution levels are increasing due to the PM_{2.5} gases which affect the heart functionalities, lung cancer and other respiratory and breathing problems. The long-term damage to the liver, kidney, brain, nerve and other organs in the human body system is affected by air pollution. The AQI is a linear feature of the pollutant concentration. At the same time particulate materials (PM_{2.5} or PM₁₀) in the air can cause several kinds of respiratory, cardiovascular diseases and blood diseases. Medical studies have shown that PM_{2.5} can be easily absorbed by the lung, and high concentrations of PM_{2.5} can lead to respiratory disease or even blood diseases. Air pollution has both acute and chronic effects on human health, affecting a number of different systems and organs.

It ranges from minor upper respiratory irritation to chronic respiratory and heart disease, lung cancer, acute respiratory infections in children and chronic bronchitis in adults aggravating pre-existing heart and lung disease, or asthmatic attacks. Because of direct impact on public health, it has gained a lot of attention. In last decade there is a lot of improvement in the techniques of air quality monitoring and forecasting. There are mainly two approaches in which researchers are working. In first approach monitoring of real-time Air Quality Monitoring and another is developing statistical models using historical data. This paper summarizes Air Quality Prediction studies in two categories. The first study is on prediction of PM_{2.5} / PM₁₀ concentration and on prediction of air pollutants like CO₂, O₃, NO₂ and then inferring Air Quality Index (AQI) using machine learning techniques.

1.1 OVERVIEW

The major goal of this system is to record a picture, video in real-time and determine if a human face is there using the Computer Vision approach. After identifying human faces in a picture and analysing, if the face is recognised, if the face is not recognized then the footage will be captured and will be send to Telegram mobile application.

1.2 PROBLEM STATEMENT

A common problem is the high dimensionality of the data. The common problem of all the existing systems is they are able to detect the AQI from the places in the particular dataset. This detection of the disease is done by taking the pre-existed dataset and collecting all the data from the dataset. The results may differ based on the algorithm we used. Although results are explicit they are having different accuracy scores when compared with each other.

CHAPTER 2

LITERATURE SURVEY

By 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.

Many researchers are working on real-time air quality monitoring and forecasting for getting accurate results which will help in implementing various government policies related to the environment or air pollution and for taking crucial decisions. There are many recent advancements in the air quality forecasting and monitoring techniques. Most of the techniques are Machine Learning (ML) based as it has become popular analysis tool because of its various distinctive features. This paper summarizes air quality forecasting models as well as real time monitoring tools and techniques based on real-time and historical data. It has discussed the merits and demerits of every methodology used for air quality forecasting and monitoring used in recent research along with their comparative analysis, limitations, and challenges.

This paper will be useful to understand current status, past work done and future research questions which needs to be addressed. Keywords— Pollution, air-quality, forecasting, monitoring, health. In this research paper the students have forecasted the air quality of India by using machine learning algorithms to predict the air quality index(AQI) of a given area. Air quality Index is a standard measure to determine the quality of air. Concentration of Gases such as SO_2 , NO_2 , CO_2 , SPM , PM_{10} , etc. are recorded by the agencies.

These students have developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient decent boosted multivariable regression problem. They improved the efficiency of the model by applying cost Estimation for predictive Problem. They say that this model is capable of successfully predicting the air quality index of a total county or any state or any bounded region provided with the historical data of pollutant concentration.

[1] This paper presents an integrated model using Artificial Neural Networks and Kriging to predict the level of air pollutants at various locations in Mumbai and Navi Mumbai using past data available from meteorological department and Pollution Control Board. The proposed model is implemented and tested using MATLAB for ANN and R for Kriging and the results are presented.

[2] This system has used the Linear regression and Multilayer Perceptron (ANN) Protocol for prediction of the pollution of next day. The system helps to predict next date pollution details based on basic parameters and analysing pollution details and forecast future pollution. Time Series Analysis was also used for recognition of future data points and air pollution prediction.

[3] This proposed system does two important tasks (i). Detects the levels of PM_{2.5} based on given atmospheric values. (ii) Predicts the level of PM_{2.5} for a particular date. Logistic regression is used to detect whether a data sample is either polluted or not polluted. Autoregression is employed to predict future values of PM_{2.5} based on the previous PM_{2.5} readings. The primary goal is to predict air pollution level in City with the ground data set.

[4] The major objective of this paper was to provide a snapshot of the vast research work and useful review on the current state-of-the-art on applicable big data approaches and machine learning techniques for air quality evaluation and predication. Air quality map were illustrated and visualized using data from Shenzhen, China. Artificial neural network (ANN), Genetic Algorithm ANN Model, Random Forest, decision tree, Deep belief network are the algorithms which were used and various pros and cons of the model were presented.

[5]The Air Pollution Forecasting System: Air Quality Index (AQI) is a record that gives the public the degree of contamination related with its wellbeing impacts. The AQI centers around the different wellbeing impacts that individuals may encounter dependent fair and square and long stretches of introduction to the poison concentration. The AQI values are not quite the same as nation-to-nation dependent on the air quality norm of the country. The higher the AQI level more noteworthy is the danger of wellbeing related problems.

The by and large point of this venture is to make a student calculation that will have the option to foresee the hourly contamination focus. Additionally, an Android application will be built up that will provide the clients about the constant contamination convergence of PM_{2.5} alongside the hourly forecasted value of the toxin fixation from the student calculation. The Android application will also recommend data of the less dirtied [1].

A brilliant air quality observing framework is proposed in [2] which detects the toxin gases at a specific region and transfer the information into the worker with the goal that everyone can know about the air contamination. The gathered information transferred to the worker anytime in the particular site. This strategy recognizes engines causing toxins and measures different kinds of poisons, and its level in air.

The deliberate data is shared to vehicle proprietor and specialists of the site guests control to keep toxins made through the Air quality observing with occasion based detecting is introduced in [3]. In this work, creators demonstrated that the procedure saves half of the sensor energy utilization contrasted with customary intermittent detecting techniques dependent on the contextual investigation in the city of Spain.

The air quality qualities are anticipated utilizing three twofold ML calculations are introduced in [4]. In this mistake investigation is finished with GLM, SVM and Bayes techniques. The precision of the basic ML techniques are analysed in [5] and variety in the exactness is given various sizes and information divisions. The informational index of air quality comprises of poison information of CO, O₃, NO₂, SO₂, PM₁₀, and PM_{2.5}. For the better air quality expectation, we should co-relate the toxin information with meteorological information.

The air contamination forecast is carried out in [6] based on various standard regularization and improvement calculations as an ML instruments. For pollution estimation or prediction, linear regression algorithms are suitable and for forecasting the pollution levels neural network methods and SVM based methods are preferred [7].

The air quality file is anticipated by utilizing ML calculations for the recognition of PM2.5 level utilizing strategic relapse [8]. There are applications that show the steady PM2.5 levels, while some show the gauge of a particular day. This structure mishandles ML models to perceive and figure PM2.5 levels reliant upon an educational assortment comprising of meteorological conditions in a specific city. The informational index [9] utilized in discovery of PM2.5 level comprises of Temperature, Wind speed, Dew point, Pressure, PM2.5, Concentration ($\mu\text{g}/\text{m}^3$).

2.1 INFERENCES FROM LITERATURE SURVEY

Some Many ML fashions carryout a higher whilst the Facts are typically dispensed and worse awhile the data is skewed make a few ameliorations And comparisons that convert the distribution of athe lids into a aregular adistribution. The discern under indicates that the homes of benzene, toluene, CO and xylene are very distorted. For a these in any other case extra normalized data,alogarithmic adjustments were used to lessen the impact of handing normalized variations in magnitude.

