AIR QUALITY INDEX PREDICTION IN GUWAHATI

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE AWARD OF THE DEGREE
OF

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE & ENGINEERING

Submitted By
Aicheng Jaohai(17/376)
Randwip Ghosh(17/049)
Sangramjit Dutta (17/314)

Guided By
Dr. Gunajit Kalita
(Assistant Professor)
Co Guided By
Mr. Prasenjit Saha
(Guest Faculty)

2015-21
ASSAM SCIENCE AND TECHOLOGY UNIVERSITY, GUWAHATI
ASSAM ENGINEERING COLLEGE, JALUKBARI
GUWAHATI-781013
March-2021

AIR QUALITY INDEX PREDICTION IN GUWAHATI

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE AWARD OF THE DEGREE
OF

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE & ENGINEERING

Submitted By
Aicheng Jaohai(17/376)
Randwip Ghosh(17/049)
Sangramjit Dutta (17/314)

Guided By
Dr. Gunajit Kalita
(Assistant Professor)
Co Guided By
Mr. Prasenjit Saha
(Guest Faculty)

2015-21
ASSAM SCIENCE AND TECHOLOGY UNIVERSITY, GUWAHATI
ASSAM ENGINEERING COLLEGE, JALUKBARI
GUWAHATI-781013
March-2021

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ASSAM ENGINEERING COLLEGE::JALUKBARI

GUWAHATI-781013

Forwarding Certificate

This is to certify that **Randwip Ghosh(17/049),Sangramjit Dutta (17/314),Aicheng Jaohai(17/376)**has/have carried out the project work**AIR QUALITY INDEX PREDICTION IN GUWAHATI** under the supervision of Dr. Gunajit Kalita&Mr. Prasenjit Saha has/have compiled this thesis reflecting the candidate's work in the semester long project. The candidate(s) did this project full time during the whole semester and the analysis, results, claims etc. are all related to his/her/their studies/study and works during the semester.

I/We recommend submission of this thesis as the partial fulfillment of the requirement for the degree of Bachelor of Engineering in Computer Science & Engineering of Assam Science and Techology University.

(HOD)

Name:………………...

Computer Science & Engineering

Signature:……………..

Assam Engineering College

Affiliation:……………..

Jalukbari , Guwahati - 781013

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ASSAM ENGINEERING COLLEGE::JALUKBARI
GUWAHATI-781013

Forwarding Certificate

This is to certify that **Randwip Ghosh(17/049), Sangramjit Dutta (17/314),Aicheng Jaohai(17/376)**has/have carried out the project work**AIR QUALITY INDEX PREDICTION IN GUWAHATI** under my supervision and has/have compiled this thesis reflecting the candidate's work in the semester long project. The candidate(s) did this project full time during the whole semester and the analysis, results, claims etc. are all related to his/her/their studies/study and works during the semester.

I recommend submission of this thesis as the partial fulfillment of the requirement for the degree of Bachelor of Engineering in Computer Science & Engineering of Assam Science and Techology University.

Dr. Gunajit Kalita
(Assistant Professor)
Signature:
Department of Computer Science & Engineering
Assam Engineering College, Jalukbari
Guwahati - 781013

Mr. Prasenjit Saha
(Guest Faculty)
Signature:
Department of Civil Engineering
Assam Engineering College, Jalukbari
Guwahati - 781013

Assam Engineering College

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
ASSAM SCIENCE AND TECHOLOGY UNIVERSITY

Declaration by the Candidate

I/We _____________, **Roll No________**an BE. student(s) of the Department of Computer Science & Engineering, Assam Engineering College hereby declares that I/we have compiled this thesis reflecting all my/our works during the semester long full time project as a part of my BE curriculum.

I/We declare that I/we have included the descriptions etc of my project work, and nothing has been copied/replicated from other's work. The facts, figures, analysis, results, claims etc depicted in my/our report are all related to my full time project work.

I/We also declare that the same report or any substantial portion of this project report has not been submitted anywhere else as part of any requirements for any degree/diploma etc. *(If the thesis work required to use results, fact and figures from external owner/ agencies, then due permission must be taken and the same must be mentioned in the thesis.)*

Date:
RollNo………...Name……….Signature……….

ACKNOWLEDGEMENT

We would like to express our sincere thanks to **Mr. Dinesh Shankar Pegu**, **Head of Electronics & Telecommunication Engineering Department and Computer Science Engineering Department**for providing us the golden opportunity to work on this wonderful internship.

We would like to express our sincere appreciation to our supervisors **Dr. Gunajit Kalita, Assistant professor, Department of Computer Science &**. **Engineering** & **Mr. Prasenjit Saha, Guest Faculty, Department of Civil Engineering of Assam Engineering College**, for their extensive support and encouragement throughout the project work. We are highly indebted to them for his guidance and constant supervision as well as for providing necessary information regarding the project work. Working under them has indeed been a great experience and inspiration for us. The project report would not have been possible without the help of several individuals who in one way or another contributed and extended their valuable assistance in the completion of the study.

Name	**Roll no**	**Signature**
Aicheng Jaohai	17/376	
Randwip Ghosh	17/049	
Sangramjit Dutta	17/314	

ABSTRACT

Diseases related to air pollution and Air pollution in general at large remain one of the leading causes of deaths in present world. In order to ensure the safety of general masses living in the age of industrialization it is extremely necessary for them to find a way to foresee the consequences of the breathing air and what future it holds. Hence **AQI prediction** provides a way to do exactly that for them by incorporating **Machine learning** techniques from the present state of measured air data to provide a dependable prediction to prepare for a safer future by taking early actions.**Machine learning (ML)** methods contributed highly in the advancement of prediction systems providing better performance and cost effective solutions. Various ML models are compared and considered to find the one model most suitable and with most precise prediction.

CONTENTS **Page No.**

Chapter1:

INTRODUCTION

Chapter2:

PROJECT WORK

LIST OF FIGURES

LIST OF TABLES

AIR QUALITY INDEX PREDICTION IN GUWAHATI

1. Introduction

1.1Overview

Pollution has remained a serious challenge for the general public and the government everywhere in the planet. Pollution leads to various major health issues of the population where breathing problems such as asthma, coughing and allergic wheezing, reduced lung function, heart diseases such as CHD,IHD, etc., are just the tip of the iceberg. Urban air quality monitoring has been a relentless challenge with the arrival of industrialisation leading to deteriorating air quality (which may indicate how efficient the relations of industries are with respect to environment).

Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels and industrial parameters play significant roles in air pollution. With this increasing air pollution, Were in need of implementing models which will record information about concentrations of air pollutants (so2,no2,etc) [1]. Air pollution is a global environmental problem that influences mostly the health of urban population, and repeated exposures to ambient air pollutants over a prolonged period of time increases the risk of being susceptible to airborne diseases such as cardiovascular and respiratory diseases and lung cancer [2].An air quality index (AQI) can be defined as a communication tool and a standardized summary measure of ambient air quality used to express the level of health risk related to particulate and gaseous air pollution [3].Air Quality Index is also used by the government for easier understanding of the air pollution to common people. Awareness of the daily levels of air pollution is important for the citizens, because of the diseases spreading in the air exposed by the air pollution [4].

In a country like ours, the speed at that population is increasing is one in all the largest indicators of a rise in pollution. The largest reason behind this can be the indiscriminate use of natural resources. Formerly, this drawback was confined to the cities, however currently this drawback is spreading to the villages and therefore the rural area. Industry has additionally magnified hugely thanks to the increasing population. Cyanogenic air from the trade has contaminated the air in lieu of providing employment to the folks.

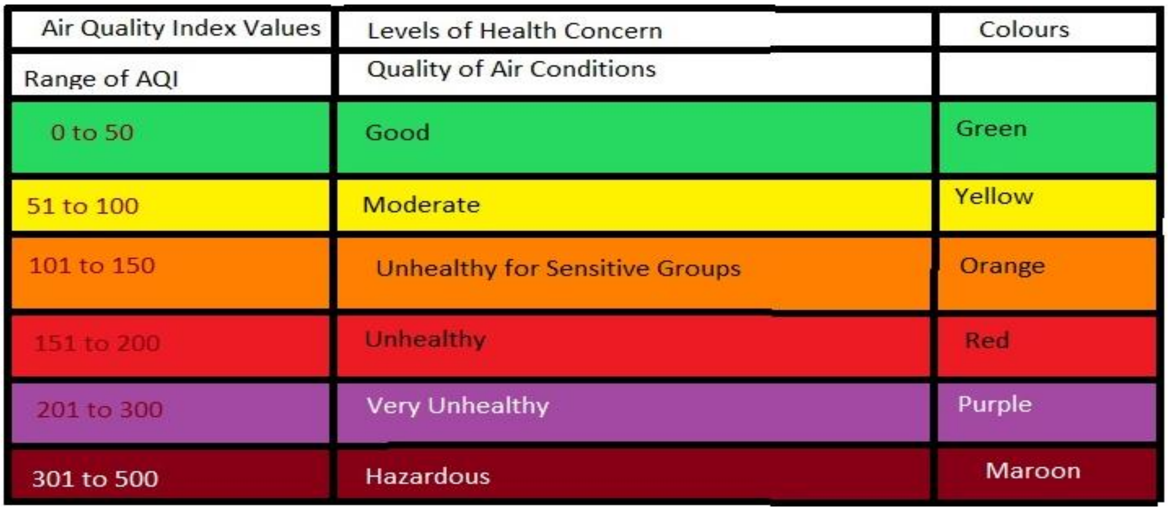
1.2Model location:

Guwahati is that the largest town within the state of province in Northeast of our country. It had been the capital town of province till 1972 once it had been moved to Dispur. As a result of its outstanding location, it's usually known as the 'Gateway to North-East India'. A few years past it had been additionally called Gauhati. The calculable population of its metropolitan space was one 1 million folks in 2020. Thanks to it being a preferred destination for those seeking work and/or education, this figure is ready to rise to one.5 million by 2035. Hence the air quality is steadily declining as population is a major factor for pollution.

In order to record and collect the data of air quality of Guwahati there are measuring stations across 6 locations in Guwahati namely Gopinathnagar,Borgaon,Bamunimaidam, Guwahati University, Khanapara, and Pragjyotish.

1.3What is AQI?

AQI or air quality index is a scale for measuring the air quality of an area within the range of 0-500, where the levels of severity of air pollution is given by good, moderate, unhealthy, very unhealthy, hazardous. The index is basically a way of determining the state of the breathable air of a particular place.



Air Quality Index Values	Levels of Health Concern	Colours
Range of AQI	Quality of Air Conditions	
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Fig 1.1

1.4How is AQI calculated?

To calculate or determine AQI of a particular area, the concentrations of constituent pollutants are measured through various measuring stations and then an average of each pollutant is considered over a standard time period to get a value that is compared to a standard AQI chart to determine the stature of the breathable air of the considered region.

Different countries use completely different purpose scales to report air quality. As an example, the USA uses a five hundred purpose scale, whereby rating between zero and fifty is taken under consideration to be good. Rating between three hundred one to five hundred rangeis deemed hazardous. Our country too follows that the five hundred purpose scale. On a daily basis concentrations of the most pollutants are monitored. These raw measurements area unit regenerate into a separate AQI value for each constituent component of polluted air (ground-level gas, particle pollution, monoxide gas , and sulphur dioxide). The very best of these AQI values / unit areais considered as a result of the AQI value for that day.

Along with categorization the air quality, AQI additionally provides recommendation on however one will improve the quality of the air one inhales. This index notably considers the folks sensitive towards pollution and advises them ways to protect themselves from completely different health risks display at numerous air quality levels.

1.5Importance of AQI:

The improvement of air quality depends on the people, i.e., and the citizens. Smart ways have to be found and sought after for mitigating the worldwide and local air pollution related problems. However, a broad summary of the high pollution concentrations or maybe the frequency of the NAAQS chance isn't ample for the people to assess the urban air quality. The overall public desires information on the pollution levels and the associated potential health risks of pollution bestowed in an easy, graspable format. There arenumerous continuous environment monitoring systems put in in massive cities showing the city’s environmental health in a range of figures and colours. However the monitored knowledge is in massive volumes and does not provide a transparent image to the decision-maker or to the common person who needs to know however safe or dangerous the air is. Thus, effective and obvious communication of air quality is vital. It is terribly crucial for those who pay most of their time operating or physical exercise outside as they'reliable to higher exposure and may have respiratory issues once pollution levels are severe.

AQI is all the additional vital in developing nations like Bharat wherever the common person is not quite familiar with the technical terminologies and measurement units (like ppm /ppb / or µg/mg3). Hence the AQI simplifies the understanding of their air quality by secret writing the standard in terms of unit less numbers and colour, with every figure and colour representing a distinct class of associated health risk.

Doctors also can use AQI in predicting once they establish a rise in cases with metabolism distress, as they decipher to be inclined once the AQI is high.

Air pollution from particulate PM2.5 and PM10 is principally the results of the burning of fossil fuels. It's thought of because the deadliest kind of pollution worldwide. Hazardous breathing conditions contribute to a lot of deaths in the long run. That is, on a mean, the average lifespan, decreases by regarding two years thanks to pollution from particulate matters suspended in polluted air. This loss of life potential is additionally devastating even when compared to the figures of communicable diseases like TB or HIV/AIDS, behavioural habits like smoke smoking, and even war. However in China and India, where pollution due to overpopulation and inefficient waste management , the usefulness of AQI I massive as the effective years added to average lifespan of the people dwelling in such countries can be increased to two to four years which in the vast majority is an impressive statistic and should be kept in one’s mind.

