AIR POLLUTION PREDICTION SYSTEM

A PROJECT REPORT

Submitted by

Aditya Sharma -20BCS9872

Aryan -20BCS9854

Radhika Patel -20BCS1212

Shikher Bhardwaj -20BCS1052

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING (B.E.)

IN

COMPUTER SCIENCE ENGINEERING

Chandigarh University

November - 2022

BONAFIDE CERTIFICATE

Certified that this project report **"AIR POLLUTION PREDICTION SYSTEM"** is the bonafide work of "Aditya Sharma (20BCS9872), Aryan(20BCS9854), Radhika Patel (20BCS1212), Shikher Bhardwaj(20BCS1052)" who carried out the project work under my/our supervision.

SIGNATURE	**SIGNATURE**
Dr. Sandeep Singh Kang	**Er. Priyanka**
HEAD OF THE DEPARTMENT	**SUPERVISOR**
Computer Science Engineering	Computer Science Engineering

Submitted for the project viva-voce examination held on _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

TABLE OF CONTENTS

List of Figures

ABSTRACT

Air pollution has become a life-threatening factor in many countries around the world in recent decades as a result of human activities, industrialization, and urbanization. Particulate Matter with a diameter of less than 2.5m (PM2.5) is a major health concern among air pollutants. It causes a variety of illnesses, including respiratory and cardiovascular problems. As a result, it is critical to accurately predict PM2.5 concentrations in order to protect citizens from the harmful effects of air pollution in advance. PM2.5 levels vary according to a number of factors, including meteorology and the concentration of other pollutants in urban areas.

We forecast India's air quality by using machine learning to predict the air quality index of a given area. The Indian Air Quality Index is a standard measure used to indicate pollutant levels (so2, no2, rspm, spm, etc.) over time. As aGradient decent boosted multivariable regression problem, we developed a model to predict the air quality index based on historical data from previous years. By using cost estimation for our predictive problem, we improve the model's efficiency. With historical pollutant concentration data, our model will be able to successfully predict the air quality index of an entire county, a state, or any bounded region.

We achieved better performance than standard regression models in our model by implementing the proposed parameter reducing formulations.

Our model predicts the air quality index for the entire country with 99% accuracy, and we determine the order of preference based on similarity to the ideal solution.

Chapter 1

INTRODUCTION

1.1 Identification of client and need

Air Pollution Prediction System is an application that uses machine learning algorithms and allows the user to predict the quality of air in his or her living premises.

What is a Prediction System?

Prediction in machine learning refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome.

Examples of Prediction in Machine Learning

Prediction in machine learning is commonly used for security, marketing, operations, risk, and fraud detection.
Here are just a few examples of how predictive analytics is utilized in different industries:

Banking and Financial Services

Predictive analytics in the banking and financial services industry are used in conjunction to detect and reduce fraud, measure market risk, and identify new business opportunities.

Security

Predictive analytics and machine learning play a critical role in security. Security institutions typically use predictive analytics to improve services and performance, but also to detect anomalies, fraud, understand consumer behavior and enhance data security.

Retail

Predictive analytics and machine learning in allow retailers to better understand consumer behavior, such as who will buy what and at what store? These questions can be readily answered with the right predictive models and data sets, helping retailers to plan ahead and stock items based on seasonality and consumer trends.

Why is machine learning used for prediction?

Machine learning increases the speed at which data is processed and analyzed. With machine learning, predictive analytics algorithms can train on even larger data sets and perform deeper analysis on multiple variables with minor changes in deployment.

1.2 Relevant Contemporary Issues

Common Pollutants

Although levels of particle pollution and ground-level ozone pollution are substantially lower than in the past, levels are unhealthy in numerous areas of the country. Both pollutants are the result of emissions from diverse sources, and travel long distances and across state lines.

Ozone can increase the frequency of asthma attacks, cause shortness of breath, aggravate lung diseases, and cause permanent damage to lungs through long-term exposure. Elevated ozone levels are linked to increases in hospitalizations, emergency room visits and premature death.

Both pollutants cause environmental damage, and fine particles impair visibility. Fine particles can be emitted directly or formed from gaseous emissions including sulfur dioxide or nitrogen oxides. Ozone, a colorless gas, is created when emissions of nitrogen oxides and volatile organic compounds react.

Climate Change

The risks to public health and the environment from climate change are substantial and far-reaching. Scientists warn that carbon pollution and resulting climate change are expected to lead to more intense hurricanes and storms, heavier and more frequent flooding, increased drought, and more severe wildfires - events that can cause deaths, injuries, and billions of dollars of damage to property and the nation's infrastructure.

Carbon dioxide and other greenhouse gas pollution leads to more frequent and intense heat waves that increase mortality, especially among the poor and elderly. Other climate change public health concerns raised in the scientific literature include anticipated increases in ground-level ozone pollution, the potential for enhanced spread of some waterborne and pest-related diseases, and evidence for increased production or dispersion of airborne allergens.

Other effects of greenhouse gas pollution noted in the scientific literature include ocean acidification, sea level rise and increased storm surge, harm to agriculture and forests, species extinctions and ecosystem damage. Climate change impacts in certain regions of the world (potentially leading, for example, to food scarcity, conflicts or mass migration) may exacerbate problems that raise humanitarian, trade and national security issues for the United States.

Toxic Pollutants

Hazardous air pollutants, also called air toxics, include 187 pollutants listed in the Clean Air Act. EPA can add pollutants that are known or suspected to cause cancer or other serious health effects, such as reproductive effects or birth defects, or to cause adverse environmental effects.

Examples of air toxics include benzene, which is found in gasoline; perchloroethylene, which is emitted from some dry-cleaning facilities; and methylene chloride, which is used as a solvent and paint stripper by a number of industries. Other examples of air toxics include dioxin, asbestos, and metals such as cadmium, mercury, chromium, and lead compounds.

