**AIR POLLUTION LIDAR SIGNALS CLASSIFICATION BASED ON MACHINE LEARNING METHODS**

Minor project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

# Computer Science and Engineering

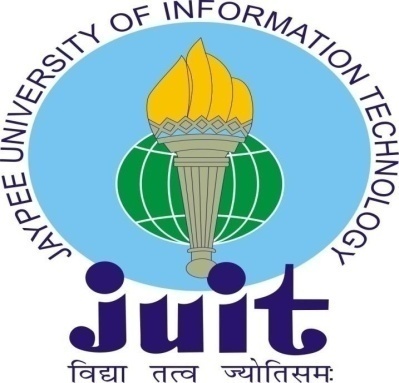
By

## HONEY GUPTA (181257)

## JYOTI SARPAL (181314)

**UNDER THE SUPERVISON OF**

Dr. VIVEK KUMAR SEHGAL



Department of Computer Science & Engineering and Information

Technology

# Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA

**DECLARATION**

I hereby declare that; this project has been done by me under the supervision of **(Dr Vivek Kumar Sehgal, Associate Professor (CSE/IT)),** Jaypee University of Information Technology.

I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

****

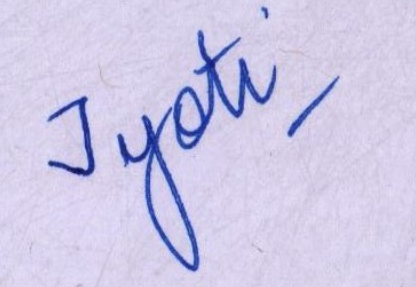
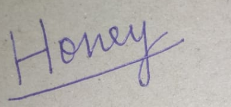
**Supervised by:**

**(Dr Vivek Kumar Sehgal)**

**Associate Professor (CSE/IT**)

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology

** **

**Submitted by:**

**(Jyoti Sarpal) (Honey Gupta)**

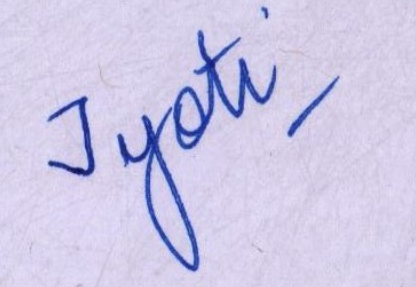
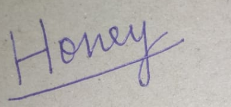
(181314) (181257)

Computer Science & Engineering Department

Jaypee University of Information Technology

**CERTIFICATE**

This is to certify that the work which is being presented in the project report titled **“AIR POLLUTION LIDAR SIGNALS CLASSIFICATION BASED ON MACHINE LEARNING METHODS**” in partial fulfilment of the requirements for the award of the degree of B Tech in Computer Science and Engineering submitted to the Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by “Jyoti Sarpal (181314)**” ,** “Honey Gupta (181257)” during the period from January 2021 to May 2021 under the supervision of **Dr Vivek Kumar Sehgal**, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

**** 

Jyoti Sarpal Honey Gupta

(181314) (181257)

The above statement made is correct to the best of my knowledge.

(**Dr Vivek Kumar Sehgal**)

**Associate Professor (CSE/IT**)

Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat,

**AKCNOWLEDGEMENT**

Firstly, I express my heartiest thanks and gratefulness to almighty God for his

divine blessing makes us possible to complete the project work successfully.

I am really grateful and wish my profound my indebtedness to Supervisor Dr Vivek Kumar Sehgal, **Associate Professor (CSE/IT**) Department of CSE Jaypee University of Information Technology, Wakhnaghat. Deep Knowledge & keen interest of my supervisor in the field of “**AIR POLLUTION LIDAR SIGNALS CLASSIFICATION BASED ON MACHINE LEARNING METHODS**” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to, Dr Vivek Kumar Sehgal Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straight forwardly or in a roundabout way in making this project a win.

In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which

have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

**TABLE OF CONTENT**

**INDEX**

**CONTENT**

**page**

**LIST OF FIGURES**

**LIST OF TABLES**

**ABSTRACT**

**1.INRODUCTION………………………………………………………………….1**

1.1. LIDAR SIGNALS……………………………………………….………….......1

1.2 OBJECTIVE OF MINOR PROJECT…………………………………………...3

1.2.1 REDUCTION OF URBAN AIR POLLUTION……………………….3

1.2.2 OBJECTIVES OF THE UPDATE…………………………………….3

1.2.3 PROBLEMS IDENTIFIED……………………………………….......4

1.3 MOTIVATION OF MINOR PROJECT…………………………………….…4

1. **MINOR PROJECT SDLC……………………………………………………..5**

2.1. FEASIBILITY STUDY ON THE MINOR PROJECT……………………...5

2.2. REQUIREMENTS OF THE MINOR PROJECT …………………………...6

2.3. USE CASE DIAGRAM…………………………………………………. …7

2.4 SEQUENCE DIAGRAM ……………………………………………………8

1. **IMPLEMENTATION OF THE MINOR PROJECT…………………..........9**

3.1. DATA SETS USED IN THE MINOR PROJECT…………………….........9

3.2. DATA SET FEATURES …………………………………...………….......9

3.2.1 TYPES OF DATA SET …………………………………………...9

3.2.2 DATA ABOUT ATTRIBUTES AND FIELDS …………………10

3.3. DESIGN OF THE PROBLEM STATEMENT…...……………………….13

3.4. ALGORITHM OF THE PROBLEM ………….………………...……......13

3.4.1 SOLVING PROBLEM USING LINEAR REGRESSION ………13

3.4.2 SOLVING PROBLEM USING LOGISTIC REGRESSION ……14

3.4.3 SOLVING PROBLEM USING SVM …………………………...15

3.4.4 SOLVING PROBLEM USING RF ………………………...…. ...16

3.5. SCREENSHOTS OF THE VARIOUS STAGES OF THE PROJECT …17

1. **RESULTS………………….............................................................................25**

4.1. DISCUSSION ON THE RESULTS ACHIEVED …………………25

4.2. CONCLUSIONS ……………………………………………...........29

4.3. APPLICATION OF THE MINOR PROJECT ……………………. 30

4.4. LIMITATION OF THE MINOR PROJECT………………………. 30

4.5 FUTURE WORK …………………………………………………… 31

**5. REFERENCES …………………………………………………………….30**

**LIST OF FIGURES**

* 1. Fig 1: TOPOGRAPHIC LIDAR …………………………………………..1
  2. Fig 2: BATHYMETRIC LIDAR ………………………………………….2
  3. Fig 3: MOBILE LIDAR …………………………………………………3

2.1 Fig 4: THE PROPOSED SYSTEM ARCHITECTURE…………...…….....9

2.2 Fig 5: THE SEQUENCE DIAGRAM FOR AIR POLLUTION SENSING **.**.

…………………………………………………………………………………10

2.3 Fig 6: NO2 WAVEFORM **………………………………………………**11

2.1 Fig 7: SO2 WAVEFORM ….…….……………………………………….11

3.1 Fig 8: DATA SET VALUES SYSTEM…….………………......................12

3.2 Fig 9: SECOND DATA SET VALUES ………………………………….13

3.3 Fig 10: BAR CHART……………………………………………………...21

3.4 Fig 11: HEAT MAP……………………………….……………………….21

3.5 Fig 12: BOX PLOT………………………………………………….……21

3.6 Fig 13: PAIR PLOT………………………………………………………..21

3.7 Fig 14: LINE PLOT………………………………………………………..21

3.8 Fig 15: DISTRIBUTION PLOT…………………………………………..22

* 1. Fig 16: XY FUNCTION PLOT…………………………………………...22

3.10 Fig 17: OBSERVATIONS ……………………………………………....23

4.1 Fig 18: LINEAR REGRESSION RESULTS……………………………. 23

4.2 Fig 19: LOGISTIC REGRESSION ALGORITHM………………………24

4.3 Fig 20: LOGISTIC REGRESSION RESUTS………………………….....24

4.4 Fig 21: SUPPORT VECTOR MACHIENE ALGORITHM ………………. …………………………………………………………………………………25

4.5 Fig 22: SUPPORT VECTOR MACHIENE RESULTS …………………. ...……………………………………………………………………………....25

4.6 Fig 23: RANDOM FOREST ALGORITHM …………...……….……... 26

4.7 Fig 24: RANDOM FOREST RESULTS ……………………………...... 26

**LIST OF TABLES**

Table 1: DATA SET DESCRIPTION **……………………………………….**13

Table 2:PCA DATA SET …………………………………………………...26

**ABSTRACT**

With the rapid development of economy, environmental problems caused by various pollutants have become increasingly prominent, and air pollution becomes a hot topic of each individual’s concern. With the development of industry, the harmful gases emissions have increased. Monitoring the amount of SO2 and NO2 in the environment can help predict the air quality in real time and track the pollution sources which requires a large-scale, pollution detecting device along with a real-time signal processing algorithm. The vehicle–mounted laser radar is an effective pollution detecting device. According to different concentrations theory, the attenuation rate of different wavelength laser signals in the atmosphere is also different, and in order to calculate the amount of SO2 and NO2 in the atmosphere, which needs to distinguish different signals from different wavelengths. There are 6 kinds of signal data, because SO2 and NO2 each need 2 kinds of signal data and other signal data to mark the signal data. Distinguishing such Lidar signals manually is time and labor consuming, there is need to find a way to automatically distinguish these original signals. With the development of machine learning methods and neural networks, we can classify data efficiently and accurately.