2.2 OPEN PROBLEMS IN EXISTING SYSTEM

A common problem is the high dimensionality of the data.The common problem of all the existing systems is they are able to detect the AQI from the places in the particular dataset.This detection of the disease is done by taking the pre-existed dataset and collecting all the data from the dataset.The results may differ based on the algorithm we used.Although results are explicit they are having different accuracy scores when compared with each other

CHAPTER-3

REQUIREMENT ANALYSIS

3.1 RISK ANALYSIS OF THE PROJECT

In The CPCB dataset blanketed within the take a look at a positive module, particle AQI and authorities organizations use this to warn people about air satisfactory in addition to to exercise its provision Researchers in the discipline suggest that lowering computational values will lessen version prices and improve forecast overall performance. In the present work, the relationship is based totally on the selection method used to determine the most function choice algorithms calculate correlations among each pair of input variable and target.

To sensitive algorithms to float,it's miles important to discover a few feature in the enter dataset that doesn't healthy the generalfashionofthatfacts.For this dataset,a statistical outlier detection method became implemented to discover outliers. For decided on essential functions, correlation analyzes of the AQI function with traits of different pollutants have been performed. Figure three beneath virtually shows that PM10, PM2.Five, CO, NO2, SO2, NOX,NO,.

3.2 SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

Programming: The way in which the software is created today has changed dramatically the last 30 years, from the childhood of personal computers in the early 80s to today's powerful devices such as Smartphones, Tablets and PCs. The Internet has also changed the way of usage in devices and software. Some still have traditional desktop applications, but Web Sites, Web Applications and so- called Apps for Smartphones, etc. are dominating the software market today.The need to find and learn Programming Languages that are suitable for the New Age of Programming. Today several thousand different Programming Languages are being in use.

Software Engineering: Software Engineering is the discipline for creating software applications. A systematic approach to the design, development, testing, and maintenance software.Planning,Requirement Analysis,Design ,Implementation,Testing,Deployment and Mainteanace are the main parts of software Engineering.

Python is highly extendable due to its high number of free available Python Packaged and Libraries. Python can be used on all platforms (Windows, macOS and Linux). Python is multi-purpose and can be used for to program Web Applications, Enterprise Applications and Embedded Applications, and within Data Science and Engineering Applications. The popularity of Python is growing fast. Python is open source and free to use. The growing Python community makes it easy to Find documentation, code examples and get help when needed.

Introduction to Python: Python is an open source and cross-platform programming language, that has become increasingly popular over the last ten years. It was First released in 1991. Latest version is 3.7.0. CPython is the reference implementation of the Python programming language. Written in C, CPython is the default and most widely-used implementation of the language.

Python is a multi-purpose programming language (due to its many extensions), examples are scientific computing and calculations, simulations, web development (using, e.g., the Django Web framework), etc. Python is an object-oriented programming language (OOP), but you can use Python in basic application without the need to know about or use the object-oriented features in Python.

MACHINE LERANING

Machine Learning is a category of algorithms that allow software applications to predict much better results without being specifically programmed. The basic premise of machine learning is to build algorithms that receive input data and use statistical analysis to predict output data while output data is updated like many input data become valid. The processes involved in machine learning are similar to the processes of data mining and predictive modelling. Both require searching for certain patterns by date, and adjusting program actions accordingly.

Many people are also familiar with machine learning from internet shopping and the advertisements that are shown to them depending on what they are buying. This is because referral engines use machine learning to customize ads that are delivered online in near real time. In addition to personalized marketing, other well-known cases in which machine learning is used are fraud detection, spam filtering, threat detection of countries in the network, maintenance, predictability, and building the flow of news.

SUPERVISED LEARNING

They require a data researcher, or data analyst, who has the knowledge of machine learning to supply the desired input and output data, in addition to delivering feedback on the accuracy of the predictions; acute during algorithm training. Data researchers determine which variables, or characteristics, should be analysed by the model and used to develop predictions. Once the training is complete, the algorithm will apply what it has learned to new data.

Supervised learning problems can be further grouped into regression and classification problems. Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”. Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively. Some popular examples of supervised machine learning algorithms are: Linear regression for regression problems. Random forest for classification and regression problems, Support vector machines for classification problems.

UNSUPERVISED ALGORITHMS

They do not need training with output data. Instead, they use a method called deep learning to review the data and come to conclusions. Unsupervised and learned algorithms, also known as neural networks, are used for more complex processes than supervised algorithms, which include image recognition, speech-to-text, and natural language generation.

These neural networks work by first combining millions of training examples with data and automatically identifying subtle correlations between multiple variables. Once trained, the algorithm can be used by associates to interpret new data. These algorithms become feasible only in the information age, because they require massive amounts of data to train. These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher.

Algorithms are left to their own devices to discover and present the interesting structure in the data. Unsupervised learning problems can be further grouped into clustering and association problems. Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y. Some popular examples of unsupervised learning algorithms are: k-means for clustering problems, Apriori algorithm for association rule learning problems.

Python Libraries that would be needed to achieve the task:

1. Numpy
2. Pandas
3. Sci-kit Learn
4. Matplotlib

CHAPTER 4

DESCRIPTION OF THE PROPOSED SYSTEM

The dataset which are entering into the data processing phase are undergoing the process of finding the data shape, data type, elimination of null values etc., the output from the data processing is obtained as a proper dataset with correct values and no repeated values. Then these are entering into the training and testing of dataset.

Here the dataset is trained to the machine and tested by machine. Then the trained dataset enter into the machine learning algorithms where many algorithms are compared for finding the best accuracy result. The supervised Machine learning algorithms are used and the final output is displayed in GUI interface

4.1 PROJECT MANAGEMENT PLAN

The predominant pollutant emissions in India are attributed to business manufacturing, avenue traffic, land and street, waste incineration, power flora, open burning of garbage, etc. This take a look at examines air pollutants facts acquired from the Central Quality Control Board. (CPCB), India.

This dataset includes observations from January 2015 to July 2020 and consists of 12 features with 29,531 samples from 23 specific cities in India. Table 2 under presents a precis of the pollutant/particulate and AQI descriptive statistics from this dataset. Analysis of some foremost air pollutants such as PM_{2.5}, PM₁₀, NO₂, CO, SO₂, O₃, etc. And AQI forecasting is the middle of ongoing work. The methodological steps of the adoption process. Better AQI accuracy, Art and desire of services, Reduce time complexity, may be accessed here, What everyone can manage to pay for and Scalable time complexity.

4.2 SYSTEM ARCHITECTURE

This is the system architecture of the project. This represents the step by step procedure of the implementation. The procedure starts with two step such as train data and test data to classify .