Most air toxics originate from manmade sources, including mobile sources such as motor vehicles, industrial facilities and small "area" sources. Numerous categories of stationary sources emit air toxics, including power plants, chemical manufacturing, aerospace manufacturing and steel mills. Some air toxics are released in large amounts from natural sources such as forest fires.

1.3 Problem Identification

Till date, the use of air quality prediction is limited. People are not in a habit of considering the changes around them at a serious note. This project will help to give an easy interface to people to be friendly with the air quality prediction and then take effective measures to protect the air and prevent air pollution around them.

1.4 Task Identification

The aim of our project is to create and give an easy interface to people to be friendly with the air quality prediction and then take effective measures to protect the air and prevent air pollution around them.
So, we have identified the different types of technology that we will be using to create the fully fledged cryptocurrency wallet:

(i) Machine Learning:

It is a branch that deals with training a machine to give suitable outcomes. We will be using Machine Learning in python to prepare out base model for prediction.

(ii) HTML/CSS:

For designing our project, we will be using HTML/CSS to create the basic structure of the cryptocurrency wallet.

(iii) Flask:

Flask is a framework that will be helpful to implement the model prepared onto a web page.

1.5 Timeline

The timeline of our project is as follows:

1. **Project selection**: Selection of Project and team members was done.

2. **Organizational Planning**: When the project was decided, it was divided into smaller tasks and each task was divided amongst the team members according to their skills and expertise.

3. **Developing a Schedule**: Once the tasks were allotted, a schedule was created, which included, allotting milestones followed by group-meetings so as to complete the whole project on time.

4. **Problem Identification:** Identifying the drawbacks of using traditional prediction system for air quality and how this project will overcome those problems.

CHAPTER 2
LITERATURE REVIEW / BACKGROUND STUDY

2.1 Timeline

Before delving into the history of the Air pollution prediction system, it is necessary to first understand the system itself. The application of science and technology to anticipate the composition of air pollution in the atmosphere for a certain location and time is known as air pollution prediction. The same as real measurements, an algorithm forecast of pollution concentrations may be turned into an air quality index.

- Air pollution prediction system is simpler to control and reduce the dangers of air pollution and guarantee a safe level of pollutant concentration in the region with an accurate technique of forecasting air pollution.
- It also aids in assessing the environmental and climatic dangers posed by inadequate air quality regulations. Accurate forecasting may also help with day-to-day planning, avoiding high-risk regions, and adopting appropriate pollution control measures.

Working

Data on air contaminants is collected via sensors, analysed, and recorded as a dataset. This dataset has been pre-processed using a number of characteristics, including attribute selection and normalisation. When the dataset is ready, it is separated into a training set and a test set. A Machine Learning method is then applied to the training dataset. The obtained results are compared to the testing dataset, and the outcomes are analyzed.

History

The consequences of ancient wood fires have been discovered in the charred lungs of mummified tissue from Egypt, Peru, and the United Kingdom. And, long before the Industrial Revolution, the Romans may have been the first to spew metallic toxins into the air.

The smoke cloud over ancient Rome was known as gravioris caeli ("heavy heaven") and infamis aer ("infamous air"). There are several complaints about its effects in classical writings.

One of the major killers of our day is air pollution. Polluted air was responsible for 6.4 million fatalities globally in 2015, with 2.8 million deaths attributed to residential air pollution and 4.2 million to ambient (outdoor) air pollution. 2,3 According to data from that year, global air pollution caused:

- Approximately 19% of all cardiovascular fatalities
- 24% of fatalities from ischemic heart disease
- 21% of stroke fatalities
- 23% of lung cancer fatalities

Air pollution has become a life-threatening factor in many countries around the world in recent decades as a result of human activities, industrialization, and urbanization.

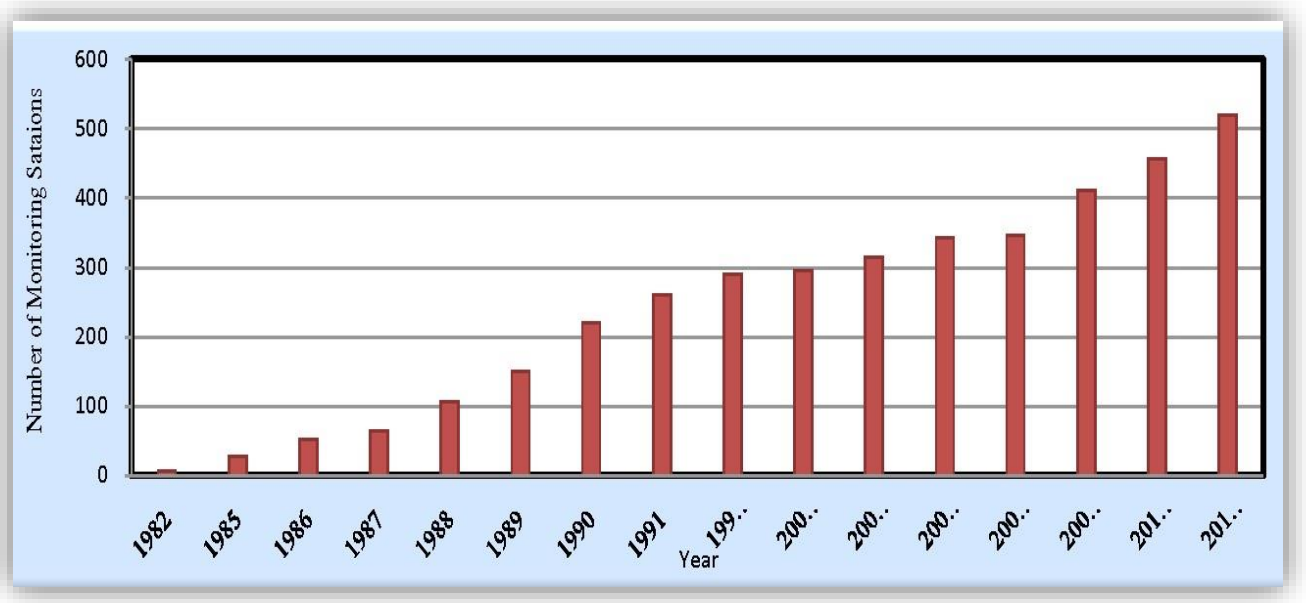

Fig. 2.1. Urban Air pollution trend in India.