PROJECT WORK

2.1 APPROACH

The main aim of our experiment is to compare the air quality index values with their corresponding dates and find the following results:

a. A relationship between the two datasets
b. A model to predict the air quality index value of a certain date
c. Compare the actual and predicted results for different stations in a city

Our aim can be considered a regression problem and thus regression analysis has been use to satisfy our approach. Regression analysis is the primary technique to use when it comes to solving regression problems. It is a predictive modelling technique that is use to analyze and find a relation between the dependent and independent sets by plotting the datasets along different axis and then creating a line that passes through the data points in such a way that the distance between the line and data points are minimum. There are different regression technique's that can be used to analyse based on the data. Here four regression models has been build and compared. They are

a. Linear regression model
b. Support vector machine
c. Random forest model
d. Decision tree regression model

The main two criteria's that has been used to compare the accuracy of the four models are

a. mean square error
b. root mean square error

Mean square error tells us how close the regression line is to the actual value data points. It first takes the distance from actual value to predicted value and squares them to remove all negative sign. And the average of all the distance is calculated. The lower the mean square error the better will be the prediction.

Mean square error = $(1/n) * \Sigma(\text{actual} - \text{forecast})^2$
Where:

- n = number of items,
- Σ = summation notation,
- Actual = original or observed y-value,
- Forecast = y-value from regression.

Root mean square error tells us the standard deviation of the prediction errors that is how far the regression line is from the actual value.

$$RMSE = \sqrt{\overline{(f - o)^2}}$$

Where:

- f = forecasts (expected values or unknown results),
- o = observed values (known results).

2.2 DATA PREPROCESSING

For our machine learning model to process the data we first have to convert the rata data into a suitable format. This process is the first and crucial step in building any machine learning model and is called data pre-processing. The main steps involved in data preprocessingare :

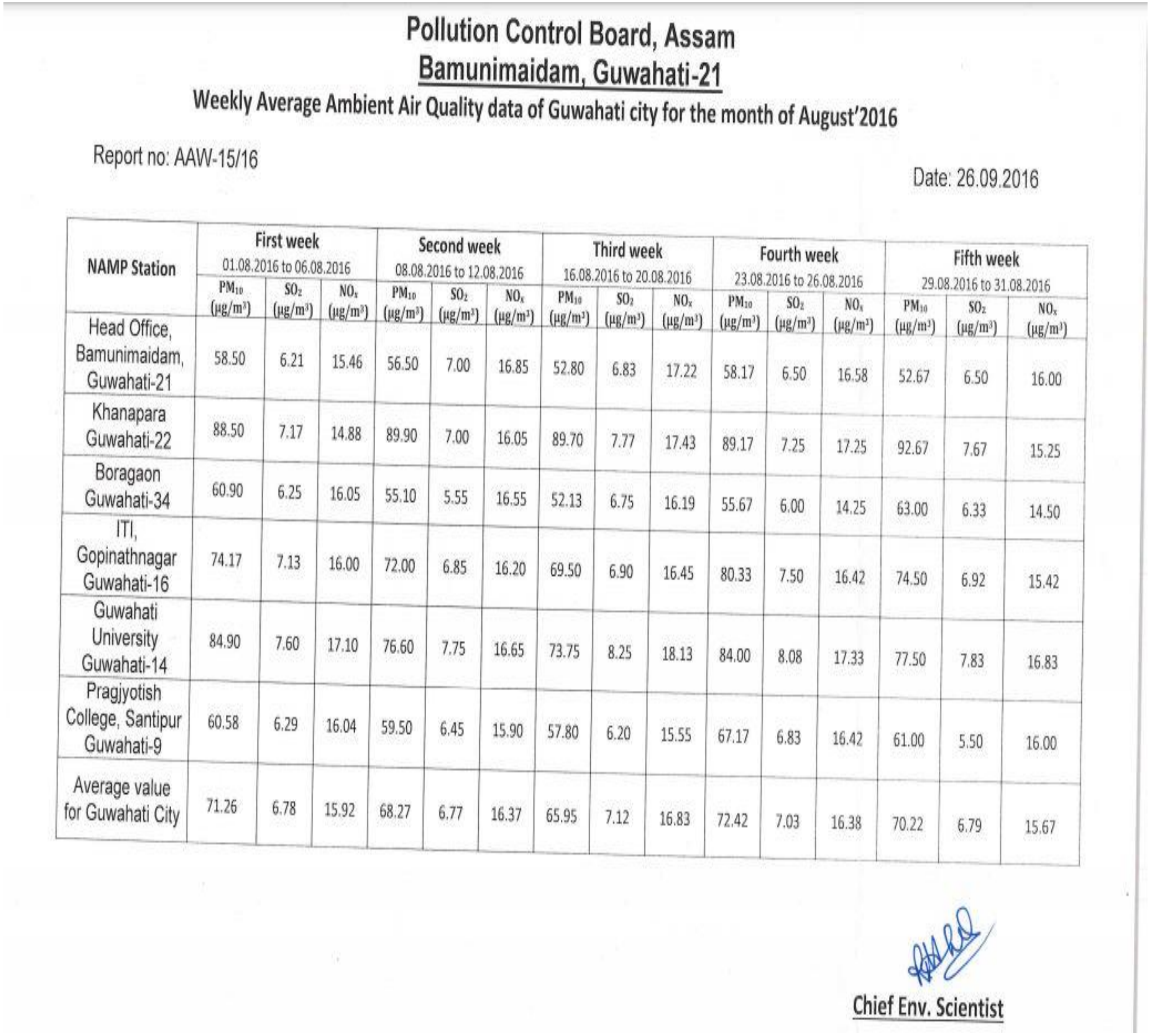
a. collecting data
b. Importing the libraries
c. Importing datasets
d. Finding missing data
e. Encoding categorical data
f. Splitting in training and test set
g. Feature scaling

2.2.1 COLLECTING DATA

The area of our interest is **Guwahati** city(formerly known as **Gauhati**). It is the largest city in the Indian state of Assam and also the largest metropolis in north eastern India. There are six main pollutants in determining air quality index according to Indian standard. Out of these six, three pollutants are observed in six stations in Guwahati out which the highest value is considered the air quality index value. The pollutants are PM10, SO2 and NOx. The stations record the weekly AQI value for these three pollutants. We have taken the records from year 2015 to 2019. The six stations are:

a. Head Office, Bamunimaidam, Guwahati-21
b. Khanapara, Guwahati-22
c. Boragaon Guwahati-34
d. ITI, Gopinathnagar, Guwahati-16
e. Pragjyotish College, Santipur Guwahati-9
f. Guwahati University, Guwahati-14

All the data has been collected from the Pollution Board of Assam website. In the raw data the weekly AQI values for each pollutant were given for each station. Thus different dataset were created for each location. There are six datasets in total. Each dataset consist of three pollutant column, month column, year column and week column. The pollutant column is considered as our dependent set. Another column has been added which is the number of day’s column. Taking 2015 as base year the column will start from value 0 and will increase by 7 since weekly AQI data is taken (1 week = 7 days). This day’s column will be our independent set. Regression models are then used to analyse the relation between the individual pollutant dependent set and the number of day’s independent set.



Pollution Control Board, Assam

Bamunimaidam, Guwahati-21

Weekly Average Ambient Air Quality data of Guwahati city for the month of August'2016

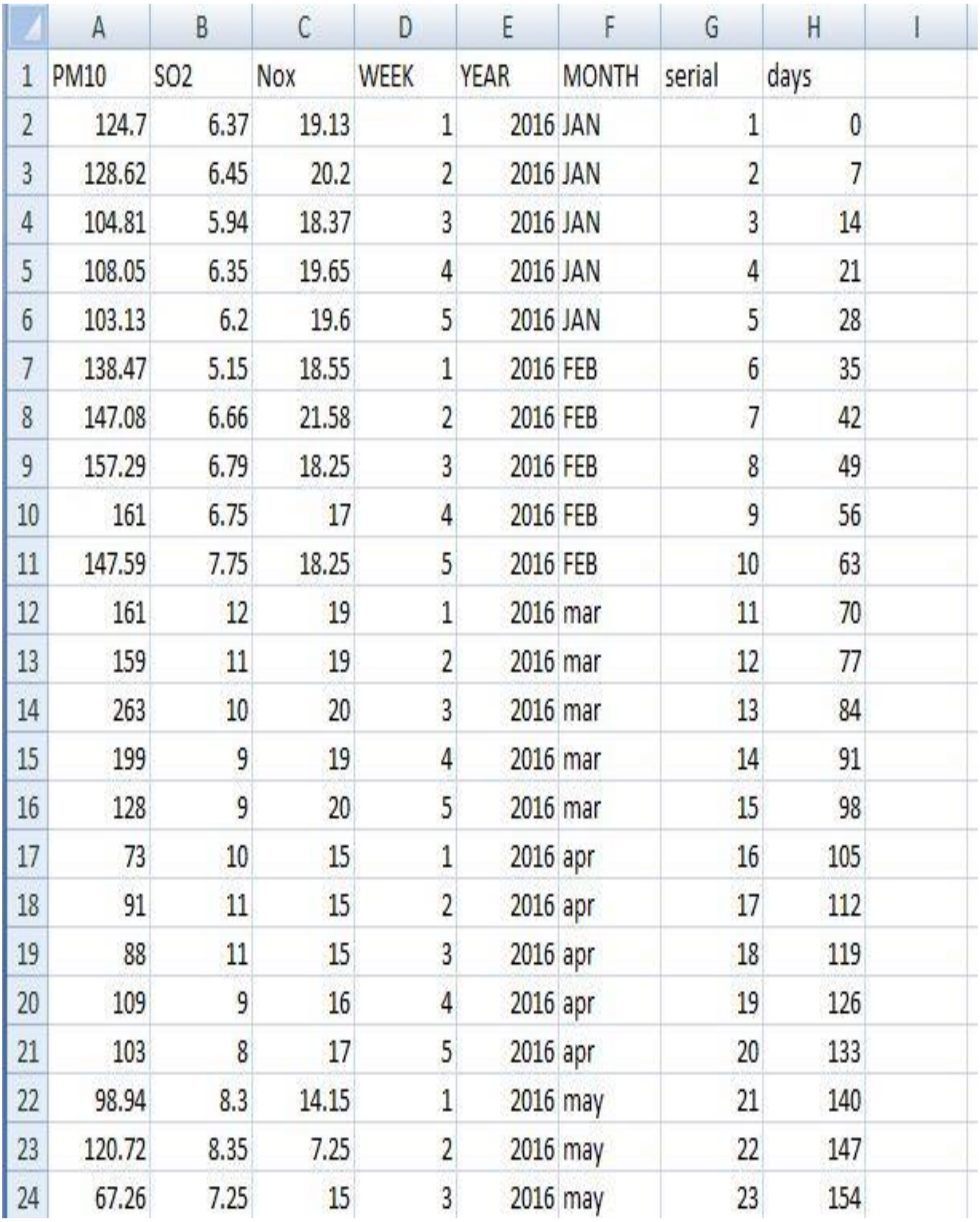
Report no: AAW-15/16

Date: 26.09.2016

NAMP Station	First week 01.08.2016 to 06.08.2016			Second week 08.08.2016 to 12.08.2016			Third week 16.08.2016 to 20.08.2016			Fourth week 23.08.2016 to 26.08.2016			Fifth week 29.08.2016 to 31.08.2016		
	PM_{10} (µg/m³)	SO_2 (µg/m³)	NO_x (µg/m³)	PM_{10} (µg/m³)	SO_2 (µg/m³)	NO_x (µg/m³)	PM_{10} (µg/m³)	SO_2 (µg/m³)	NO_x (µg/m³)	PM_{10} (µg/m³)	SO_2 (µg/m³)	NO_x (µg/m³)	PM_{10} (µg/m³)	SO_2 (µg/m³)	NO_x (µg/m³)
Head Office, Bamunimaidam, Guwahati-21	58.50	6.21	15.46	56.50	7.00	16.85	52.80	6.83	17.22	58.17	6.50	16.58	52.67	6.50	16.00
Khanapara Guwahati-22	88.50	7.17	14.88	89.90	7.00	16.05	89.70	7.77	17.43	89.17	7.25	17.25	92.67	7.67	15.25
Boragaon Guwahati-34	60.90	6.25	16.05	55.10	5.55	16.55	52.13	6.75	16.19	55.67	6.00	14.25	63.00	6.33	14.50
ITI, Gopinathnagar Guwahati-16	74.17	7.13	16.00	72.00	6.85	16.20	69.50	6.90	16.45	80.33	7.50	16.42	74.50	6.92	15.42
Guwahati University Guwahati-14	84.90	7.60	17.10	76.60	7.75	16.65	73.75	8.25	18.13	84.00	8.08	17.33	77.50	7.83	16.83
Pragjyotish College, Santipur Guwahati-9	60.58	6.29	16.04	59.50	6.45	15.90	57.80	6.20	15.55	67.17	6.83	16.42	61.00	5.50	16.00
Average value for Guwahati City	71.26	6.78	15.92	68.27	6.77	16.37	65.95	7.12	16.83	72.42	7.03	16.38	70.22	6.79	15.67

Chief Env. Scientist

Table 1.1



	A	B	C	D	E	F	G	H	I
1	PM10	SO2	Nox	WEEK	YEAR	MONTH	serial	days	
2	124.7	6.37	19.13	1	2016	JAN	1	0	
3	128.62	6.45	20.2	2	2016	JAN	2	7	
4	104.81	5.94	18.37	3	2016	JAN	3	14	
5	108.05	6.35	19.65	4	2016	JAN	4	21	
6	103.13	6.2	19.6	5	2016	JAN	5	28	
7	138.47	5.15	18.55	1	2016	FEB	6	35	
8	147.08	6.66	21.58	2	2016	FEB	7	42	
9	157.29	6.79	18.25	3	2016	FEB	8	49	
10	161	6.75	17	4	2016	FEB	9	56	
11	147.59	7.75	18.25	5	2016	FEB	10	63	
12	161	12	19	1	2016	mar	11	70	
13	159	11	19	2	2016	mar	12	77	
14	263	10	20	3	2016	mar	13	84	
15	199	9	19	4	2016	mar	14	91	
16	128	9	20	5	2016	mar	15	98	
17	73	10	15	1	2016	apr	16	105	
18	91	11	15	2	2016	apr	17	112	
19	88	11	15	3	2016	apr	18	119	
20	109	9	16	4	2016	apr	19	126	
21	103	8	17	5	2016	apr	20	133	
22	98.94	8.3	14.15	1	2016	may	21	140	
23	120.72	8.35	7.25	2	2016	may	22	147	
24	67.26	7.25	15	3	2016	may	23	154	

Table 1.2

2.2.2 IMPORTING THE LIBRARIES

We used the python language for preprocessing of the data. One of the main advantages of using python language is that there are already predefined libraries present which we can use to pre-process the data. There are three specific libraries that we use for our data. They are

Numpy :Numpy library is use to solve any type of mathematical operation and scientific calculation in python. We can also create multidimensional arrays and matrices. We use this library while plotting independent variable vs dependent variable graph. We create a number of days matrices which has a series starting from lowest value in the independent variable to the highest value where the value increases by 0.01. this is then used by the machine learning model as an input and we get our predicted output. This is then plotted in a graph. This is used to view the predicted curve of our model in much more higher definition.

