AIR QUALITY INDEX

Submitted for partial fulfilment of the requirements for

the award of

BACHELOR OF TECHNOLOGY in

COMPUTER SCIENCE & ENGINEERING

by

S. JAHNAVI – 18BQ1A05H2

P. HAMPI – 18BQ1A05F2

P. VISWAKSENA – 18BQ1A05E0

P. CHANDRA SEKHAR – 18BQ1A05F7

Under the guidance of

V. KOTESWARA RAO

Assistant Professor, Dept of CSE, VVIT, Guntur

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

(B. Tech Program is Accredited by NBA)

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Permanently Affiliated to JNTU Kakinada, Approved by AICTE

Accredited by NAAC with 'A' Grade, ISO 9001:2008 Certified

NAMBUR(V), PEDAKAKANI(M), GUNTUR-522 508

Tel no: 0863-2118036, url: www.vvitguntur.com

July 2022

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY
Permanently Affiliated to JNTU Kakinada, Approved by AICTE
Accredited by NAAC with 'A' Grade, ISO 9001:2008 Certified
Nambur, Pedakakani (M), Guntur (Dt) - 522508
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
B.Tech Program is Accredited by NBA

CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Ms. SATRAM JAHNAVI, Ms. PENUMOLU HAMPI, Mr. PALAPARTHI VISWAKSENA, Mr. POTLA CHANDRA SEKHAR** bearing **Reg.No. 18BQ1A05H2, 18BQ1A05F2, 18BQ1A05E0, 18BQ1A05F7** who had carried out the project entitled "**Air Quality Index**" under our supervision.

Project Guide	**Head of the Department**
Mr. V. Koteswara Rao Asst.prof	Dr. V. Ramachandran

Submitted for Viva voce Examination heldon _____

External Examiner

DECLARATION

We, **Ms. S. JAHNAVI**, **Ms. P. HAMPI, Mr. P. VISWAKSENA, Mr. P. CHANDRA SEKHAR** here by declare that the Project Report entitled "**Air Quality Index**" done by us under the guidance of **Mr. V. KOTESWARA RAO**, **Assistant Professor**, **CSE** at Vasireddy Venkatadri Institute of Technology is submitted for partial fulfillment of the requirements for the award of Bachelor of Technology in Computer Science& Engineering. The results embodied in this report have not been submitted to any other University for the award of any degree.

DATE :

PLACE :

SIGNATURE OF THE CANDIDATE(S)

(S. JAHNAVI - 18BQ1A05H2)

(P. HAMPI - 18BQ1A05F2)

(P. VISWAKSENA - 18BQ1A05E0)

(P. CHANDRA SEKHAR – 18BQ1A05F7)

ACKNOWLEDGEMENT

I take this opportunity to express my deepest gratitude and appreciation to all those people who made this project work easier with words of encouragement, motivation, discipline, and faith by offering different places to look to expand my ideas and helped me towards the successful completion of this project work.

First and foremost, I express my deep gratitude to **Mr. Vasireddy VidyaSagar**, Chairman, Vasireddy Venkatadri Institute of Technology for providing necessary facilities throughout the B.Tech programme.

I express my sincere thanks to **Dr. Y. Mallikarjuna Reddy**, Principal, Vasireddy Venkatadri Institute of Technology for his constant support and cooperation throughout the B.Tech programme.

I express my sincere gratitude to **Dr. V. Ramachandran**, Professor & HOD, Computer Science & Engineering, Vasireddy Venkatadri Institute of Technology for his constant encouragement, motivation and faith by offering different places to look to expand my ideas.

I would like to express my sincere gratefulness to my guide **Mr. V. Koteswara Rao Asst. Professor**for his insightful advice, motivating suggestions, invaluable guidance, help and support in successful completion of this project.

I would like to take this opportunity to express my thanks to the **teaching and non-teaching** staff in Department of Computer Science & Engineering, VVIT for their invaluable help and support.

SATRAM JAHNAVI (18BQ1A05H2)
PENUMOLU HAMPI (18BQ1A05F2)
PALAPARTHI VISWAKSENA (18BQ1A05E0)
POTLA CHANDRA SEKHAR (18BQ1A05F7)

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Prediction of pollution is an increasingly important problem. Due to human activities, industrialization and urbanization air is getting polluted. The major air pollutants are CO, NO, C6H6, etc. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. The challenge of predicting the Air Quality Index (AQI), with the aim to minimize the pollution before it gets adverse. Traditional air pollution prediction methods have limitations. Earlier techniques such as Probability, Statistics etc. were used to predict the quality of air, but those methods are very complex to predict, Machine learning provides one approach that can offer new opportunities for prediction of air pollution. With the need to predict air relative humidity by considering various parameters such as CO, Tin oxide, non-metallic hydrocarbons, Benzene, Titanium, NO, Tungsten, Indium oxide, Temperature etc, approach uses Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest Method (RF) to predict the Relative humidity of air and uses Root Mean Square Error to predict the accuracy.The results show improvement of the prediction accuracy and suggest that the model can be used in other smart cities as well.

CHAPTER-1

INTRODUCTION

Air is perhaps the most fundamental characteristic assets for the endurance and presence of each and each life on this planet. All types of life including plants and creatures rely upon air for their fundamental endurance and presence. Air contamination can influence our wellbeing and climate from numerous points of view. A huge number of untimely passing happens each year because breathing in a high level of contamination focuses like PM10, PM2.5, CO, Nitrogen Oxides (NO+NO2).Air pollution is one of the major hazards among the environmental pollution. As each living organism needs fresh and good quality air for every second. None of the living things can survive without such air.

Our system takes past and current data and applies them to our model to predict air pollution. This model reduces the complexity and improves the effectiveness and practicability and can provide more reliable and accurate decisions for environmental protection departments for smart cities. Worldwide, air pollution is responsible for around 1.3 million deaths annually according to the World Health Organization (WHO) . All the above mentioned raises an urgent need to anticipate and plan for pollution fluctuations to help communities and individuals better mitigate the negative impact of air pollution. To do so, air quality evaluation plays a significant role in monitoring and controlling air pollution. The Environmental Protection Agency (EPA) tracks the commonly known criteria pollutants, i.e., ground-level ozone (O3), Sulphur dioxide (SO2), particulates matter (PM10 and PM2.5), carbon monoxide (CO), carbon dioxide (CO2), and nitrogen dioxide (NO2). Air quality has been studied for the last three decades in the United States (US) since the creation of the Clean Air Act program. Although this program has entailed an improvement in air quality over the years, air pollution is still a problem in this situation.Total combustion emissions in the US are accountable for about 200,000 premature deaths per year due to the concentration of pollutants such as particulate matter 2.5 ($PM_{2.5}$) and 10,000 deaths per year due to ozone concentration changes.

The American Lung Association estimated that air pollution-related illnesses cost approximately 37 billion dollars each year in the US, with California alone hitting $15 billion .In the face of increasingly serious environmental pollution problems, scholars have conducted a significant quantity of related research, and in those studies, the forecasting of air pollution has been of paramount importance. Thus, in full knowledge of the increasing pollution derived problems, the importance of accurately forecasting the levels of air pollutants has increased, playing an important role in air quality management and population prevention against pollution hexes. The study aims to build models for hourly air quality forecasting for the state of California, using one of the most powerful existing machine learning (ML) approaches.

Namely, a variant of support vector machines (SVMs), called support vector regression (SVR). The Environment is nothing but everything that encircles us. The environment is getting polluted due to human activities and natural disaster, very severe among them is air pollution. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. If the humidity is more, we feel much hotter because sweat will not evaporate into the atmosphere. Urbanization is one of the main reasons for air pollution because, increase in the transportation facilities emits more pollutants into the atmosphere and another main reason for air pollution is Industrialization. The major pollutants are Nitrogen Oxide (NO), Carbon Monoxide (CO), Particulate matter (PM), SO2 etc. Carbon Monoxide is produced due to the deficient Oxidization of propellant such as petroleum, gas, etc. Nitrogen Oxide is produced due to the ignition of thermal fuel; Carbon monoxide causes headaches, vomiting; Benzene is produced due to smoking, it causes respiratory problems; Nitrogen oxides causes dizziness, nausea; Particulate matter with a diameter 2.5 micrometer or less than that affects more to human health.Measures must be taken to minimize air pollution in the environment.

Earlier classical methods such as probability, statistics were used to predict the quality of air, but those methods are very complex to predict the quality of air. Due to advancement of technology, now it is very easy to fetch the data about the pollutants of air using sensors. Assessment of raw data to detect the pollutants needs vigorous analysis. Convolution Neural networks, Recursive Neural networks, Deep Learning, Machine learning algorithms assures in accomplishing the prediction of future AQI so that measures can be taken appropriately.Machine learning which comes under artificial intelligence has three kinds of learning algorithms, they are the Supervised Learning, Unsupervised learning, Reinforcement learning. In the proposed work we have used supervised learning approach. There are many algorithms under supervised learning algorithms such as Linear Regression, Nearest Neighbor, SVM, kernel SVM, Naive Bayes and Random Forest. Compared to all other algorithms Random forest gives better results.so our approach selects Random Forest to predict the accurate air pollution. Particulate matter can be either human-made or naturally occur.Some examples include dust, ash and sea-spray. Particulate matter (including soot) is emitted during the combustion of solid and liquid fuels, such as for power generation, domestic heating and in vehicle engines.