In this project, we try to classify lidar signals to evaluate and find the concentration percentages of nitrogen dioxide and sulfur dioxide by comparing several machine learning methods like SVM, Random Forest (RF), Logistic Regression (LR), and Linear Regression.

By comparing the classifying performance of SVM, LR, RF and Linear Regression, we can find that the RF performs best on the original dataset. Thus, in the end, by the comparison, we can conclude that the original data extraction characteristics to build a data set, run on the above model leads to good results while Random Forest helps in constructing individual data set in order to achieve good results.

Keywords: [air pollution lidar signals classification](https://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:air%20pollution%20lidar%20signals%20classification&newsearch=true), [environmental problems](https://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:environmental%20problems&newsearch=true), [real-time signal processing algorithm](https://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:real-time%20signal%20processing%20algorithm&newsearch=true), [nitrogen dioxide](https://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:nitrogen%20dioxide&newsearch=true), [sulfur dioxide](https://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:sulfur%20dioxide&newsearch=true), [machine learning methods](https://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:machine%20learning%20methods&newsearch=true),

**CHAPTER-1**

**INTRODUCTION**

**LIDAR** uses electromagnetic (EM) waves. It is an active sensor which means that it sends out an EM wave and receives the reflected wave signal back. **LIDAR** uses a pulsed laser machine device to calculate an object's variable distances from the earth surface. These light pulses gather together with the information collected by the airborne system and generate accurate 3D information and image about the earth surface and the target object.

**1.1 LIDAR SIGNALS:**

DIFFERENT TYPES OF LIDAR SIGNALS:

There are two types of lidars as defined below:

## Airborne:

With airborne lidar, system is in either a fixed-wing aircraft or a helicopter. The laser light is emitted toward the ground and reflected back to the moving airborne lidar sensor. There are two types of airborne sensors:

### **Topographic lidar**

Topographic lidar can be used in many applications, like forestry, hydrology, geomorphology, urban planning, coastal engineering, and volumetric calculations.

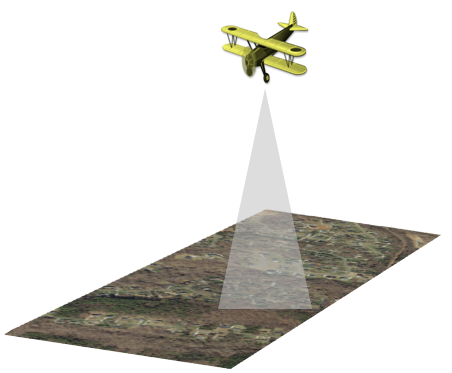


Fig 1: **TOPOGRAPHIC LIDAR**



Page 1

### **Bathymetric lidar**

Bathymetric lidar is a type of airborne device which is water penetrating. Helpful for collecting elevation and water depth one after the other, which provides an airborne lidar survey of the land-water interface. In it the infrared light (traditional laser system) is reflected from the device from the land and water surface, along with some additional green laser lights which travel through the water column. Used in harbors, and places which are near shores and banks.

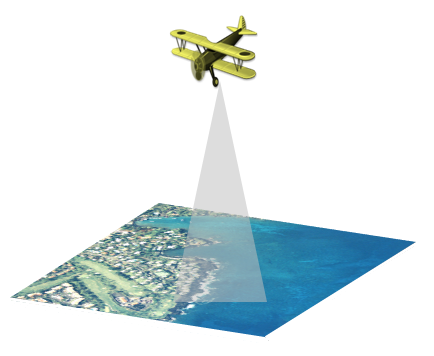


Fig 2: **BATHYMETRIC LIDAR**

## Terrestrial:

There are two main types of terrestrial lidars as defined below: mobile and static.

Terrestrial lidar collects highly calculated accurate points, which allows in-depth identification of objects. These dense point clouds can be used for managing facilities, conducting highway and rail surveys, and even creating 3D city models for house designs.

### **Mobile**

In the case of mobile acquisition, the lidar system is placed on top of a moving vehicle. They can include any number of lidar mounted on the vehicle which is moving. Can be mounted on vehicles, trains, boats etc. They consist of a sensor, cameras, GPS (Global Positioning System) widely used, and an INS (inertial navigation systems). Used for analyzing road infrastructure locating overhead wires, light poles, and road signs.



Page 2

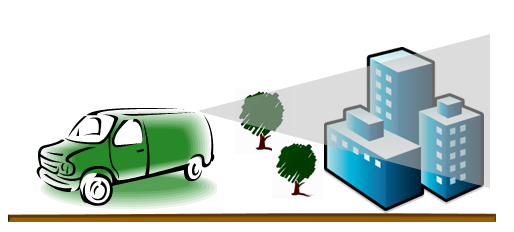


Fig 3: **MOBILE LIDAR**

### **Static Lidars:**

In the case of static acquisition, the lidar system is typically is placed on top of a tripod or stationary device. It is the collection of lidar points from a static (not moving) location. The lidar sensor is mounted on top of a tripod mount and is a fully portable, laser-based ranging system. These systems can collect lidar point clouds which are found inside buildings as well as outside a building. Applications include usage in engineering, mining, and by archaeologists.

**1.2 OBJECTIVE OF MINOR PROJECT**

Key objective: Pollution Prevention

**1.21** **Reduction of urban air pollution**

Planned actions:

a) Optimizing the standards, accounting for environment related problems and combined health impacts.

b) Implementation of Best Available Algorithms and best practices.

c) Development of a mechanism for accounting for state-level conditions and technological capacity in the course of setting emission limit values.

**1.22 Objectives of the update**

a) reduction of population exposure to air pollution

b) address air quality standards

c) air quality monitoring

Page3 

**1.23 Problems identified**

a) Health impacts of urban air pollution

b) Weakness of air quality control systems

c) Excessively strict ambient air quality standards

d) Weak technological capacity, resulting in higher emissions

e) Lack of economic incentives to reduce emissions

f) Inadequacies of regulation of road transport emissions

**1.3 MOTIVATION OF MINOR PROJECT**

Air pollution is all around us. It affects us all, whether we recognise this fact or not. We have taken undue advantage of the air that we breathe.

But the current recent research has started to throw light on some rather worrisome aspects of what the air around us really contains, and how it affects our bodies and life. And the more we learn, the more we come to realize that this essential source of life for the planet needs some serious care. Without air there can be no life but breathing polluted air leads to a life of disease and early death.

**Polluted air is creating a health emergency**:

There is no doubt today that air pollution is a global public health emergency. It affects everyone from unborn babies to children to women to elderly.

Outside the street and inside the house, the air pollution sources can be very different, yet their effects are equally contagious: asthma, other respiratory illnesses and heart disease are all caused by inhaling polluted air. According to the World Health Organization, every year around 8 million unborn deaths are due to air pollution and almost 800 people every hour or 13 every minute. Overall, air pollution is responsible formore deaths than many other risk diseases like malnutrition, alcohol use and physical inactivity.

**Children are most at risk:**

Globally, 93 per cent of all children inhale air that contains higher concentrations of pollutants than the desired World Health Organization (WHO) rate according to which is safe to human health. As a result,600,000 children die prematurely each year because of pollution. Along with it there comes the after effects which include, distortion of brain development, leading to cognitive and motor impairments, and a greater risk for chronic disease later in life.

Page 4

Air pollution is very harmful to women and children. About 60 per cent of household air pollution-related deaths are seen in women and children, and more than half of all pneumonia deaths are seen in children aged under five which can be attributed to indoor air pollution.

**Pollution and poverty go hand in hand:**

Air pollution affects poor people drastically. In household systems, air pollution comes mostly from fuels and cooking systems. Clean cooking and heating fuels and resources are out of hand reach for low-income families, so polluting alternatives are the only way. About 3 billion people are dependent on the burning of solid fuels like coal or kerosene to meet energy requirements and 3.8 million of the deaths each year are from exposure to these harmful gases. A lack of awareness about the risks along with breathing polluted air also the reason to the problem, which affects the cost and difficulty to access healthcare.

Crowded cities and trafficked suburbs are hubs for outdoor air pollution. According to the World Health Organization, 97 per cent of cities more than 100,000 people do not meet the minimum air quality levels. Around 4 million people die from air pollution-related diseases every year in the Asia-Pacific region.

