A

Industry Oriented Mini Project Report

On

AIR QUALITY PREDICTION BASED ON MACHINE LEARNING

(Submitted in partial fulfilment of the requirements for the award of Degree)

Bachelor of Technology

in

COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

By

K. Durga Prasad (217R1A6796)

Under the guidance of

Dr.A.MAHENDAR

Assosciate Professor

Department of Computer Science & Engineering (Data Science)

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

Accredited by NBA & NAAC with 'A' Grade

Approved by AICTE, New Delhi and JNTU, Hyderabad.

2024-2025

Department of Computer Science & Engineering (Data Science)

CERTIFICATE

This is to certify that the project entitled **"Air Quality Prediction Based on Machine Learning"** being submitted by **K. Durga prasad (217R1A6796)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering (Data Science) to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by them under our guidance and supervision during the year 2024-25.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dr. A. Mahendar
Associate Professor
Internal Guide

Dr. K. Murali
Associate Professor & HOD

External Examiner

Submitted for viva-voce Examination held on: _____

ACKNOWLEDGMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.

We take this opportunity to express my profound gratitude and deep regard to my guide **"Dr.A.MAHENDAR"**, Designation for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to the Project Review Committee (PRC) **Mr. Sanjib Kumar Nayak, Mr. Ramji, Mr. Shankar Nayak, Ms. Sandhyarani** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Murali**, Head, Department of Computer Science and Engineering (Data Science) for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy**, Director for being cooperative throughout the course of this project. We also express our sincere gratitude to **Sri. Ch. Gopal Reddy**, Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

K. Durga Prasad (217R1A6796)

ABSTRACT

One major basic right is clean air that is integral to the concept of citizenship and it's while not a doubt, the responsibility of every subject to try to do his/her half to stay the air clean. Air quality prognostication has been looked into because the key answer of early warning and management of pollution. During this paper, we tend to propose an Associate in nursing air quality prediction system supported by a machine learning framework known as the sunshine GBM model, to predict air quality. This model, trained victimization lightweight GBM classifier, take meteorology knowledge jointly of sources for predicting the air quality thereby increasing the prediction accuracy by creating full use of obtainable abstraction data. The prevailing air quality observance stations and satellite meterologic knowledge offer period air quality observance info that is employed to predict the trend of air pollutants within the future. The projected system was found to administer Associate in nursing accuracy of ninety-two.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SCREENSHOTS

Chapter 1
INTRODUCTION

1.1 Introduction

The Environment describe about the thing which is everything happening in encircles the Environment is polluted by human daily activities which include like air pollution, noise pollution. If humidity is increasing more than automatically environment is going hotter. Major cause of increasing pollution is increasingday by day transport and industries. There are 75 % NO or other gas like CO, SO2 and other particle is exist in environment.. The expanding scene, vehicles and creations square measure harming all the air at a feared rate. Therefore, we have taken some attributes data like vehicles no., Pollutants attributes for prediction of pollution in specific zone of Delhi .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 with of the particle).

1.2 Project Scope

Air quality is a critical factor affecting public health and the environment. There is a need to predict air quality levels based on various environmental factors such as temperature, humidity, wind speed, and historical air quality data. The objective of this project is to develop a machine learning model that can accurately predict air quality index (AQI) values for the region. By leveraging historical data and real-time environmental parameters, the model aims to provide timely predictions that can inform residents and local authorities about potential air quality issues, enabling them to take proactive measures to mitigate health risks associated with poor air quality.

1.3 Project Purpose

The objective of the air quality prediction project is to develop a machine learning model that can accurately forecast air quality levels in a specific region.

1.4 Project Features

The limitations include reliance on data quality and availability, the complexity of environmental factors, risk of overfitting, and challenges with temporal and spatial

variability. Additionally, many models can be hard to interpret, making it difficult to understand predictions. Resource intensity for development and maintenance can also pose challenges, especially for smaller organizations.

Chapter 2
SYSTEM ANALYSIS

2.1Introduction

Air pollution is a global health concern, especially in fast-developing countries like India and China, with both short- and long-term exposure causing severe health impacts. The project aims to develop a machine learning-based prediction system to forecast air pollution levels and provide early warnings to protect human health. Advanced machine learning algorithms, including convolutional neural networks and deep learning, are used to predict the Air Quality Index (AQI).The system compares different models to find the most accurate solution for predicting future air pollutant concentrations and helping reduce health risks.

2.2Existing System

The existing system aims to predict air quality using machine learning techniques. It utilizes historical air pollutant data collected from sensors and preprocesses the data by cleaning and normalizing it. The dataset is divided into training and testing sets, and feature selection is applied to identify relevant attributes. Various supervised machine learning algorithms, including Linear Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest, are employed to predict air quality levels. The system's architecture involves data extraction, pre processing, feature selection, training and testing using machine learning algorithms, and prediction. The goal is to develop a model that accurately predicts air quality levels for improved pollution management and public health protection.

2.2.1Disadvantages of Existing System

- Less accuracy.
- Lack of data quality and short life cycles.
- The accuracy of air quality predictions heavily relies on the quality and coverage of sensor data.

2.3Proposed System

The proposed system aims to enhance air quality prediction by incorporating advanced machine learning algorithms like Random forest, decision tree , svm. It includes ensemble learning, time-series analysis, and real-time monitoring for

adapting to changing data. The system integrates external factors, improves model interpretability, and features a user-friendly interface.