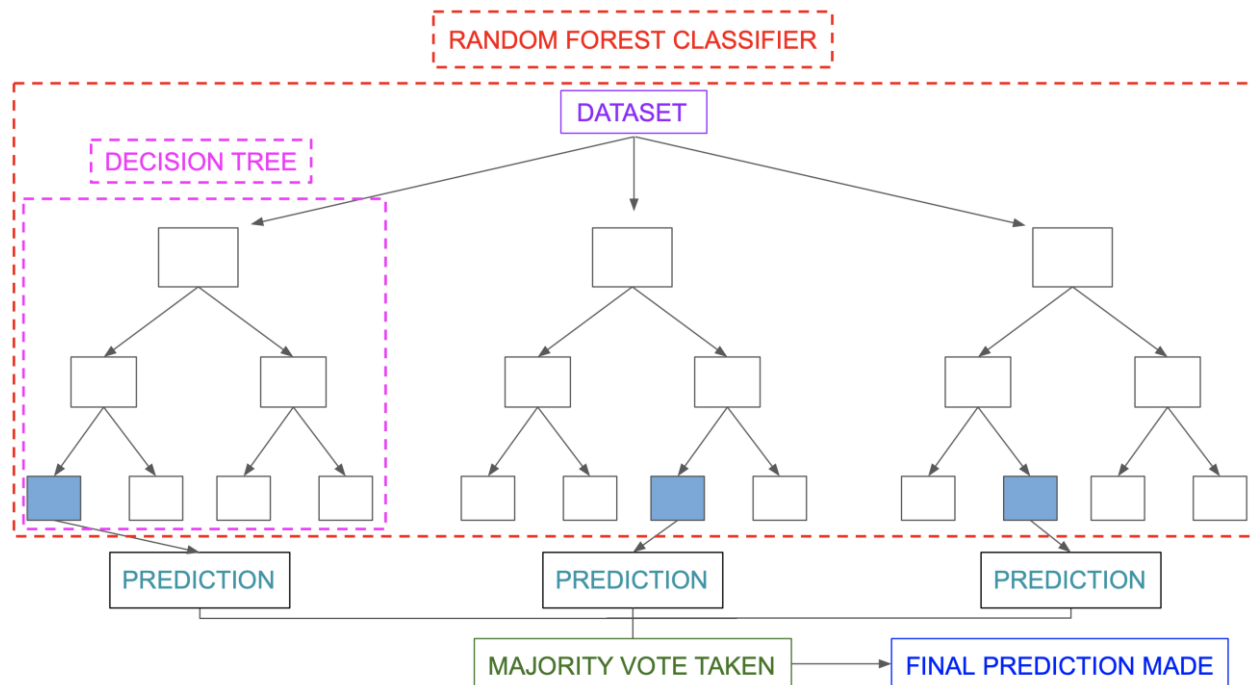


Fig 4.2.1 System Architecture

4.3 SELECTED METHODOLOGY

In this, the data was gathered from the 3 Central Pollution Control Board (CPCB) monitoring stations in Chennai city. The stations are situated at Manali, Velachery and Alandur. The investigative variables collected from these places are relative humidity (RH), PM2.5 values, atmospheric pressure (BP), wind speed (WS) and wind degree (WD). The collected data are presented in 15 min intervals intended for the time of 00:00, 01 May 2019 to 23:59, 30 April 2020 and all station yielded a dataset containing 35039 data rows which total a 105117 data row.

The PM2.5 missing values were 78.28 % approximately. The data was then processed for removing any rows that had empty columns and the information was then limited to rows that had levels of PM 2.5 below 250 $\mu\text{g}/\text{m}^3$. This data left reduced to 22827 data rows as certain elements were not there in all the other rows. The meteorological features like PM2.5, NO₂, SO₂, CO and Ozone were gathered for every 15 min from 3 stations all through Chennai.

The collected data for one year from a period 01 May 2019 to 30 April 2020, along with each station that contributes 35039 information rows and therefore 490546 rows totally in the Dataset. The details of the Dataset which are used for analysis. The implemented Data sets can hold information to be used by a program running on the system. Data sets are also used to store information needed by applications or the operating system itself, such as source programs, macro libraries, or system variables or parameters.

FEATURE EXTRACTION:

Feature extraction approach is computed in which the features are extracted from the data. In the analysis of statistical texture, texture features are evaluated from the numerical distribution of the observed intensity combinations in a particular place on comparing each one in the data. As per the number of intensity pixels (points) in all grouping, data are categorised as first-order, second order and higher-order statistics. The method correlation matrix is a technique of extracting a second-order statistical texture feature. This has been employed in a wide range of applications.

Higher and Third-order textures are regarded as the relationships between three or more pixels. These are possible hypothetically but not implemented commonly because of interpretation difficulty and calculation time. Hence, the feature extraction based on correlation matrix is an effective approach for extracting features. The features can be extracted by using the correlation matrix. The correlation matrix approach is a way to remove statistical texture characteristics of second order.

The technique has been used in many applications. In contrast, third-and higher-order textures find the interaction of three or more pixels. The correlation matrix is an arithmetic purpose, which normally can effectively remove the artefacts. The picture quality can also be clearly distinguished. The picture can be isolated for the process of study. Use the correlation matrix with this to remove the feature's functionality. In an exact differential area, correlation matrix can specify the frequency of the pixels. The single pixel is to be questioned here and another pixel is to be known as the route l and the adjacent value detachment of m . Usually, m obtains a single value, and can benefit directionally. Then the obtained directional value can remove the attributes of the images used for the segmentation process. The correlation matrix process may be set.

MODEL TRAINING:

According to Isakndaryan et al. [19], the most used metrics are RMSE (root mean squared error) and MAE (mean average error), calculated based on the difference between the prediction result and the true value, while another metric, R^2 (R-squared) is essential to explain the strength of the relationship between predictive models and target variables [20]. These three metrics provide a baseline for comparative analysis across different parameter settings for each model and across different methods.

However, performance validation leads to a bias when the data set is split, trained, and tested only one time. This also means the result drawn from the testing dataset may no longer be valid after the testing subset is changed. To overcome this problem, each model is re-built 20 times using different random subsets of training and testing samples. The splitting proportion remains the same. All metrics report only a single value from the average performance of 20 identical models validated into 20 different subsets of testing instances.

TESTING MODEL

Random forest regression is used for the training data and testing data to assess the performance of AQI for SO₂. In Fig. 8, the original Dataset is shown in red for the SO₂ versus AQI index value, the prediction of the AQI value is shown in green for the SO₂. It is seen that the model performs a good job in fitting both training and test data. The MSE value and R² values are obtained for SO₂ as 1.005 and -0.005, respectively. The AQI values for the observed data and predicted data for the SO₂ is depicted in Fig. 9. The proposed model is used for the training data and testing data to assess the performance of AQI for CO. In Fig. 10, the original Dataset is shown in red for the CO versus AQI index value, the prediction of the AQI value is shown in green for the CO. It is seen that the model performs a good job in fitting both training and test data. The MSE value is 0.920 and R² value is 0.080. The AQI values for the observed.

PERFORMANCE EVALUATION

The predictions are generated using the random forest regression for both the training and testing data to assess the performance of the model. the original Dataset is shown in red colour for the PM_{2.5} versus AQI index value, the prediction of the AQI value is shown in green colour for the PM_{2.5}. It is seen that the model performs as a good job in fitting both training and test data. The seasonal factors dependent upon the time was depicted (a). The continuous monitoring of the seasonal factors was compulsory for effective prediction of the quality of the air. The RMSE & R² values are obtained and described. Training algorithms are maintained in the suggested classifier and a few steps are required for training and prediction computation.

The first step is data entry of inputs stage, and the second step is creating classifier system with membership function. Training of given input data with selection of learning algorithm is the third step, and after learning, performing iterative calculations for test data and train score. The random forest regression model for Regression. The result predicts that the model has an average error of 7.804 µg/m³ for the training set and 10.995 µg/m³ for the test set. The value of R² obtained with training and testing set is 0.632 and 0.570, respectively, which indicates that the model is the best fit for PM_{2.5} prediction.

The performance analysis of the PM2.5 for observed data is shown in red and predicted AQI value for PM2.5 is shown in green and plotted (b). The random forest regression method predicts the AQI value for PM2.5 by applying the proper training of the model and predicts the correct values. The Mean squared Error and R2 value are measured for PM2.5 are 0.179 and 0.821, respectively.

PREDICTION

Machine learning techniques are used to aid in this process, and the accuracy level of AQI for each city is estimated. Machine learning prediction, or prediction in machine learning, refers to the output of an algorithm that has been trained on a historical dataset. The algorithm then generates probable values for unknown variables in each record of the new data. The purpose of prediction in machine learning is to project a probable data set that relates back to the original data. This helps organizations predict future customer behaviours and market changes. Essentially, prediction is used to fit a shape as closely to the data as possible. With machine learning predictions, organizations use proactive decisions to avoid predicted user churn. To gain the most success with prediction in machine learning, organizations need to have infrastructure in place to support the solutions, and high-quality data to supply the algorithm.