**The cheaper the fuels, the higher the costs**

When people get sick, the entire community suffers. Ill people require medical care and medicine, children also tend to skip school and adults miss their office, as a result of their own poor health, or to care for someone in the family. According to the World Bank, air pollution costs more than US$5 trillion every year in terms of welfare costs and $225 billion in almost lost in income. A 2016 study by the Organisation for Economic Cooperation and Development predicts that, if the situation remains the same, by 2060 or more the welfare costs of premature deaths from air pollution would be 18-25 trillion, with the costs of pain and suffering from acute illness and diseases is estimated at around 2.2 trillion.

Ground-level ozone is expected to reduce by 26 per cent by 2030, creating issues like security of food and nutrition challenges. Air pollution also degrades materials, decreasing their useful potential and generating costs for more maintenance like cleaning, repair and replacement.

**Right to clean air: A human right**

The right to a healthy environment enjoys constitutional status of a right in more than 100 countries. At least 155 states are obligated, by the constitutions and legislation, to respect, protect the right to a healthy environment.

Page 5

**1.4 LANGUAGES USED**

* PYTHON
* MACHINE LEARNING METHODS.
  1. **TECHNICAL REQUIREMENTS(HARDWARE/SOFTWARE)**
* JETSON NANO
* JUPYTER NOTEBOOK
* PYTHON IDE
  1. **DELIVERABLES OF THE PROJECT**

Since our project is based on how to detect air pollution using machine learning methods, thus it plays a vital role in detecting air pollution and to find out the concentrations of SO2 and NO2, we can detect the quality of air.

In this work, we mainly solve the problem of fast classification of lidar signals for measuring air pollution.

By comparison, it can be concluded that the classification accuracy can be improved by introducing the

statistical features. By comparing the classification performances of SVM, LR, RF, Linear Regression, we can find that RF performs best on the PCA dataset.

Page 6

**CHAPTER 02**

**MINOR PROJECT SDLC**

**2.1 FEASIBILITY STUDY ON MINOR PROJECT**

* Classification: Here we would like to provide some reviews on classification with time-series data from various aspects to show the systematic research in this area. A lot of evidence shows that the data set with more features has better performance in the model. It uses the discrete wavelet transform to extract the features from physiological signal. So, in this paper, we use 2 different data sets with different features to fit. The first data set uses original signal data of the states and the second data set uses the features with PCA.
* When constructing a classifier based on clean training data for a specific test environment, the effects of environmental noise and channels should be considered. They proposed a kernel method to take into account the test environment when using and training an SVM from training

data. The proposed SVM is the best for SNRs greater than 0 db. At lower SNR all classifiers had a

poor performance.

So here we use two different data sets with features to fit. The first uses original signal data and the

second data uses the features with PCA and some additional features to it.

When constructing a classifier based on clean training data for a specific test environment noise and

channels should be considered.

Decision tree classification provide a fast and efficient classified data set. Although there are many

calculation methods for improving the structure of decision tree, these methods may be affected by

changes in the training data set. An evolutionary method has been proposed that provides flexibility

for the decision tree using a common evolutionary method have been proposed that provides

flexibility for decision tree using the common evolutionary competition between decision trees

and training data sets. Not only do evolutionary methods have the ability to develop the structure of

a decision tree classification system they can also manipulate a set of values and data types. This is

a major improvement over many other classification systems. A radar signal classification algorithm

based on auto correction function and directed graph model.

Page 7

**2.2 REQUIREMENTS ON MINOR PROJECT**

**Tools and technologies to be used:**

* **Tensor flow:** TensorFlow is an open source plus end-to-end platform for machine learning problems. It has variety of a comprehensive, flexible tools and multi-purpose libraries along with community resources that help researchers to develop ML models and easily build and deploy Machine Learning bases powered

applications.

* **Tensor board:** Tensor Board provides the necessary visualization tools which are needed for machine learning model building. Also helps us in:

1. Keep track and visualize metrics like loss and accuracy
2. Model the graph and its various layers.
3. Develop histograms, boxplots, line graphs, bar charts, and other tensors which change over (i.e., include new and additional features) from time to time.
4. Projecting the model to a lower dimensional space and effectively study time axis relation.

* **Pytorch:** PyTorch is an open source plus end-to-end platform for machine learning problems. Used for:

1. Highlevel tensorcomputing which is generally done using python libraries like NumPywith strong acceleration via graphics processing units found in processors.
2. **PyTorch** being easy to learn and work with is relatively **better** for ML based projects and building

prototypes.

Page 8

**2.3 USE CASE DIAGRAM**



Fig 4: **THE PROPOSED SYSTEM ARCHITECTURE**

The lidars mounted at the top of a vehicle or a building has several basic tasks including gathering data from air pollution sensors, detection of the location, process of communication with cloud, and reformatting and sending data. The first step is to receive pollution data from the installed sensors. The sensors sense the particulate matter of NO2, and SO2 of the air, and we can receive the signals containing the information concerning the local level of pollution though lidars. In the next step, the coordinate of the lidar is obtained through GPS. It helps to detect where the pollution data has been sensed (longitude and latitude wise coordinates). The geographical location of constantly kept under observation by GPS. Since the system is used in for large scale, pollution data is locally processed and transformed into AQI format. AQI facilitates the easier understanding of surrounding pollution.

As it is communication is both time and energy consuming process, the data is not continuously sent to the cloud. Instead, an automatic mode is used, which changes location module. According to this module, the current position is continuously compared with the last position and in case of a change, the location change is then detected. In case, no location change is detected, the data is sent in a longer period of time.

Page 9

**2.4 SEQUENCE DIAGRAM**

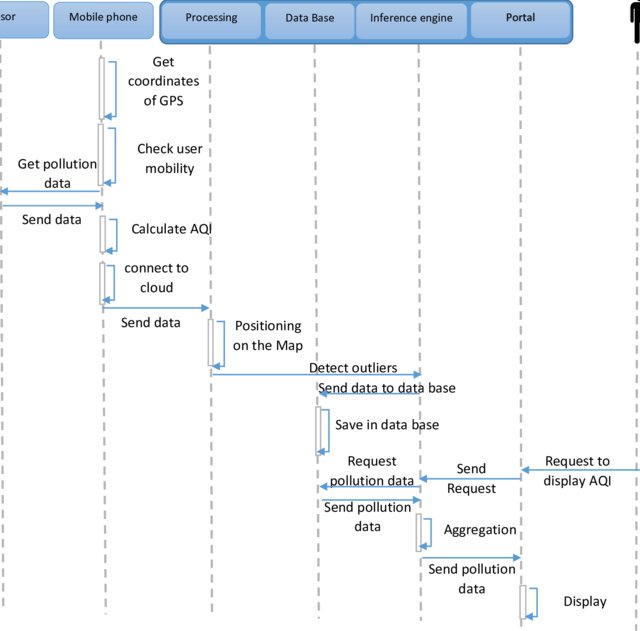


Fig 5: **THE SEQUENCE DIAGRAM FOR AIR POLLUTION SENSING**

In the proposed system, only those much-polluted areas can be warned in real time in order to protect and safeguard the people living in those areas to leave the open area. Since AQI is calculated and displayed after sending pollution data by the user, it plays a very important role in pollution detection overall. Fig. 5 shows the sequence diagram for this part of scenario by the proposed system.

Page 10

**CHAPTER 03**

**IMPLEMENTATION OF THE MINOR PROJECT**

**3.1 DATA SETS USED IN THE MINOR PROJECT**

The signal data collected by lidar were stored in bin file and was originally in binary format. The file holds the data of 753 pulses. Each pulse was originally an ON or a OFF signal. Both pollutants NO2 and SO2 have been separately stored in our data set each having 753 data values. The mean value of the flag decided whether the

Pulse belonged to ON or OFF signal.