Mathplotlib: this is a 2d-plotting library. We mainly use its sub library pyplot to plot charts. We first create a scatter plot using the dependent and independent variables as data points and then plot the predicted curve over the scatter plot. We plot both for predicted values and actual values to compare the graph. There are 4 types of graphs that has been created for both actual and predicted values. They are:

a. Graph for the predicted curve for a single pollutant in a specific station. There will be 18 graphs in total
b. Graph for the actual curve for a single pollutant in a specific station. There will be 18 graphs in total
c. Graph for the predicted curve of a single pollutant for all the stations. There will be 3 graphs in total
d. Graph for the actual curve of a single pollutant for all the stations. There will be 3 graphs in total
e. Graph for the predicted curve of all the pollutants for a single stations. There will be 6 graphs in total
f. Graph for the actual curve of all the pollutants for a single stations. There will be 6 graphs in total

Panda: the panda library is used to import and manage the datasets. We use the panda library to extract the data from the dataset and also creating the dependent and independent variable.

2.2.3 IMPORTING THE DATASETS

We have total of 6 CSV files. We then use the panda library to extract the data from the CSV file to an appropriate dataset. There will be six datasets in total, one for each location. We then use the iloc function of the panda library to create the independent and dependent variable. The independent variable here will be the number of day's column. This is will same for all six locations. The independent variables will be the pollutant columns. We extract each pollutant in different variables that is one variable consist of one pollutant value. Thus we analyze each pollutant separately and not all three as a whole.

2.2.4 HANDLING MISSING DATA

The next step in data preprocessing is the handling of missing data. Often in our datasets there will be bound to be missing data which creates error for the machine learning model. Thus it is crucial to handle the missing data. There are mainly two ways of handling missing data. The first way is to completely deleting the row consisting of missing data. This method of handling null values is not how one should be tackling this problem as it is can lead to data loss resulting in an output with poor accuracy. Also we only have one row in our variable, so deleting it will lead to no variable at all. Thus we are going for the second method which is to calculate the mean of the entire column and then replace all the null values with the mean value. This method is especially useful for variables having numeric data type. Our independent variable does not contain any missing values. Only our dependent variable has missing values. We use the imputer function from the preprocessing sub library of scikit-learn.

2.2.5 ENCODING CATEGORICAL DATA

Categorical data are non-numeric data which consists of categories such as name of countries and yes or no. Machine learning model only works on numeric data and cannot understand categorical data. Thus we have to convert the categorical data to a suitable numerical format. This is mainly down by using the label encoder. Since our dataset do not consist of any categorical data we skip this step and proceed to the next data preprocessing step.

2.2.6 SPLLITING INTO TRAINING AND TEST SET

This is one of the most crucial step in data pre-processing as it can determine the accuracy of the model. We spilt our dataset into two sets. One is the training set and other is the test set. The training set is used first by the model. Here the input and the output is both known to the model and the model analyze the correlation between the two variables and builds a relation. Then we used the test set where the input is known and the model tries to predict the output. We then compare the predicted and actual values in order to determine the performance of the model. when we train a model with a dataset and then bring in new dataset the model will find it harder to understand the correlation between the variables. If the training set is too much the model tends to overfit leading to bad performance for other datasets. Thus we must find that perfect balance of data in training and test set such that we get desirable output for both sets. Here we split the data using the train_test_split function that takes in four arguments that is the X and Y variable, the test size and random_state. Here we use 80% of the data as training data and rest 20% as test data. If we input a number for random_state then everytime the program is run the result will always be the same else the result will vary. We use the integer 42 as input for random_state.

2.2.7 FEATURE SCALING

Feature scaling is use to standardize the independent variables to a specific range. This is the final step of data preprocessing. Machine learning models are models mostly based on Euclidean distance. If a independent variable has values very low as compared to other variables then this independent variable can almost be considered as null values as it won’t have any effect and thus will not be considered in the final computation. This can lead to issues in the model. Thus feature scaling is use to scale the values to a specific range. StandardScalar is mostly use for feature scaling. Since our data has only one independent variable we do not need feature scaling. Thus our data preprocessing is done. Now we use data in the four regression models and then compare the accuracy of each models to find the best suited model.

2.3 COMPARING DIFFERENT MODELS

We build four models for our air quality index prediction and compared the results.

2.3.1 LINEAR REGRESSION MODEL

Linear regression is one of the most basic regression model. The model consist of independent variable being linearly related to the dependent variables. A linear regression model with more than one independent variables is called a multi regression model. The equation for the linear regression model is y=mx+c where m is the slope of the line and c is the intercept.

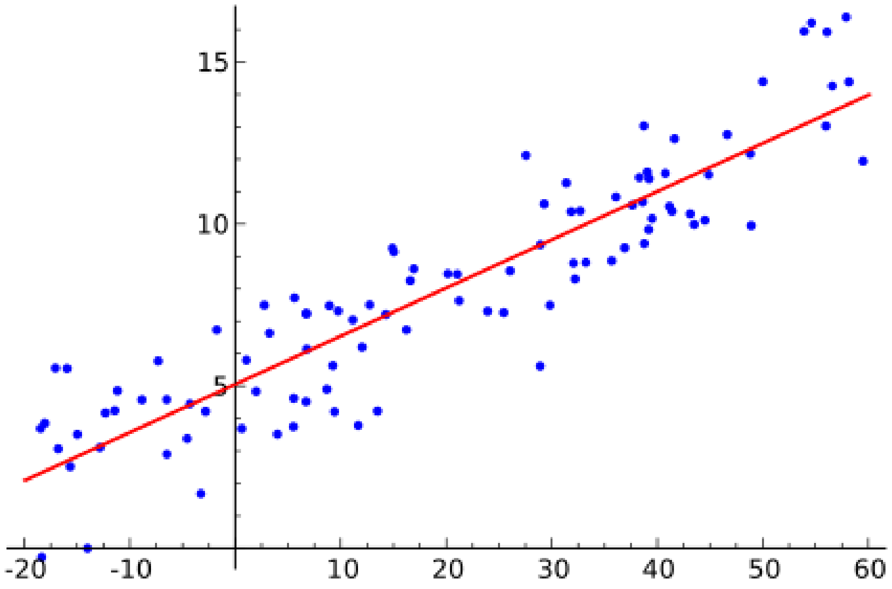
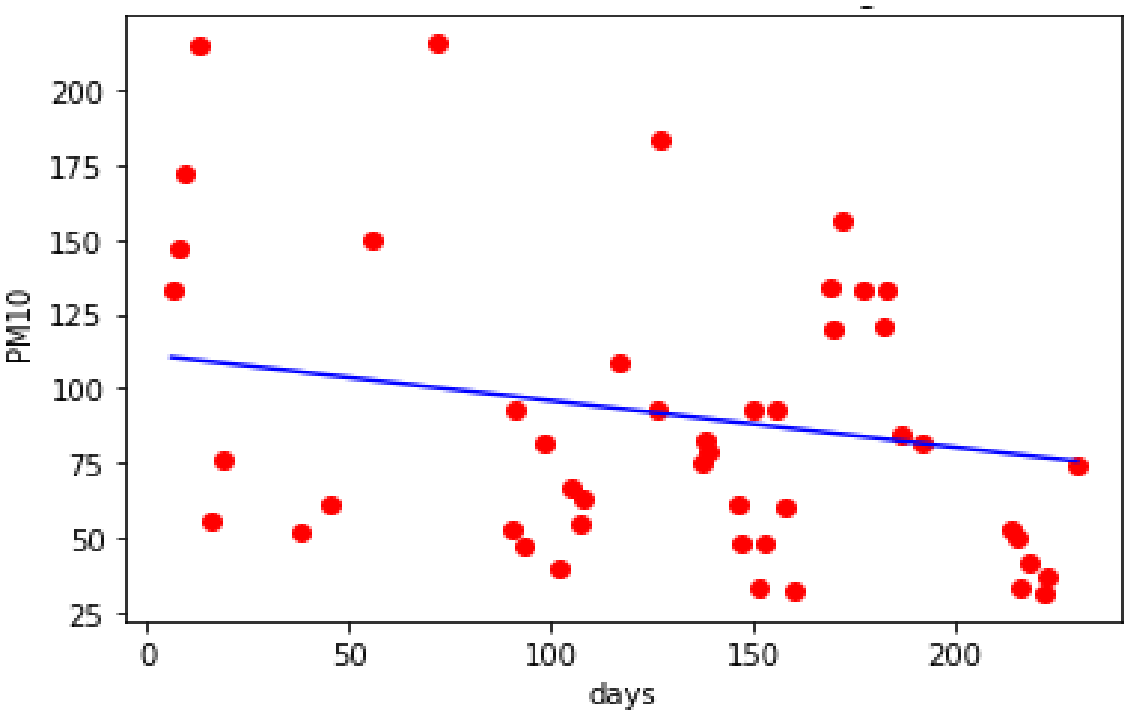


Fig 2.1

The best fit line is found by altering the slope and intercept value. The best fit line would be the line where the total distance error of predicted and actual value is minimum. However the linear regression model is underfit for the PM10 dataset as the data is very high in variance can thus cannot find a linear relation model between the two variables. It has the worse accuracy compared to other models. For SO2 and NOx although the accuracy is far better than the PM10 data its accuracy is still low compared to other models.


PM10
200
175
150
125
100
75
50
25
0
50
100
150
200
days

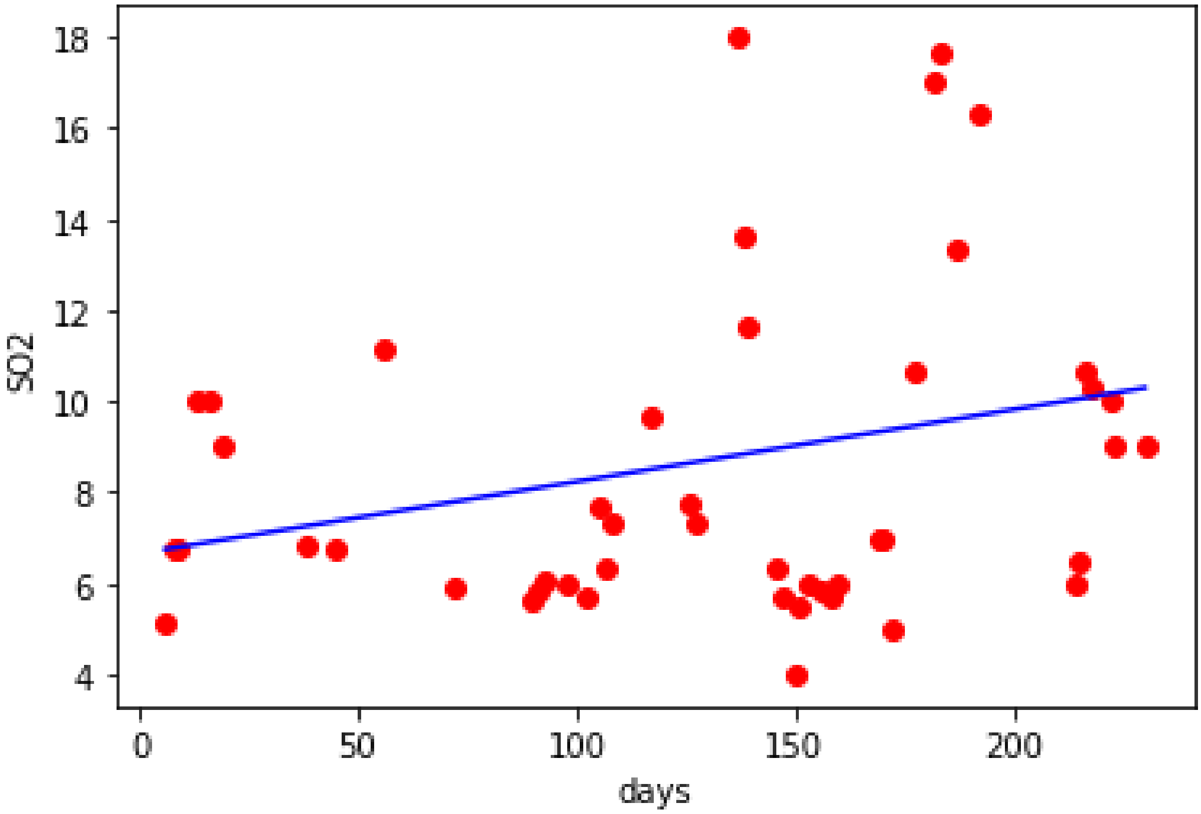

SO2
18
16
14
12
10
8
6
4
0
50
100
150
200
days

Fig 2.2

2.3.3 SUPPORT VECTOR MACHINE

Support vector machine creates a hyperplane in a N-dimensional space where N is the number of features that classifies the data points distinctly. The are many possible ways of creating a hyper plane but the best hyperplane would be the one where the margin is maximum that is the distance between the data points is maximum for both classes.