2.2 Air Pollution Prediction System

An effective system for monitoring and forecasting air pollution is critical for human health and government decision-making. The mechanism and process of PM2.5 generation, on the other hand, are exceedingly complicated due to the complexity of their features, such as nonlinear properties in time and space, which have a substantial influence on forecast accuracy.

As a result, it needs careful evaluation. Furthermore, the air quality data is tightly tied to time, implying that it is a time series with an apparent periodicity. Because of the data's timeliness, time projections have become critical themes that academics and scholars must pay close attention to. This demonstrates the importance of time series analysis in a variety of fields, including economics, medicine, and astronomy. Machine learning prediction models have been found to be more trustworthy and consistent. Because of advancements in technology, data collection is now simple and precise.

Sensors and technology Only machine learning (ML) algorithms can manage the rigorous analysis required to create accurate and efficient predictions from such large amounts of environmental data. The KNN method, which is better suited for prediction tasks, is used to forecast air pollution.

2.3 Proposed Solutions

- More reliable and consistent
- Easy to use
- Accurate
- User friendly

2.4 Bibliometric Analysis

This program will introduce new features to the area of this prediction system. Key features of this application linked to local weather conditions and pollution emissions are all of these are explained in this chapter later. Right now, the drawback of this application could be that sometimes this application may result less accurate result. But in near future, we will upgrade this.

2.5 Review Summary

Predicting air pollution is a beneficial investment on several levels: individual, community, national, and global. Predictability enables individuals to prepare ahead, reducing the consequences on health

and the expenses connected with them.

People may take action to reduce pollution if they are aware of differences in the quality of the air they breathe, the influence of pollutants on health, as well as concentrations likely to have detrimental consequences. Furthermore, because individuals seek knowledge about air quality, there is a better possibility of inspiring improvements in both individual behaviour and governmental policy.

Measuring has proven to be a big effort since air pollution has increased on a daily basis. Continuous air quality monitoring at a certain site determines the level of pollution in that region. The data collected by the sensors displays the source and concentration of pollutants in the region. Using that knowledge and the ML model, measures to reduce pollution levels may be implemented.

2.6 PROBLEM DEFINITION

Until now, the use of air quality prediction has been limited. People are not in the habit of taking the changes around them seriously.

This project will assist in providing people with an easy interface to be friendly with air quality prediction and then take effective measures to protect the air and prevent air pollution around them.

The features are:

1) Flexible: Building a forecasting system based on specific pollutant concentration levels that can estimate air quality hourly will make the Air pollution Prediction system more flexible and beneficial for the population's health. As a result, systems that can create alerts based on air quality are required and crucial for the public. They may play an important role in health alerts when air pollution levels may exceed specified levels; additionally, they may integrate existing emission control programs, such as by providing environmental regulators with the option of "on-demand" emission reductions, operational planning, or even emergency response.
2) Accurate: Accurate forecasting may also aid in daily planning, avoiding high-risk areas, and implementing appropriate pollution control measures.

2.7 Goals and Objectives

The main objective of our project - Controlling atmospheric emissions, as well as understanding pollutant dispersion and monitoring emission levels, i.e. concentration in ambient air, are all part of protecting the atmospheric environment. Air Quality Monitoring Networks are in place to keep track of these values.

Our project's goal is to design and provide a simple interface for people to be friendly with air quality prediction and then take appropriate steps to protect the air and avoid air pollution in their surroundings.

This system is to record the concentration levels of atmospheric pollutants in order to define air quality limits and develop action plans in the event that high amounts of contamination are identified.

• Identifying contaminated issue regions and comprehending associated spacetime alterations.

• Informing residents on the current state of local air pollution.

Provides Safety

It becomes easier to manage and mitigate the risks of air pollution and ensure a safe level of pollutant concentration in the region. It also helps assess risks to the environment and the climate caused by poor air quality standards.

Accurate

Accurate forecasting may also help with day-to-day planning, avoiding high-risk regions, and adopting appropriate pollution control measures.

Easy to use:

With its simple and intuitive user interface, even those who are just starting out in this field can easily adapt and use our app.

Chapter 3

DESIGN FLOW/PROCESS

3.1 Evaluation & Selection of Specifications/Features

Accuracy and user interface were the major issue faced in the previous models and projects that were predicting the air quality by using the content of gases as input.

In this project we have used following features:

(i) **Python**: We have used Python language to preform machine learning algorithms. Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

(ii) **Model Accuracy**: We have used random forest classifier algorithm as the main algorithm to predict the values as it has the maximum accuracy amongst all the other algorithm and is best suited for this project. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap = TRUE (default), otherwise the whole dataset is used to build each tree.

(iii) **Webpage**: Developed a webpage that is user-friendly and looks great and is easy to handle. It makes it easy for the user to work with the model and hence the user can

predict the quality of air.

(iv) **Flask**: This is a framework in Python. We have used flask in our project to connect the model that we have prepared to the webpage and the input taken

from the user on the webpage is given to the model and the predicted output is shown to the user on the webpage.

3.2 Design Constraints

This is a software-based project and is designed in following steps:

1) Choosing the dataset and preforming EDA on it.

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set.

Data sets can also consist of a collection of documents or files.