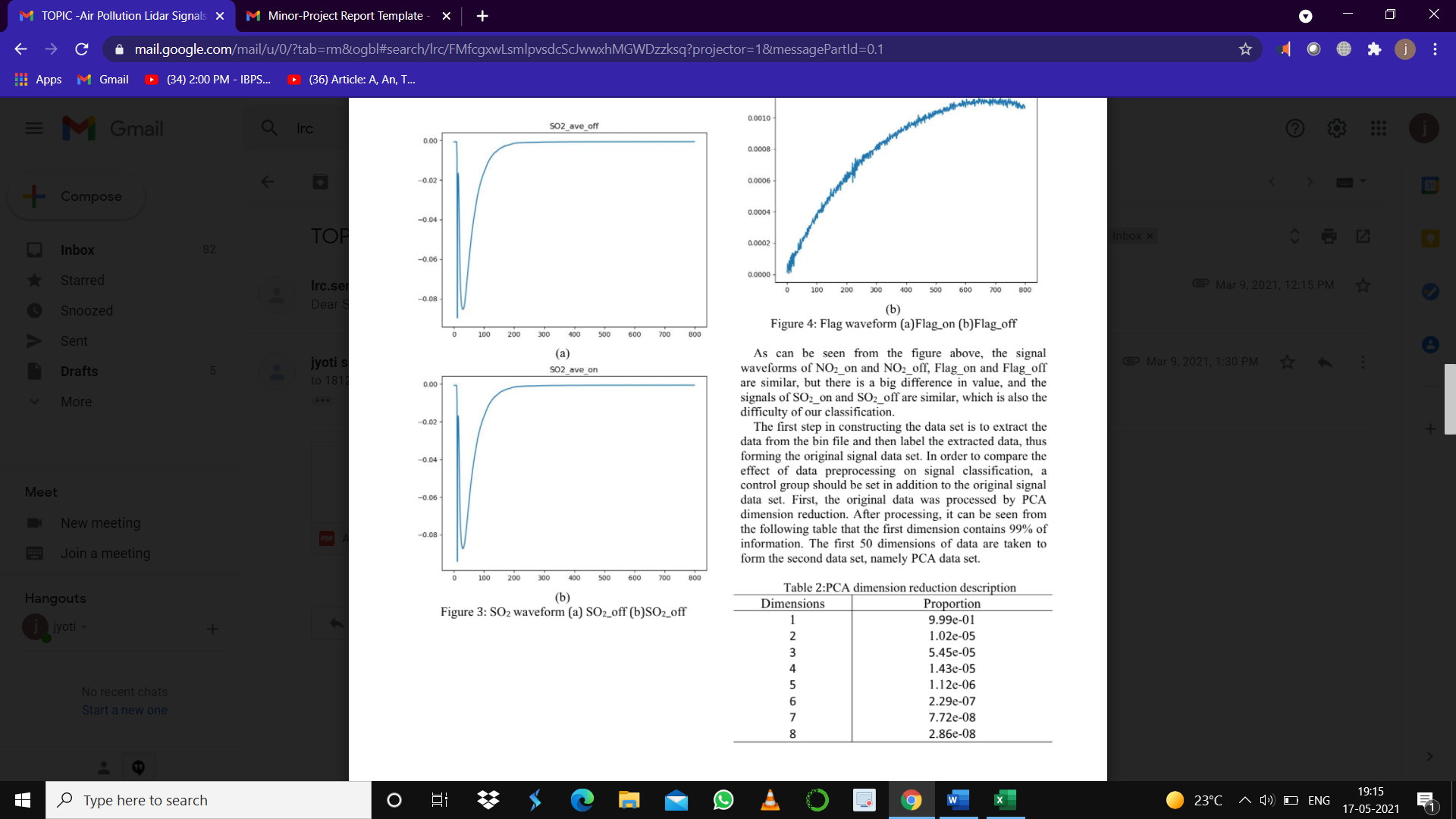


Fig 6: **NO2 WAVEFORM** Fig 7: **SO2 WAVEFORM**

As it can be seen from the above figures that, the signal waveforms of NO2 are a bit similar, but the difference lies in value, which is also the difficulty of our classification.

The first step in constructing the data set is extraction of data from the bin file, labelling the extracted data and then forming the original data set.

**3.2 DATA SET FEATURES**

**3.2.1 Types of Data Set**

Air is what keeps us humans alive. Monitoring it and having an understanding air quality is of immense importance to our well-being. The dataset contains air quality data of SO2 and NO2 at and daily level of various stations across multiple cities in India.

Page 11

**3.2.2 Data about Attributes and Fields**

Cities

**Ahmedabad, Aizawl, Amaravati, Amritsar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram,**

**Bengaluru, Bhopal, Brajrajnagar, Guwahati, Hyderabad, Jaipur, Shillong, Talcher, Thiruvananthapuram, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna,Visakhapatnam**

**Column attributes include**

**City of origin, the date as per the value is recorded, and values of both SO2 and NO2 air pollutants.**

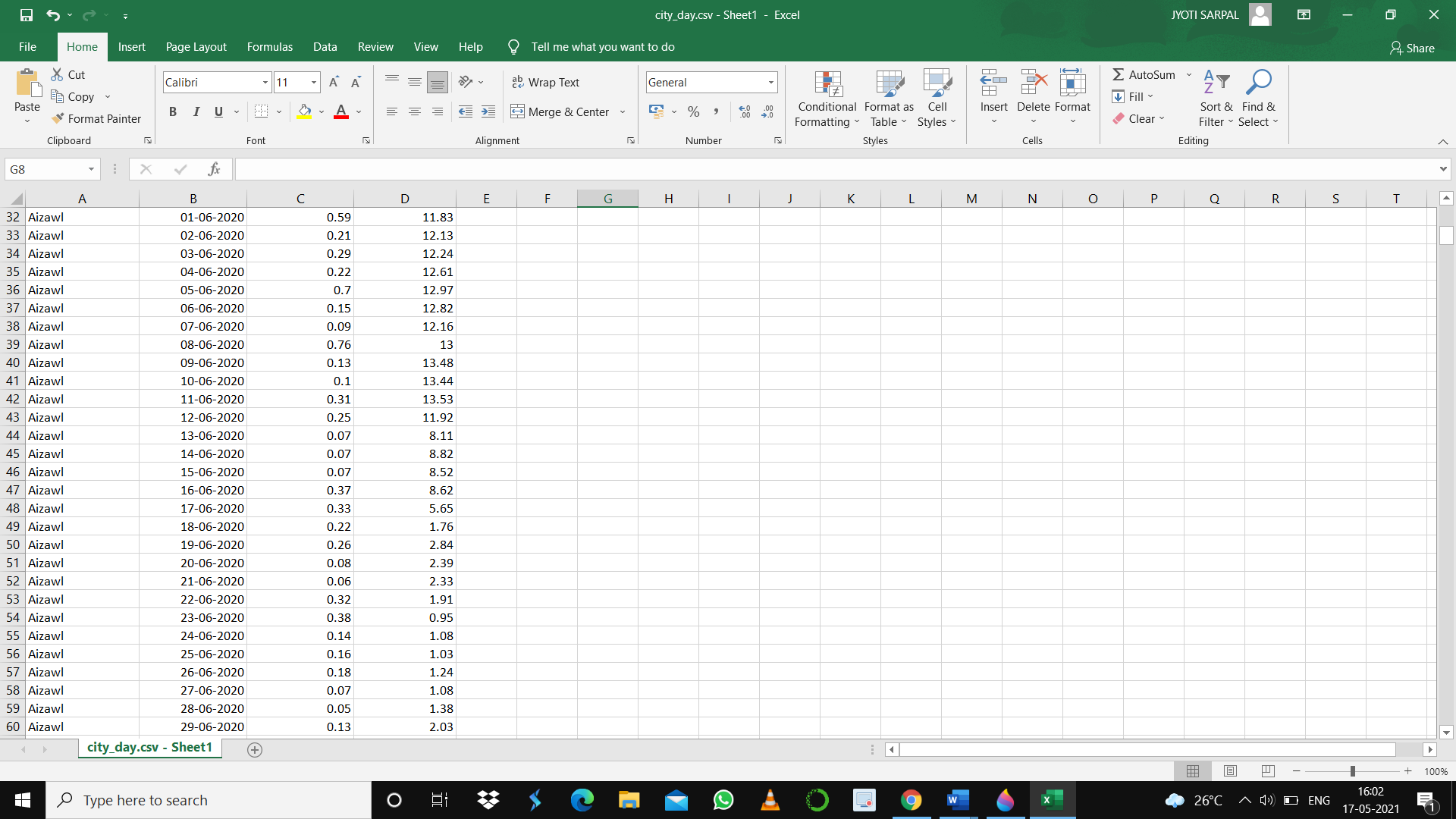
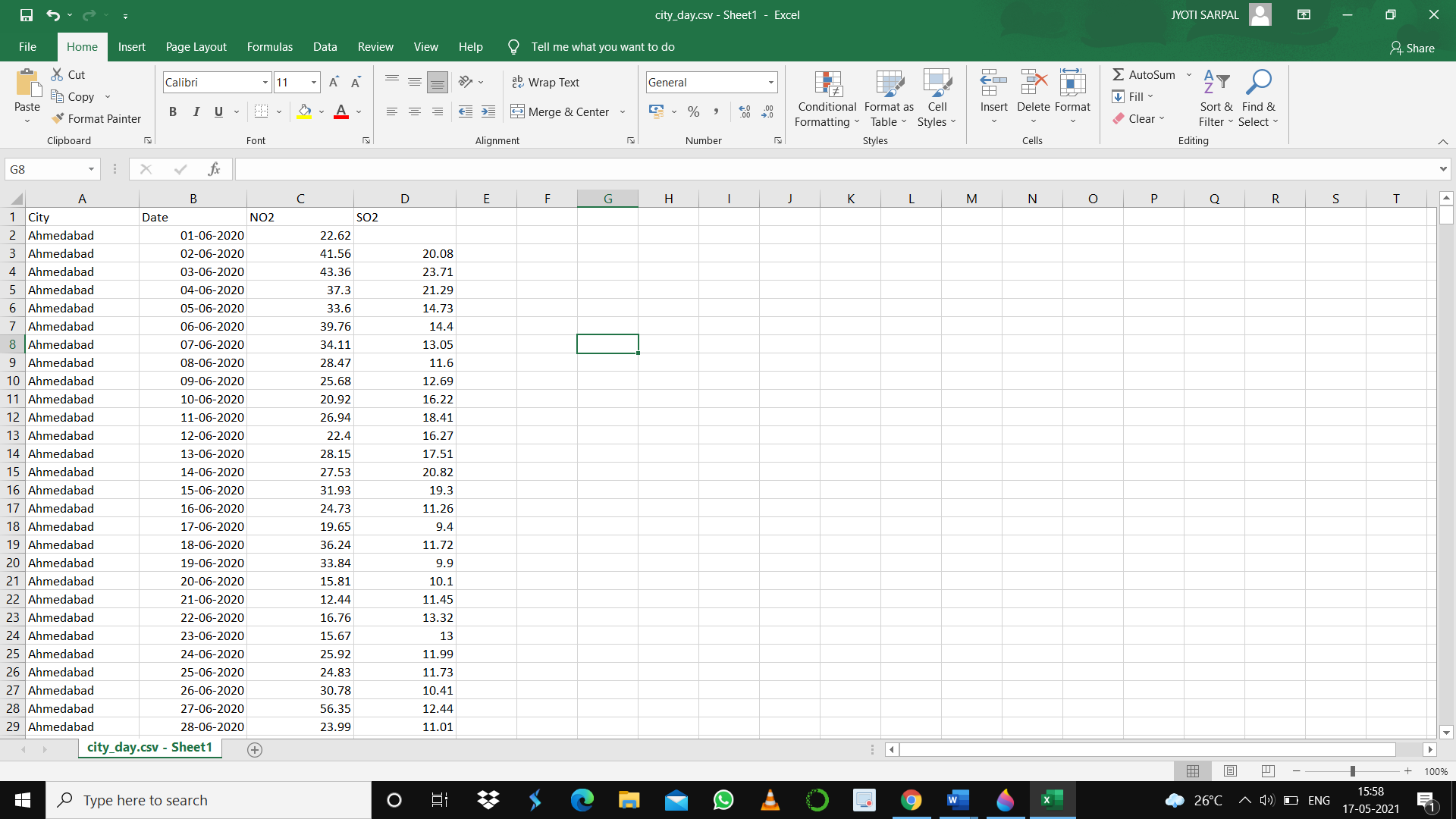