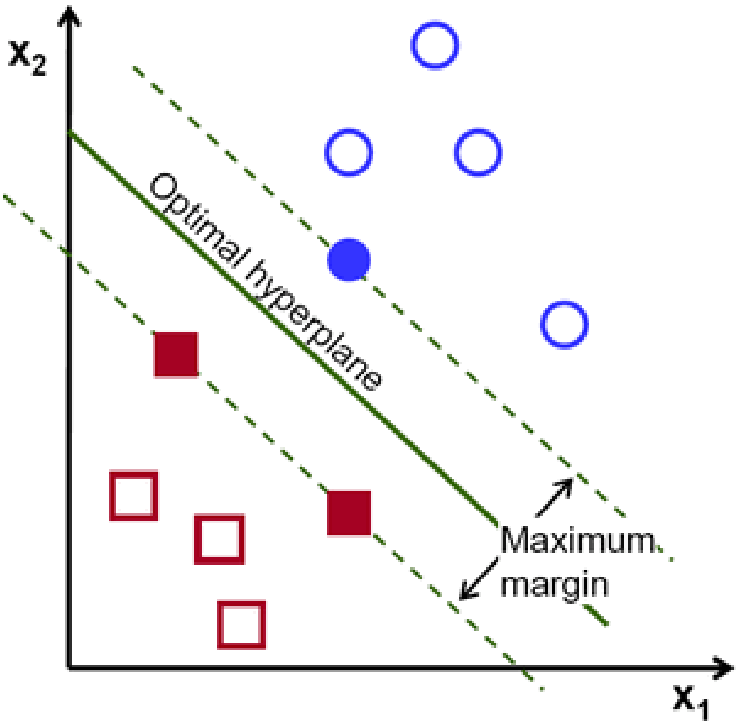


Fig 2.3

Datapoints closer to the hyperplane are called support vectors can influence the position and orientation of the hyperplane. We maximize of the margin of the classifier by using these classifiers.

For our PM10 dataset its accuracy was very low compared to random forest model and decision tree model but was better than linear regression model. One of the reasons for such low accuracy is due to the high variance in data whereas the predicted output was only either of two values. For NOx the accuracy was good but compared to random forest model and decision tree model it was abit low. It was better than linear regression model. For SO2 it is the same result as NO2 however for two location its accuracy was abit higher compared to all the models.

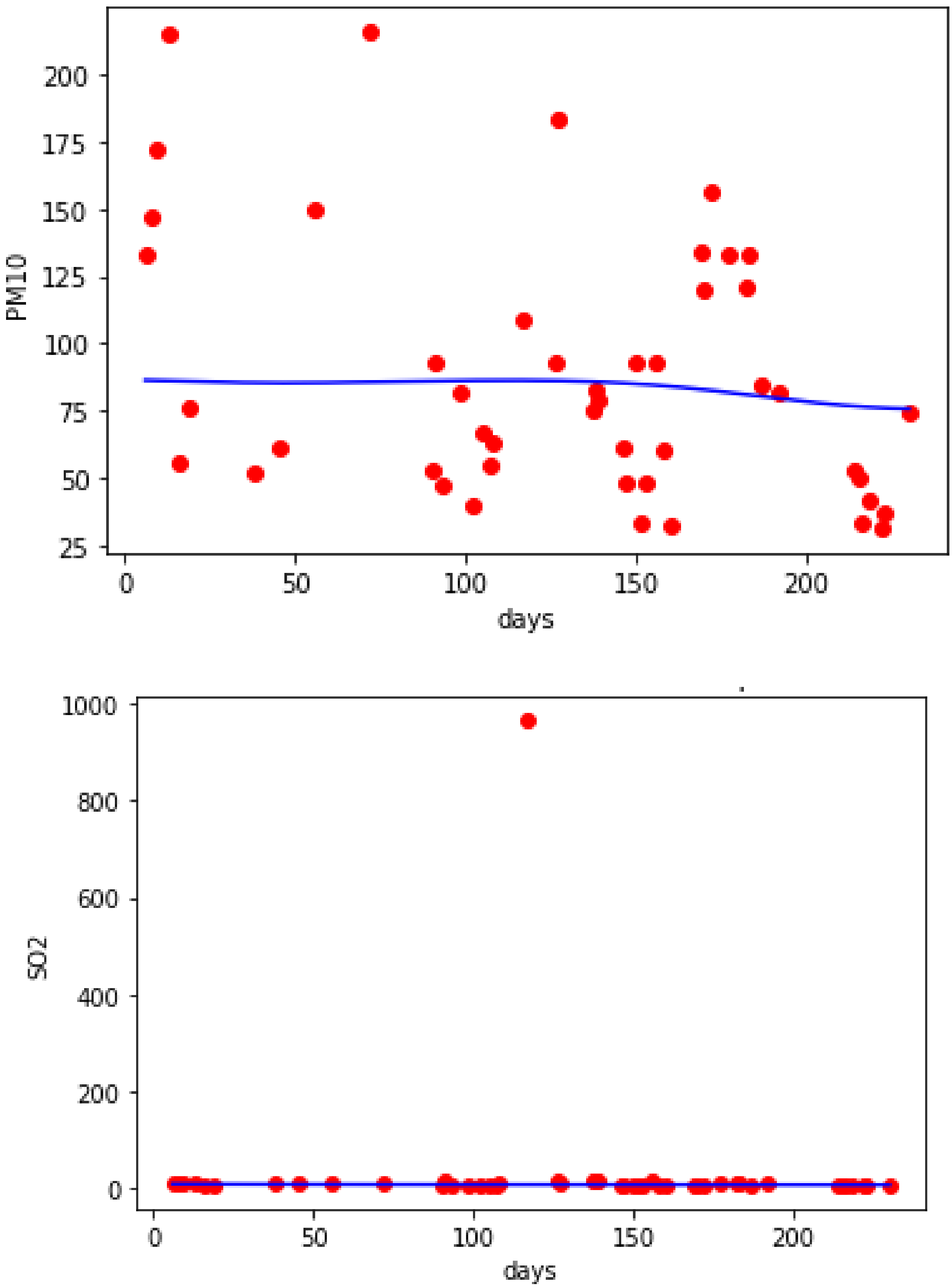

PM10
200
175
150
125
100
75
50
25
0
50
100
150
200
days
SO2
1000
800
600
400
200
0
0
50
100
150
200
days

Fig 2.4

2.3.4 DECISION TREE MODEL

A decision tree classifies data items by posing a series of questions about the features associated with the items. Each question is contained in a node, and every internal node points to one child node for each possible answer to its question. The questions thereby form a hierarchy, encoded as a tree. In the simplest form we ask yes-or-no questions, and each internal node has a 'yes' child and a 'no' child. An item is sorted into a class by following the path from the topmost node, the root, to a node without children, a leaf, according to the answers that apply to the item under consideration. An item is assigned to the class that has been associated with the leaf it reaches [7].

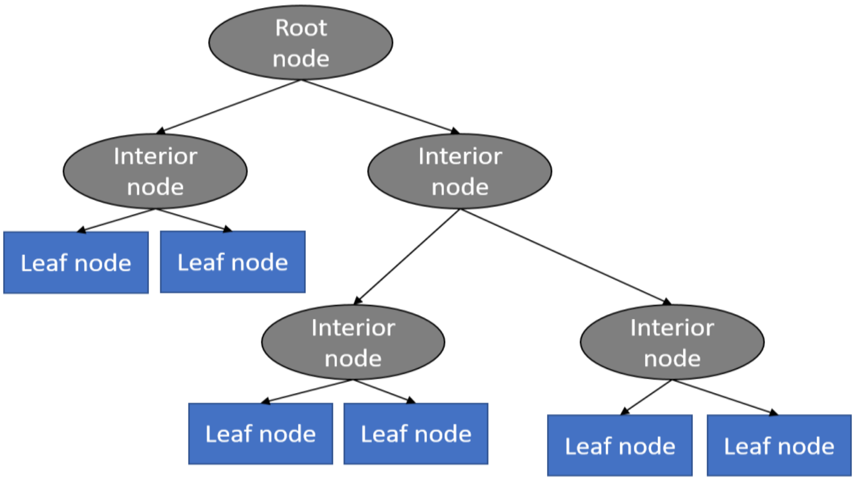


Fig 2.5

For PM10 data the decision tree model has the best accuracy out of the three models and alil bit higher than random forest. For NOx also decision tree has the highest accuracy. However for SO2 support vector machine has better accuracy for two location but other than that it has the highest accuracy.

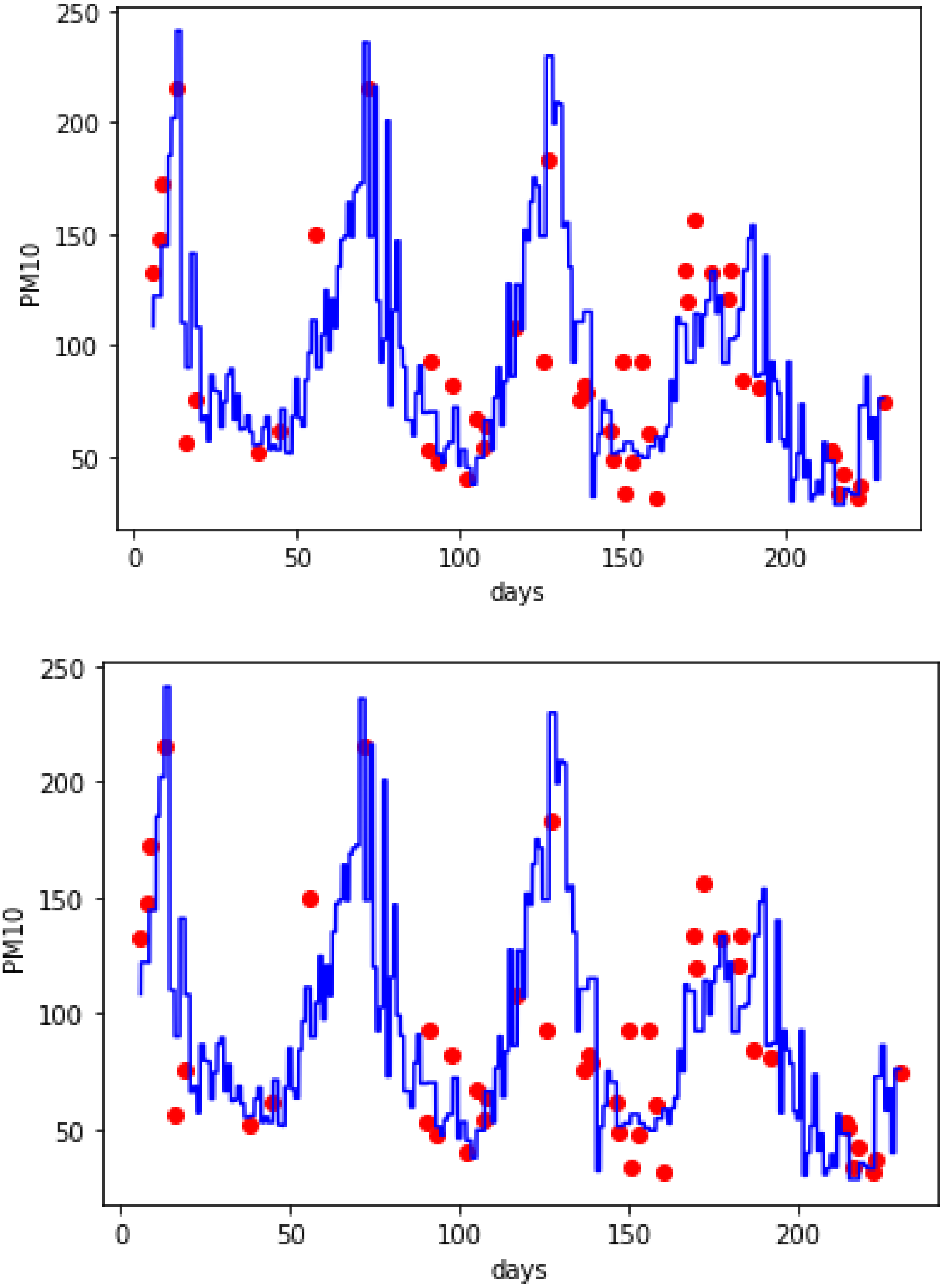

PM10
250
200
150
100
50
0
50
100
150
200
days
PM10
250
200
150
100
50
0
50
100
150
200
days

Fig 2.6

2.3.2 RANDOM FOREST

In Random forest classifier the dataset is divided into multiple sets and for each set a decision tree classifier works on that subset. Then the results are taken from each decision tree and the final output is then predicted based on majority votes of prediction. It solves the overfitting problem that decision tree model tend to have.