Particulate matter varies in size (i.e. the diameter or width of the particle). PM2.5 refers to the mass per cubic meter of air of particles with a size (diameter) generally less than 2.5 micrometers PM2.5 is also known as fine particulate matter (2.5 micrometers is one 400th of a millimeter). Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. Different machine learning models have been applied to detect air pollution and predict PM2.5 levels based on a data set consisting of daily atmospheric conditions.Naive Bayes classification and support vector machine algorithms were applied by Dan to get the minimum error with respect to prediction of the air quality in Beijing city. A fuzzy inference system was introduced by José Juan Carbajal to perform parameter classification using a reasoning process and integrating them into an air quality index .

There are applications that display the real time PM2.5 levels, while some show the forecast of a particular day. However, PM2.5 levels for dates after a week is not forecasted. This system exploits machine learning models to detect and predict PM2.5 levels based on a data set consisting of atmospheric conditions in a specific city. The proposed system does two tasks (i). Detects the levels of PM2.5 based on given atmospheric values. (ii) Predicts the level of PM2.5 for a particular date. 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. The primary goal is to predict air pollution level in City with the ground data set. As the largest growing industrial nation, India is producing record amount of pollutants specifically Co2, pm2.5 etc and other harmful aerial contaminants. Air quality of a particular state or a country is a measure on the effect of pollutants on the respected regions, as per the Indian air quality standard pollutants are indexed in terms of their scale, these air quality indexes indicates the levels of major pollutants on the atmosphere.

There are various atmospheric gases which causes pollution on our environment. The Environment is nothing but everything that encircles us. The environment is getting polluted due to human activities and natural disaster, very severe among them is air pollution. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. If the humidity is more, we feel much hotter because sweat will not evaporate into the atmosphere. Urbanization is one of the main reasons for air pollution because, increase in the transportation facilities emits more pollutants into the atmosphere and another main reason for air pollution is Industrialization. The major pollutants are Nitrogen Oxide (NO), Carbon Monoxide (CO), Particulate matter (PM), SO2 etc. Carbon Monoxide is produced due to the deficient Oxidization of propellant such as petroleum, gas, etc. Nitrogen Oxide is produced due to the ignition of thermal fuel; Carbon monoxide causes headaches, vomiting; Benzene is produced due to smoking, it causes respiratory problems Natural sources Natural pollution sources are natural phenomena that discharge harmful substances or have harmful effects on the environment.

Natural phenomena, such as volcanic eruptions and forest fires, will result in air pollutants, including SO2, CO2, NO2, CO, and sulfate. Anthropogenic (man-made) sources Man-made sources such as the burning of fuels, discharges from industrial production processes, and transportation emissions are the main sources of air pollution. There are many kinds of pollutants emitted by man-made pollution sources, including hydrogen, oxygen, nitrogen, sulfur, metal compounds, and particulate matter. With the increasing world population and the developing world economy, the demand for energy in the world has increased dramatically.

The large-scale use of fossil energy globally has also led to a series of environmental problems that have received much attention due to their detrimental effects on human health and the environment .Air pollution is a fundamental problem in many parts of the world, with two important concerns: the impact on human health, such as cardiovascular diseases, and the impact on the environment, such as acid rain, climate change, and global warming . These environmental impacts are described below. Climate change Some chemicals released into the atmosphere by human activities, such as CO2, CH4, N2O, and chlorofluorocarbons (CFCs, exemplified byFreon-12), cause a greenhouse effect. The burning of fossil fuels and other human activities increase the concentration of greenhouse gases, leading to global warming. This also leads to a rise in sea level, more extreme weather, and melting glaciers and ice caps. More alterations to the environment are inevitable as temperatures continue to climb. The studies have indicated that the rate of sea level increase was the fastest in the twentieth century, and data have proven this point of view. The sea level has risen 14 cm in the twentieth century. A study shows that the sea level will rise by 28 cm and is expected to reach a total of 131 cm in 2100 while average global temperature will increase by 3.6 ◦F to 8.1 ◦F (2 ◦C to 4.5 ◦C). Ozone Hole The ozone layer is a relatively high level concentration of ozone in the stratosphere, and its main function is to absorb ultraviolet radiation. It has many useful functions for Earth, and the most important of those functions is to protect human beings, animals, and plants from short wave ultraviolet radiation [10]. It also protects against the heating effect, as ozone absorbs the Sun’s ultraviolet rays and converts it to heat energy that heats the atmosphere.

Freon, a halohydrocarbon, and N2O can produce the greenhouse effect and can also react with stratospheric ozone, resulting in the depletion of the ozone layer and creation of holes in the ozone layer. The decline of the stratospheric ozone level from anthropogenic source is internationally recognized as one of the Earth's most important environmental issues . The ozone hole is affecting human health and the environment negatively and can cause severe diseases, such as skin cancer, eye damage, and genetic mutations . Research results show that if stratospheric ozone concentrations decreased by 1%, the amount of ultraviolet radiation will be increased by 2%, and the cataract rate will increase 0.2–0.6%. Moreover, the depletion of the ozone layer seriously harms the human body, crops, and forests, even destroying natural biosphere generation and the marine ecological balance.In recent years, scientists discovered that the phenomenon of ozone reduction occurs in both the Antarctic and Arctic. In the spring of 2011, ozone column loss had reached 40%. According to the observations of Chinese atmospheric physics and meteorology over the Qinghai-Tibetan Plateau, the ozone layer is being reduced at a rate of 2.7% per 10 years. Int. J. Environ. Res. Public Health 2018, 15, 780 3 of 44 Particulate matter pollution Atmospheric particulate matter consists of solid or liquid granular substances in the atmosphere. Thick smog along with particulate matter (PM) occurs and covers most cities of world frequently .

According to medical research, PM causes different degrees of harm to human respiratory, cardiovascular, and central nervous, and immune systems and to genes . China, as the largest developing country, has attracted great attention from all over the world for its rapid economic development and its air pollution. In 2015, China's air pollution situation was very serious with most cities' air quality exceeding the China National Standard. Moreover, some cities in China have been selected as the 10 most polluted cities in the world . In recent years in China, high concentrations of particulate matter have received increasing attention . Generally, air pollutants do not just harm the local or regional environment. They can also cause damage on a global scale. Certain man-made chemicals have damaged the planet's protective ozone layer, allowing more harmful solar radiation to strike the Earth's surface. Although the use of these chemicals is being phased out, their destructive effects will linger for many more decades.

Control of air pollution and improving air quality are presently concern of scientists globally . As one of the important results of urban air pollution control, urban air pollution forecasting has established an urban air pollution alarm system, effectively reducing the cost of air pollution control. The establishment of a reasonable and accurate forecasting model is the basis for forecasting urban air pollution. Forecasting is a requisite part of in the science of big data and can be used to infer the future development of an object relative to previous information.

So “pollution forecasting” can be understood as estimation of pollutant concentration at specified future date. Since the 1960s, with the development of air pollution control and research, it has become urgent for people to understand the influence of air pollution and the trends of pollution. Therefore, forecasting air pollution began. Forecasting pollution using different patterns of performance can be divided into three types: potential forecasts, statistical models, and numerical models. For different elements, it is divided into pollution potential forecasting and concentration forecasting.

Statistical methods and numerical modelling methods result in concentration forecasts. A potential forecast is mainly based on the meteorological conditions for atmospheric dilution and diffusion capacity. When the weather conditions are expected to be in line with the standards for possible serious pollution, a warning will be issued. A concentration forecast will forecast the concentration of pollutants in a certain area directly, and the forecast results are quantitative. These air pollutions forecasting models can be divided into parametric and nonparametric models, or deterministic and nondeterministic models. It is easy to distinguish the parametric models from nonparametric models, and deterministic models from nondeterministic models, but it is difficult to differentiate the parametric models from deterministic models. The most significant difference between parametric models and deterministic models is that for a deterministic model, the output can be determined, as long as inputs are fixed, regardless of the number of trials; while the parametric model is to determine the parameters of equations in the known model, and its output is uncertain. For example, the diffusion models in this paper belong to the deterministic model, and they are based on physical equations, driven by the chemistry and the transport of pollutants, requiring many accurate input data models based on large amounts of historical data, such as regression, principal component analysis, etc., are usually parametric models. The most popular statistical method uses artificial intelligence (AI) models.

CHAPTER-2

AIM AND SCOPE

EXISTING SYSTEM :

In existing system we analysed air pollution prediction, in few algorithms like Support vector machine, Linear regression, Random forest. They took individual pollutant samples and predicted the accuracy values, Likewise they have taken other pollutant samples and calculated the accuracy values by using all the algorithms, among these Random forest gave the best result.

Disadvantage

1. The score in support vector algorithm given a negative result, which is impossible to determine air pollution, because the values must be positive
2. By considering individual samples it is very time-consuming process.

PROPOSED SYSTEM:

In the Proposed System the prediction models proposed based on different city air pollution data record collecting can also be adapted to other weather reports, while might require some changes. But before considering model adaptations, we need to be clear on whether the three geolocation problems on cities, i.e., prediction of home location on pollution rate, city location and mentioned location, are applicable to the target platform or not. For example, based on city mentioned location prediction on some image and video sharing platforms and other weather reporting platforms may not be applicable to base on prediction of the weather reports. We have considered Decision Tree which gave best accuracy.