Fig 8: **DATA SET VALUES**

After being processed by PCA dimension reduction the first 25 dimensions of the data are taken to form the second data set, namely PCA data set. Secondly in addition to dimensionality reduction, some data features are also added. PCA dimensionality reduction was retained in the previous three-dimensional form, and some new data features were added.



Page  12

**Maximum x(max) = max{x} x= x1, x2, x3 ---------**

**Minimum x(min) = min{x} x= x1, x2, x3 ---------**

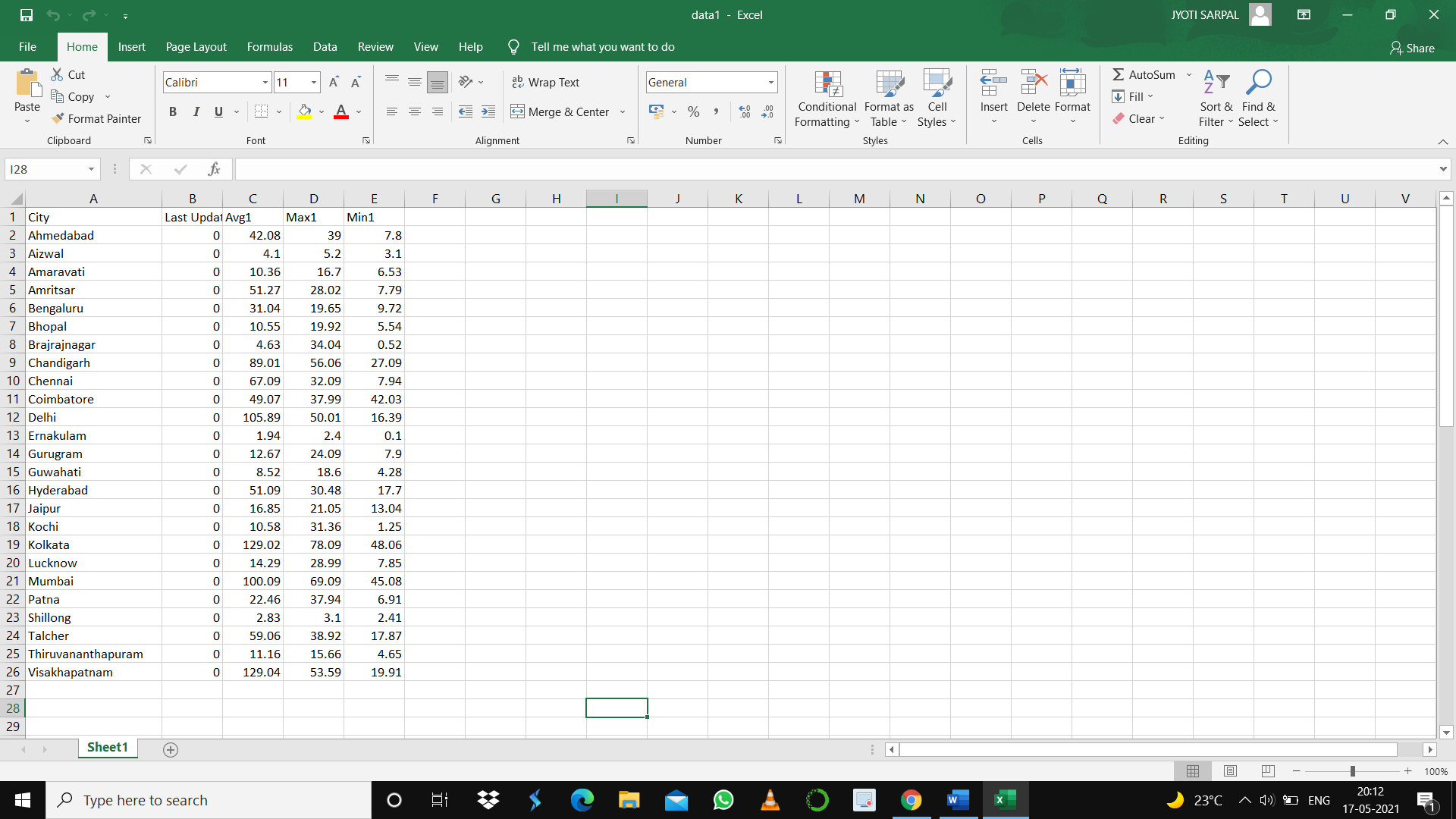


Fig 9: **SECOND DATA SET VALUES**

Table 1: **DATA SET DESCRIPTION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **DATA SETS** | **LENGTH** | **CLASSES** | **TRAIN** | **TEST** |
| Original | **752** | **4** | **452** | **302** |
| PCA | **25** | **5** | **15** | **10** |

Page 13

**3.3 DESIGN OF PROBLEM STATEMENT**

In this project, we try to classify lidar signals to evaluate and find the concentration percentages of nitrogen dioxide and sulfur dioxide by comparing several machine learning methods like Linear Regression, SVM, Random Forest (RF), Logistic Regression (LR).

**3.4 ALGORITHM / PSEUDO CODE OF THE PROJECT PROBLEM**

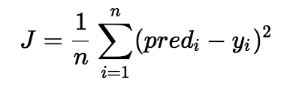
There are various methods for classification, such as SVM, RF, Logistic Regression, Linear Regression.

So here we use above four methods to classify our data sets and compare the different data sets of different classification methods. Below are the various methods used.