Perform EDA on the data set and remove all the null values and get a finalized dataset that could be used to train the model.

2) Preparing the model using Random Forest Classifier Algorithm.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning. 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.

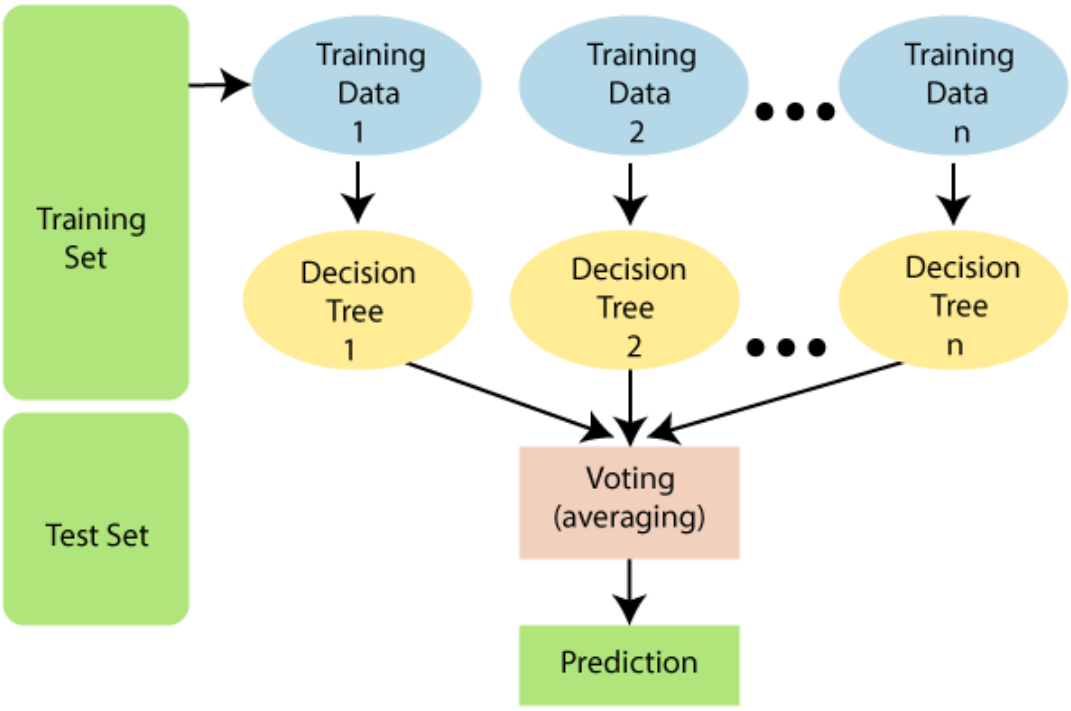

Fig 3.1 Random Forest Classifier Technique

3) Prepare a webpage and embed the model in it Using flask.

A webpage is a document written in HTML and can be viewed on any web browser. It is contained within the web server, which can be accessed by entering the URL for that web page, and once it is loaded, it appears on the user's web browser. Each webpage is linked with a unique URL; hence two pages cannot have the same URL.

3.3 Analysis and Feature finalization subject to constraints

User-Friendly: A simple webpage that is visually simple and is easy to use. Everything is clear and now point of confusion can arise.

Accurate: Using a highly accurate machine algorithm to prepare the model. Gives accurate results

and are 99.98% correct.

3.4 Design Flow

As we are living in the era of new emerging technology, machine learning is one of them. There are already several air quality prediction model available on the internet, and each one is different from the other in some ways.

Since, the background logic behind every air quality prediction model is the same, that is, at the application layer, there is a regulation mechanism, at the contract layer, there is dataset and predictions, consistency and scalability is dealt with, there is privacy preservation and P2P interface.