Advantage

1. It doesn't take much time.
2. It is less cost
3. High Performance

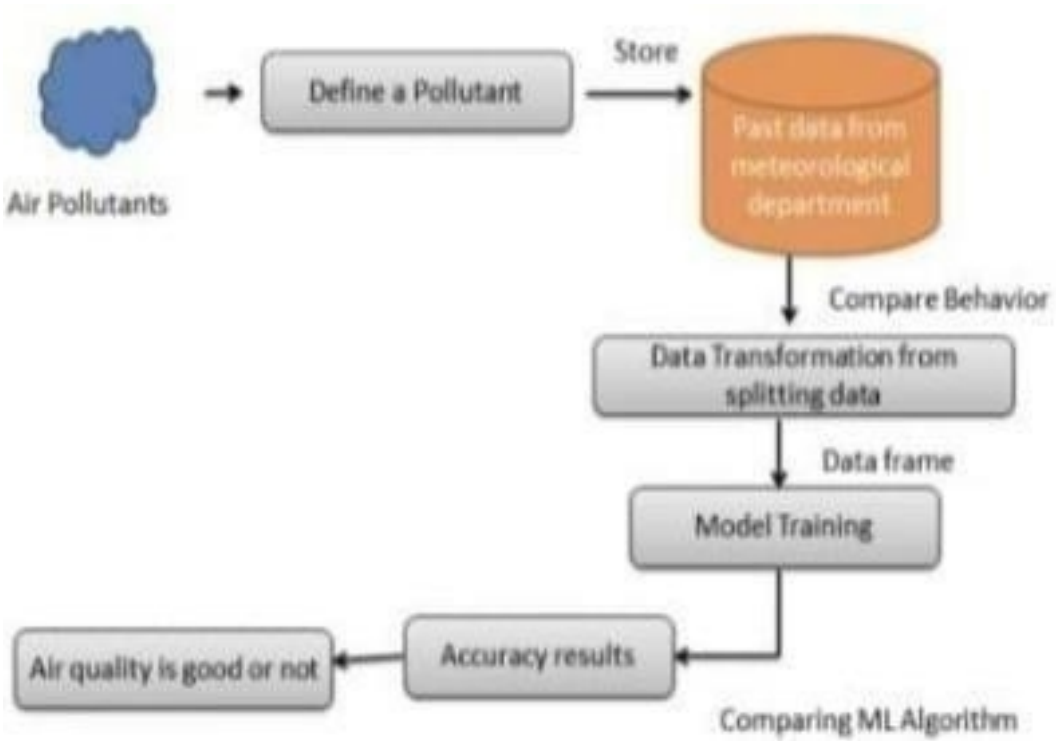

Fig 1.System Architecture of Air Pollution

Software and Hardware Requirements:

Hardware:

RAM4GB (minimum) Processor
Core 2 duo (minimum) hard disk -
250 GB (minimum)

Software:

Python
Anaconda
Win 7 or above

CHAPTER-3

REVIEW OF LITERATURE

K. B. Shaban, A. Kadri, and E. Rezk, "Urban Air Pollution Monitoring System With Forecasting Models" IEEE Sensors Journal, vol. 16, no. 8, pp. 2598–2606, Apr. 2016

A system for monitoring and forecasting urban air pollution is presented in this paper. The system uses low-cost air-quality monitoring motes that are equipped with an array of gaseous and meteorological sensors. The focus of this paper is on the monitoring system and its forecasting module. Three machine learning (ML) algorithms are investigated to build accurate forecasting models for one-step and multi-step ahead of concentrations of ground-level ozone (O_3), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). These ML algorithms are support vector machines, M5P model trees, and artificial neural networks (ANN). Two types of modeling are pursued: 1) univariate and 2) multivariate. The performance evaluation measures used are prediction trend accuracy and root mean square error (RMSE).

Li S., and Shue L., "Data mining to aid policy making in air pollution management," Expert Systems with Applications, vol. 27, pp. 331-340, 2004. In this paper we are applied mining process to extract knowledge from weather dataset. Data collected from DAV BDL public school, Bhanur, Medak weather station. Data set includes four years period [2011-2015] of daily weather observations. We are trying to apply data mining techniques clustering, Association and classification. We collected Weather data parameters temperature, pressure, humidity and dew point, wind speed rain falls and wind direction. These parameters one is related to another parameter. For outlier detection, data analysis and experimental results we are used weka data mining tool and graphs from Excel tool provide a very useful and accurate knowledge in a form of tables and graphs. This knowledge can be used to obtain decision making for different areas like Agriculture, Air pollution; Disaster Management and also for prediction. Our future work includes building an automatic, efficient and accurate system to predict weather.

Gu, Ke, JunfeiQiao, and Weisi Lin. "Recurrent Air Quality Predictor Based on Meteorology and Pollution Related Factors." IEEE Transactions on Industrial Informatics (2018).

PM-10 is one of major air pollutants which affect on human health. Since PM-10 comes from various emission sources and its level of concentration is largely dependent on meteorological and geographical factors of the local region, the forecasting of PM-10 concentration is of great interest to protect daily human health. In this study, the dependent variables on PM-10 concentration were derived from the correlation analysis between PM-10 and meteorological as well as environmental factors based on the observations at the monitoring stations. Using the potential variables on the PM-10 level, the neural network model was developed and tested. The root mean square errors of the prediction in test runs were 0.064 to 0.077 and the test results implied that the system could be used in real forecasting within 10% error rates.

García Nieto, P.J., Sánchez Lasheras, F., GarcíaGonzalo, E. et al. "Estimation of PM10 concentration from air quality data in the vicinity of the major steel works site in the metropolitan area using machine learning techniques " Stoch Environ Res Risk Assess (2018).

Atmospheric particulate matter (PM) is one of the pollutants that may have a significant impact on human health. Data collected over 7 years from the air quality monitoring station is analyzed using four different mathematical models: vector autoregressive moving-average, autoregressive integrated moving-average (ARIMA), multilayer perceptron neural networks and support vector machines with regression. Measured monthly, the average concentration of pollutants (SO2, NO and NO2) and PM10 is used as input to forecast the monthly average concentration of from one to 7 months ahead. Simulations showed that the ARIMA model PM10 performs better than the other models when forecasting 1 month ahead, while in the forecast from one to 9 months ahead the best performance is given by the support vector regression.

Hu, Ke, Ashfaqur Rahman, Hari Bhrugubanda, and Vijay Sivaraman. "Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors." IEEE Sensors Journal 17, no. 11 (2017): 3517-3525. Metropolitan air pollution is a growing concern in both developing and developed countries. Fixed-station monitors, typically operated by governments, offer accurate but sparse data, and are increasingly being augmented by lower fidelity but denser measurements taken by mobile sensors carried by concerned citizens and researchers. In this paper, we introduce HazeEst-a machine learning model that combines sparse fixed-station data with dense mobile sensor data to estimate the air pollution surface for any given hour on any given day in Sydney. We assess our system using seven regression models and tenfold cross validation. Our results can be visualized using a Web-based application customized for metropolitan Sydney. We believe that the continuous estimates provided by our system can better inform air pollution exposure and its impact on human health.

CHAPTER-4

PROPOSED METHOD

PURPOSE OF THE PROJECT

The principal focal point of this paper is to investigate the reasonable AI methods that will help in better anticipating air contamination fixation. The information is gathered from CPCB (Central Pollution Control Board) online information and sensors over the objective district. At that point the dissemination of suspended particles like PM10, PM2.5, SO2, and NO2 contaminated climate air are recognized.

We additionally considered some meteorological variables to anticipate the combination of air like Temperature, Least Temperature, Maximum Temperature, Wind speed and Relative Humidity and a few different highlights. This will help in the forecast of air quality in metropolitan also, modern regions of Ghaziabad and this could fill in as a significant reference for government offices in assessing present and contriving future air contamination approaches. Our examination centres around the expectation of air contamination level of a specific or explicit locale. In air contamination forecasts, model exactness, productivity and flexibility are key contemplations. Satellite observation of air pollution allows for wider geographical scope, and in doing so can facilitate studies of air pollution's effects on natural capital and ecosystem resilience. Many air pollution-related aspects of the sustainability of development in human systems are not being given their due attention. Opportunities exist for air pollution monitoring to attend more to these issues. Improvements to the resolution and scale of monitoring make these opportunities realizable. Rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The results show that machine learning models (logistic regression and autoregression) can be efficiently used to detect the quality of air and predict the level of PM2.5 in the future.

The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities. The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system.

ALGORITHMS

Machine Learning is a system that can learn from example through self- improvement and without being explicitly coded by programmer. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results in it.

A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation. Machine learning is also used for a variety of task like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on. Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in theindustry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain .Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example.

When we give the machine a similar example, it can figure out the outcome. However, like a human, if its feed a previously unseen example, the machine has difficulties to predict.The core objective of machine learning is the learning and inference. First of all, the machine learns through the discovery of patterns. This discovery is made thanks to the data. One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a feature vector. You can think of a feature vector as a subset of data that is used to tackle a problem.The machine uses some fancy algorithms to simplify the reality and transform this discovery into a model. Therefore, the learning stage is used to describe the data and summarize itinto a model. For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant. This is the model Inferring.When the model is built, it is possible to test how powerful it is on never-seen- before data. The new data are transformed into a features vector, go throughthe model and give a prediction. This is all the beautiful part of machine learning.

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.

4.2.1 Random forest classifier 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. For prediction, each test data point is passed through every decision tree in the forest. 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) shows the illustration of a random forest that consists of m number of trees.