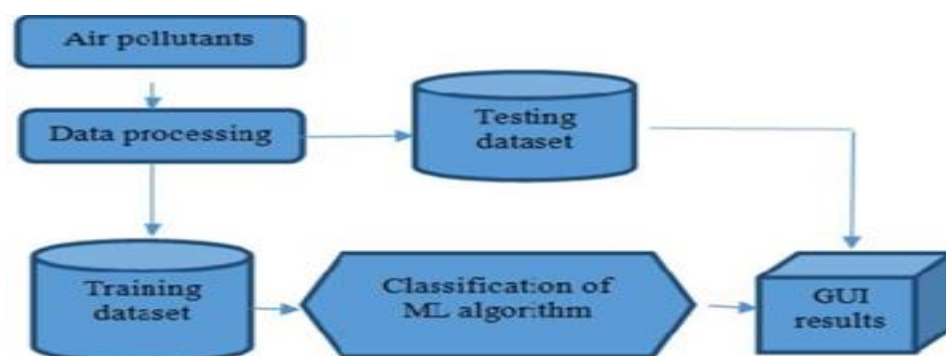


FIG 4.3.1:PROPOSED MODEL

4.4 DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION

The proposed work of this project involved collecting Information approximately air quality is extracted from the sensors, that are processed in a single scheme and stored for dependable facts. This dataset is pre- processed the usage of diverse capabilities together with normalization, characteristic choice, and discretization. Once the charts are geared up, it's far split right into a training dataset and a take a look at dataset. In addition, device getting to know algorithms had been implemented to the training dataset. The consequences acquired are in comparison with the test information and the evaluation results. Reborn 1 describes the proposed architectural model. Air pollutants prediction the use of the desired machine gaining knowledge of method considers 4 device learning algorithms inclusive of LR, SVM, DT, and RF.

4.5 TRANSITION/ SOFTWARE TO OPERATIONS PLAN

Our Main Objective of the Project is to predict the air quality index based on machine learning algorithms for better accuracy. So we used multiple various libraries like numpy, matplotlib, pandas, scikit, jinja and other software features like flask, gunicorn, markupsafe etc.. Using Random Forest Regression algorithm train and test data be analyzed for better prediction. Model and main app folder be created to store and predict values for air quality index. Web based application have been created to predict the air quality index. A Base bgsript first created then using various img web page can be created/modified based on user interest. Easy to navigate button has been created to start/begin the process of predicting. For before getting result, essential values related to air needed to be entered first. Then based on collective information stored values the model application able to make validation process and predict the accurate air quality index.

CHAPTER 5

DESIGN AND DEVELOPMENT

5.1 DEVELOPMENT AND DEPLOYMENT SETUP

This Each network attempts to are expecting the outcome as appropriately as possible. The accuracy of the cost within the network effects from a cost characteristic that attempts to punish the community whilst it fails. The exceptional performance is the least value. In this study, for all networks, we applied MSE (Mean Square Error) to the price feature. Repetition steps in training commonly work with a division of the education data, known as the batch size. The variety of samples in each batch is a hyper parameter this is normally acquired through trial and mistakes. The cost of this parameter in all models is 24, 32, 64 and 128, respectively, as proven in this examine. At every generation stage, the price feature is calculated as the average of these MSE found and expected awareness patterns *PM2.5*. The quantity of generation steps for a neural community is referred to as an epoch; In every epoch, the time series of the river runoff from the community is sampled as soon as. As in different regulations, neurons or layers of the recurrent network may be decided on at will. In our study, to compare the models with each other, the structures of all recurrent fashions are created in the equal way.

5.2 DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION

The Information approximately air quality is extracted from the sensors, that are processed in a single scheme and stored for dependable facts. This dataset is preprocessed the usage of diverse capabilities together with normalization, characteristic choice, and discretization. Once the charts are geared up, it's far split right into a training dataset and a take a look at dataset. In addition, device getting to know algorithms had been implemented to the training dataset. The consequences acquired are in comparison with the test information and the evaluation results. Reborn 1 describes the proposed architectural model. Air pollutants prediction the use of the desired machine gaining knowledge of method considers 4 device learning algorithms inclusive of LR, SVM, DT, RF.

5.3 DECISION TREE CLASSIFIER

The structure of a choice tree includes the root, inner and leaf nodes. The choice tree version depends on the extraordinary collections of rows inside the dataset. The type tree decision approach is executed in steps: tree creation and tree pruning. A tree is made from pinnacle to backside with appreciate to the target variable. The version of the falling tree analyzes the records primarily based on three nodes, particularly: The root node - that is the principle node from which all of the others carry out their node - This node handles the country of based variables. Leaf node - the final result is transferred to the leaf node.

5.4 LOGISTIC REGRESSION

The Logistic regression is a classification method used to assess studying the opportunity of the target being variable. When the character of the goal or established variable is dichotomous, simplest two training are feasible. The dependent variable is binary and given either 1 (bypass/sure) or zero (fail/no). A logistic regression version mathematically predicts $P(Y = 1)$ as a feature of X . This is a totally primary system studying technique. Can be used to resolve various class problems along with junk mail detection, diabetes prediction; cancer diagnostics and extra Below is an instance of a logistic equation.

5.5 RANDOM FOREST

Use Another prominent machine learning method, random forest, a supervised learning ensemble algorithm, combines multiple decision trees to form a forest and the bagging concept, that latter adds the randomness into the model building. The random selection of features is used to split the individual tree while the random selection of instances is used to create training data subset for each decision tree

At each decision node in every tree, the variable from the random number of features is considered for the best split. If the target attribute is categorical, random forests will choose the most frequent as its prediction. On the other hand, if it's numerical, the average of all predictions will be chosen. Similar to SVM, the random forest can tackle both classification and regression case.

The trees then vote on an outcome and the prediction is produced from a majority vote among the models and henceforth resulting in a stronger and more robust single learner. Random forests can overcome the prediction variance that each decision tree has, in the way that the prediction average will approximate the ground truth (classification) or true value (regression).

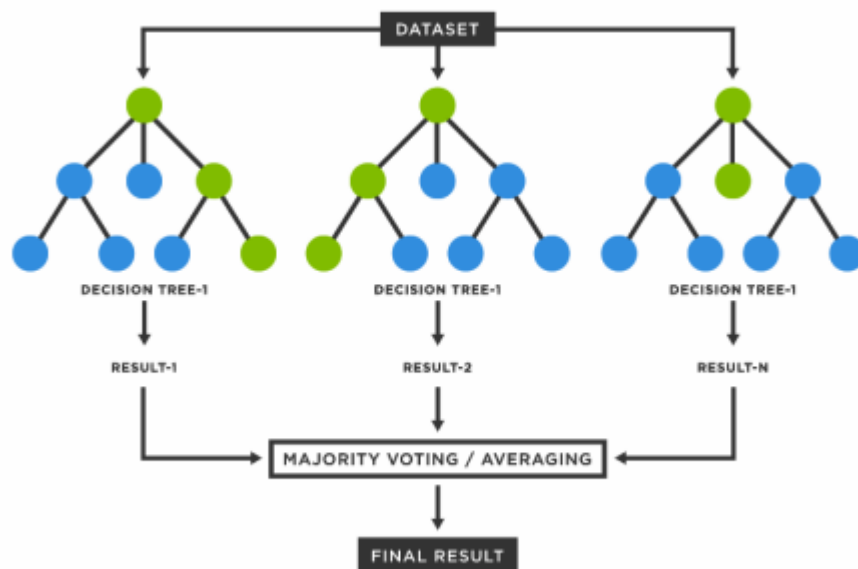


Fig 5.5.1 RANDOM FOREST DECISION TREE CHART

5.6 RANDOM FOREST REGRESSION

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or "the random subspace method", generates a random subset of features, which ensures low correlation among decision trees. This is a key difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset.