2.3.1Advantages of Proposed System

- Improved prediction accuracy.
- Real-time monitoring and forecasting.
- Identification of key pollutant contributors.
- Enhanced decision-making for policy and regulation**.**

2.4Feasibility Study

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

♦ Economical Feasibility

♦ Technical Feasibility

♦ Social Feasibility

2.4.1Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified.

2.4.2Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client.

2.4.3Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

Chapter 3

HARDWARE & SOFTWARE REQUIREMENTS

3.1Introduction

It is crucial to outline the necessary hardware and software requirements to ensure successful implementation. The hardware components typically include sensors for data collection, powerful computing resources for processing and training machine learning models, and storage solutions for managing large datasets. On the software side, a variety of programming languages, libraries, and frameworks are essential for data analysis, model development, and visualization. Understanding these requirements helps in setting up an efficient environment that supports accurate predictions and effective analysis of air quality data.

3.1.1Hardware Requirements

- System : Intel Core i7.
- Hard Disk : 1TB.
- Monitor : 15'' LED
- Input Devices : Keyboard, Mouse
- Ram : 8GB.
-

3.1.2Sofware Requirements

- Operating system : Windows 10.
- Coding Language : Python
- Tool : PyCharm, Visual Studio
- Database : SQLite

Chapter 4

SYSTEM DESIGN

4.1Introduction

Architecture defines the components, modules , interfaces and data for a system to satisfy specified requirements. One should see as the applications of the systems theory to product development**.**

4.2Architecture

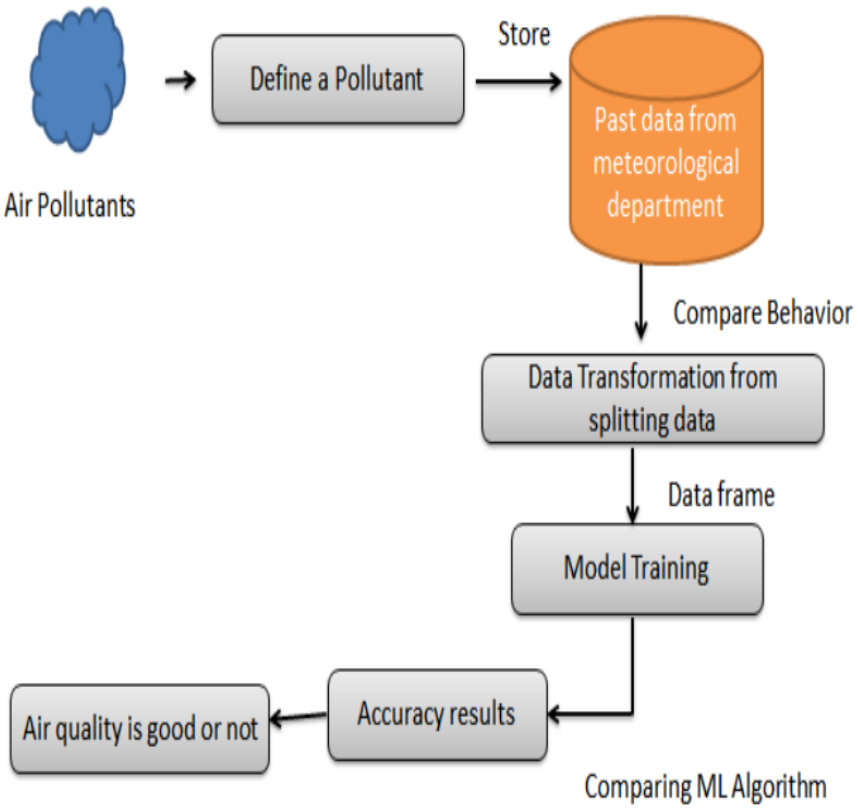

Figure 4.1 Proposed Architecture

4.3Unified Modelling Language (UML)

UML stands for Unified Modelling Language, UML id a standardized several modelling language in the field of object-oriented software engineering. The standard is managed and was created by, the Object Management Group. The goal is for UML to become a regular dialect for making fashions of item arranged PC programming. In its gift frame UML is contained two noteworthy components a Meta-show and documentation, later on, a few type of method or system can also likewise be brought to, or related with, UML. The Unified Modelling Language is a popular dialect for indicating. Visualization. Constructing and archiving the curios of programming framework, and for business demonstrating and different non- programming frameworks. The UML speaks to an accumulation of first-rate building practices which have verified fruitful in the showing of full-size and complicated frameworks. The UML is an essential piece of creating gadgets located programming and the product development method. The UML makes use of commonly graphical documentations to specific the plan of programming ventures.

GOALS OF UML:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.