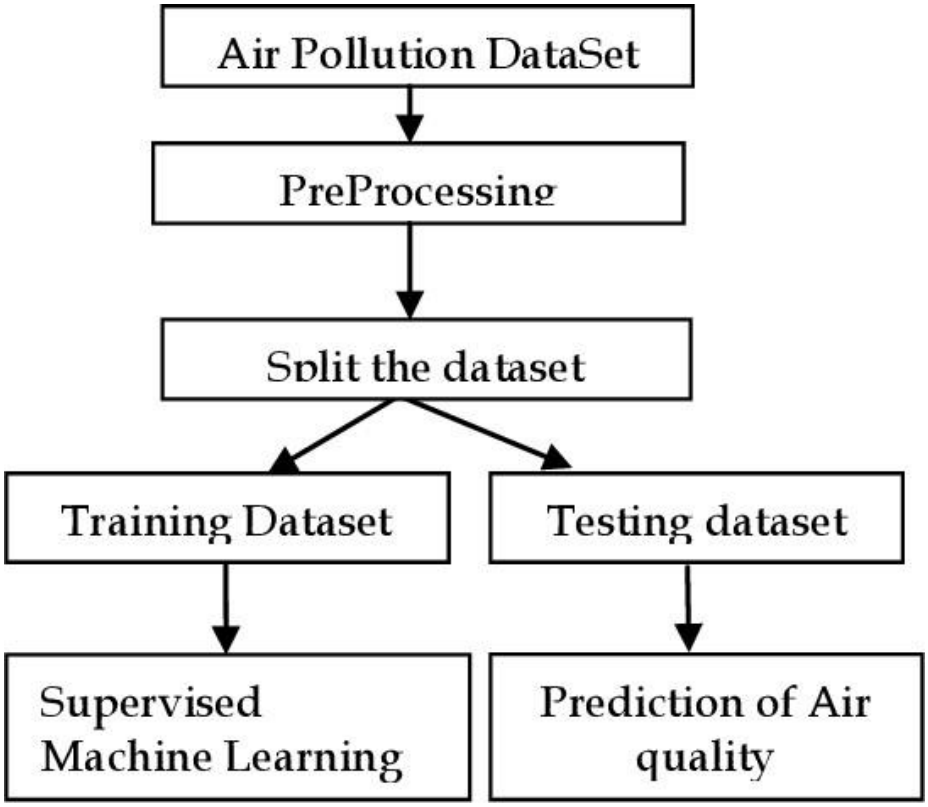

Fig 3.2: Flow-Chart Design

3.5 Implementation plan/methodology

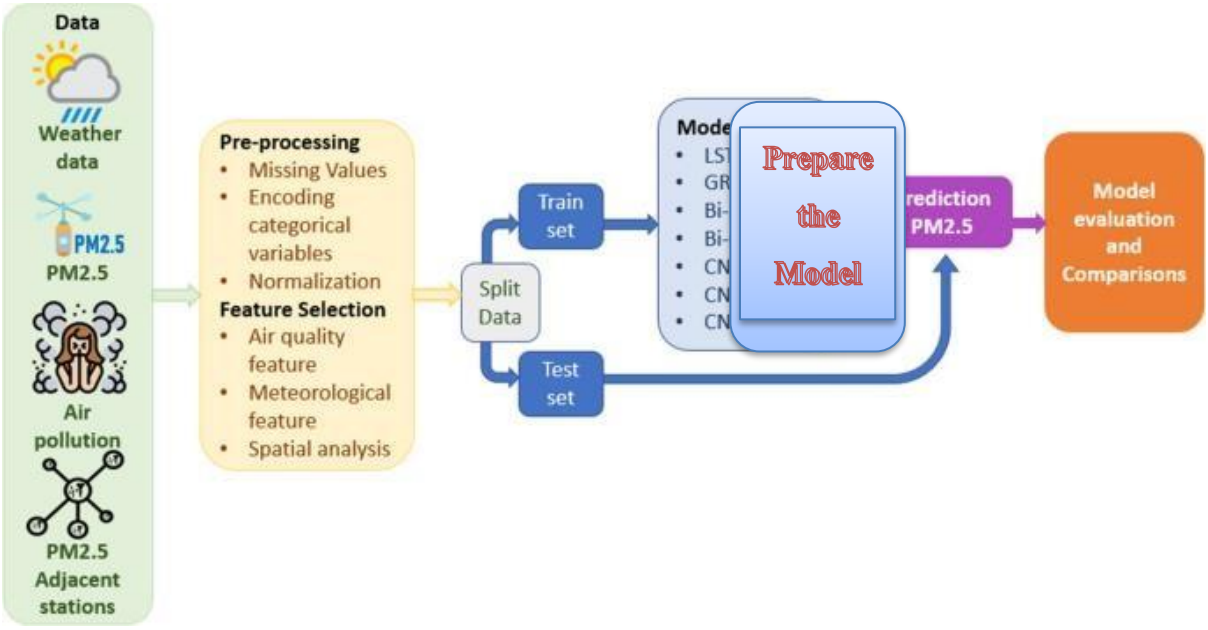

Fig 3.3. Architecture of Air Quality Predicting Model

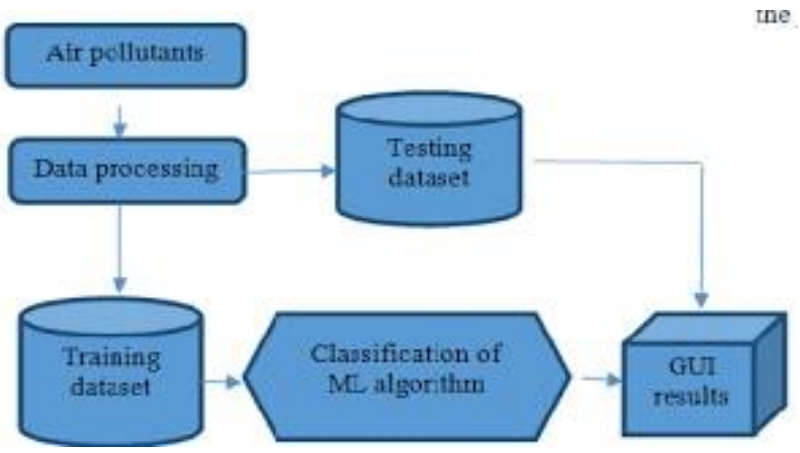

Fig 3.4: Basic flow diagram of Project

Chapter 4

RESULTS ANALYSIS AND VALIDATION

4.1 Modern Engineering Analysis

Predicting changes in air pollutant concentrations due to human and nature drivers are critical and challenging, particularly in areas with scant data inputs and high variability of parameters. Immoderate amounts of gases, particles, or molecules that are brought into Earth's atmosphere certainly cause Air Pollution. Potential air pollution impacts are estimated using air quality computational simulation models. We try to create one such Machine learning Model through our project, to help determine the Air Quality Index (AQI) using historical data and certain other relevancies like density, pollutants, etc. The model uses various modern Machine Learning Algorithms for the analysis of the data and interpreting accurate results as much as possible. The Machine Learning Algorithms used for analysis are briefly explained below:

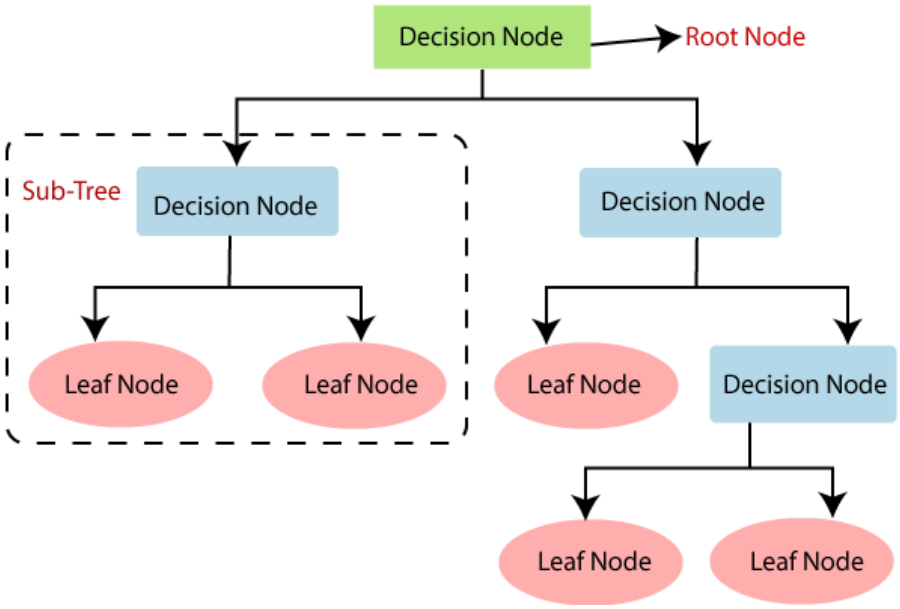

Fig 4.1 - Decision Tree workflow

Decision Tree: The decision tree is considered among the most effective and common algorithms for classification and future prediction. DT is a conceptual tree alike model, where each internal node represents a feature that best split the data into subsets using statistical measures such as information gain and gain ration, the process of splitting data is a recursive process until reaching the leaf (normally one of class labels). In machine learning, DT is one of the supervised learning algorithms.

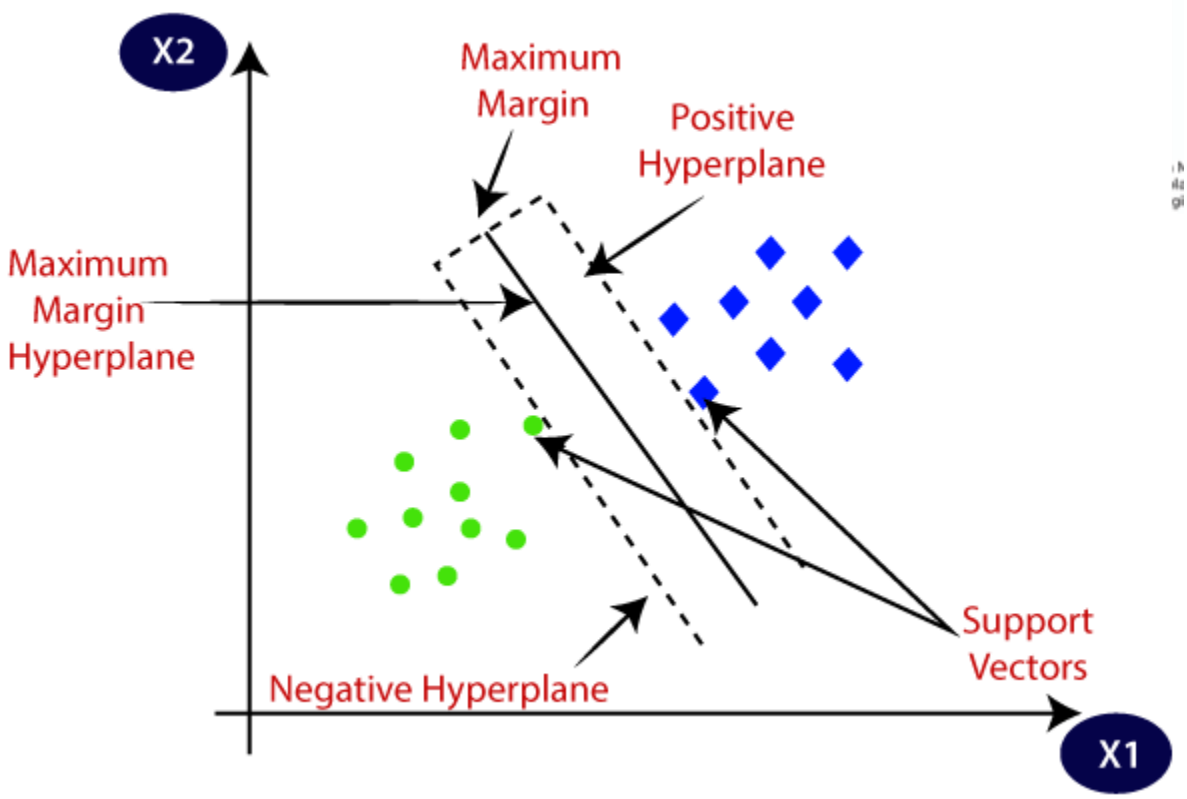

Fig 4.2 – Support Vector Machine (SVM)

Support Vector Machine (SVM): Support Vector Machine (SVM)is a well-known machine learning algorithm for classification and prediction purposes. SVM label each instance to a certain and given target class, by making it a non-probabilistic binary linear classifier. The model will focus on instances at the edge of each cluster and use the middle point between clusters as threshold, then allocate each new instance in accordance to its distance to threshold.