CHAPTER 6

RESULTS & DISCUSSION

6.1 RESULT

The primary goal of the project is to develop a reliable On continuing with our Air quality prediction project we recently making this first page description how we predict the air quality AQI of particular city and after getting the AQI of particular parameter it's shows its level for our health or not and then we decided to survive or not in this city properly. The air quality forecast has aroused attention from governments and scientists for improving environmental quality of citizens and upgrading environmental sensing studies. Because numerous scientific studies have linked PM2.5 (particulate matter) to health problems, this study proposes a machine learning approach based on GBM to predict the PM2.5 concentration.

By extracting temporal sequences from air monitoring and weather stations into features, the proposed mechanism efficiently predicts the air quality in terms of the PM2.5 concentration for the next 24 h at individual air monitoring stations. This study investigated the geographical and meteorological divergence of forecasting results and measurements at regional monitoring areas. We observe that city has a more complex environment due to domestic pollution sources, overseas pollution sources, and geographic topography; however, it exhibits a similar prediction performance. This may be because cities have similar topography and are financial centres without domestic pollution sources. Finally, because domestic industrial pollution makes prediction difficult, this study proposes an enhanced fusion version of the prediction model by incorporating additional features.

6.2 DISCUSSION

After considering the industrial impact, the proposed method achieves significant improvement in the values and considerably decreases the compared with the conventional approach for City. The studied method produced a suitable model of the hourly atmospheric pollution, allowing us to obtain, generally, good accuracy in modelling pollutant concentrations. Since our model is capable of predicting the current data with more accuracy it will successfully predict the upcoming air quality index of any particular data within a given region.

With this model we can forecast the AQI and alert the respected region of the country also it a progressive learning model it is capable of tracing back to the particular location needed attention provided the time series data of every possible region needed attention. The essential perspectives that should be viewed as with regards to gauging of the poison focus are its different sources alongside the components that impact its fixation. With good results for air quality prediction system, it offer the best performance of target predictions based on datasets. At the same time, being able to model, predict, and monitor air quality is becoming more and more important, especially in urban areas, due to the observed critical impacts of air pollution for populations and the environment.

CHAPTER 7

CONCLUSION & FUTURE WORK

7.1 CONCLUSION

In conclusion, utilizing supervised ML methods, our system successfully .Air quality forecasting is a complicated venture due to the dynamic nature, volatility and excessive spatial and temporal variability of pollutants and debris. This paintings supplied a random wooded area examine to predict pollutant and particle levels and correctly become aware of AQI. To be greater particular, randomization offers the exceptional results, at the same time as random forest offers the great result. Only offers tremendous consequences for predictions. The effects additionally confirm that the system mastering strategies used in this have a look at are popular techniques within the literature, such as SVM, logistic regression, and ANN.

In other words, random woodland can be taken into consideration as a new and advanced opportunity for AQI prediction. This have a look at also shows that overall performance prediction varies by area. Comparing the outcomes of the datasets of the three areas shows the best outcomes for AQI prediction, where the impact of the decrease with the boom of the time step is less. In addition, ninety five% self belief intervals are calculated, respectively. Compared to a single price for prediction, it could offer a better reference for the choice maker. For instance, the organizer of the event can decide whether to proceed with actions based totally on air nice forecasts with higher self assurance.

Predicting the air quality is a complex task due to the dynamic nature, volatility, and high variability in space and time of pollutants and particulates. This work presented a study of random forest to forecast pollutants and particulates' levels and to correctly identify the AQI. To be more specific, the stacking ensemble delivers the best results, while random forest provides the best results. only provides meaningful results for predictions. The results also confirm that the machine learning methods, employed in this study can outperform popular methods in the literature, such as SVM, logistic regression and ANN. In other words, random forest can be considered new and superior alternatives for AQI forecast.

This study also indicates that prediction performance varies over different regions. Comparing results from datasets sourced from three different regions displays best results for AQI prediction where performance decay with increased time step is less. Also, 95% confidence intervals forecast are calculated, respectively. Compared to the single value prediction, can provide a better reference to the decision-maker. For example, an event planner can decide if the outdoor activities can go on based on the air quality forecast with better confidence.

7.2 FUTURE WORK

This Future work should focus on improving performance using stacking ensemble, random forest with hyperparameter optimization, particularly for predictions with larger time steps.

7.3 RESEARCH ISSUES

During the Research of this project, some issues have to be addressed to ensure the success of the project. Information or reference papers regarding the subject were not widely available. In preview of that it was also found that Lesser data of aqi and air pollutants available and not upto date. But here there has been a new thought including latest available data of previous years of indian cities to get trained and tested to predict the accurate results of aqi .so adding multi layers of data and algorithms helps to get this project implemented.

REFERNCES

- [1] 'Blacksmith Institute Press Release'. (October 21, 2008). [Online]. Available: <http://www.blacksmithinstitute.org/the-2008-top-ten-list-of-world-s-worst-pollution-problems.html>
- [2] V. M. Niharika and P. S. Rao, "A survey on air quality forecasting techniques," International Journal of Computer Science and Information Technologies, vol. 5, no. 1, pp.103-107, 2014.
- [3] NAAQS Table. (2015). [Online]. Available: <https://www.epa.gov/criteria-air-pollutants/naaqs-table>
- [4] E. Kalapanidas and N. Avouris, "Applying machine learning techniques in air quality prediction," in Proc. ACAI, vol. 99, September 1999.
- [5] Questioning smart urbanism: Is data-driven governance a panacea? (November 2, 2015). [Online]. Available: <http://chicagopolicyreview.org/2015/11/02/questioning-smart-urbanism-is-data-driven-governance-a-panacea/>
- [6] D. J. Nowak, D. E. Crane, and J. C. Stevens, "Air pollution removal by urban trees and shrubs in the United States," Urban Forestry & Urban Greening, vol. 4, no. 3, pp. 115-123, 2006.
- [7] T. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations," presented at 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), IEEE, 2016.
- [8] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in Proc. the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2267-2276, August 10, 2015.
- [9] J. A. Engel-Coxa, C. H. Hollomanb, B. W. Coutantb, and R. M. Hoffc, "Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality," Atmospheric Environment, vol. 38, issue 16, pp. 2495–2509, May 2004.
- [10] J. Y. Zhu, C. Sun, and V. Li, "Granger-Causality-based air quality estimation with spatio-temporal (ST) heterogeneous big data," presented at 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2015.
- [11] C. J. Wong, M. Z. MatJafri, K. Abdullah, H.S. Lim, and K. L. Low, "Temporal air quality monitoring using surveillance camera," presented at IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2007.
- [12] S. Y. Muhammad, M. Makhtar, A. Rozaimie, A. Abdul, and A. A. Jamal, "Classification model for air quality using machine learning techniques," International Journal of Software Engineering and Its Applications, pp. 45-52, 2015.
- [13] A. Sarkar and P. Pandey, "River water quality modelling using artificial neural network technique," Aquatic Procedia, vol. 4, pp. 1070-1077, 2015.
- [14] E. Kalapanidas and N. Avouris, "Applying machine learning techniques in air quality prediction," Sept. 1999.