2. Provide extendibility and specialization mechanisms to extend the core concepts.

3. Be independent of particular programming languages and development process.

4. Provide a formal basis for understanding the modelling language.

4.3.1Use Case Diagram

A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

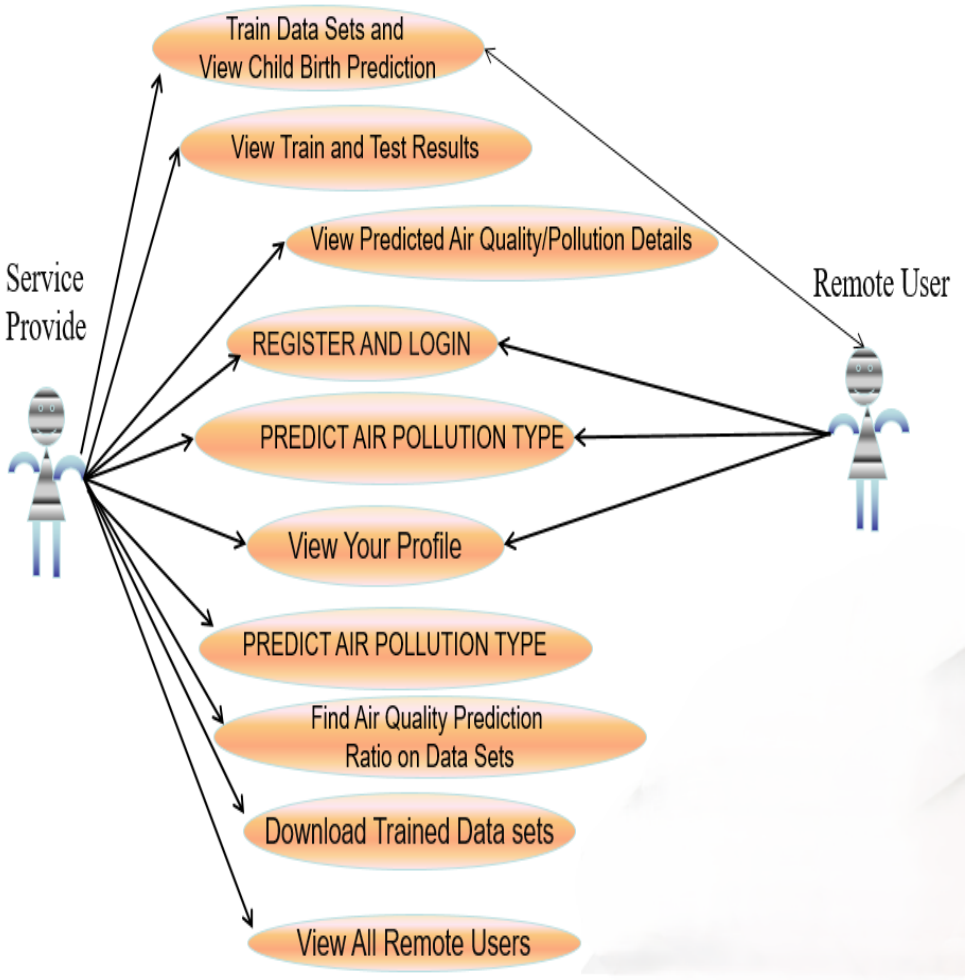

Figure 4.2 Use Case Diagram

4.3.2 Class Diagram

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

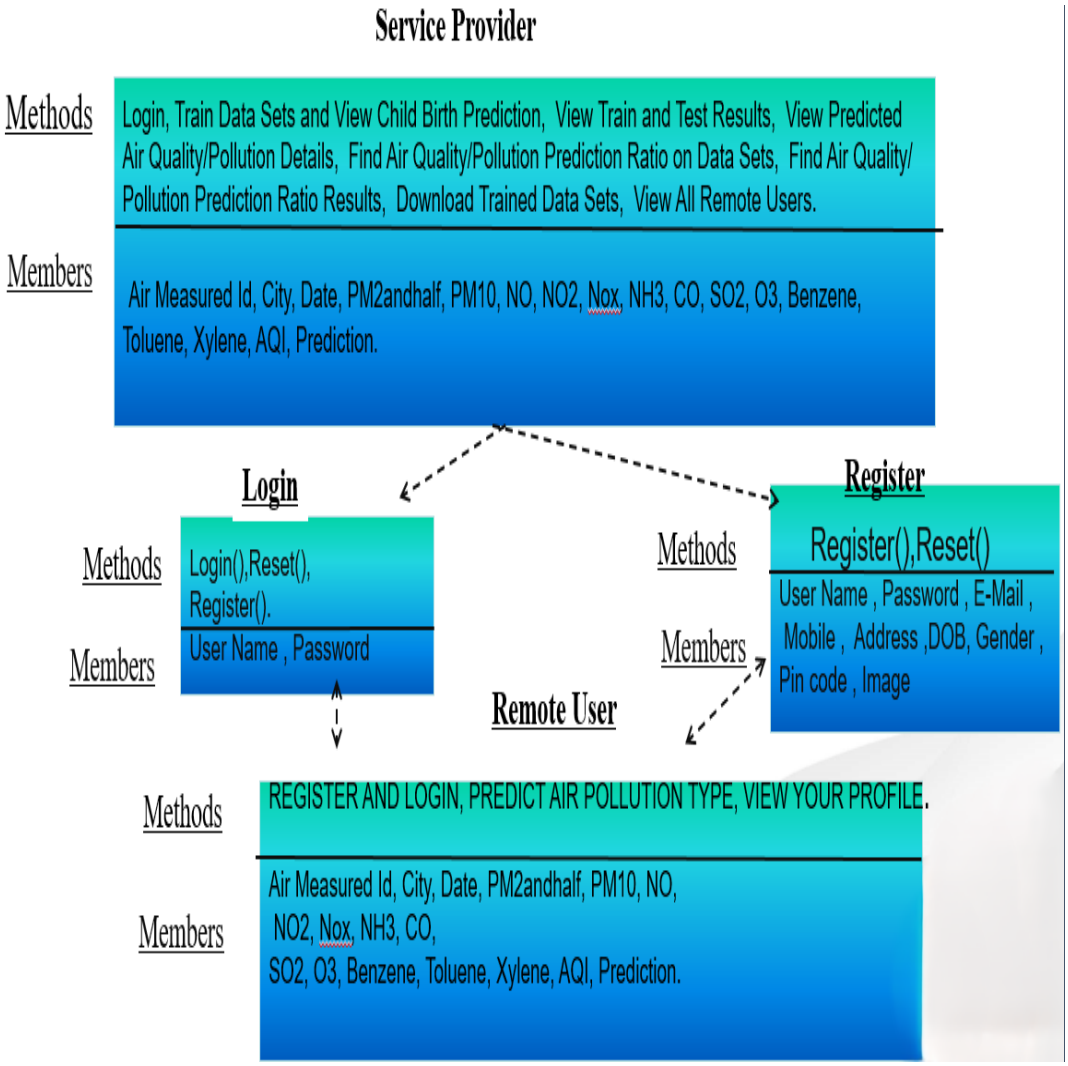