**3.4.1 SOLVING OUR PROBLEM USING LINEAR REGRESSION:**

**Linear Regression** is a machine learning algorithm which is used for **supervised learning**. It performs best on a task which is related to **regression**. These models predict a target value based on variables which are independent. Used for finding out the relationship between different values of different variables. Cost function

Of Linear Regression is given by

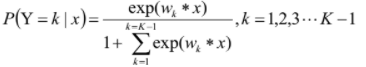


|  |
| --- |
| X= df.drop(labels='class',axis=**1**)  y=df.loc[:,'class']  # division into training and test sets.  **from** **sklearn.model\_selection** **import** train\_test\_split  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y,test\_size=**0.2**, random\_state=**0**)  regressor = LinearRegression()  regressor.fit(X\_train, y\_train)  **print**("Training complete.")  # Testing data - In Hours  **print**(X\_test)  # Predicting the scores  y\_pred = regressor.predict(X\_test)  # Comparing Actual vs Predicted  df = p.DataFrame({'Actual': y\_test, 'Predicted': y\_pred})  **import** **sklearn.metrics** **as** **sm**  **print**("Mean absolute error =", round(sm.mean\_absolute\_error(y\_test, y\_pred), **2**))  **print**("Mean squared error =", round(sm.mean\_squared\_error(y\_test, y\_pred), **2**))  **print**("Median absolute error =", round(sm.median\_absolute\_error(y\_test, y\_pred), **2**))  **print**("Explain variance score =", round(sm.explained\_variance\_score(y\_test, y\_pred), **2**))  **print**("R2 score =", round(sm.r2\_score(y\_test, y\_pred), **2**)) |

Page 14

**3.4.2 Solving our problem using Logistic Regression (LR):**

LR is a model used to predict probability. The model for multiple classification is as follows:

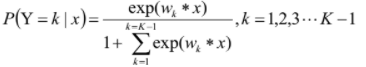


|  |
| --- |
| X= df.drop(labels='class',axis=**1**)  y=df.loc[:,'class']  #spliting our data to training sets and test sets  **from** **sklearn.model\_selection** **import** train\_test\_split X\_train,X\_test,y\_train,y\_test= train\_test\_split(X,y, test\_size=**0.2**, random\_state=**1**,stratify=y)  **from** **sklearn.model\_selection** **import** cross\_val\_score  **from** **sklearn.linear\_model** **import** LogisticRegression  logR= LogisticRegression() #object of a class  logR.fit(X\_train,y\_train) #training for our model using fit method  predictR = logR.predict(X\_test) #ready to make predictions by calling predict method  **print**("")  **print**("Classification report of Logistic Regression Results:")  **print**("")  **print**(classification\_report(y\_test,predictR  accuracy = cross\_val\_score(logR ,X, y, cv=**7**)  **print**("Cross validation test results of accuracy:")  **print**(accuracy)  **print**("")  # get mean of every field  **print**("Accuracy result of Logistic Regression is:",accuracy.mean()\***100**)  **print**("")  cm1=confusion\_matrix(y\_test,predictR)  **print**("Confusion Matrix result of Logistic Regression is:**\n**",cm1)  **print**("")  sensitivity1= cm1[**0**,**0**]/(cm1[**0**,**0**]+cm1[**0**,**1**])  **print**("Senstivity :",sensitivity1) # Cases actual positive cases that got predicted as true  **print**("")  specificity1 = cm1[**1**,**1**]/(cm1[**1**,**0**]+cm1[**1**,**1**])  **print**("specificity :",specificity1) # Cases actual positive cases that got predicted as false  **print**("")  TN = cm1[**0**][**0**]  FN = cm1[**1**][**0**]  TP = cm1[**1**][**1**]  FP = cm1[**0**][**1**]  **print**("True positive :",TP)  **print**("True negative :",TN)  **print**("False positive :",FP)  **print**("False negative :",FN)  **print**("")  TPR = TP/(TP+FN)  TNR = TN/(TN+FP)  FPR = FP/(FP+TN)  FNR = FN/(TP+FN)  **print**("")  **print**("True positive Rate :",TPR)  **print**("True negative Rate :",TNR)  **print**("False positive Rate:",FPR)  **print**("False negative Rate :",FNR) |
|  |

**3.4.3 Solving our problem using Support Vector Machine (SVM):**

It is an extended support vector classification method to solve the classification problem. It uses the nonlinear

Transformation to transform the problem into a linear problem, and then solves the original nonlinear problem by solving the linear problem after the transformation. Common kernel functions are as follows:

****

|  |
| --- |
| X= df.drop(labels='class',axis=**1**)  y=df.loc[:,'class']  #spliting our data to training sets and test sets  **from** **sklearn.model\_selection** **import** train\_test\_split X\_train,X\_test,y\_train,y\_test= train\_test\_split(X,y, test\_size=**0.2**, random\_state=**1**,stratify=y)  from sklearn.svm import SVC  s = SVC()  s.fit(X\_train,y\_train)  predicts = s.predict(X\_test)  print(" ")  print("Classification report of Support Vector Machines Report:")  print("")  print(classification\_report(y\_test,predicts))  **print**(classification\_report(y\_test,predictR  accuracy = cross\_val\_score(logR ,X, y, cv=**7**)  **print**("Cross validation test results of accuracy:")  **print**(accuracy)  **print**("")  # get mean of every field  **print**("Accuracy result of Support Vector Machines is:",accuracy.mean()\***100**)  **print**("")  cm1=confusion\_matrix(y\_test,predictR)  **print**("Confusion Matrix result of Support Vector Machines is:**\n**",cm1)  **print**("")  sensitivity1= cm1[**0**,**0**]/(cm1[**0**,**0**]+cm1[**0**,**1**])  **print**("Senstivity :",sensitivity1) # Cases actual positive cases that got predicted as true  **print**("")  specificity1 = cm1[**1**,**1**]/(cm1[**1**,**0**]+cm1[**1**,**1**])  **print**("specificity :",specificity1) # Cases actual positive cases that got predicted as false  **print**("")  TN = cm1[**0**][**0**]  FN = cm1[**1**][**0**]  TP = cm1[**1**][**1**]  FP = cm1[**0**][**1**]  **print**("True positive :",TP)  **print**("True negative :",TN)  **print**("False positive :",FP)  **print**("False negative :",FN)  **print**("")  TPR = TP/(TP+FN)  TNR = TN/(TN+FP)  FPR = FP/(FP+TN)  FNR = FN/(TP+FN)  **print**("")  **print** ("True positive Rate :",TPR)  **print**("True negative Rate :",TNR)  **print**("False positive Rate:",FPR)  **print**("False negative Rate :",FNR) |

**3.4.4 Solving our problem using Random Forest (RF):**

An important ML method, which can be used for classification, regression and other problems. By generating M decision tress, a random forest is formed. The classification results of the new data depend on the scores generated by the classification tree voting. Its essence is to improve the decision tree algorithm of multiple decision tree combination. The capability of a single tree maybe small, but after randomly generating a large no of decision trees, a test sample can be selected according to classification results of each tree to select the most likely classification.

|  |
| --- |
| X= df.drop(labels='class',axis=1)  y=df.loc[:,'class']  #spliting our data to training sets and test sets  from sklearn.model\_selection import train\_test\_split X\_train,X\_test,y\_train,y\_test= train\_test\_split(X,y, test\_size=0.2, random\_state=1,stratify=y)  from sklearn import model\_selection  from sklearn.ensemble import RandomForestClassifier  rfc = RandomForestClassifier()  rfc.fit(X\_train,y\_train)  predicts = rfc.predict(X\_test)  print(" ")  print("Classification report of Random Forest:")  print("")  print(classification\_report(y\_test,predicts))  print(classification\_report(y\_test,predictR)  accuracy = cross\_val\_score(logR ,X, y, cv=7)  print("Cross validation test results of accuracy:")  print(accuracy)  print("")  # get mean of every field  print("Accuracy result of Random Forest is:",accuracy.mean()\*100)  print("")  cm1=confusion\_matrix(y\_test,predictR)  print("Confusion Matrix result of Random Forest is:\n",cm1)  print("")  sensitivity1= cm1[0,0]/(cm1[0,0]+cm1[0,1])  print("Senstivity :",sensitivity1) # Cases actual positive cases that got predicted as true  print("")  specificity1 = cm1[1,1]/(cm1[1,0]+cm1[1,1])  print("specificity :",specificity1) # Cases actual positive cases that got predicted as false  print("") |

**3.5 SCREEN SHOTS OF THE VARIOUS STAGES OF THE PROJECT**

**#STEP 1: Data validation and Pre-processing**

|  |
| --- |
| **import** **pandas** **as** **p**  **import** **numpy** **as** **n**  **import** **matplotlib.pyplot** **as** **plt**  **import** **seaborn** **as** **s**  **import** **warnings**  warnings.filterwarnings("ignore")  data = p.read\_excel("city\_day.xlsx") #reading the dataset file using pandas  data.head(**10**) # give first 10 row command name attribute and variables  data.tail() # give last 5 rows of the dataset  data.shape #shows number of rows and columns in the dataset  data.columns  data.describe() # Generate descriptive statistics  data.info()  #information about dataset  # Cleaning up the data (Cleansing the data)  # Converting raw data to clean dataset  df= data.dropna() #used to remove rows and columns with Null/NaN  df.duplicated() #Checking for duplicate values and eliminating them one by one  df.isnull().sum() #Checking for missing values  num=df.\_get\_numeric\_data() #Checking for numeric data if negative values are there need to be replaced  num[num<**0**]=**0**  data  **print**(“Minimum value of NO2 is:”,df.NO2.min())  **print**(“Minimum value of SO2 is:”,df.SO2.min())  **print**(“Maximum value of SO2 is:",df.SO2.max())  **print**("Maximum value of NO2 is:",df.NO2.max()) |
|  |