4.2.1 Support vector machine Support vector machine, a supervised learning method for classification, regression, and outlier detection, constructs the hyperplane that acts as a boundary between distinct data points and thus the output can be deduced hereafter [11]. Two distinctive versions of SVM are shown in Figure 1. For classification problem in Figure 1a, data points that lie at the edge of an area closest to the hyperplanes are considered as support vectors. The space between these two regions is the margin between the classes.). Nevertheless, data points inside the boundaries will be exempted. Since support vectors represent the data points located near these boundary lines Finally, since most realistic problems aren't linear, the kernel trick iscommonly performed by mapping training data onto the high-dimensional feature space. Kernel functions,

Hence, the additional parameter, known as the ε-insensitive loss is introduced to tolerate some deviations that lie inside the ε region tube The boundary lines (dashed lines) across the hyperplane (solid line) in SVR (stands for support vector regression) are defined with regards to parameter ε, in which the resulting lines are the shifted function in the amount of –ε and +ε from the hyperplane (assume it is a straight line with an equation of +b). The SVR uses a penalty concept introduced by parameter C (cost factor) for output variables outside the boundaries either above (ξi) or below ($\xi *$ i). Nevertheless, data points inside the boundaries will be exempted. Since support vectors represent the data points located near these boundary lines (see Figure 1b), if the ε moves further from the hyperplane, the number of support vectors decreases; otherwise, the number of support vectors increases as the ε approximates towards the hyperplane. Finally, since most realistic problems aren't linear, the kernel trick is commonly performed by mapping training data onto the high-dimensional feature space. Kernel functions,

E.g., linear, polynomial, radial basis function (RBF), sigmoid, hyperbolic tangent, etc., are used to convert the once inseparable input data into the separable ones. The parameter ε has brought a couple of advantages, yet is sometimes difficult to tune. Hence, scholars from Australian National University proposed the substitution of parameter ε into parameter ν (hereinafter referred to as ν-SVM) to avoid such a tedious parameter tuning process for regression [13]. Moreover, parameter ν is also applicable for classification, where it becomes the replacement for cost factor C [14]. Values of parameter ν with the upper bound of training margin errors and lower bound for the support vectors are recommended from 0 to 1 so that the ν-SVM can offer a more meaningful parameter interpretation

4.2.2 Linear regression Linear regression is probably the method where most of the academicians started their first machine learning experience. Its main working principle lies behind the fitting of one or more independent variables with the dependent variable into a line in n dimensions. n usually denotes the number of variables within a dataset. This line is supposedly created as it would be minimizing the total errors when trying to fit all the instances into the line.

Under machine learning, linear regression is equipped with the capability to learn continuously by optimizing the parameters in the model. These parameters are including w0, w1, w2, . . . , wm (as illustrated in Figure 4). Most commonly, optimization is carried out by a method called gradient descent. It works by partially deriving the loss function and all parameters will be updated by subtracting the previous value with the derivative times a specified learning rate. The learning rate can be tuned by the simplest way, which is rule of thumb (trial and error), or a more sophisticated rule, e.g., meta-heuristic. Another parameter that is left for tuning is the amount of generalization added to the model. Regularization is undergone as an effort to lessen the chance of overfitting and increase the robustness of the model. Two types of regularization used in linear regression are lasso and ridge regression. Lasso regularization will eliminate less important feature by letting the feature's coefficient to zero, and retain another more important one. Ridge regularization on the other hand will not try to eliminate a feature, but instead, tries to shrink the magnitude of coefficients to get a lower variance in the model.

4.2.3 Decision Tree A decision tree is a flowchart-like structure in which each of internal node represents a test on a feature (e.g. whether an coin flip comes up heads or tails)each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. They paths from root to leaf represent classification rules. Decision tree is one of them are predictive modelling approaches used in statistics, data mining and machine learning. A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules. Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non- parametric supervised learning method used for both classification and regression tasks.

Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is general term for this. Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

Here the decision or the outcome variable is Continuous**,** e.g. a number like 123. Working Now that we know what a Decision Tree is, we'll see how it works internally. There are many algorithms out there which construct Decision Trees, but one of the best is called as ID3 Algorithm. ID3 Stands for Iterative Dichotomiser 3. Before discussing the ID3 algorithm, we'll go through few definitions. Entropy, also called as Shannon Entropy is denoted by H(S) for a finite set S, is the measure of the amount of uncertainty or randomness in data. Intuitively, it tells us about the predictability of a certain event. Example, consider a coin toss whose probability of heads is 0.5 and probability of tails is 0.5. Here the entropy is the highest possible, since there's no way of determining what the outcome might be. Alternatively, consider a coin which has heads on both the sides, the entropy of such an event can be predicted perfectly since we know beforehand that it'll always be heads. In other words, this event has no randomness hence it's entropy is zero. In particular, lower values imply less uncertainty while higher values imply high uncertainty. Information gain is also called as Kullback-Leibler divergence denoted by IG(S,A) for a set S is the effective change in entropy after deciding on a particular attribute A. It measures the relative change in entropy with respect to the independent variables. Alternatively, where IG(S, A) is the information gain by applying feature A. H(S) is the Entropy of the entire set, while the second term calculates the Entropy after applying the feature A, where P(x) is the probability of event x.

4.3 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage Anaconda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, mac OS and Linux. In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages, and use multiple environments to separate these different versions. The command line program conda is both a package manager and an environment manager, to help data scientists ensure that each version of each package has all the dependencies it requires and works correctly. Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages and update them, all inside Navigator.

The following applications are available by default in Navigator:

JupyterLab

Jupyter Notebook

QTConsole Spyder

VSCode

CHAPTER-5

PROPOSED METHODOLOGY

5.1 Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a test on afeature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules. Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non- parametric supervised learning method used for both classification and regression tasks. Decision Trees, but one of the best is called as ID3 Algorithm. ID3 Stands for Iterative Dichotomiser 3. Before discussing the ID3 algorithm, we'll go through few definitions. Entropy Entropy, also called as Shannon Entropy is denoted by H(S) for a finite set S, is the measure of the amount of uncertainty or randomness in data. Intuitively, it tells us about the predictability of a certain event. Example, consider a coin toss whose probability of heads is 0.5 and probability of tails is 0.5. Here the entropy is the highest possible, since there's no way of determining what the outcome might be. Alternatively, consider a coin which has heads on both the sides, the entropy of such an event can be predicted perfectly since we know beforehand that it'll always be heads.

CHAPTER-6

DESIGN AND IMPLEMENTATION

MODULES

Dataset Selection:

The main pollutant emissions in Taiwan are due to energy production industry, traffic, waste incineration and agriculture. In Taiwan, six pollutants (O3, PM2.5, PM10, CO, SO2, and NO2) are monitored and controlled based on their concentration time-series. Types of data used as predictors to perform analysis involve AQ: air quality data, MET: meteorological data, and TIME: the day of the month, day of the week, and the hour of the day. From 1 January 2008 to 31 December 2018, air quality data are collected from several monitoring stations across Taiwan and reported via the EPA'swebsite [18]. With the same timeframe, meteorological data are provided in 1-h intervals by Taiwan's Central Weather Bureau (CWB) from three air monitoring stations: Zhongli (Northern Taiwan), Chuanghua (Central Taiwan), and Fengshan (Southern Taiwan). The datasets represent different environmental conditions related to air pollutant concentration.

Data Processing:

The number of raw data points for the Zhongli, Changhua, and Fengshan monitoring stations includes 91,672, 94,453, and 94,145, respectively. The analysis of these readings begins with a crucial phase – data preprocessing. Various preprocessing operations precede the learning phase. At any particular time, one invalid variable will not affect the whole data group, and thus it will just be either marked blank or, where available, replaced by a value sourced from the CWB, without eliminating the full row. The missing values are treated by imputation to recover the corresponding values.

Feature Engineering:

In regard to selecting features in the predictive models, the hourly AQI readings with the highest index out of 6 pollutants: O3, PM2.5, PM10, CO, SO2, and NO2 are selected. To convert the time- window-specific concentration of 6 pollutants, the AQI Taiwan Guidelines [18] are adopted and the AQI is manually calculated using the following Equations (1) and (2), where index values of O3, PM2.5, and PM10 are needed to define AQI in Taiwan, and the lack of one or more of these values will significantly reduce the accurate assessment of current air quality.

The concentration of the specific pollutant using categories of good, moderate, unhealthy which includes specific groups, unhealthy, very unhealthy, and hazardous. The data transformation defines the time-window-specific concentration to calculate Ii values. For example, based on the AQI from Taiwan's EPA website [18], the concentration valueO3 = 0.06 ppm will fall in the interval with lbO3 = 0.055 ppm and ubO3 = 0.070 ppm corresponding to the "moderate" pollutant level with LBmoderate = 51 and UBmoderate= 100. The valueO3 is defined bymatching either of two conditions: if the 8-h average concentration is more precautionary for a specific site and is also below 0.2 ppm, then this value is used; otherwise, the 1-h average concentration will be considered. Both valuePM2.5 and valuePM10 are the moving average values which consider two time-windows, i.e., the last 12 h and 4 h (see Table 1). Other variables,such as valueCO and valueNO2 only account for a single time window, i.e., last 8 h and 1 h, respectively. Meanwhile, valueSO2 emphasizes the 24-h average concentration if the 1-h average concentration exceeds 185 ppb; otherwise, the 1-h average value will be used. The AQI mechanism introduces several new variables to train the prediction model (Table 1). For several pollutants, time windows other than hourly are more sensitive in determining AQI; therefore, the prediction interval related to the accuracy of long-term predictions is under investigation to clarify the time dependency between consecutive data points. As the AQI calculation is already established, the future value of the AQI readings in three different time intervals will be regarded as target variables and are summarized.

Performance Evaluation:

According to Isakndaryan et al. , 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 . 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.

Given the lack of spatial proximity of the readings to the original monitoring stations, the missing values are imputed for relative humidity, temperature, and rainfall, without using wind speed or wind direction. The next imputation process used the k-NN algorithm to substitute the rest of the invalid or missing data that did not qualify for the previous imputation process. Note that the percentage of missing values is lower than 1.3% in all three-station datasets. Then, input and target data are normalized to eliminate potential biases; thus, variable significance won't be affected by their ranges or their units. All raw data values are normalized to the range of [0, 1] Inputs with a higher scale than others will tend to dominate the measurement and are consequentlygiven greater priority. Normalization not only improves the model learning rate, but also supports k-NN algorithm performance because the imputation is decided by the distance measure.