- [15] H. Zhao, J. Zhang, K. Wang, et al., "A GA-ANN model for air quality predicting," IEEE, Taiwan, 10 Jan. 2011.
- [16] V. Jagannath. [Online]. Available: <https://community.tibco.com/wiki/random-forest-template-tibco-spotfire-wiki-page>
- [17] R. Yu, Y. Yang, L. Yang, G. Han, and, O. A. Move, "RAQ—A random forest approach for predicting air quality in urban sensing systems," Sensors, vol. 16, no. 1, p. 86, 2016.
- [18] Machine learning with decision trees. [Online]. Available: <https://blog.knoldus.com/2017/08/14/machine-learning-with-decisiontrees/>
- [19] S. Deleawe, J. Kuszniir, B. Lamb, and D. J. Cook, "Predicting air quality in smart environments," J Ambient Intell Smart Environ., pp. 145-152, 2010.
- [20] W. F. Ip, C. M. Vong, J. Y. Yang, and P. K. Wong, "Least squares support vector prediction for daily atmospheric pollutant level," in Proc. 2010 IEEE/ACIS 9th International Conference on Computer and Information Science (ICIS), pp. 23-28, IEEE., August 2010.
- [21] L. Xiang, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," Environmental Science and Pollution Research, vol. 23, no. 22, pp. 22408-22417, 2016. [22] J. Gao, C.-L. Xie, and C.-Q. Tao, "Big data validation and quality assurance - issues, challenges, and needs," IEEE Symposium on Service-Oriented System and Engineering, IEEE Computer Society, Oxford, UK, April, 2016.

APPENDIX

A. SOURCE CODE

Model.py:

```
# Importing the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RandomizedSearchCV
import pickle
from sklearn.linear_model import LinearRegression

df=pd.read_csv("Real_Combine.csv")

df=df.dropna()
X=df.iloc[:, :-1] ## independent features
y=df.iloc[:, -1] ## dependent features

## Check Null values
X.isnull()
y.isnull()

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.3,random_state =
0)

regressor= RandomForestRegressor()
regressor.fit(X_train,y_train)

score=cross_val_score(regressor,X,y,cv=5)
score.mean()
prediction=regressor.predict(X_test)
RandomForestRegressor()

n_estimators = [int(x) for x in np.linspace(start = 100, stop = 1200, num = 12)]

# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 100, stop = 1200, num = 12)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(5, 30, num = 6)]
```

```

# max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10, 15, 100]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 5, 10]
# Method of selecting samples for training each tree
# bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf}

# Use the random grid to search for best hyperparameters
# First create the base model to tune
rf = RandomForestRegressor()
# Random search of parameters, using 3 fold cross validation,
# search across 100 different combinations
rf_random = RandomizedSearchCV(estimator = rf, param_distributions =
random_grid, scoring='neg_mean_squared_error', n_iter = 100, cv = 5, verbose=2,
random_state=42, n_jobs = 1)
rf_random.fit(X_train,y_train)
rf_random.best_score_
predictions=rf_random.predict(X_test)

# open a file, where you ant to store the data
file = open('random_forest_regression_model.pkl', 'wb')

# dump information to that file
pickle.dump(rf_random, file)
print(model.predict([[7.4,9.8,4.8,1017.6,93,0.5,4.3,9.4]]))

regressor = LinearRegression()

#Fitting model with trainig data
regressor.fit(X, y)

# Saving model to disk
pickle.dump(regressor, open('model.pkl','wb'))

# Loading model to compare the results
model = pickle.load(open('model.pkl','rb'))
print(model.predict([[2, 9, 6]]))

```

RANDOM FOREST REGESSOR.PY :

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import seaborn as sns
from sklearn.ensemble import ExtraTreesRegressor
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RandomizedSearchCV
import sklearn.metrics as metrics
from sklearn import metrics
import pickle
import pandas as pd
import sklearn

df=pd.read_csv("Real_Combine.csv")

df.head()

sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')

df=df.dropna()

X=df.iloc[:, :-1] ## independent features
y=df.iloc[:, -1] ## dependent features

## Check Null values
X.isnull()

y.isnull()
sns.pairplot(df)

df.corr()

#get correlations of each features in dataset
corrmat = df.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))
#plot heat map
g=sns.heatmap(df[top_corr_features].corr(),annot=True,cmap="RdYlGn")

corrmat.index

model=ExtraTreesRegressor()
model.fit(X,y)
```

```

X.head()

print(model.feature_importances_)

feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(5).plot(kind='barh')
plt.show()
sns.distplot(y)

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.3,random_state =
0)

regressor= RandomForestRegressor()
regressor.fit(X_train,y_train)

print("Coefficient of determination R^2 <-- on train set:
{}".format(regressor.score(X_train, y_train)))

print("Coefficient of determination R^2 <-- on train set:
{}".format(regressor.score(X_test, y_test)))

score=cross_val_score(regressor,X,y,cv=5)

score.mean()

# ### Model Evaluation

prediction=regressor.predict(X_test)

sns.distplot(y_test-prediction)

plt.scatter(y_test,prediction)

RandomForestRegressor()

n_estimators = [int(x) for x in np.linspace(start = 100, stop = 1200, num = 12)]
print(n_estimators)

#Randomized Search CV

n_estimators = [int(x) for x in np.linspace(start = 100, stop = 1200, num = 12)]

max_features = ['auto', 'sqrt']

max_depth = [int(x) for x in np.linspace(5, 30, num = 6)]

```

```

min_samples_split = [2, 5, 10, 15, 100]

min_samples_leaf = [1, 2, 5, 10]

# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf}

print(random_grid)

rf = RandomForestRegressor()

rf_random = RandomizedSearchCV(estimator = rf, param_distributions =
random_grid, scoring='neg_mean_squared_error', n_iter = 100, cv = 5, verbose=2,
random_state=42, n_jobs = 1)

rf_random.fit(X_train,y_train)

rf_random.best_score_

predictions=rf_random.predict(X_test)

sns.distplot(y_test-predictions)

plt.scatter(y_test,prediction)

print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))

print('MAE:', metrics.mean_absolute_error(y_test, prediction))
print('MSE:', metrics.mean_squared_error(y_test, prediction))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, prediction)))

# open a file, where you ant to store the data
file = open('random_forest_regression_model.pkl', 'wb')

# dump information to that file
pickle.dump(rf_random, file)

print(model.predict([[7.4,9.8,4.8,1017.6,93,0.5,4.3,9.4]]))

```

App.py:

```
#pip install Flask==2.1.0
# pip install Werkzeug==2.0.3
# pip install markupsafe==2.0.1
import numpy as np
from flask import Flask,render_template,url_for,request,jsonify
import pandas as pd

import pickle

# load the model from disk
model=pickle.load(open('random_forest_regression_model.pkl', 'rb'))
app = Flask(__name__)

@app.route('/',methods=['GET'])
def Home():
    return render_template('home.html')

@app.route('/predict',methods=['POST'])
@app.route("/predict", methods=['POST'])
def predict():

    if request.method == 'POST':
        T= float(request.form['T'])
        TM= float(request.form['TM'])
        Tm= float(request.form['Tm'])
        SLP= float(request.form['SLP'])
        H= int(request.form['H'])
        VV= float(request.form['VV'])
        V= float(request.form['V'])
        VM= float(request.form['VM'])
        prediction=model.predict([[T,TM,Tm,SLP,H,VV,V,VM]])
        output=round(prediction[0],2)

        return render_template('home.html', prediction_text='Air Quality Index is
{}'.format(output))
    else:
        return render_template('home.html')