**#STEP 2: Exploration of data and it's visual analysis**

|  |
| --- |
| p.crosstab(df.City,df.NO2) #finding relationship between city and no2 concentration values  p.crosstab(df.City,df.SO2) #finding relationship between city and so2 concentration values  fig ,ax = plt.subplots(figsize=(**16**,**10**)) # subplots  ax = s.barplot(x='City',y='NO2',data=df)  ax.set(ylabel='Different NO2 concentrations',title = 'NO2 concentration across different cities')  plt.show()  fig ,ax = plt.subplots(figsize=(**16**,**10**))  ax = s.barplot(x='City',y='SO2',data=df)  ax.set(ylabel='Different SO2 concentrations',title = 'SO2 concentration across different cities')  plt.show()  # Heatmap : two-dimensional graphical representation of data  #correlation heatmap  plt.figure(figsize=(**7**,**6**))  correlation=df.corr()  s.heatmap(correlation,annot=True,cmap='OrRd')  df.boxplot(column='NO2', by='City') # Boxplot for finding No2 concentration values among different cities  df.corr() # Correlation of SO2 and NO2 usually values between -1 and +1  s.pairplot(df,hue = 'SO2') #Pairplot  s.pairplot(df,hue = 'NO2')  fig,axs=plt.subplots(nrows=**2**,ncols=**2**,figsize=(**8**,**6**));  s.countplot(df['SO2'],ax=axs[**0**][**0**])  s.countplot(df['NO2'],ax=axs[**0**][**1**])  s.countplot(df['City'],ax=axs[**1**][**0**])  axs[**0**][**0**].set\_title('SO2',fontsize=**20**)  axs[**0**][**1**].set\_title('NO2',fontsize=**20**)  axs[**0**][**1**].set\_title('City',fontsize=**20**)  plt.tight\_layout()  #Line plot  plt.figure(figsize=(**10**,**4**))  s.lineplot(x='City',y='SO2',data=df)  plt.legend('SO2 Concentration across different Cities')  fig, axs = plt.subplots(ncols=**2**, nrows = **2**, figsize = (**10**,**10**))  s.distplot(df['NO2'], color = 'blue', ax = axs[**0**][**0**])  s.distplot(df['SO2'], color = 'orange', ax = axs[**0**][**1**])  axs[**0**][**0**].set\_title('NO2 Distribution', fontsize = **18**)  axs[**0**][**1**].set\_title('SO2 Distribution', fontsize = **18**)  plt.show()  fig,ax=plt.subplots(figsize=(**20**,**8**)) #scatterplots  ax.scatter(df['NO2'],df['City'])  ax.set\_xlabel('NO2')  ax.set\_ylabel('City')  plt.show()  plt.figure(figsize=(**20**,**10**))  plt.plot(df['SO2'], color='red')  plt.grid(alpha=**0.2**) |