UML DIAGRAMS

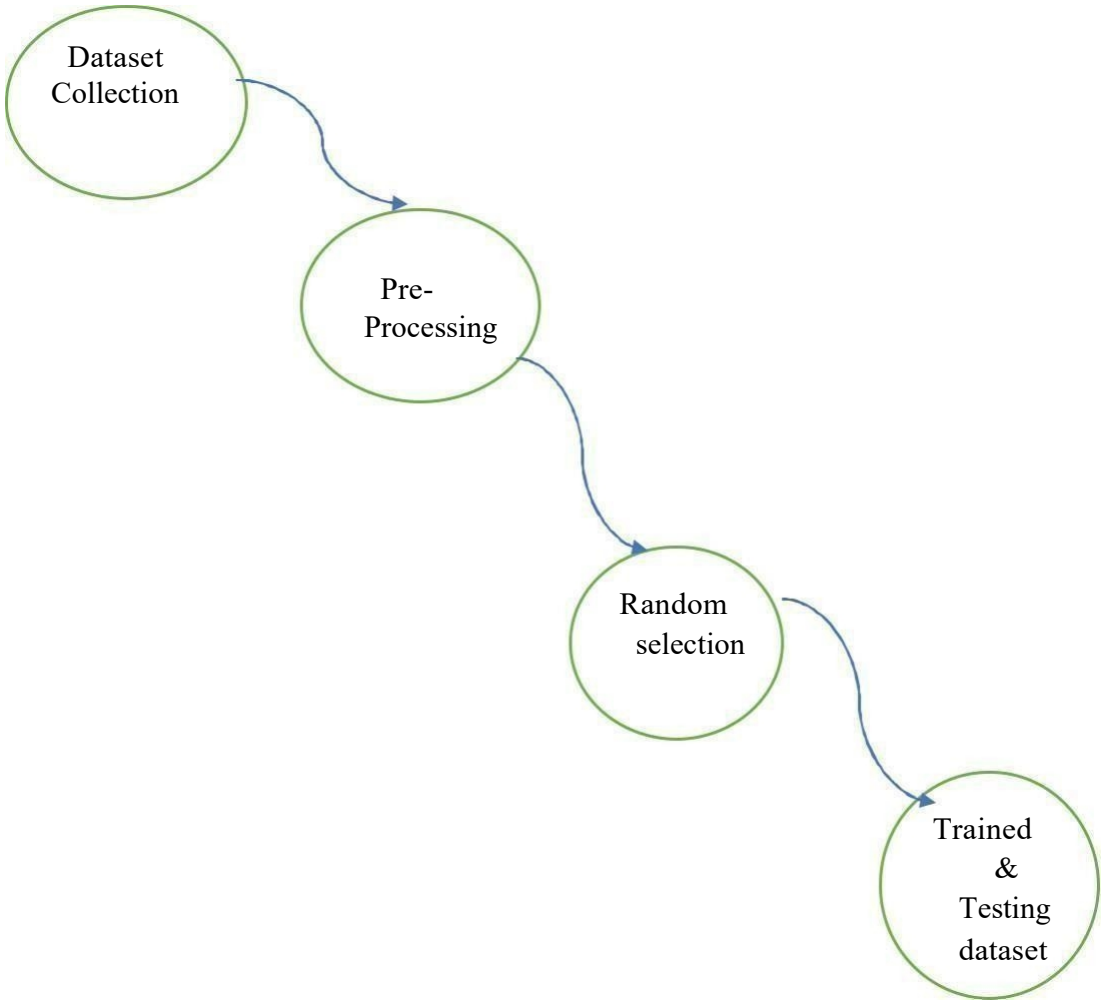

Fig 6.2.1 Data Flow Diagram level 0 for Air Pollution

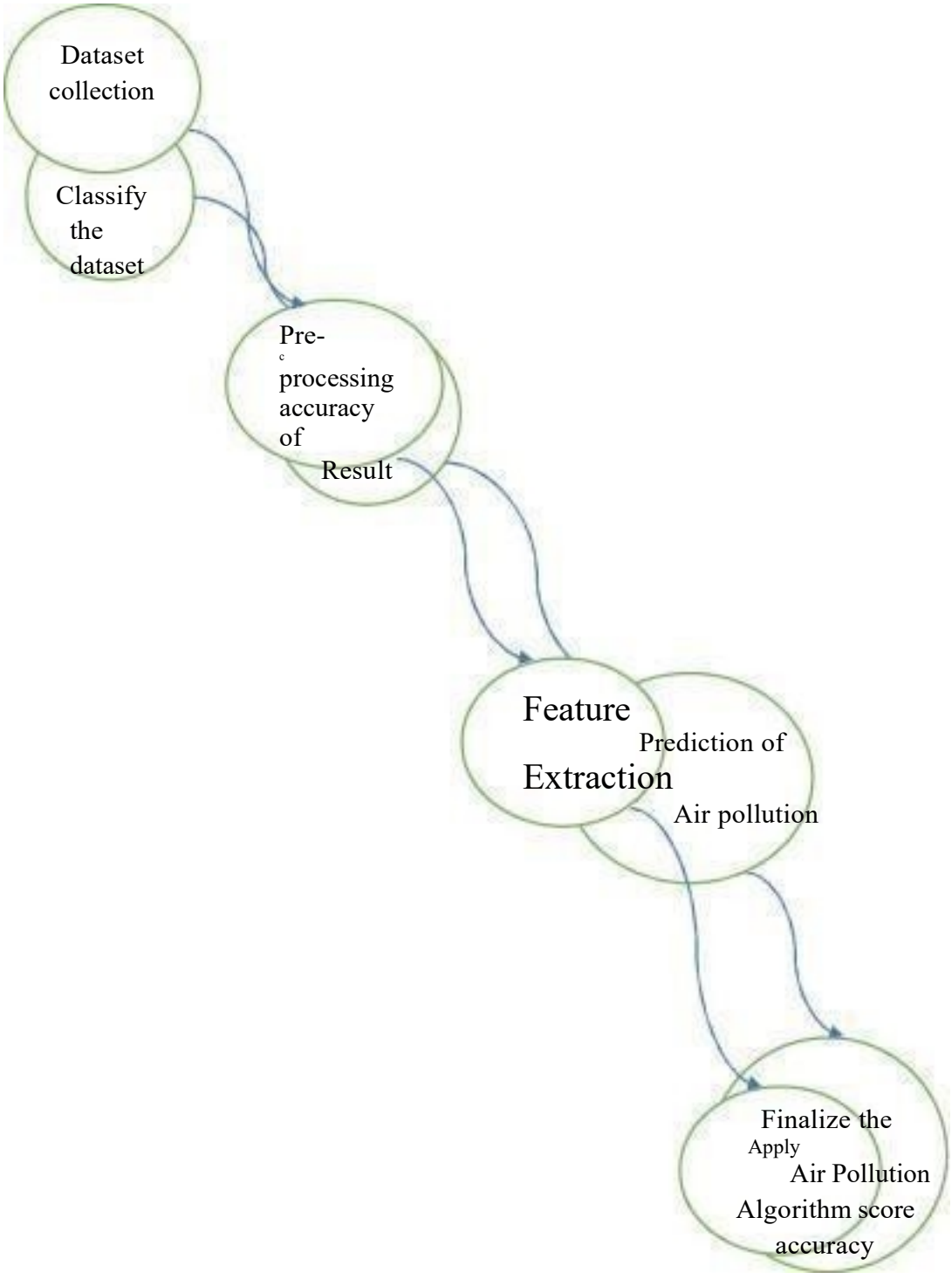

Fig 6.2.2.Data flow diagram level 1&2 for Air Pollution

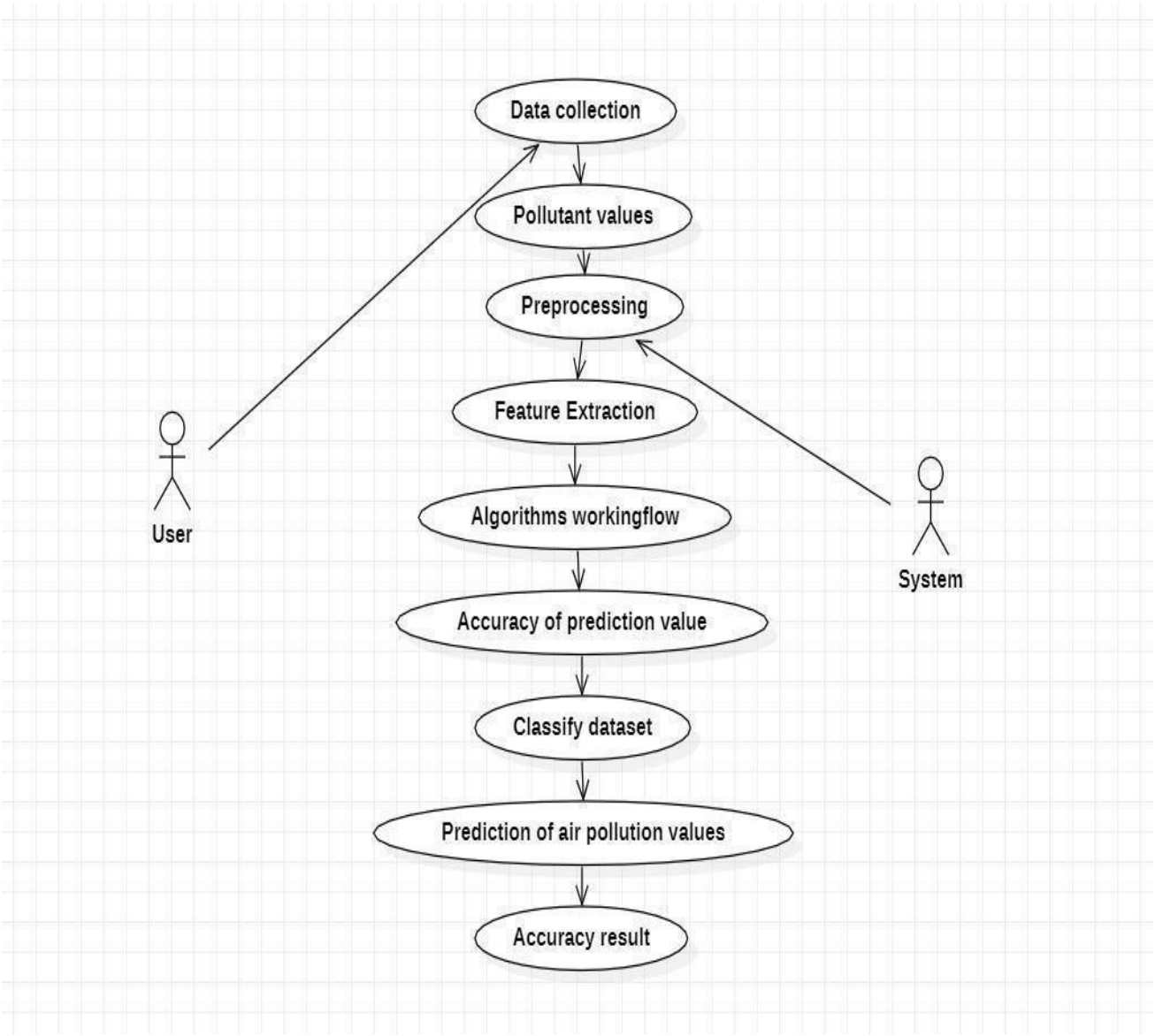

Fig 6.2.3.Use Case Diagram for Air Pollution

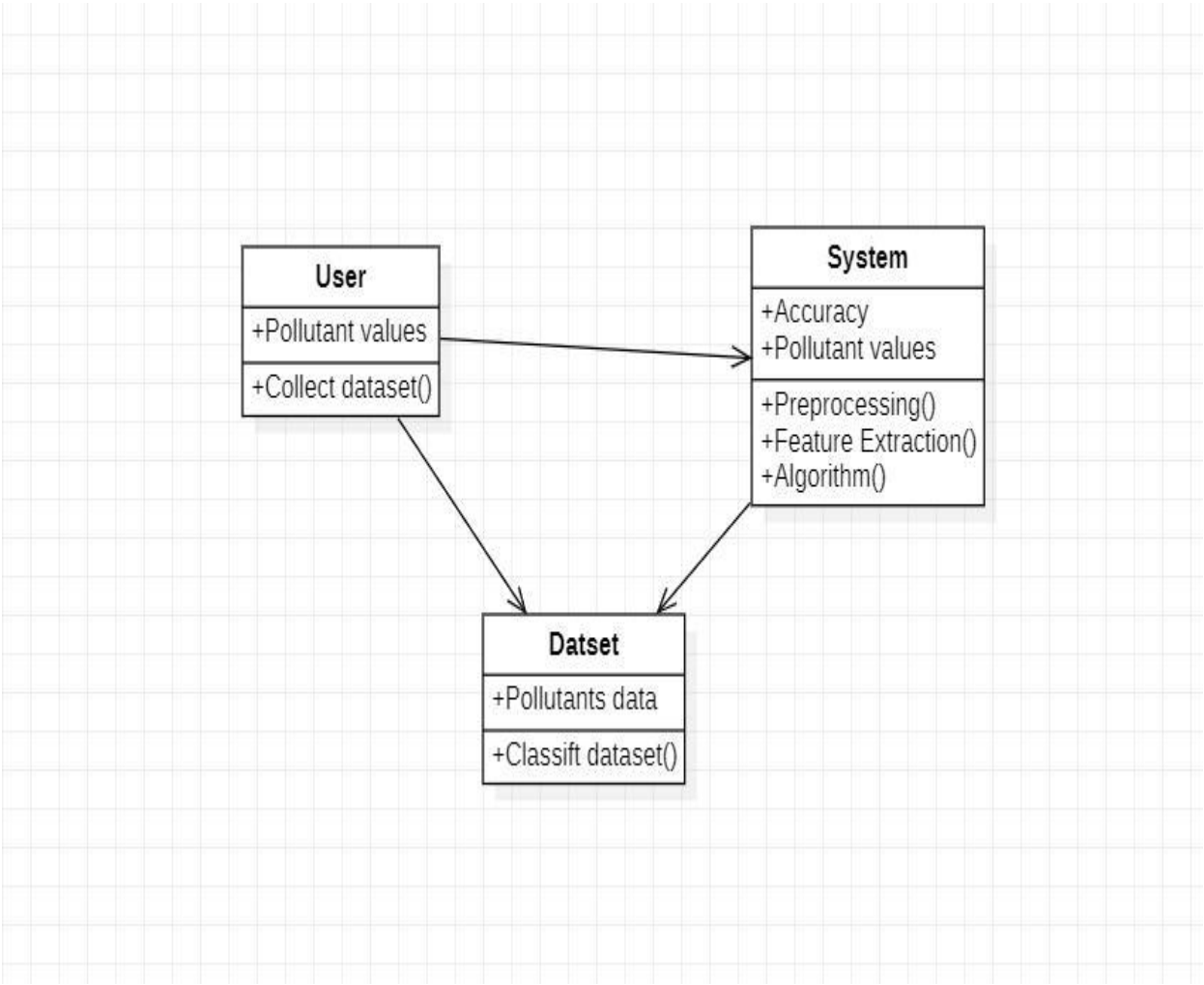

Fig 6.2.4 Class Diagram for Air Pollution

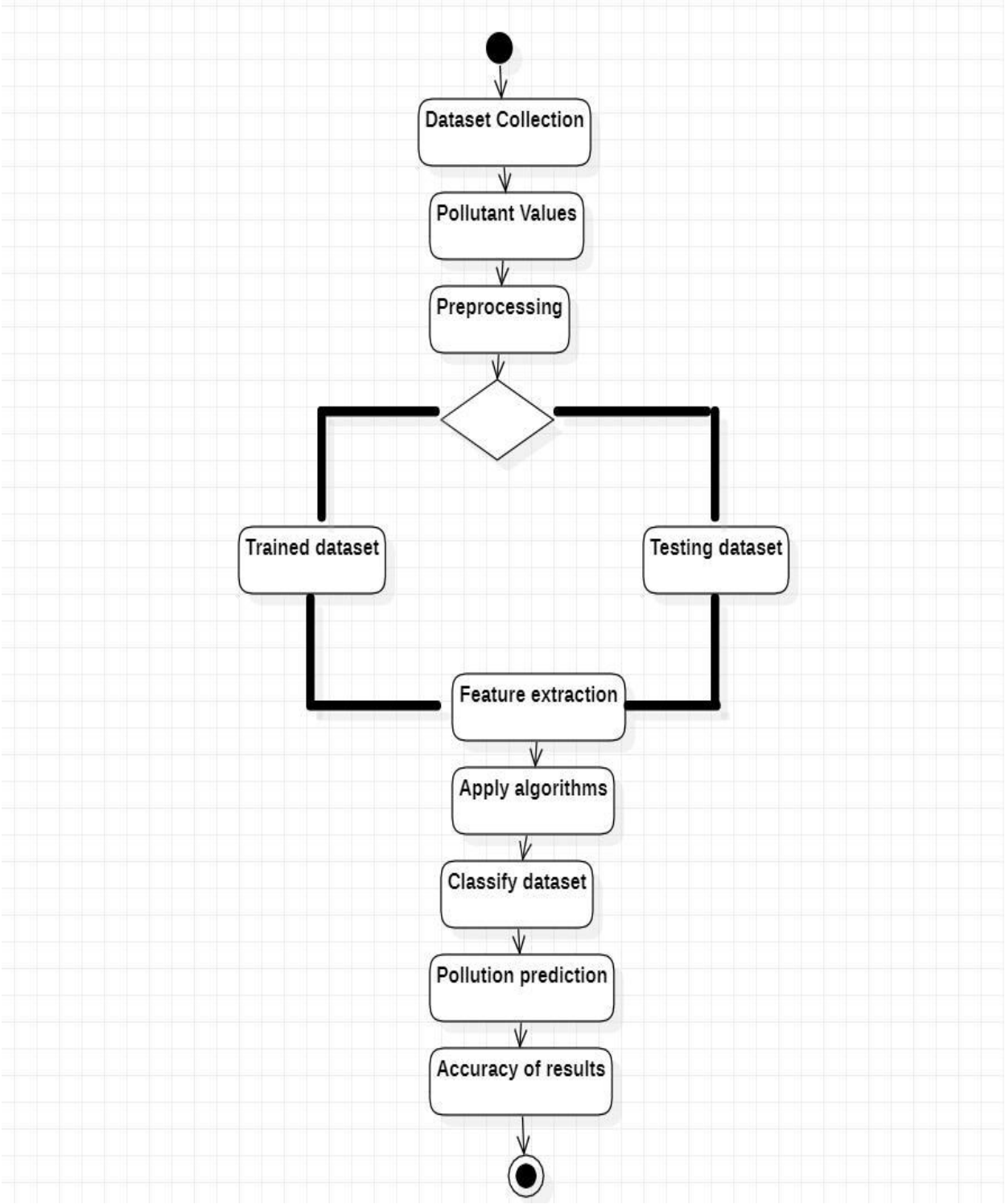

Fig 6.2.5 Activity Diagram for Air Pollution

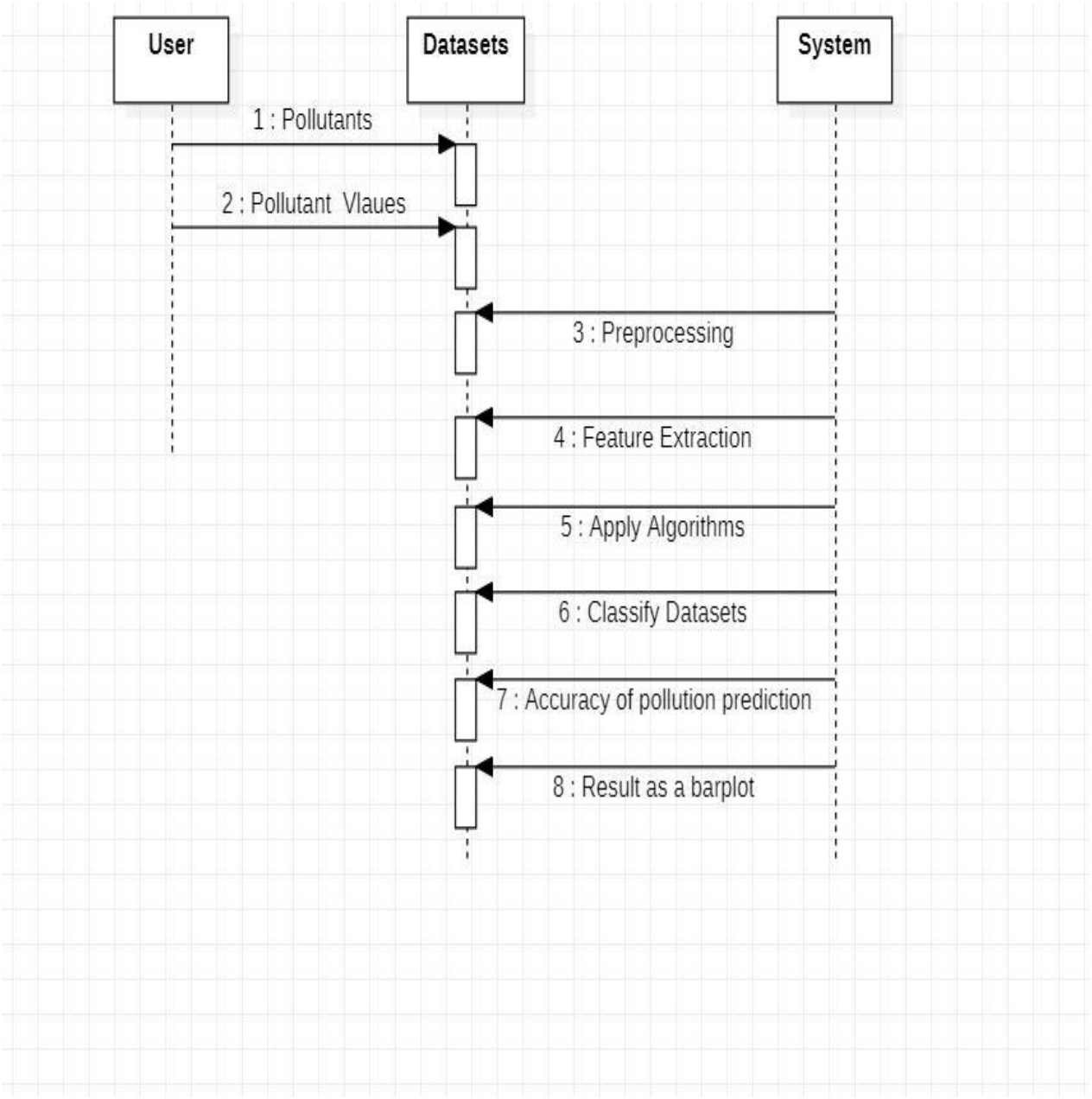

Fig 6.2.6 Sequence Diagram for Air Pollution

PYTHON OVERVIEW

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages. Python is Interpreted: Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP. Python is Interactive**:** You can actually sit at a Python prompt andinteract with the interpreter directly to write your programs. Python is Object-Oriented: Python supports Object-Oriented style or technique of programming that encapsulates code within objects. Python is a Beginner's Language: Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

Python Library

Three common data pre processing steps are :

- Formatting: The data you have selected may not be in a format that issuitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a textfile.
- Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the dataentirely
- Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the wholedataset.

PANDAS

Pandas is quite a game changer when it comes to analyzing data with Python and it is one of the most preferred and widely used tools in datamunging/wrangling if not THE most used one. Pandas is an open source .What's cool about Pandas is that it takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software (think Excel or SPSS for example. People who are familiar with R would see similarities to R too). This is so much easier to work with in comparison to working with lists and/or dictionaries through for loops or list comprehension.

In order to "get" Pandas you would need to install it. You would also need to have Python 2.7 and above as a pre-requirement for installation. It is also dependent on other libraries (like NumPy) and has optional dependancies (like Matplotlib for plotting). Therefore, I think that the easiest way to get Pandas set up is to install it through a package like the Anaconda distribution, "a cross platform distribution for data analysis and scientific computing". Importing a library means loading it into the memory and then it's there for you to work with. In order to import Pandas all you have to do is run the following code:

```
import pandas as pd

import numpy as np
```