Figure 4.3 Class Diagram

4.3.3 Sequence Diagram

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

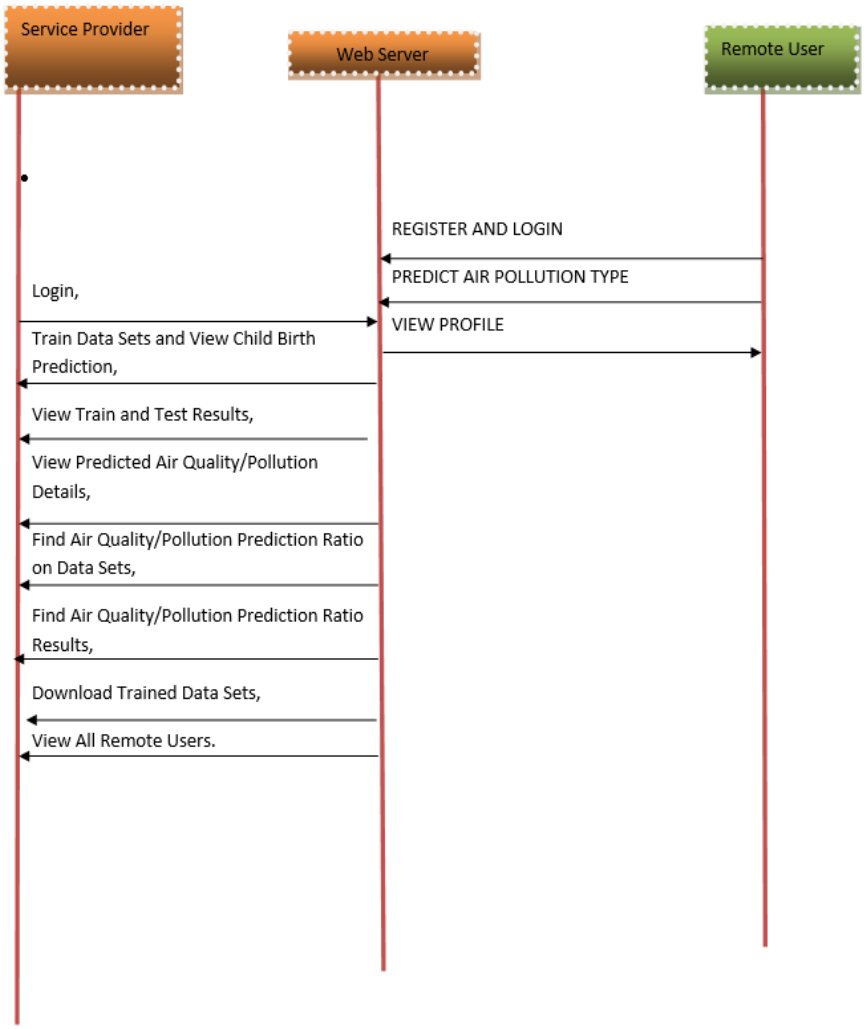

Figure 4.4 Sequence Diagram

4.3.4 Flowchart Diagram

A flowchart diagram is a separate diagramming technique used to represent the flow of process or algorithm. They use various shapes and arrows to illustrate the sequence of steps or decisions in a process. Flowcharts are commonly used in software development, business processes and problem solving.