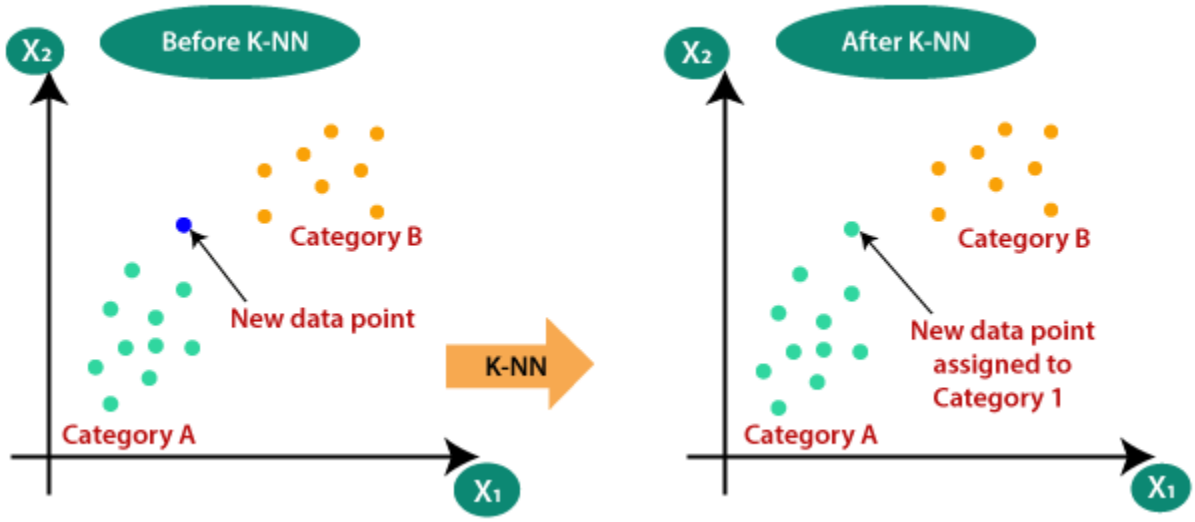

Fig 4.3 – K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN): The K-Nearest Neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another. For classification problems, a class label is assigned on the basis of a majority vote—i.e., the label that is most frequently represented around a given data point is used. While this is technically considered "plurality voting", the term, "majority vote" is more commonly used in literature. The distinction between these terminologies is that "majority voting" technically requires a majority of greater than 50%, which primarily works when there are only two categories. When you have multiple classes—e.g., four categories, you don't necessarily need 50% of the vote to make a conclusion about a class; you could assign a class label with a vote of greater than 25%.

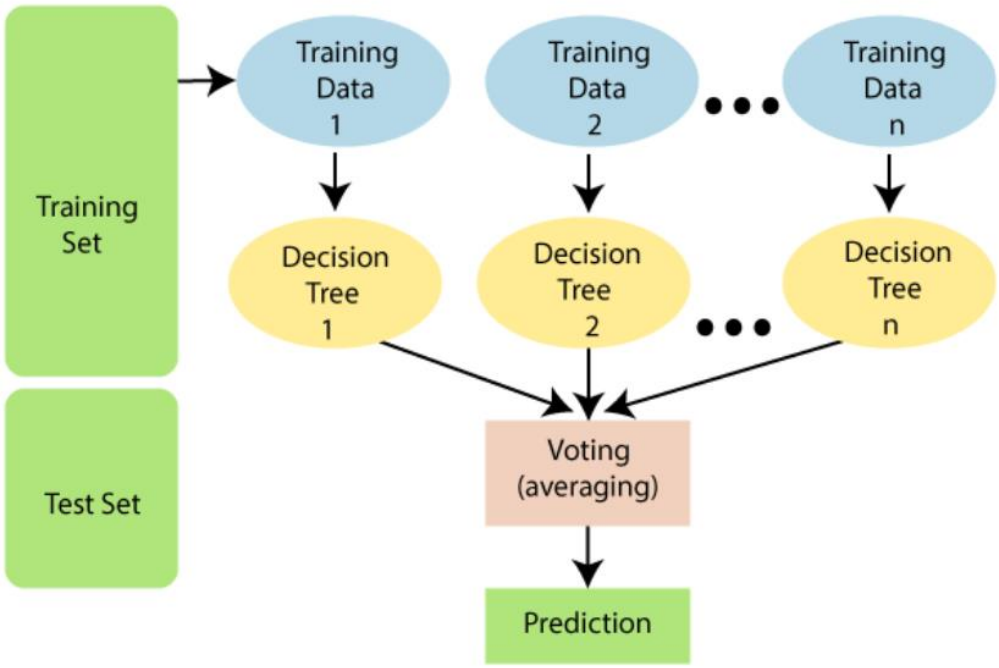

Fig 4.4 – Random Forest

Random Forest: Random Forest (RF) is one of the most popular machine learning algorithms for regression and classification tasks. RF creates a number of decision trees called forest trees to enhance the prediction process and produce higher accuracy. Building RF tree is similar to decision tree (DT) using information gain or other measures. Since RF is a set of DTs; each tree obtains a certain output and RF will choose the majority output produced by DTs or the mean in case of regression problem. RF is used over DT because of its ability to handle missing and solve the overheating problem.

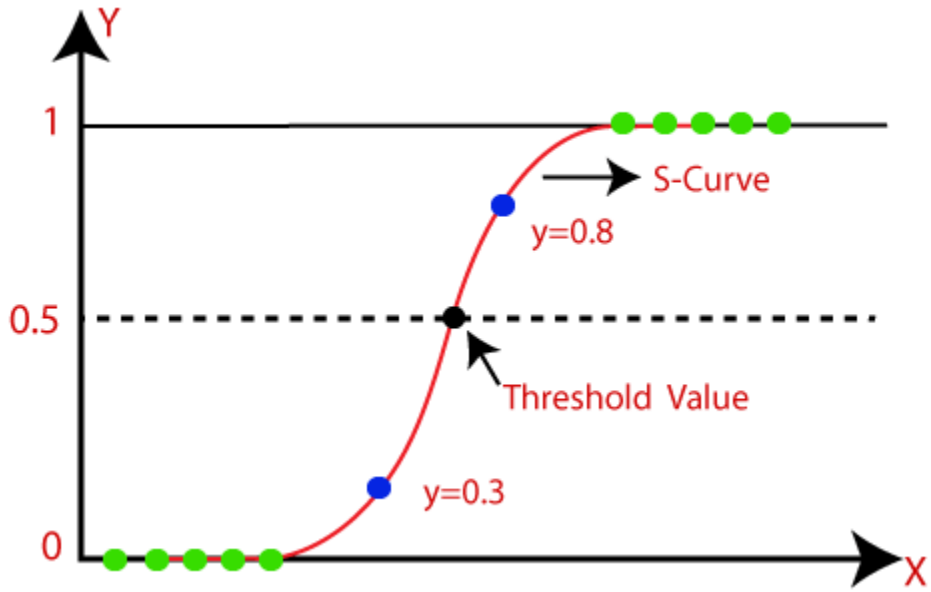

Fig 4.5 – Logistic Regression

Logistic Regression: Logistic Regression is a popular statistical machine learning algorithm for classification problems, the prediction of output is performed using nonlinear function such as sigmoid and logit functions. This algorithm is applied on Categorical Dependent Variable.