Usually you would add the second part ('as pd') so you can access Pandas with 'pd.command' instead of needing to write 'pandas.command' every time you need to use it. Also, you would import numpy as well, because it is very useful library for scientific computing with Python.

Now that you've loaded your data, it's time to take a look. How does the data frame look? Running the name of the data frame would give you the entire table, but you can alsoget the first n rows with df.head(n) or the last n rows with df.tail(n). df.shape would give you the number of rows and columns. df.info() would give you the index, datatype and memory information. The command s.value_counts(dropna=False) would allow you to view unique values and counts for a series (like a column or a few columns). A very useful command is df.describe() which inputs summary statistics. It is also possible to get statistics on the entire data frame or a series (a column etc):

df.mean() Returns the mean of all columns

df.corr() Returns the correlation between columns in a data frame

df.count() Returns the number of non-null values in each data

df.max()Returns the highest value in each column

df.min()Returns the lowest value in each column

df.median()Returns the median of each column

df.std()Returns the standard deviation of each column

Data set collection

One of the things that is so much easier in Pandas is selecting the data you want in comparison to selecting a value from a list or a dictionary. You can select a column (df[col]) and return column with label col as Series or a few columns (df[[col1, col2]]) and returns columns as a new DataFrame. You can select by position (s.iloc[0]), or by index (s.loc['index_one']) . In order to select the first row you can use df.iloc[0,:] and in order to select the first element of the first column you would run df.iloc[0,0] . These can also be used in different combinations, so I hope it gives you an idea of the different selection and indexing you can perform in Pandas.

Data Cleaning

Data cleaning is a very important step in data analysis. For example, we always check for missing values in the data by running pd.isnull() which checks for null Values, and returns a 35oolean array (an array of true for missing values and false for non-missing values).In order to get a sum of null/missing values, run pd.isnull().sum(). Pd.notnull() is the opposite of pd.isnull(). After you get a list of missing values you can get rid of them, or drop them by using df.dropna() to drop the rows or df.dropna(axis=1) to drop the columns. A different approach would be to fill the missing values with other values by using df.fillna(x) which fills the missingvalues with x (you can put there whatever you want) or s.fillna(s.mean()) to