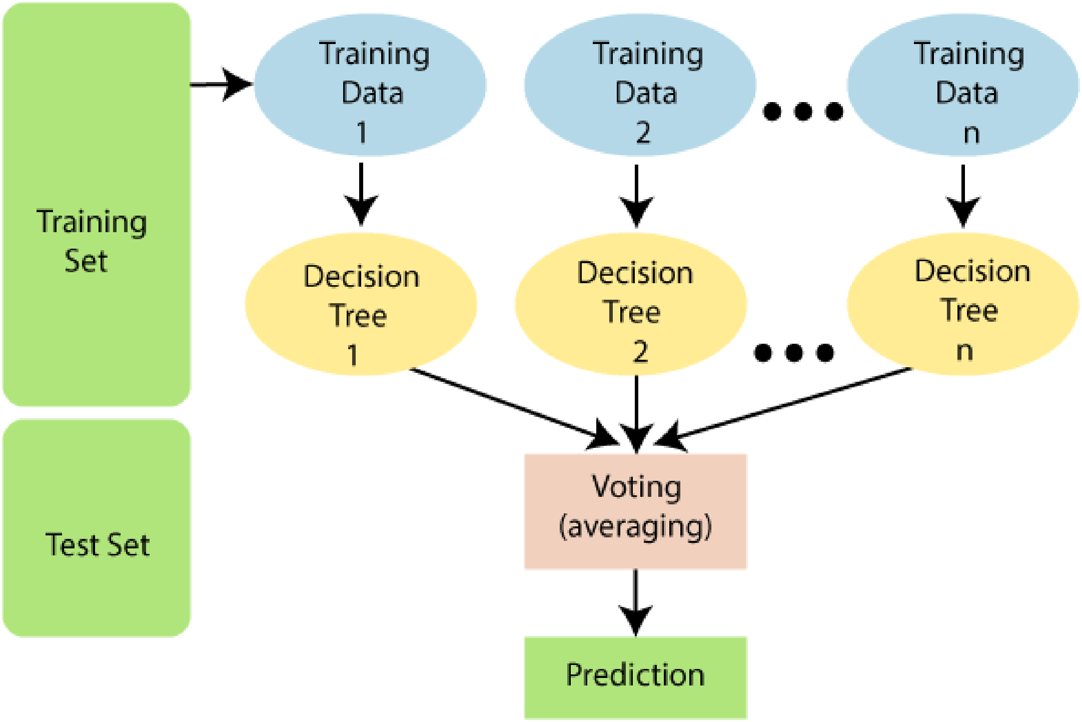


Fig 2.7

For random forest it has better accuracy for PM10 and NO2 compared to support vector machine and linear regression. It s accuracy is slightly lower than decision tree model. For SO2 also its accuracy is better than linear regression and support vector machine however for 2 locations support vector was better than random forest and also its accuracy is lower than decision tree also.

Thus after comparing the average accuracy of the four models it is found that the model are rank in the order decision tree > random forest > support vector machine > linear regression. Thus we choose decision tree model as our machine learning model for air quality index prediction.

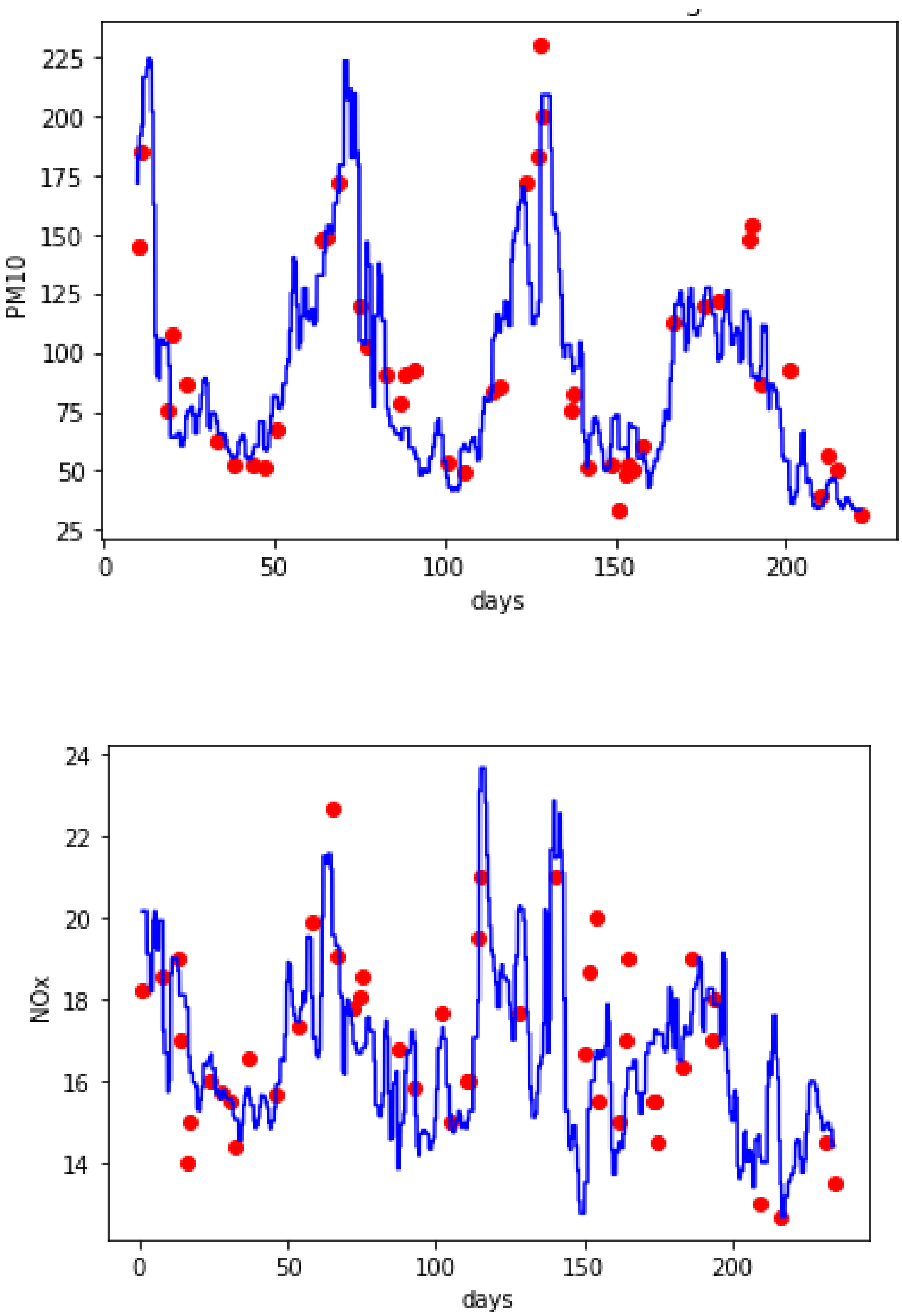

PM10
225
200
175
150
125
100
75
50
25
0
50
100
150
200
days
NOx
24
22
20
18
16
14
0
50
100
150
200
days

Fig 2.8

2.4 WORKING MODEL

The number of day's column is taken as X input variable. Taking 2015 as base year the column will start from value 0 and will increase by 7 since weekly AQI data is taken (1 week = 7 days). The pollutant value is taken as Y output variable. The dataset has been split into 80% training set and 20% test set. The decision tree model is then trained using this 80% training set with the number of days and its corresponding air quality index value is taken as input. After training our model the 20% test set is introduce as a new data set to the model where the number of day's is the input and the air quality index of the pollutant is the output. After prediction is complete we plot the predicted and actual values and compare their graphs. We plot two types of graph

1. 1. The AQI value of a single pollutant for all locations in one graph.
2. 2. The AQI value of all the pollutants in one location.

Also we build another model where date is taken as input. The date is then subtracted by the base date which is 2015 and is converted to number of days which is use as input by the model to predict the pollutant value.

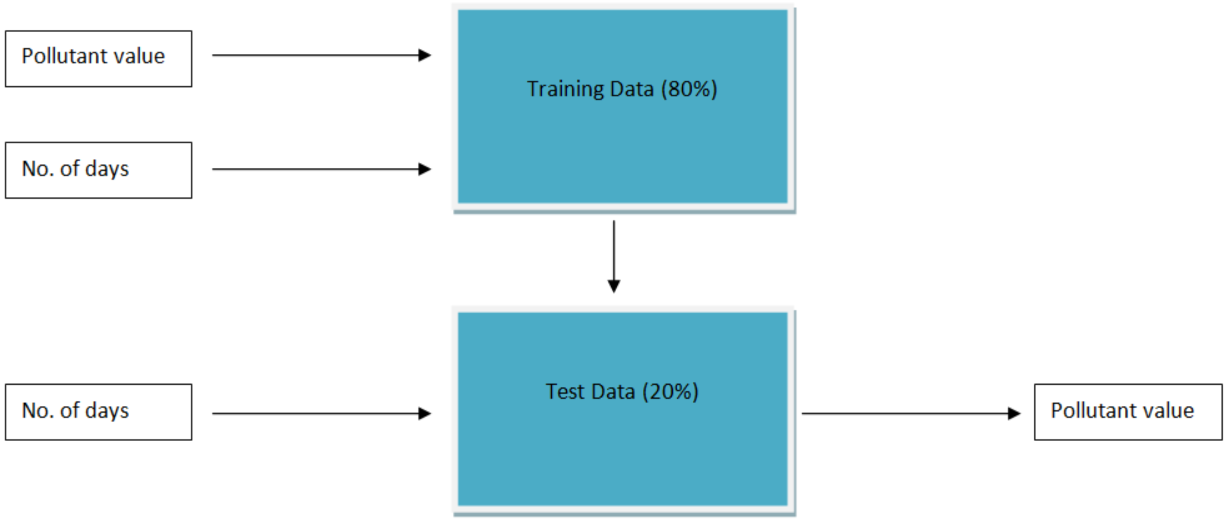


Fig 2.9.1

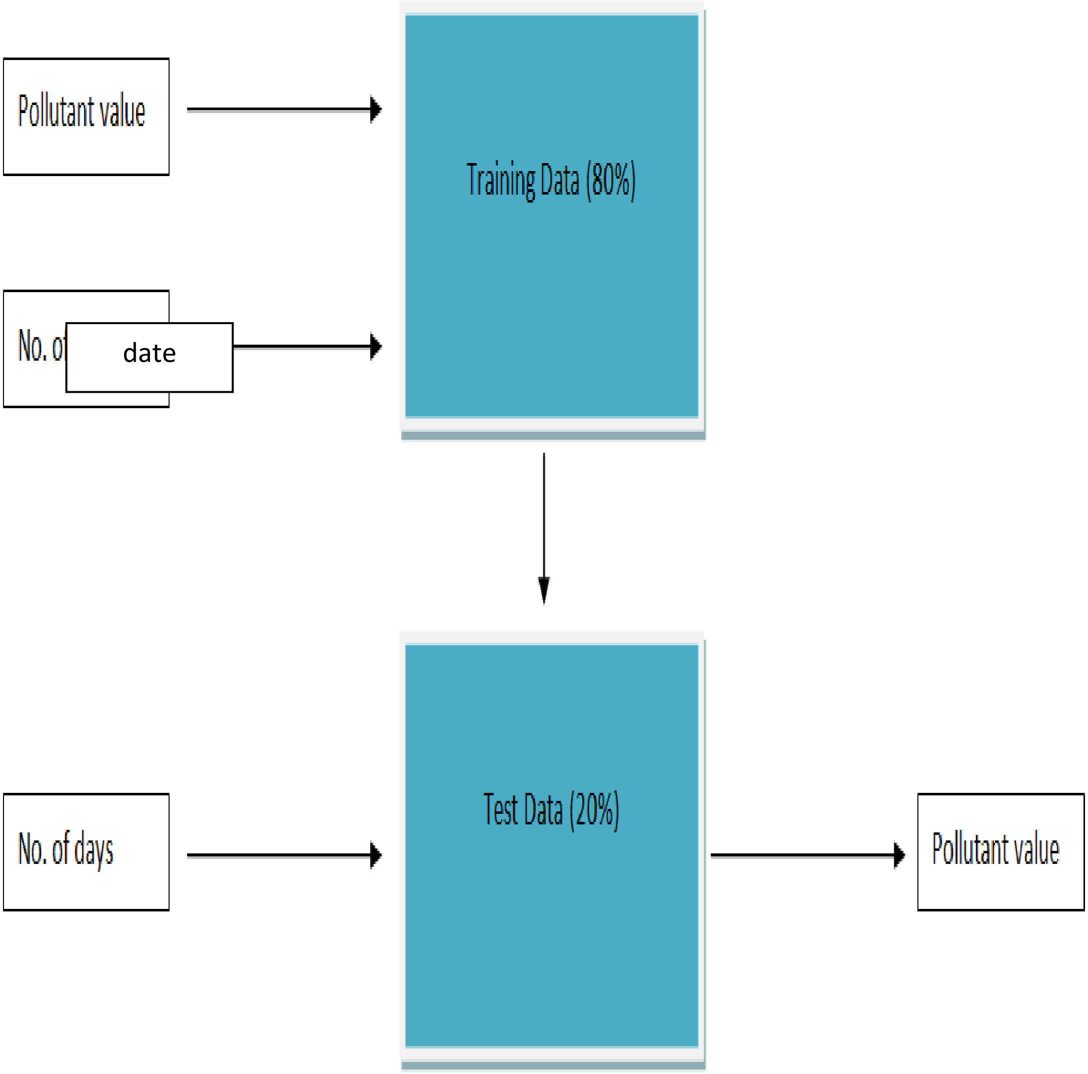

Pollutant value
No. of
date
Training Data (80%)
Test Data (20%)
No. of days
Pollutant value

Fig 2.9.2

CHAPTER 3: RESULTS AND DISCUSSION

Two types of graphs were plotted -

1. The AQI value of a single pollutant for all locations in one graph.

2. The AQI value of all the pollutants in one location.

3. The pollutants were coloured as follows

- PM10-Red
- SO2-Blue
- NOx-Black

Comparison has also been done between actual and predicted values.