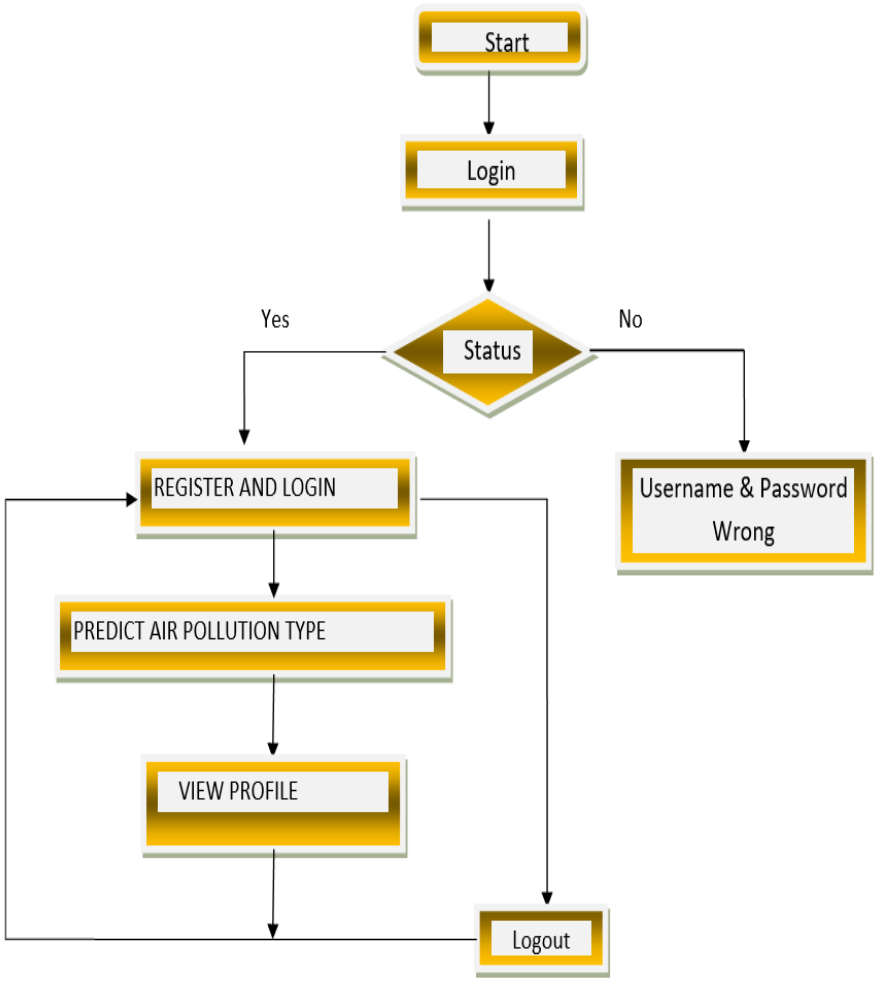

Figure 4.5 Flowchart Diagram for Remote user

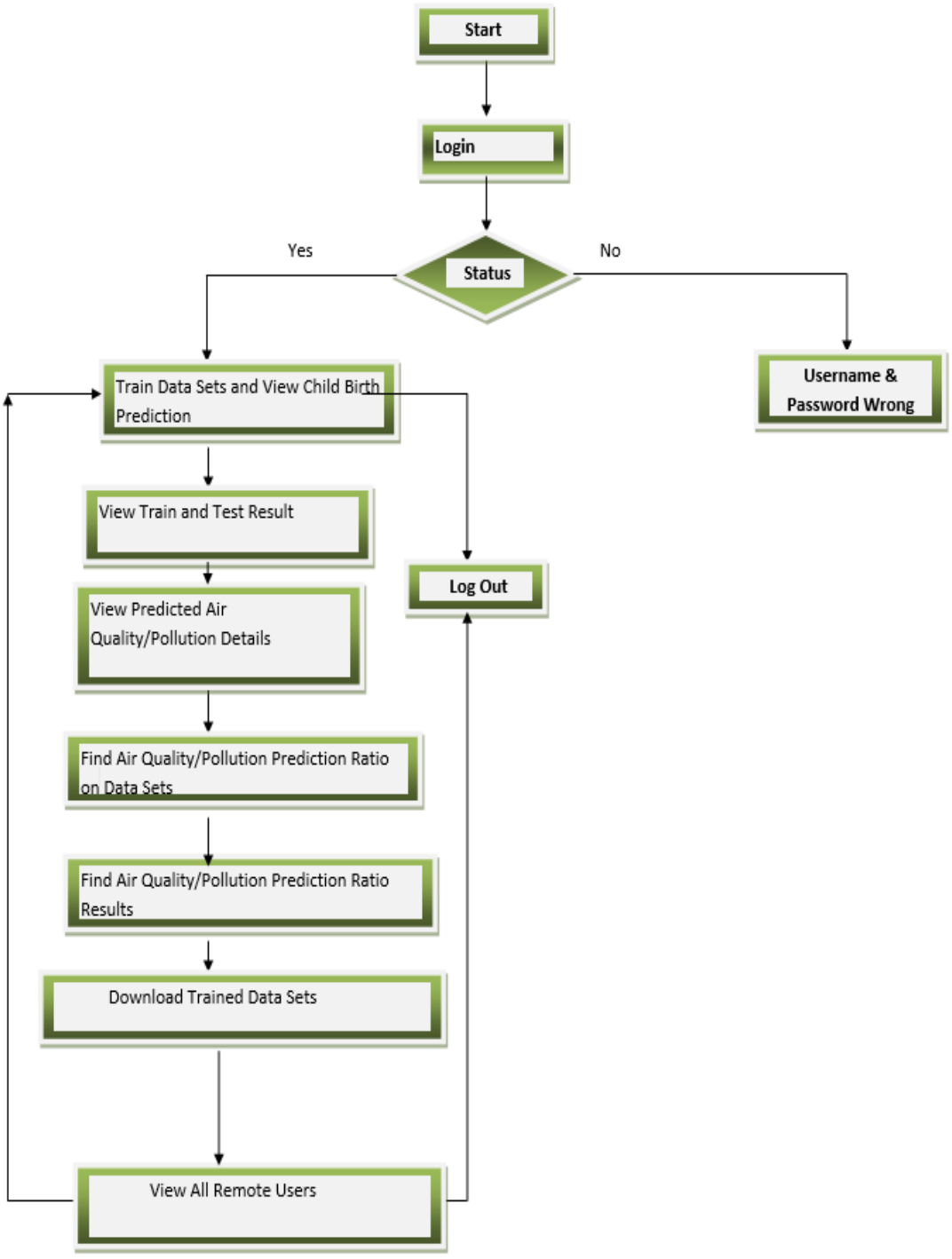

Figure 4.6 Flowchart Diagram for Service Provider

Chapter 5

IMPLEMENTATION

5.1Machine Learning Algorithms

5.1.1Decision Tree Algorithm

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck.