if __name__ == '__main__':
    app.run()
```

B. SCREEN SHOTS

AIR QUALITY INDEX PREDICTION

Temperature:

Maximum Temperature:

Minimum Temperature:

Sea Level Pressure:

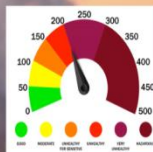
Rainfall/Snowmelt:

Visibility:

WindSpeed:

Max Sustain WindSpeed:

Want to know India Air Quality Index (AQI)?



The gauge shows a needle pointing to approximately 150, which is in the yellow 'MODERATE' range. The scale ranges from 0 to 500 with color-coded segments: Green (0-50), Yellow (50-100), Orange (100-150), Red (150-200), Purple (200-250), and Dark Purple (250-300).

25°C Partly cloudy

AIR QUALITY INDEX PREDICTION

Temperature:

Maximum Temperature:

Minimum Temperature:

Sea Level Pressure:

Rainfall/Snowmelt:

Visibility:

WindSpeed:

Max Sustain WindSpeed:

Want to know India Air Quality Index (AQI)?



The gauge shows a needle pointing to approximately 85, which is in the yellow 'MODERATE' range. The scale ranges from 0 to 500 with color-coded segments: Green (0-50), Yellow (50-100), Orange (100-150), Red (150-200), Purple (200-250), and Dark Purple (250-300).

Air Quality Index is 85.85

HIGHER ACCURACY OF AIR QUALITY INDEX PREDICTION USING MACHINE LEARNING ALGORITHM

S. GOKUL KRISHNA
UG SCHOLAR
Dept. OF IT
Sathyabama Institute of
Science and Technology
Chennai, India
gokulkrisshna@gmail.com

DR.P.AJITHA
ASSOCIATE PROFESSOR
Dept. OF IT
Sathyabama Institute of
Science and Technology
Chennai, India
ajithaponnupilai@gmail.com

Abstract – Now-a--days Human Air quality includes harmful gases and first-class particulate be counted that have an effect on air high-quality. It has mostly Ineffective essential query of medical studies. Therefore, many experts and students from various studies centers, universities and overseas are worried in a large number of PM quality prediction studies. In this situation, the authors proposed different machine getting to know models consisting of linear regression, random jump, KNN, ridge and lagged models, XG Boost and Ada Boost fashions to are expecting PM pollutants in polluted cities. This take a look at is done using the Jupyter code book in Python . From the outcomes for MAE, MAPE, and RMSE rankings. Between XG Boost, Ada Boost, Random Forest, and KNN fashions more reliable models are considered. The variety of pollutant Concentrations observed in those fashions. The fashions can both predict pollutant and higher predict air high-quality ranges. When comparing the numerous present models and the proposed fashions, it became found that the proposed models predict pollutants with better outcomes and a lower blunders price than the present one. The purpose of the article is to use an web based predictive application on Air quality index and to indentify the accurate air level of the readings predicted by machine learning algorithms .

This project focuses on insatnt and accurate air quality prediction which can be accessed on web pages on the gadgets with supports good internet connectivity.

Keywords- AirQuality, DecisionTree, KNN, Linear Regression, Machine Learning, Random Forest , SVM, AdaBoost, XGBoost, MAPE .

I. INTRODUCTION

When a At gift, accurate estimation of air pollutants is hard and critical because of the growth of air pollutants, that is a essential problem in lots of parts of the arena. Quality is typically divided into two kinds: (1) herbal pollutants due to volcanic eruptions and wooded area fires, resulting in emissions of SO₂, CO₂, CO, NO₂ and sulfates as air pollutants, and (2) anthropogenic quality due to a few such human sports. Which include oil burning, industrial emissions and delivery emissions that include PM_{2.5} as predominant air quality that receive a lot of attention because of their devastating outcomes on human health, different species, and the surroundings. Various research display that air pollutants leads to respiration and cardiovascular illnesses main to the dying of animals and plant life, acid rain, weather exchange, worldwide warming, etc., which reasons financial damage and makes it hard for human society to live on. . Regarding the impact of PM_{2.5} studied during the last 25 years the use of machine learning techniques analysis. Ameer et al. Envisioned that about 4.2 million humans died due to long-

time period publicity to PM2.5 inside the ecosystem, and any other 250,000 people died from ozone publicity.

II. LITERATURE SURVEY

A literature review is a body of writing that seeks to summarize the key ideas in the field's current understanding and/or methodological approaches to a certain subject. It's secondary sources that talk about stuff that's already been published in a given field, and sometimes it talks about stuff that's already been published in a certain field during a certain time frame. Its primary purpose is to provide the reader with a summary of the most current findings in a particular area of study, and it also lays the groundwork for other goals, such as subsequent research. An annotated bibliography might serve as this step's equivalent. It has a format and often incorporates synthesis and summary. Difference between a summary and a synthesis is that the latter just restates the source's most important points, while the former reorganizes and restructures the information. It may provide a new perspective on established findings, combine new and old ideas, or trace the evolution of the field's thinking across time, highlighting important debates along the way. Depending on the context, the literature review may provide an evaluation of the available sources and recommendations on which are the most useful to the reader.

One author named gopal getting to know best in extraordinary regions He positioned locations where in records become unavailable. The creator has evolved an internet software to predict air quality in any area of the town. Sanjeev examined the dataset between pollutant concentrations. The writer analyzed and anticipated air best and high-quality itsfar overfitting. Another author .Maduri observe that windvelocity and course, humidity and temperature play an vital rolein the retention of air pollutants. The authors used set of rules whichshowed minimum category errors. Author moni sri gathered air quality records from diverse sources and attempted to broaden a blended version for air prediction. The authors argue that the proposed versionis designed to research the population small cities expecting airfirst-rate. The authors analyzed statistics accumulated over a duration of 28 . They found that the version they proposed turned into quite accurate for the maximum polluted regions. Patil supplied

numerous literature works on numerous device studying methods for AIR modeling and prediction. castelli attempted to expect air pleasant at the pollutant and particulate stages using the ML vector regression set of rules. The authors claimed to have evolved a brand new approach for modeling atmospheric pollutants hours. The author Doreswami investigated predictive system mastering models to predict air PM concentration. The authors investigated six years of huge-scale air satisfactory statistics in Taiwan and applied current fashions. They asserted that the aforementioned goods and the goods themselves had been very near every different. Liang studied the performance of six machine learning classifiers to record Taiwan AQI based on research statistics. The authors mentioned that Adapting Boost and are most suitable for air satisfy prediction, however the overall across geographies. Madan as compared 20 distinctive literatures across pollutants studied, applied observe algorithms, and their characteristics. The authors observed that a few of the papers blanketed meteorological statistics such as humidity, wind pace, and temperature to extra appropriately are expecting They discovered a that neural networks and version schooling are the notable studying algorithms.