Variation of the AQI parameters

The matplotlib library of python was used to plot the pollutants against the days the pollutants was measured.Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.Matplotlib was originally written by John D. Hunter. Python 3 support started with Matplotlib 1.2. Matplotlib 1.4 is the last version to support Python 2.6[8]

In the above graph we can see that all the pollutants have been plotted for the station Khanapara. The pollutants have been plotted in the y-axis against the days the pollutants have been measured , plotted in the x-axis. The pollutants have been labelled as follows: PM10 has been represented with orange colour, SO2 has been represented by blue and Nox has been represented with black colour.

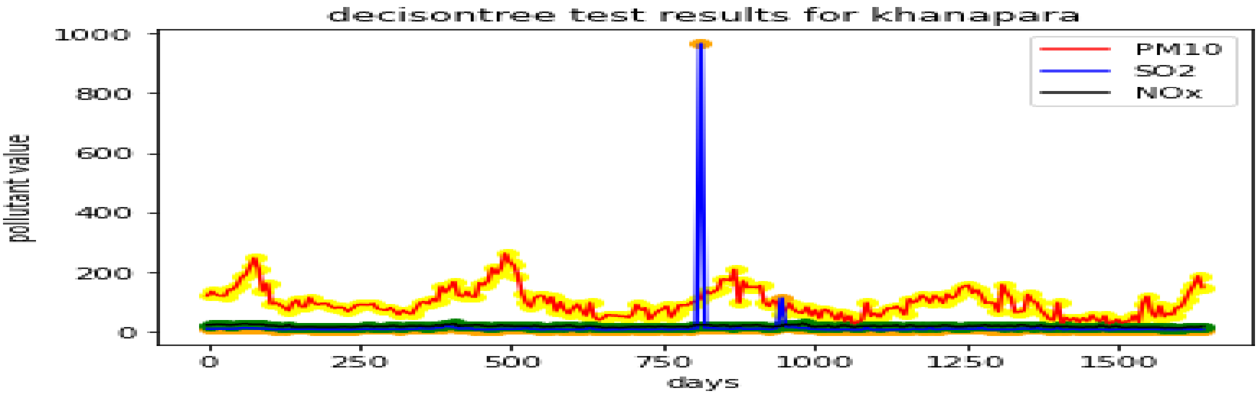


Figure 3.1

Dealing with Anomalies

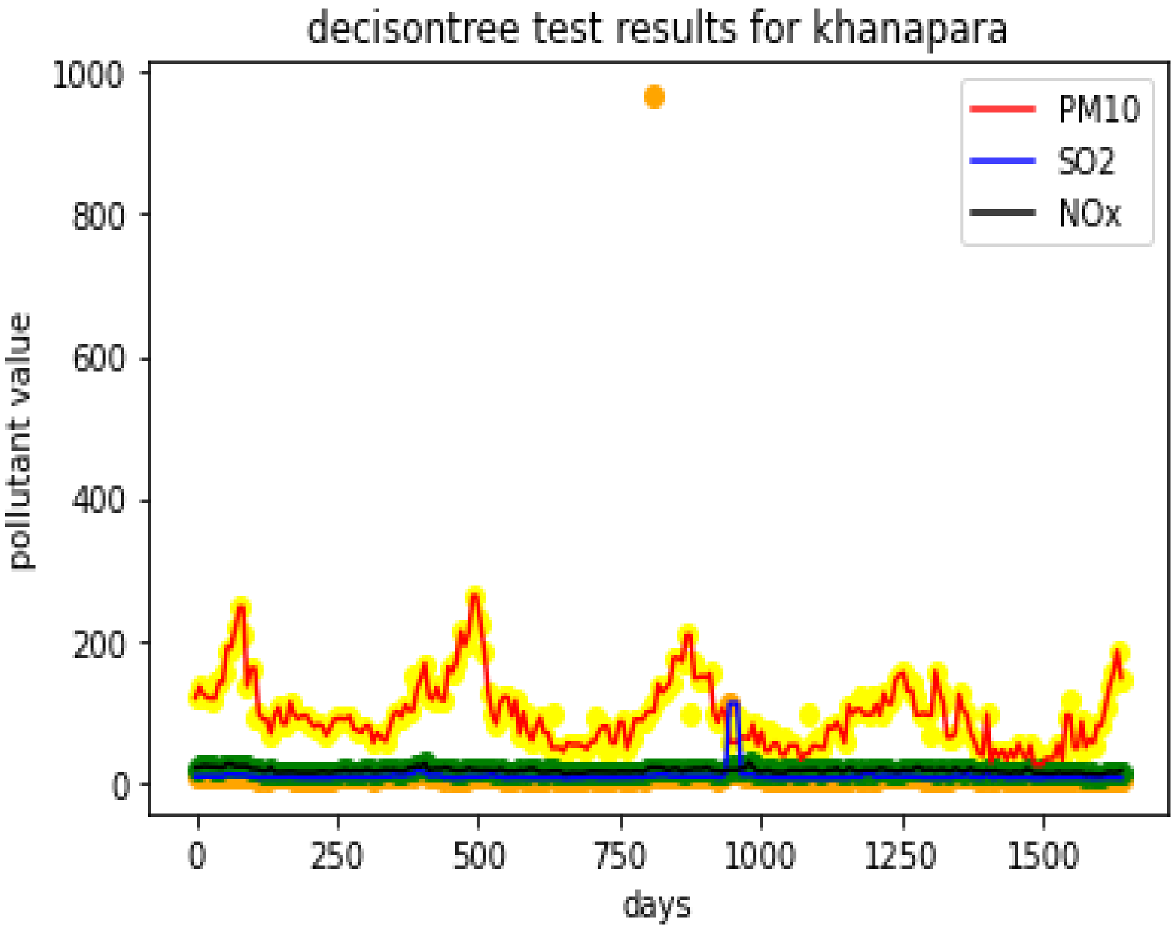


Figure 3.2

In the above graph ,we can also see an abnormal spike in the SO2 value for a day. It has occurred due to an error in the sensor at that location. We chose to take that anomaly because we wanted to keep the collected data as much unaltered as possible. In the prediction from our model , the value was seen to be balanced. The results for predicted values for the khanapara are given below

3.1 COMPARISON OF VALUES ACTUAL VS PREDICTED IN ALL THE STATIONS OF GUWAHATI

- Total six number of graphs were plotted. The graphs were plotted for the three pollutants separately and included all the stations' data.
- The graphs were generated using the matplotlib library of Python.
- Each graph contains the data of a single pollutant for a period of about 300 day.
- The pollutant measuring stations were labelled as following colors:
  - Borgaon-Red

- Pragjyotish-Blue
- Bamunimaidan-Green
- Guwahati University-Yellow
- Khanapara-Black
- Gopinathnagar-Violet

1.For pollutant Nox

Decision tree graph results for Actual measured values of pollutant NOx

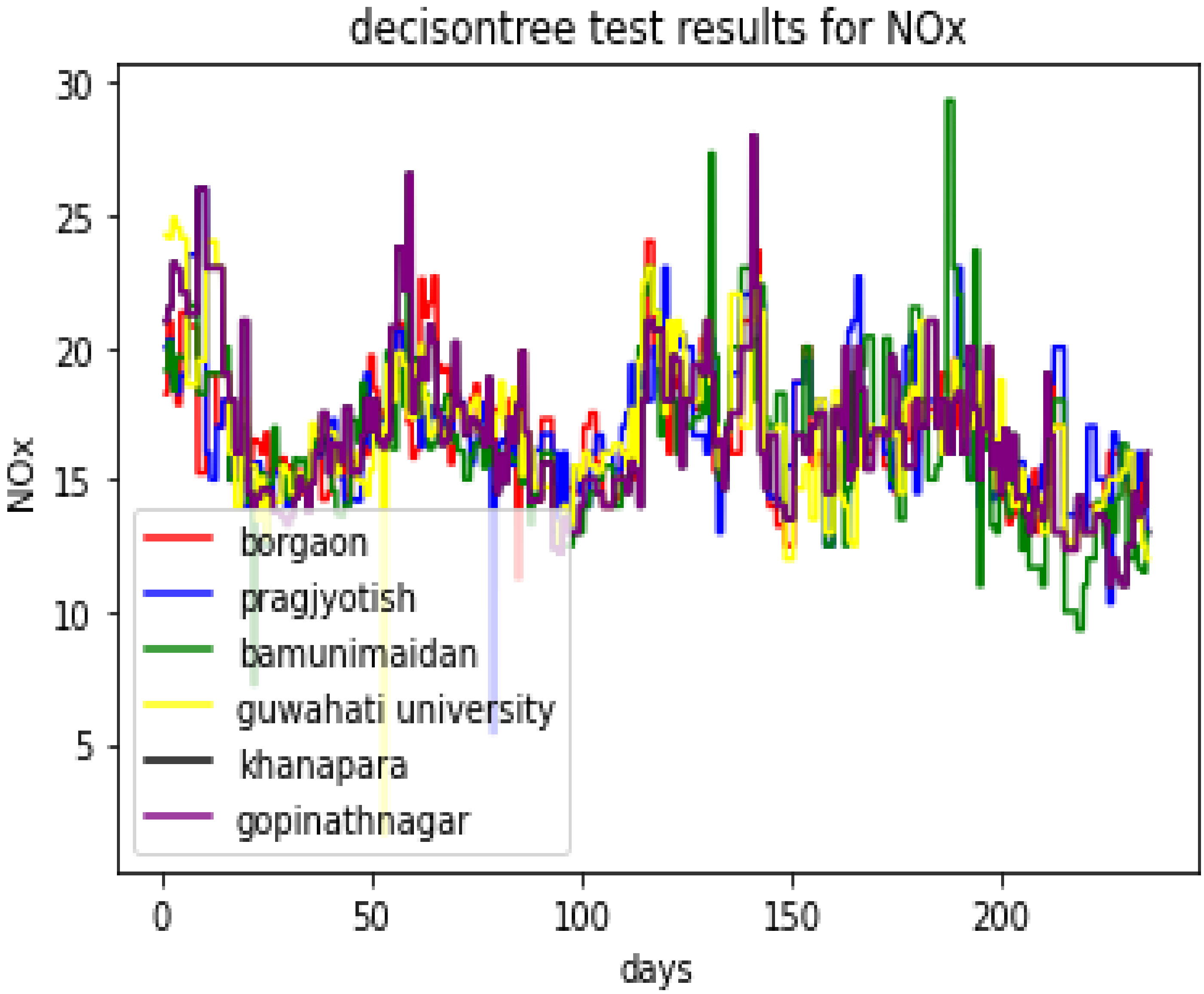


Figure 1.3

In this graph, the pollutant NOx has been plotted. The x-axis represents the days on which the data was obtained from the sensors from each location i.e. Borgaon, Pragjyotish,Bamunimaidan,GuwahatiUniversity,Khanapara, Gopinathnagar. The y-axis contains the NOx value obtained from the sensors at each location.

Decision tree graph results for predicted values of pollutant NOx

In this graph, the pollutant NOx has been plotted. The x-axis represents the days on which the data

was obtained from the sensors from each location i.e. Borgaon, Pragjyotish,Bamunimaidan,GuwahatiUniversity,Khanapara, Gopinathnagar. The y-axis contains the NOx value obtained from the predicted values of the decision tree.