5.1.2KNN Algorithm

Simple, but a very powerful classification algorithm Classifies based on a similarity measure. It is non-parametric and also a lazy learning .Does not “learn” until the test example is given .Whenever we have a new data to classify, we find its K-nearest neighbors from the training data.

5.1.3 Random Forest Algorithm

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

5.1.4 SVM Algorithm

In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches.

5.2Data set

Dataset/Source: Ahmedabad Govt. Website .Structured/Unstructured data taken Structured Data in CSV format. DatasetDescription: Pollutants data is collected from Ahmedabad state. So we had 20383 records. This dataset consist of 13 data values listed below.

1) station code
2) sampling data date
3) state
4) location
5) agency name
6) type R/I
7) so2(sulfur dioxide)
8) no2
9) rspm
10) spm
11) location monitoring station
12) pm2.5
13) date

Information fields and their depiction: -

1.The Air Quality Index (AQI) [9]: is define about an inventory which contain record of the ordinaryair greatness. Polluted and clean air data arises here every day by some parameter.

2.STATION: station define where we have enrolled vehicle in different place in Maharastra.

5.3 Sample Code

```
 import pandas as pd
import matplotlib.pyplot as plt
def avg_data_2013():
    temp_i=0
    average=[]
    for rows in pd.read_csv('Data/AQI/aqi2013.csv',chunksize=24):
        add_var=0
        avg=0.0
        data=[]
            df=pd.DataFrame(data=rows)
        for index,row in df.iterrows():
            data.append(row['PM2.5'])
        for i in data:
            if type(i) is float or type(i) is int:
                add_var=add_var+i
            elif type(i) is str:
                if i!='NoData' and i!='PwrFail' and i!='---' and i!='InVld':
                    temp=float(i)
                    add_var=add_var+temp
                        avg=add_var/24
                temp_i=temp_i+1
                average.append(avg)
    return average
def avg_data_2014():
    temp_i=0
    #creating average varibale with empty list
    average=[]
    #Iterating 24 data Values from the 2013 data set that store in rows
    for rows in pd.read_csv('Data/AQI/aqi2014.csv',chunksize=24):
        #creating Varibales with Empty Value and empty list
        add_var=0
        avg=0.0
        data=[]
```

```
        #Covertinng Rows values into dataframe
        df=pd.DataFrame(data=rows)
        #Accesing
        for index,row in df.iterrows():
            data.append(row['PM2.5'])
        for i in data:
            if type(i) is float or type(i) is int:
                add_var=add_var+i
            elif type(i) is str:
                if i!='NoData' and i!='PwrFail' and i!='---' and i!='InVld':
                    temp=float(i)
                    add_var=add_var+temp
        avg=add_var/24
        temp_i=temp_i+1
                average.append(avg)
    return average
def avg_data_2015():
    #Creating temp_i varibale With 0 Value
    temp_i=0
    #creating average varibale with empty list
    average=[]
    #Iterating 24 data Values from the 2013 data set that store in rows
    for rows in pd.read_csv('Data/AQI/aqi2015.csv',chunksize=24):
        #creating Varibales with Empty Value and empty list
        add_var=0
        avg=0.0
        data=[]
        #Covertinng Rows values into dataframe
        df=pd.DataFrame(data=rows)
        #Accesing
        for index,row in df.iterrows():
            data.append(row['PM2.5'])
        for i in data:
        if type(i) is float or type(i) is int:
```

```
            add_var=add_var+i
        elif type(i) is str:
            if i!='NoData' and i!='PwrFail' and i!='---' and i!='InVld':
                temp=float(i)
                add_var=add_var+temp
    avg=add_var/24
    temp_i=temp_i+1
        average.append(avg)
  return average
def avg_data_2016():
  #Creating temp_i varibale With 0 Value
  temp_i=0
  #creating average varibale with empty list
  average=[]
  #Iterating 24 data Values from the 2013 data set that store in rows
  for rows in pd.read_csv('Data/AQI/aqi2016.csv',chunksize=24):
    #creating Varibales with Empty Value and empty list
    add_var=0
    avg=0.0
    data=[]
    #Covertinng Rows values into dataframe
    df=pd.DataFrame(data=rows)
    #Accesing
    for index,row in df.iterrows():
      data.append(row['PM2.5'])
    for i in data:
      if type(i) is float or type(i) is int:
        add_var=add_var+i
      elif type(i) is str:
        if i!='NoData' and i!='PwrFail' and i!='---' and i!='InVld':
          temp=float(i)
          add_var=add_var+temp
    avg=add_var/24
    temp_i=temp_i+1
```

```
        average.append(avg)
    return average
if __name__=="__main__":
    lst2013=avg_data_2013()
    lst2014=avg_data_2014()
    lst2015=avg_data_2015()
    lst2016=avg_data_2016()
    #lst2017=avg_data_2017()
    #lst2018=avg_data_2018()
    plt.plot(range(0,365),lst2013,label="2013 data")
    plt.plot(range(0,364),lst2014,label="2014 data")
    plt.plot(range(0,365),lst2015,label="2015 data")
    plt.plot(range(0,365),lst2016,label="2016 data")
    plt.xlabel('Day')
    plt.ylabel('PM 2.5')
    plt.legend(loc='upper right')
    plt.show()
```