4.2 Model Drawing / Schematics

The left part of the Schematic Flowchart shows the working of non-parametric Machine learning algorithm, i.e., K-Nearest neighbors (KNN). While the right part of the model schematic is depicting the working of parametric Machine Learning algorithms used in analysis of Air Quality, namely, Decision Tree, Support Vector Machine (SVM), Random Forest and Logistic Regression.

All these algorithms are made to work together and their accuracy in analysing the Air Quality is been compared.

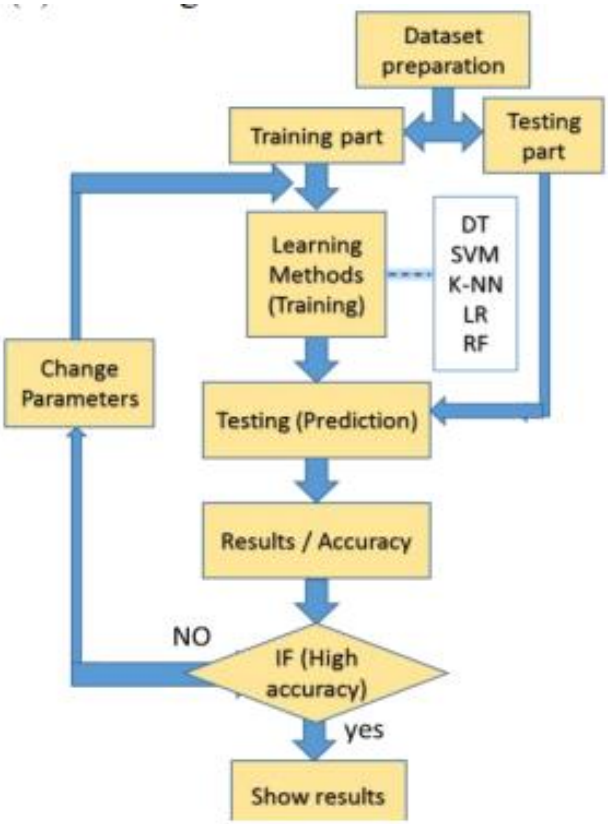

Fig 4.6 Working Flowchart

4.3 Testing / Data Validation

Algorithm	Accuracy
Decision Tree	99.97%
Support Vector machine (SVM)	99.72%
K-Nearest Neighbours (KNN)	99.67%
Random Forest	99.98%
Logistic Regression	82.71%

Fig 4.7 Accuracy Percentage

Testing and Validation has been done on various machine Learning Algorithms as mentioned above. The algorithm with the highest accuracy of 99.98%, Random Forest was chosen as a result.

Chapter 5
CONCLUSION AND FUTURE WORK

5.1Conclusion

Since our model can predict current data with 99% accuracy, it can successfully predict the upcoming air quality index of any specific data within a given region. With this model, we can forecast the AQI and alert the appropriate regions of the country. Because it is a progressive learning model, it is capable of tracing back to the specific location that requires attention if time series data from every possible region is available. The air quality data used in this paper is from the India air quality testing and investigation stage, and it includes the normal daily fine particulate issue (PM2.5), inhalable particulate issue (PM10), ozone (O3), CO, SO2, NO2 fixation, and air quality record (AQI). The most important aspects to consider when gauging the poison focus are its various sources as well as the factors that influence its fixation.

5.2Future work

There are many things that can be done in this domain of air pollution prediction system. Increasing the dataset and adding more data can give the model a training to predict more deviating and vivid values that are required.

It can also be integrated with hardware and be an application of IoT (Internet of Things) by which it can predict the real time air quality of a particular place. It is a very project that can be useful in day to day life of a person(People with respiratory disorders) by giving air quality to an individual who can prepare himself well in advance to face the air outside.

REFERENCES

1. Urban population (% of total population). https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS Accessed 20 Oct 2021
2. Department of Economic and Social Affairs: Urban Population Change; 2018. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html. Accessed 20 Oct 2021.
3. Nada Osseiran, Christian Lindmeier: 9 out of 10 people worldwide breathe polluted air, but more countries are taking action; 2018. https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action Accessed 20 July 2021.
4. Ailshire JA, Crimmins EM. Fine particulate matter air pollution and cognitive function among older US adults. Am J Epidemiol 2014;180(4):359–66. https://doi.org/10.1093/aje/kwu155. https://academic.oup.com/aje/article-pdf/180/4/359/8640802/kwu155.pdf.
5. Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, Balakrishnan K, Brunekreef B, Dandona L, Dandona R, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. Lancet. 2017;389(10082):1907–18.
6. Nallakaruppan, M. K., and Harun SurejIlango. "Location Aware Climate Sensing and Real Time Data Analysis." Computing and Communication Technologies (WCCCT), 2017 World Congress on. IEEE, 2017.
7. Singh Kunwar P., Shikha Gupta and Premanjali Rai, " Identifying pollution sources and prediction urban air quality using ensemble learning methods", Atmospheric environment80 (2013): 426-437.