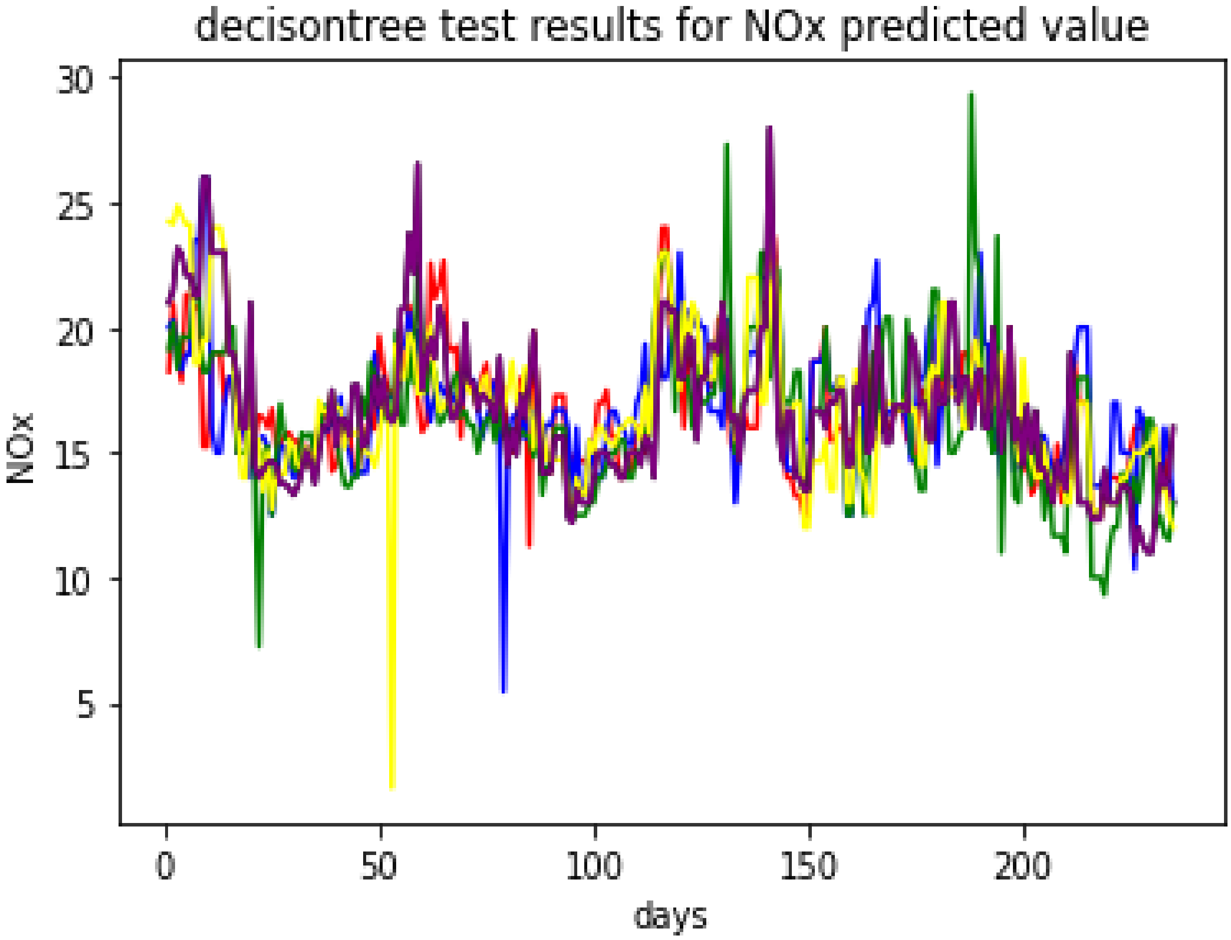


Figure 3.4

2.For Pollutant PM10

Decision tree graph results for Actual measured values of pollutant PM10.In this graph, the pollutant PM10 have been plotted. The x-axis represents the days on which the data was obtained from the

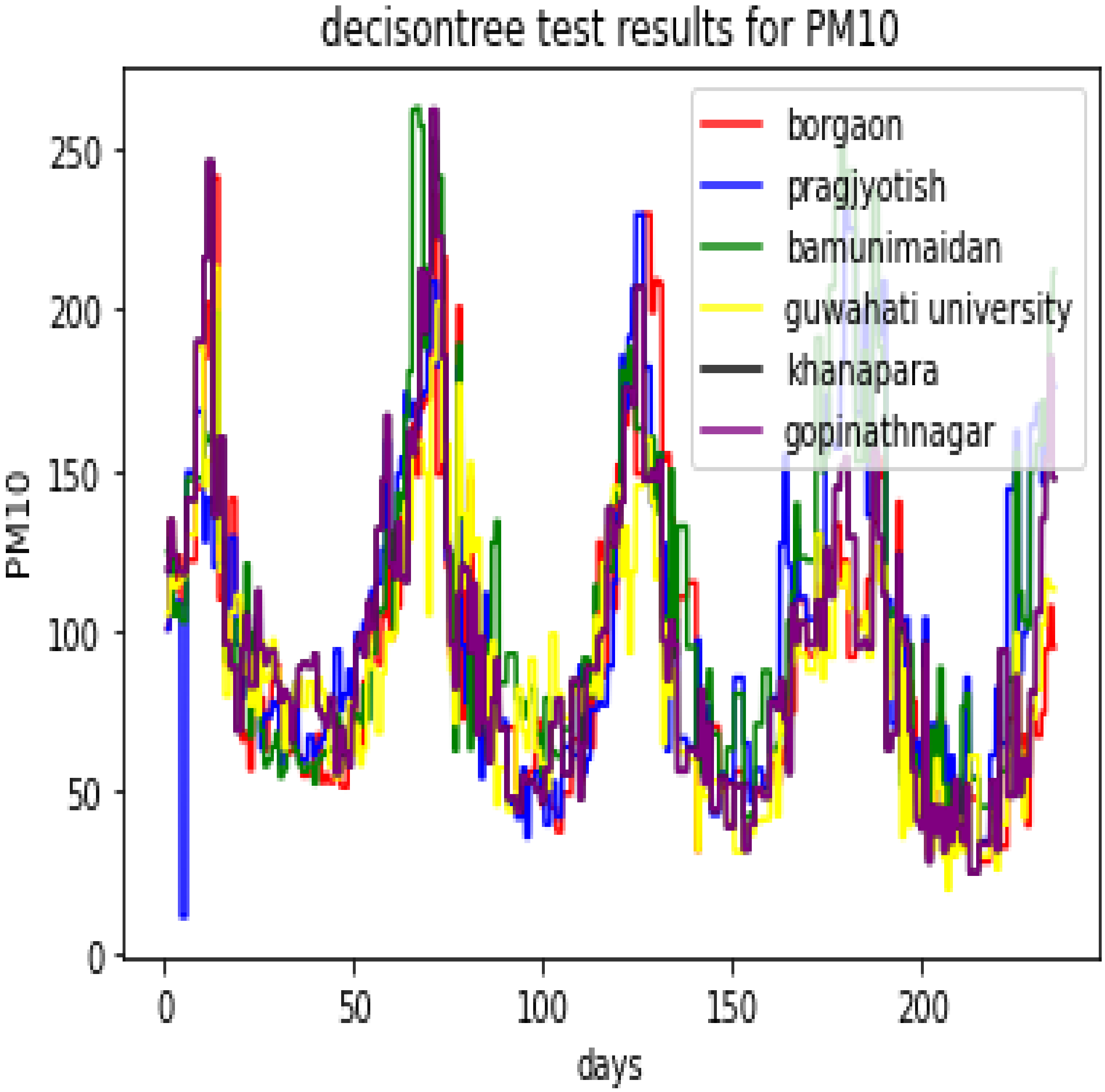


Figure 3.5

sensors from each location i.e. Borgaon, Pragjyotish,Bamunimaidan,GuwahatiUniversity,Khanapara, Gopinathnagar. The y-axis contains the PM10 value obtained from the sensors at each location.

In this graph, the pollutant PM10 has been plotted. The x-axis represents the days on which the data was obtained from the sensors from each location i.e. Borgaon, Pragjyotish,Bamunimaidan,GuwahatiUniversity,Khanapara, Gopinathnagar. The y-axis contains the PM10 value obtained from the predicted values of the decision tree.

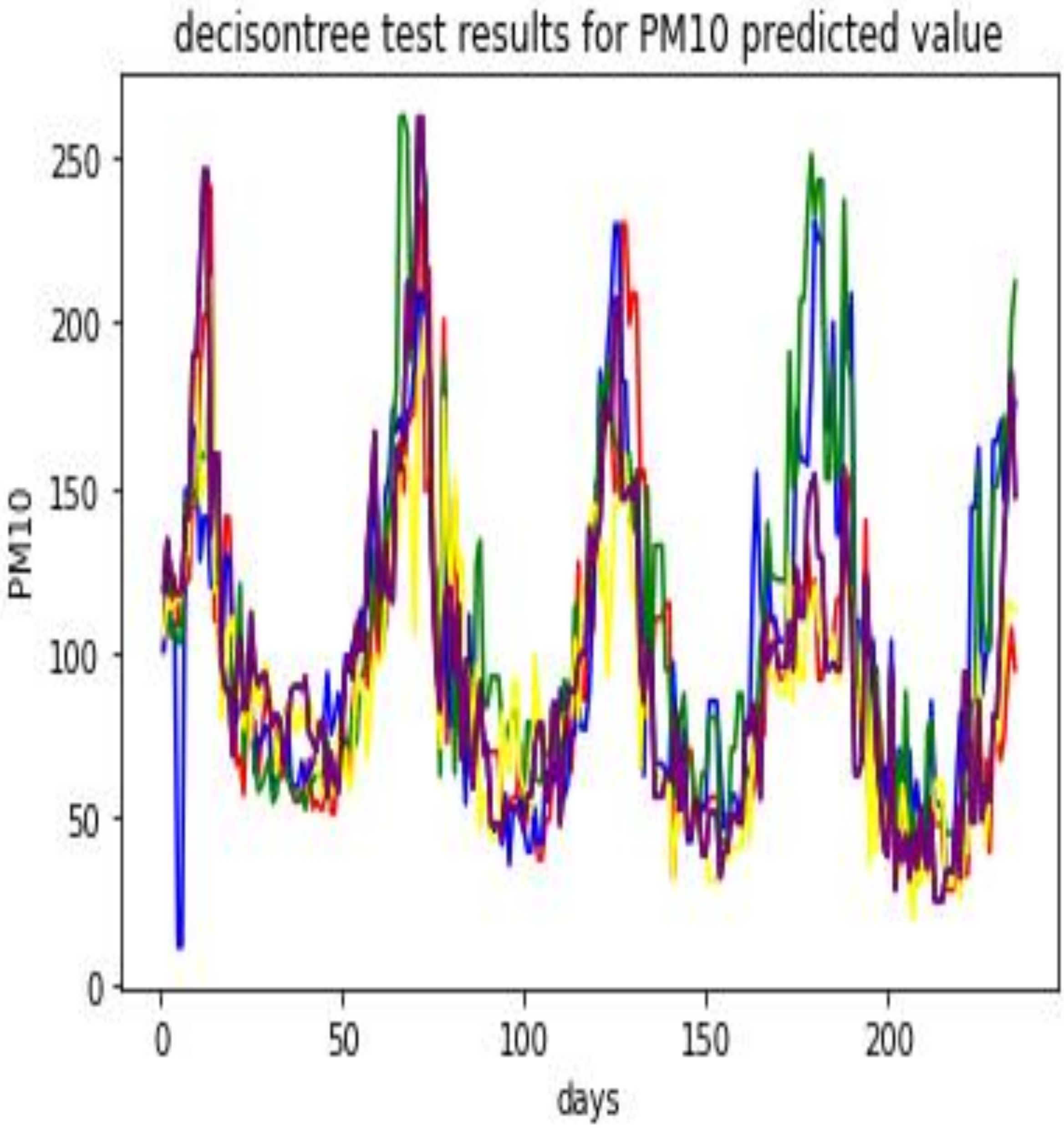


Figure 3.6

3.For pollutant SO2

Decision tree graph results for Actual measured values of pollutant SO2

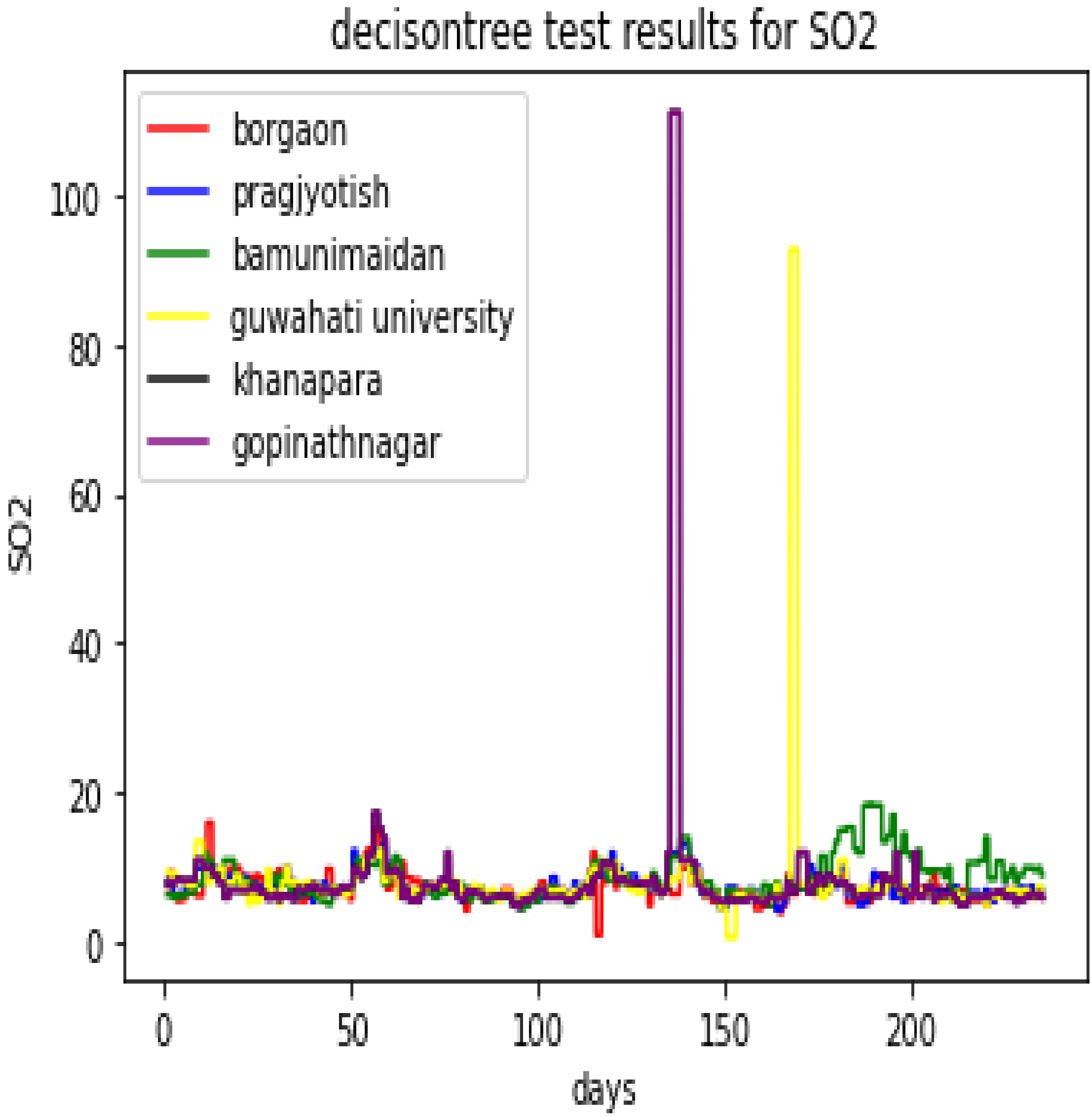


Figure 3.7

In this graph, the pollutant SO2 has been plotted. The x-axis represents the days on which the data was obtained from the sensors from each location i.e. Borgaon, Pragjyotish,Bamunimaidan,GuwahatiUniversity,Khanapara, Gopinathnagar. The y-axis contains the SO2 value obtained from the sensors at each location.