Chapter 6

SCREENSHOTS

6.1Home Page

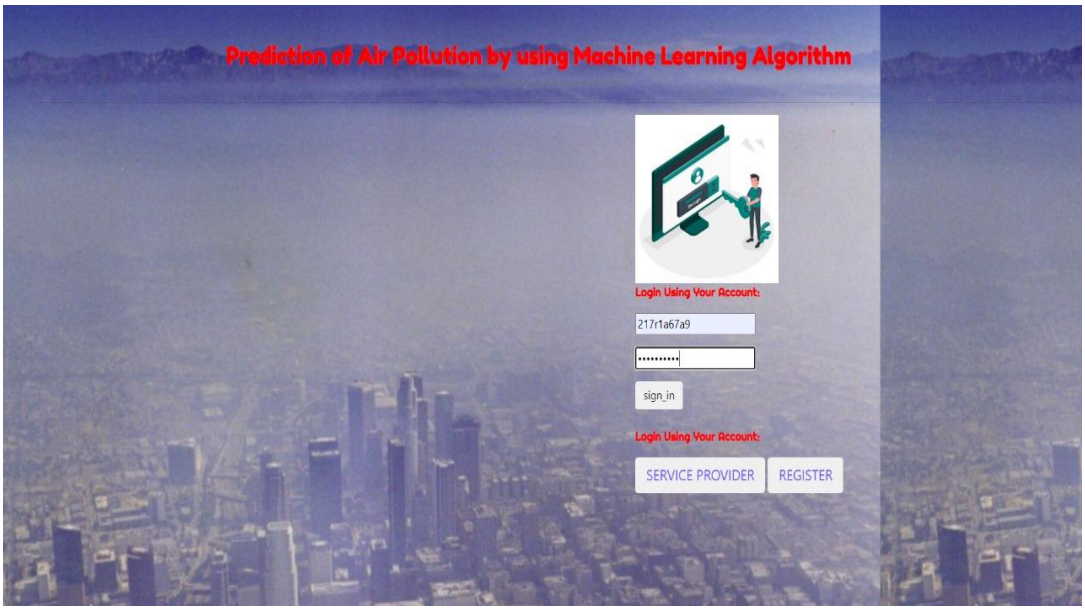

Figure 6.1 Home Page

(It is the page were login credentials need to be provided to view the result)

6.2 Predicted Data Set

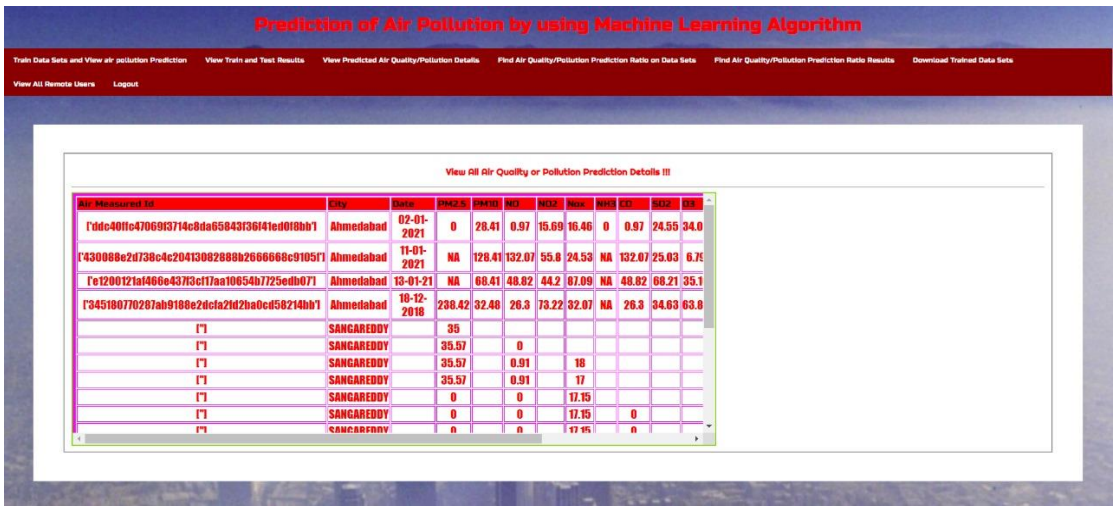

Figure 6.2 Predicted Data Set

(It is the dataset that we consider to check air purity)