III. EXISTING SYSTEM

A common problem is the high dimensionality of the data. The common problem of all the existing systems is they are able to detect the AQI from the places in the particular dataset. The results may differ based on the algorithm we used. Although results are explicit they are having different accuracy scores when compared with each other.

IV. PROPOSED SYSTEM

The proposed work of this project involved collecting Information approximately air quality is extracted from the sensors, that are processed in a single scheme and stored for dependable facts. This dataset is pre- processed the usage of diverse capabilities together with normalization, characteristic choice, and discretization. Once the charts are geared up, it's far split right into a training dataset and a take a look at dataset. In addition, device getting to know algorithms had been implemented to the training dataset. The consequences acquired are in comparison with the test information and the evaluation results. Reborn 1 describes the

proposed architectural model. Considering the failure of all of the methods used inside the previous structures, the maximum commonplace query become selecting the great algorithm. For one data set, if one algorithm offers better accuracy, for any other data set, another information set offers higher accuracy. So I came up with a strategy to keep away from guide dataset and testing. Here we're going to take good attributes from the user

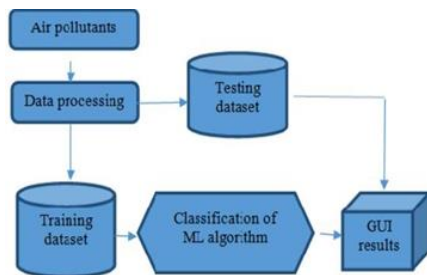


FIG:1.PROPOSEDARCHITECTURE

Every attempts are expecting outcome as appropriately. Accuracy cost within the effects from a cost characteristic that attempts to punish the community whilst it fails. The exceptional performance least value. Repetition steps in training commonly work with a division of the education data, known as the batch size. The variety of samples in each batch is a hyper parameter this is normally acquired through trial and mistakes. At every generation stage, the price feature is calculated as the average of these MSE found and expected awareness patterns *PM2.5*. The quantity of generation steps for a neural community is referred to as an epoch; In every epoch, from the community is sampled as soon as. As in different regulations, neurons or layers of the recurrent network may be decided on at will. In our study, to compare the models.

V. WORKING

Some Indian cities are within the listing of maximum polluted air quality is growing worse every day. Worse air first-rate in India is now considered a first-rate fitness difficulty and a primary impediment to economic increase. Air pollutants in India reasons annual losses of up to 7 million rupees

(\$ninety five billion) (Dalberg 2019), in keeping with a brand new take a look at published mutually via the British non-earnings management business enterprise Dalberg Advisors and the Industrial Corporation. The predominant air substance created based on industry manufacturing,avenue traffic ,land ,street and also burning waste such as non degradable and electronic waste plays a part etc..so after collecting the required data of air substance from the database ,we need to analyze the data whenever its accurate or not and then use it with our usecase to find required amount of index prediction.Data Source - The dataset used to expect coronary heart sickness is from the UCI machine mastering . UCI is a hard and fast of databases used for device mastering algorithms. The statistics published here is actual information. The dataset consists of 300 times facts with 14 parameters. A clinical facts set approximately assessments that are carried out in connection with coronary heart conditions, which includes blood pressure levels, chest ache.Additionally,description dataset has been incorporated into the model which includes 40 rows.

Decision Tree Classifier

The structure of a choice tree includes the root, inner and leaf nodes. The choice tree version depends on the extraordinary collections of rows inside the dataset. The type tree decision approach is executed in steps: tree creation and tree pruning. A tree is made from pinnacle to backside with appreciate to the target variable. The version of the falling tree analyzes the records.

Logistic Regression

Logistic regression is a classification method used to assess studying the opportunity of the target being variable. When the character of the goal or established variable is dichotomous, simplest two training are feasible. The dependent variable is binary and given either 1 (bypass/sure) or zero (fail/no). A logistic regression version mathematically predicts P as a feature of X. This is a totally primary system studying technique.Can be used to resolve various class problems along with junk mail detection,diabetes prediction;cancer diagnostics.

Random Forest

A random wooded area is a supervised classification system gaining knowledge of set of rules that makes use of an ensemble technique. A random woodland consists of many selection timber and allows to remedy the problem of evaluating selection trees. These choice trees are built randomly to randomly choose characters from a given set of data. A random forest makes a selection or prediction based totally at the maximum wide variety of votes obtained from the selection timber. The result received the most quantity of times through multiple selection trees is taken into consideration a random woodland end result.

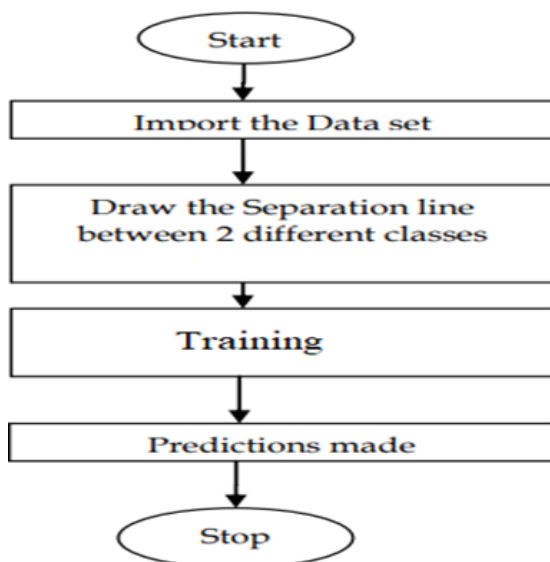


FIG 2. FLOW DIAGRAM OF PROCESS

VI. FUTURE WORK

working on predicting various available datasets of the air index pollutants and other gases of different environments and climate to the respective region to needed to be predicted. Various easy understandable User Experience(UX) needs to be implemented in the web application implementation of project.

VII. RESULT

The primary goal of the project is to develop a reliable air quality index prediction system that can be used to analyze various conditions of weather and atmosphere to predict accurate air quality in the everyday life. Prediction has been made using various machine various algorithms, graphical data and datasets. creating an simplified web based application for user convience.

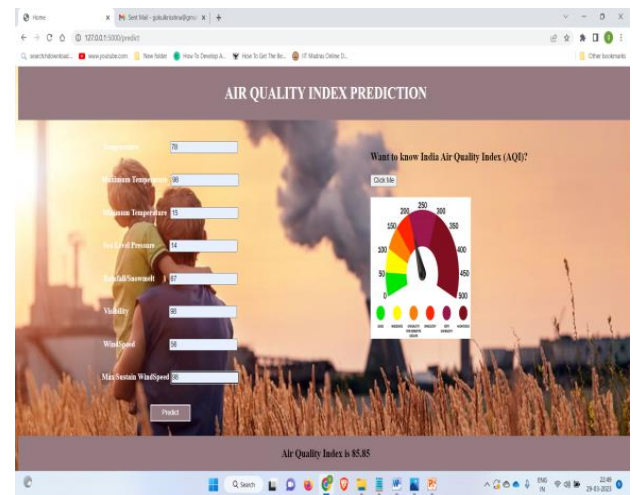


Fig.3.Output Screenshot of Prediction

GRAPH ANALYSIS

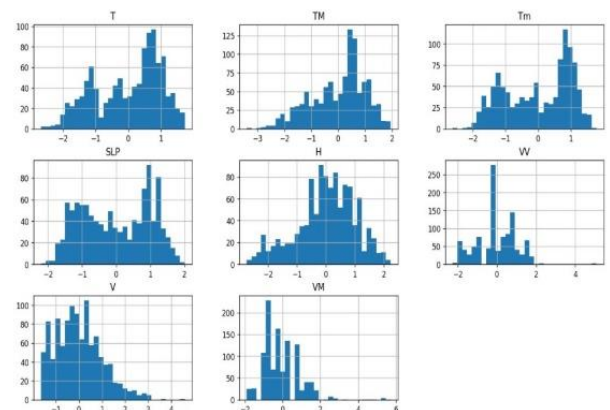


Fig.4.Graph analysis of air pollutants



Fig.5.Readings of Air pollutants

VIII. REFERENCES

- [1] Castelli ,A Machine Learning Approach to Predict Air Quality in California[2020]
<https://doi.org/10.1155/2020/8049504>
- [2] Doreswamy, Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models[2020]
<https://doi.org/10.1016/j.procs.2020.04.221>
- [3] Liang, Machine Learning-Based Prediction of Air Quality [2020]
<https://doi.org/10.3390/app10249151>
- [4] Madan, Air Quality Prediction using Machine Learning Algorithms [2020]
<https://doi.org/10.1109/ICACCN51052.2020.9362912>
- [5] .K. B. Shaban et al., Urban air quality amonitoring system with forecasting models, IEEE Sensors Journal, vol. 16, no. 8, app. 2598 2606, April 2017.
- [6] V. Singh et al., Acokrigingbasedapproach to reconstruct air quality maps, processing measurement station concentrations and deterministic model simulations, Environmental Modelling and Software, vol. 26, no. 6, pp.778786,2019.
- [7] K. Hu et al., Design and evaluation of a metropolitan air quality sensing system, IEEE Sensors Journal, vol. 16, no.5,pp.14481459,March 2018
- [8] V. Sivaraman et al., Hazewatch: A participatory sensor system for monitoring air quality in sydney, in Local Computer Networks Workshops (LCN Workshops), 2017 IEEE 38th Conferenceon,Oct2013,pp.5664