Decision tree graph results for predicted values of pollutant SO2

In this graph, the pollutant SO2 has been plotted. The x-axis represents the days on which the data

was obtained from the sensors from each location i.e. Borgaon, Pragjyotish,Bamunimaidan,GuwahatiUniversity,Khanapara, Gopinathnagar. The y-axis contains the SO2 value obtained from the predicted values of the decision tree.

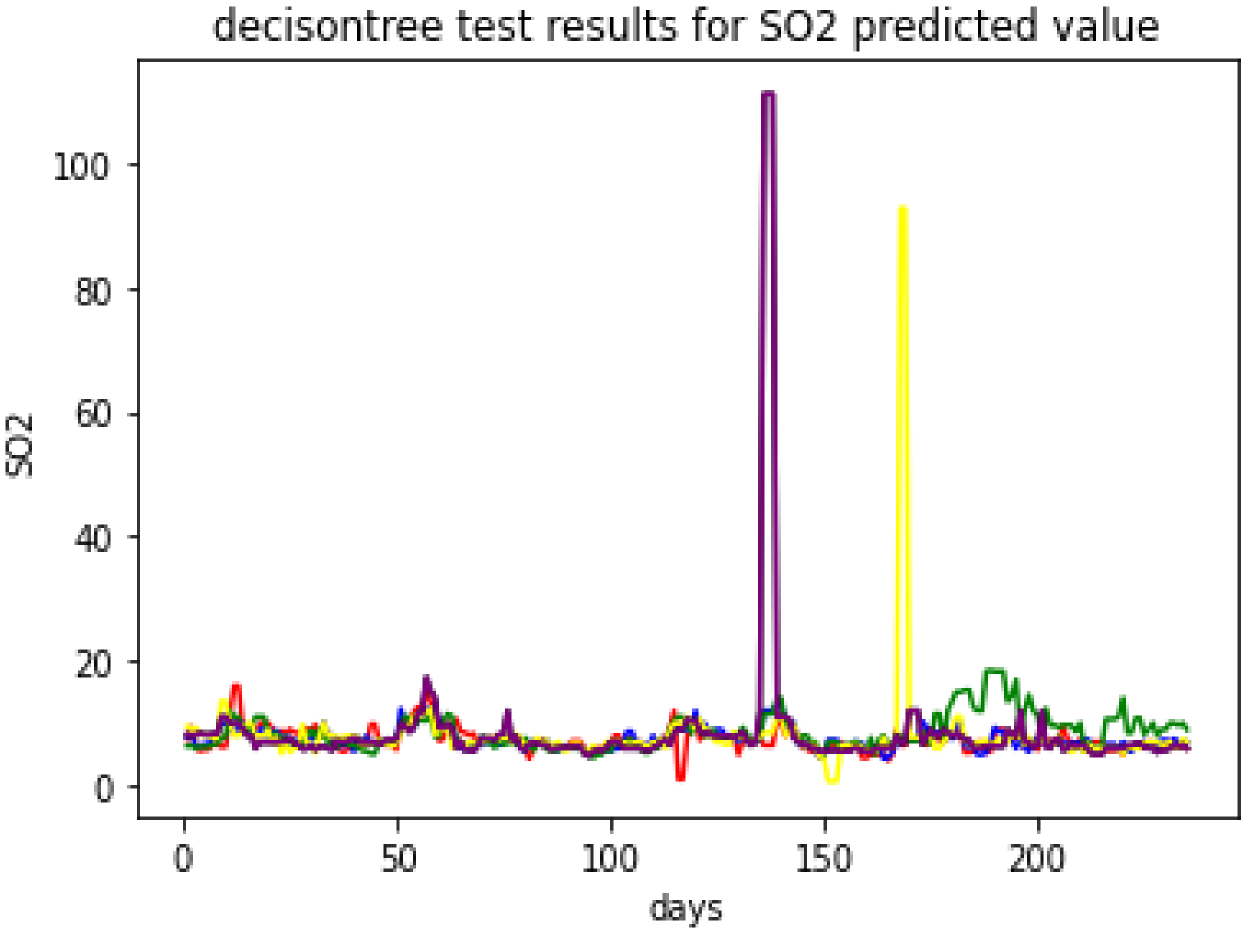
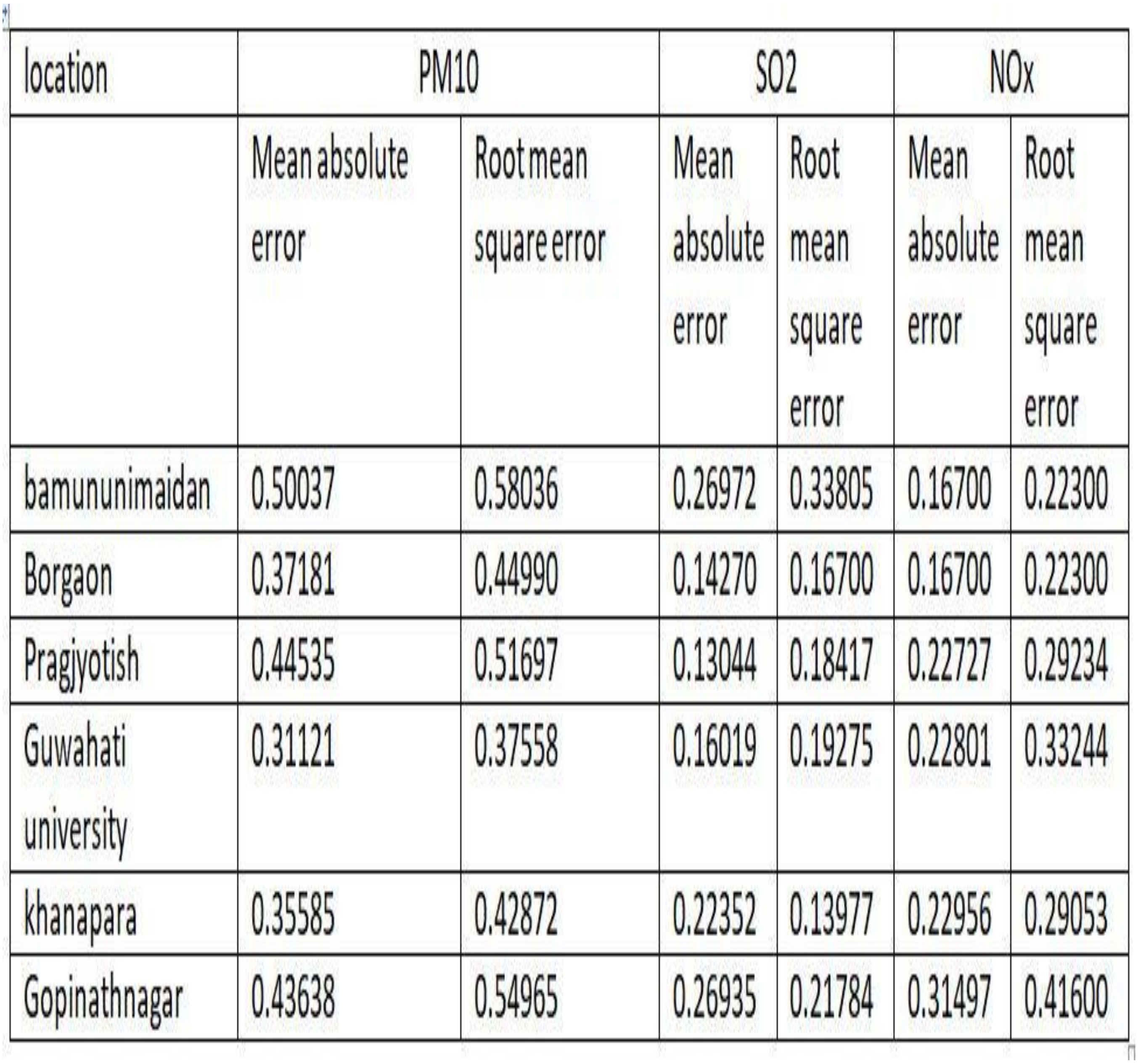


Figure 3.8

3.2 ACCURACY ANALYSIS

The following mean absolute error and root mean square errors were found



location	PM10		SO2		NOx	
	Mean absolute error	Root mean square error	Mean absolute error	Root mean square error	Mean absolute error	Root mean square error
bamununimaidan	0.50037	0.58036	0.26972	0.33805	0.16700	0.22300
Borgaon	0.37181	0.44990	0.14270	0.16700	0.16700	0.22300
Pragjyotish	0.44535	0.51697	0.13044	0.18417	0.22727	0.29234
Guwahati university	0.31121	0.37558	0.16019	0.19275	0.22801	0.33244
khanapara	0.35585	0.42872	0.22352	0.13977	0.22956	0.29053
Gopinathnagar	0.43638	0.54965	0.26935	0.21784	0.31497	0.41600

Table3.1

CHAPTER 4: CONCLUSION AND FUTURE SCOPE

4.1 CONCLUSION

1. From the results obtained we can see that out of all the considered algorithms, the Decision Tree algorithm gives the highest accuracy, hence is most precise in predicting the AQI values of different locations of Guwahati.

2. The PM10 Values were higher than that of SO2 and NOx in all the locations.

3. The actual and predicted values of pollutants are almost linearly comparable.

4. All the stations showed similar accuracy in predicting the pollutant values.

4.2 FUTURE SCOPE

In the future, we can have multiple additions and improvements to this project. The data input can be continuous data input with new data being fed into the prediction model as we keep obtaining it from the sensors. Other machine learning and deep learning models can be implemented and their accuracy can be checked, such as RNNs, VGG based models. Hybrid models can also be tested and implemented which consists of combination of two or more than two prediction models. The models can be implemented using historical data as well as continuous data.

The Prediction or future forecasting of Air Quality indexes can be done using IOT based systems, Website, progressive web apps and mobile apps as well. Proper warning can be if the AQI prediction is too much alarming. More pollutants and air particles can also be added based of the availability of sensors.

REFERENCES

1. Title: Air Quality Prediction using Machine Learning Algorithms

International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–8656

Pooja Bhalgat , Sejal Pitale , Sachin Bhoite

2. WHO (2009) Global health risks: mortality and burden of diseases attributable to selected major risks. World Health Organization, Geneva, Available online at http://www.who.int/healthinfo/global_burden_ disease/GlobalHealthRisks_report_full.pdf
3. Kowalska M, Osrodka L, Klejnowski K, Zejda JE, Krajny E, Wojtylak M (2009) Air quality index and its significance in environmental health risk communication. Arc Environ Prot 35(1):13–21, ISSN: 0324-8461
4. Title : analysis of air quality index

Coimbatore Institute of Technology

Nikila Varshini.E1 , Sreeha.MR2 , Lhavanya Roobini. VN3 ,Vijayarangam.J4 , Sujithra.M5

5. Air Quality Index Prediction Using Simple Machine Learning Algorithms

International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)

Volume 7, Issue 1, January - February 2018 ISSN 2278-6856

Kostandina Veljanovska1 , Angel Dimoski2

6. Study and Analysis of Decision Tree Based Classification Algorithms

International Journal of Computer Sciences and Engineering

Vol.-6, Issue-10, Oct. 2018 E-ISSN: 2347-2693

Harsh H. Patel , Purvi Prajapati

7. NIH Public Access Author Manuscript

Nat Biotechnol. Author manuscript; available in PMC 2009 June 24

Published in final edited form as:

Nat Biotechnol. 2008 September ; 26(9): 1011–1013. doi:10.1038/nbt0908-1011.

8. https://en.wikipedia.org/wiki/Matplotlib