6.3 Predicted Data Set Result

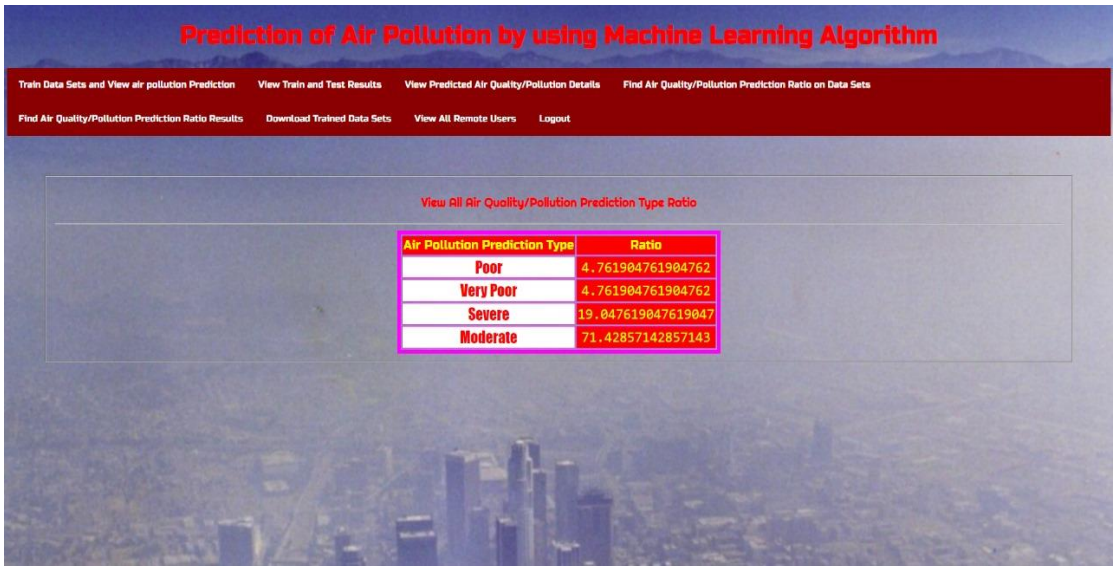

Figure 6.3 Predicted Data Set Result

(It gives the ratio of different aspects to give a proper conclusion)

6.4 Comparison Results

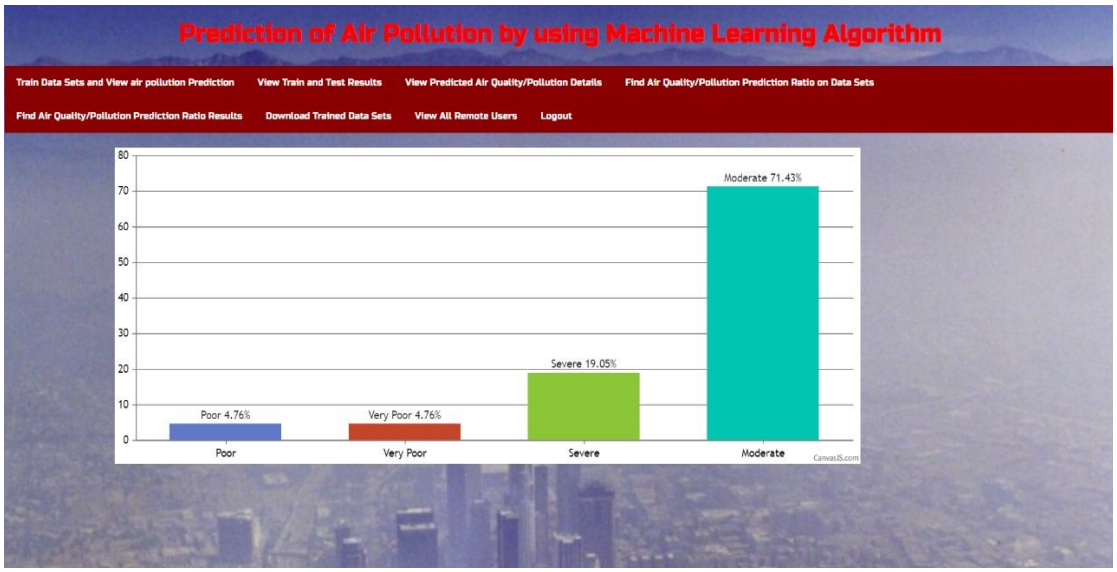

Figure 6.4 Comparison Results

(It shows which area has poor, very poor, severe, Moderate air pollution)

Chapter 7

TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/are finished product it is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail unacceptably. There are various types of test. Each test type addresses a specific.

7.1 Types of Testing

7.1.1 Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

7.1.2 Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

7.1.3 Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centre on the following items: Valid input identified classes of valid input must be accepted Invalid input identified classes of invalid input must be rejected. Functions identified functions must be exercised output identified classes of application outputs must be exercised. Systems/procedures interfacing systems or procedures must be invoked. Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

7.1.4 System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System this is based on process descriptions and flows, emphasizing pre-driven process links and integration.

7.1.5 White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached.

7.1.6 Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, asmost other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document.

Chapter 8

CONCLUSION & FUTURE SCOPE

8.1 Conclusion

Precision of our model is very acceptable. The anticipated AQI has a precision of 96%. Future upgrades incorporate expanding the extent of district and to incorporate whatever number locales as could be allowed as of now this venture targets foreseeing the AQI estimations of various areas of close by New Delhi. Further, by utilizing information of various urban areas the extent of this venture can be exhausted to anticipate AQI for different urban communities also.

8.2 Future Scope

It includes enhanced data collection through IOT devices, integration with other data sources for better accuracy, personalized predictions for individuals, support for policy-making, and the development of public awareness tools. As machine learning techniques advance, we can expect more accurate and reliable models, ultimately improving public health and environmental sustainability.

REFERENCES

1. G. Corani and M. Scanagatta, "Air pollution prediction via multi-labelclassification," Environ. Model. Softw., vol. 80, pp. 259-264,2016.
2. Ni, X.Y.; Huang, H.; Du, W.P. “Relevance analysis and short-termprediction of PM 2.5 concentrations in Beijing based on multi-sourcedata.” Atmos. Environ. 2017, 150, 146-161.
3. RuchiRaturi, Dr. J.R. Prasad .“Recognition Of Future Air Quality Index using Artificial Neural Network”. International Research Journal of Engineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN:2395-0072 Volume: 05 Issue: 03 Mar-2018
4. Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu .” Detection and Prediction of Air Pollution using Machine
5. Learning Models”. International Journal o f Engineering Trends and Technology (IJETT) - volume 59 Issue 4 - May 2018
6. Suhasini V. Kottur , Dr. S. S. Mantha. “An Integrated Model Using Artificial Neural Network