replace all null values with the mean (mean can be replaced with almost any function from the statistics section).Itis sometimes necessary to replace values with different values. For example, s.replace(1,'one') would replace all values equal to 1 with 'one'. It's possible to do it for multiple values: s.replace([1,3],['one','three'])would replace all 1 with 'one' and 3 with 'three'. You can also rename specific columns byrunning: df.rename(columns={'old_name': 'new_ name'})or use df.set_index('column_one') to change the index of the data frame.

NUMPY

Numpy is one such powerful library for array processing along with a large collection of high-level mathematical functions to operate on these arrays. These functions fall into categories like Linear Algebra, Trigonometry, Statistics, Matrix manipulation, etc. NumPy'smain object is a homogeneous multidimensional array. Unlike python's array class which only handles one-dimensional array, NumPy'sndarray class can handle multidimensional array and provides more functionality. NumPy's dimensions are known as axes. For example, the array below has 2 dimensions or 2 axes namely rows and columns. Sometimes dimension is also known as a rank of that particular array or matrix.

Sklearn

In python, scikit-learn library has a pre-built functionality under sklearn. Pre processing. Next thing is to do feature extraction Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally our models are trained using Classifier algorithm.. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered . The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre processed data. The chosen classifiers were Decision tree , Support Vector Machines and Random forest. These algorithms are very popular in text classificationtasks.

Data Visualization

Data visualization is the discipline of trying to understand data by placing it in a visual context, so that patterns, trends and correlations that might not otherwise be detected can be exposed. Python offers multiple great graphing libraries that come packed with lots of different features. No matter if you want to create interactive, live or highly customized plots python has a excellent library for you. To get a little overview here are a few popular plotting libraries: Matplotlib: lowlevel, provides lots of freedom .Seaborn**:** high-level interface, great default styles In this article, we will learn how to create basic plots using Matplotlib, Pandas visualization and Seaborn as well as how to use some specific features of each library. This article will focus on the syntax and not on interpreting the graphs.

Matplotlib

Matplotlib is the most popular python plotting library.

1. To install Matplotlib pip and conda can be used.
2. pip install matplotlib
3. conda install matplotlib
4. Matplotlib is specifically good for creating basic graphs like line charts.
5. import matplotlib.pyplot as plt

8. Bar Diagram

A bar-chart can be created using the bar method to do this. The bar-chart is useful categorical data thatdoesn't have a lot of different categories .

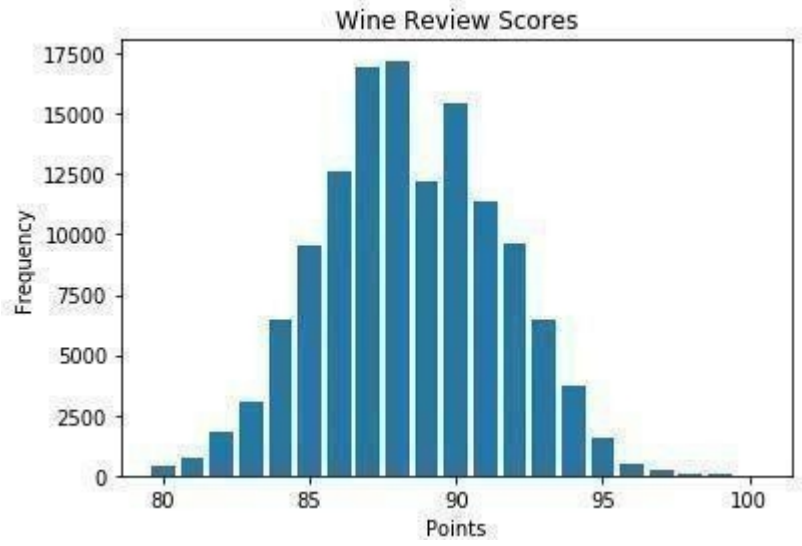

Fig 8. Bar Diagram for Air Pollution

9. Matplotlib:

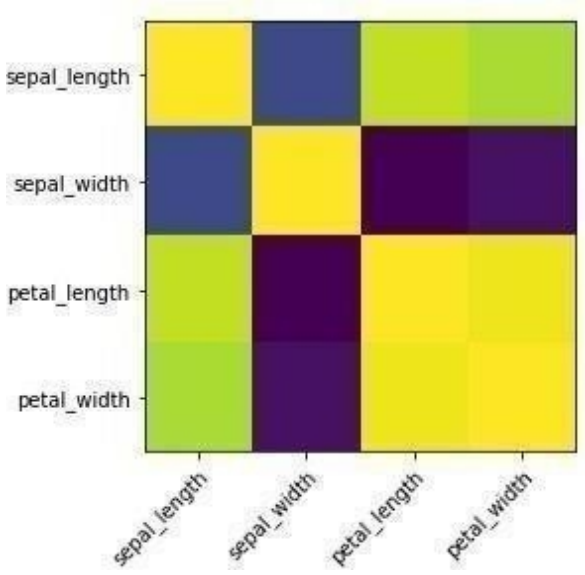

Fig 9. Matplotlib Diagram for Air pollution

CHAPTER 7

RESULTS

Score and Prediction of Random Forest:

In the Random Forest Algorithm score and prediction values of Random Forest algorithm of score 65% andprediction accuracy is 39%.

```
In [58]: a3
Out[58]: 0.6558501769205956

In [59]: y_pred=regr.predict(x_test)

In [60]: print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
         print('MSE:', metrics.mean_squared_error(y_test, y_pred))
         print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

         MAE: 22.887287905329835
         MSE: 1683.5403694614397
         RMSE: 41.030968419736816

In [61]: regr.predict([[2.03,2.25,63.0,25.36,25.0,14.6,3.25,6.32,4.25,96.0,5.36,0.02,0]])
Out[61]: array([39.08111814])
```

Fig 10.Output of RF Algorithm

Score and Prediction of Linear Regressor:

In the Linear Regressor Algorithm score and prediction values of Random Forest algorithm of score 70% andprediction accuracy is 73%.

```
In [37]: a1
Out[37]: 0.7091356400658614

In [40]: x.keys()
Out[40]: Index(['PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO2', 'O3', 'Benzene',
                'Toluene', 'Xylene', 'AQI', 'AQI_Bucket'],
               dtype='object')

In [41]: lr.predict([[129.06,1.26,26.00,14.85,10.28,0.14,26.96,117.44,0.22,7.95,0.08,197.00,0.00]])
Out[41]: array([73.83446935])
```

Fig 11. Output of LR Algorithm

Score and Prediction of Support Vector Regression:

In the SVM Algorithm score and prediction values of Random Forest algorithm of score -13% and predictionaccuracy is -62%.

```
In [82]: a4
Out[82]: -1.3562041533090432

In [83]: print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
         print('MSE:', metrics.mean_squared_error(y_test, y_pred))
         print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

         MAE: 80.49091761329773
         MSE: 11129.714840268816
         RMSE: 105.49746366746841

In [84]: regressor.predict([[2.03,2.25,63.0,25.36,25.0,14.6,3.25,6.32,4.25,96.0,5.36,0.02,0]])
Out[84]: array([-0.62529401])
```

Fig 12. Output of SVR Algorithm

Score and Prediction of Decision Tree:

In the Decision Tree Algorithm score and prediction values of Random Forest algorithm of score 99% andprediction accuracy is 64%.

```
In [51]: a2
Out[51]: 0.9985369671318264

In [52]: print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
         print('MSE:', metrics.mean_squared_error(y_test, y_pred))
         print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

         MAE: 16.980956064822067
         MSE: 1292.8952094759823
         RMSE: 35.956852051813186

In [53]: regressor.predict([[104.09,2.56,28.07,17.01,11.42,0.09,19.00,138.18,0.17,5.02,0.07,188.00,0.00]])
Out[53]: array([64.18])
```

Fig 13. Output of DT Algorithm

Comparison of algorithms

Comparison of Accuracy levels for all the algorithms

```
In [87]: df2
```

Out[87]:

	Algorithm	Accuracy
0	Linear Regression	0.709136
1	Decision Tree	0.997802
2	Random Forest	0.648891
3	SVR	-1.357571

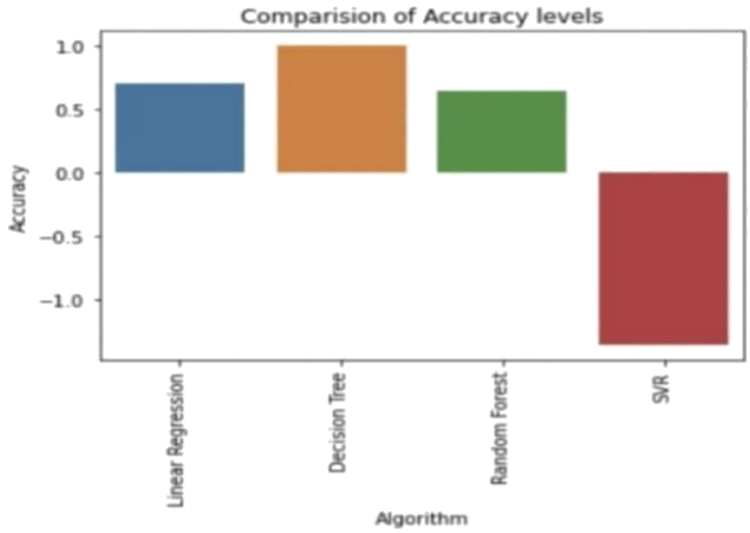

Fig 14. Comparing Algorithms

CHAPTER 8

CONCLUSION

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. 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.This work presented a study of Decision Tree to forecast pollutants and particulates levels and to correctly identify the Score. The studied method produced a suitable model of the hourly atmospheric pollution, allowing us to obtain, generally, good accuracy in modelling pollutant

concentrations like O3, CO, and SO2.

As it is a regression model, we have taken four algorithms to predict air pollution. They are Linear Regression, Random Forest method, Support vector Machine, Decision Tree, The model which gives highest score is the best model. From our results we got decision tree value as the best. So we conclude Decision tree is the best compared to other algorithms.The proposed system will definitely help in improving the prediction of air pollution in our smart city. Multivariate Multistep Time Series Prediction Using Random Forest technique improve the the performance and reduce the complexity of the air pollution prediction model. Also here we are using feature selection technique which makes our prediction even better.

With the advancement of Machine Learning Techniques Real-time air quality prediction and evaluation is desirable for future smart cities. Our study focuses on prediction of air pollution level of a particular or specific region. In air pollution prediction, model accuracy are key considerations.

CHAPTER 9

REFERENCES

1.K. B. Shaban, A. Kadri, and E. Rezk, “Urban Air Pollution Monitoring System With Forecasting Models” IEEE Sensors Journal, vol. 16, no. 8, pp. 2598–2606, Apr. 2016

2. Li S., and Shue L., "Data mining to aid policy making in air pollution management," Expert Systems with Applications, vol. 27, pp. 331-340,2004.

3. Gu, Ke, Junfei Qiao, and Weisi Lin. "Recurrent Air Quality Predictor Based on Meteorology and Pollution Related Factors." IEEE Transactions on Industrial Informatics (2018).

4. García Nieto, P.J., Sánchez Lasheras, F., GarcíaGonzalo, E. et al. “Estimation of PM10 concentration from air quality data in the vicinity of the major steel works site in the metropolitan area using machine learning techniques”Stoch Environ Res Risk Assess (2018), https://link.springer.com/article/10.1007/s00477-018-1565-6.

5. Hu, Ke, Ashfaqur Rahman, Hari Bhrugubanda, and Vijay Sivaraman. "Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors." IEEE Sensors Journal 17, no. 11 (2017): 3517-3525