**Different visualizations for our data**

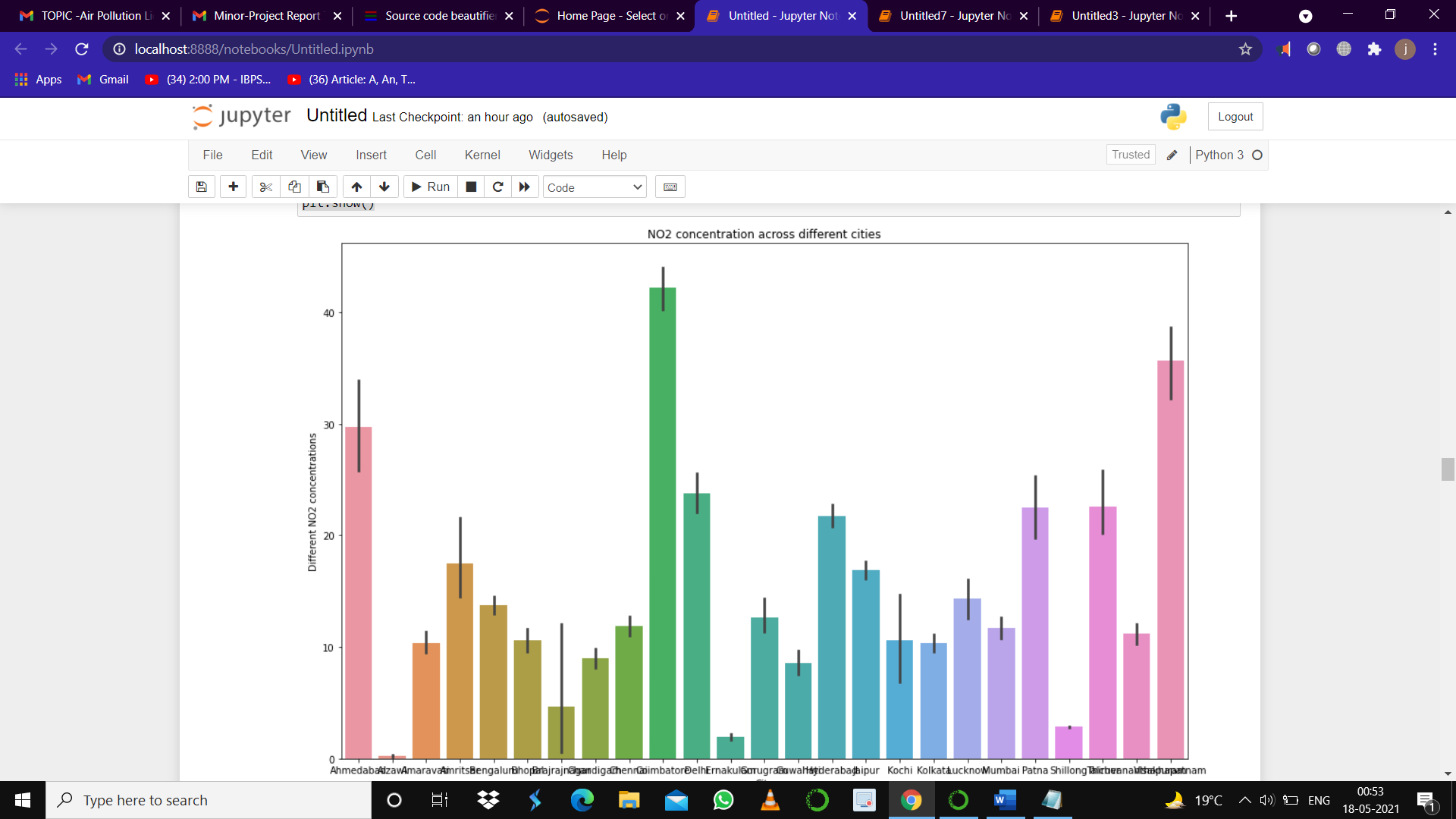


Fig 10: **BAR CHART**

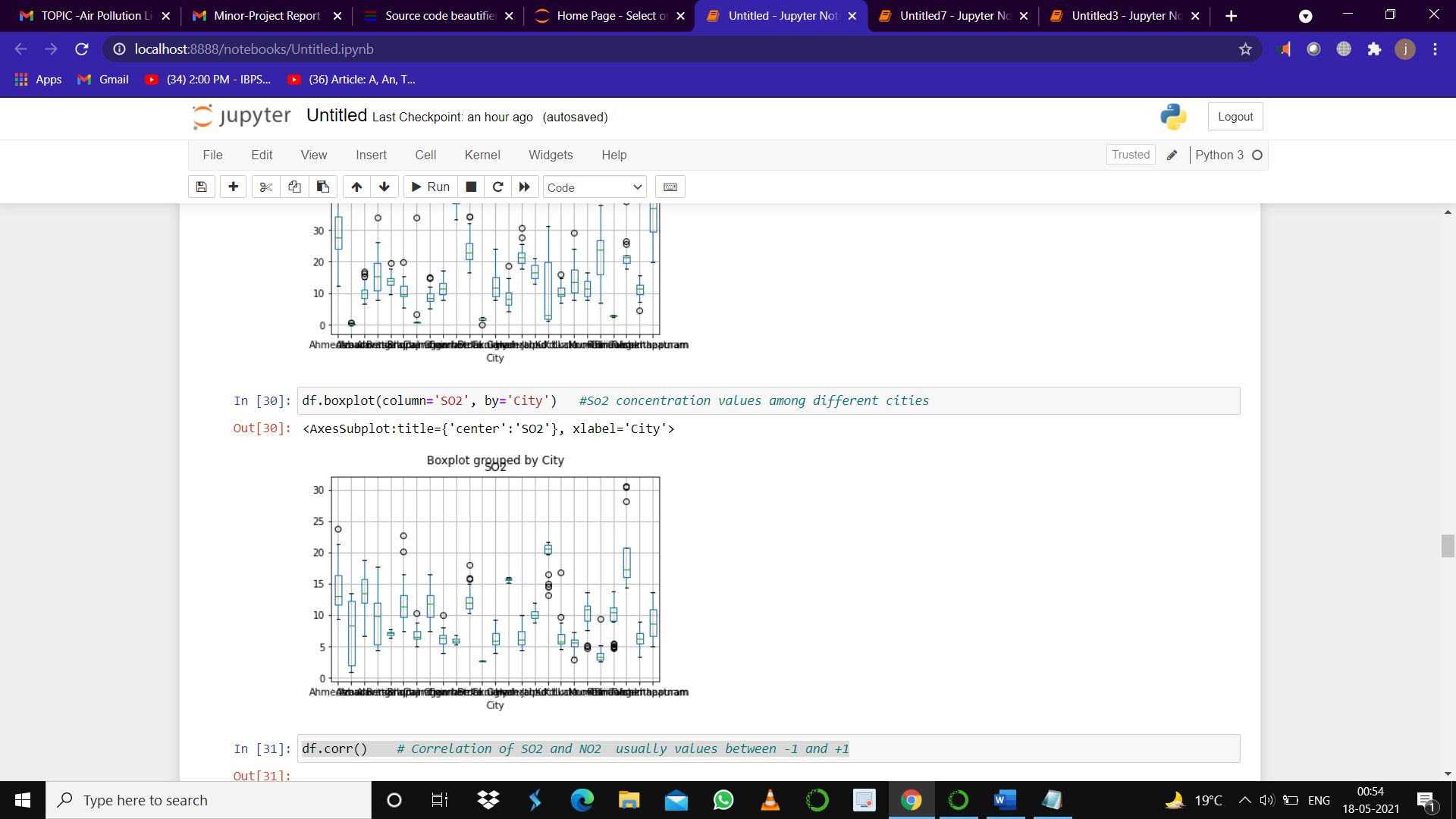
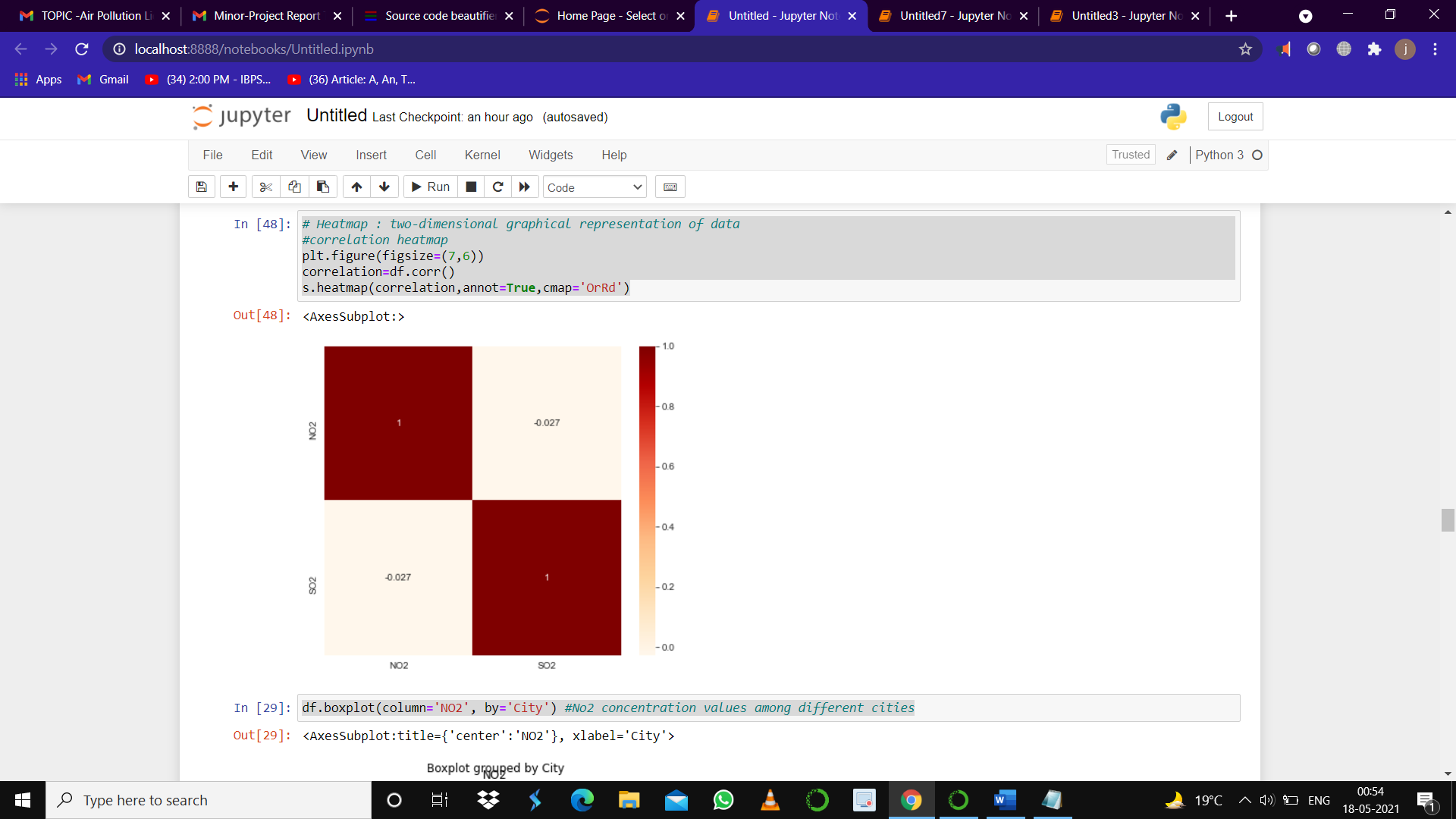


Fig 11: **HEAT MAP** Fig 12: **BOX PLOT**

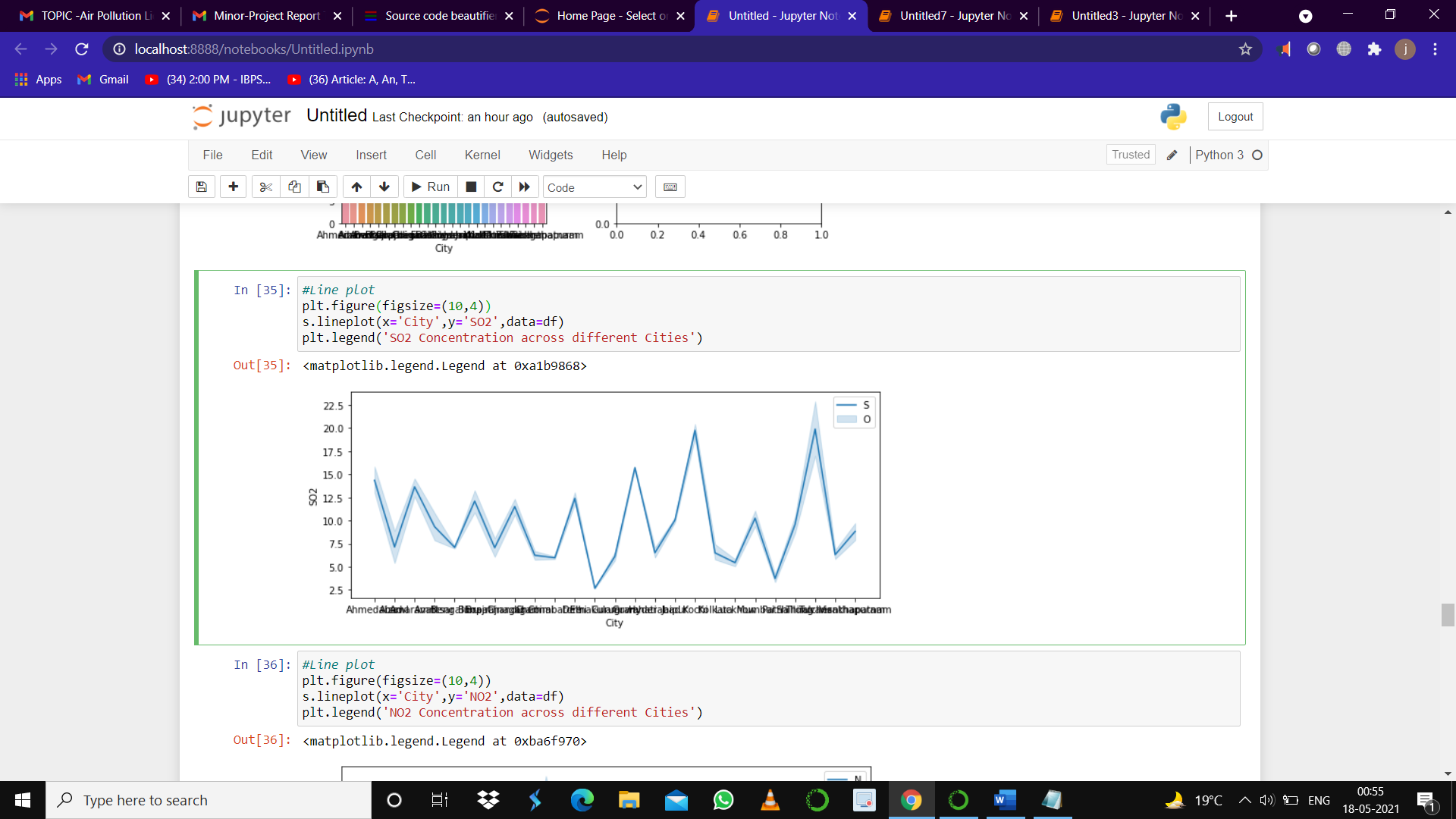
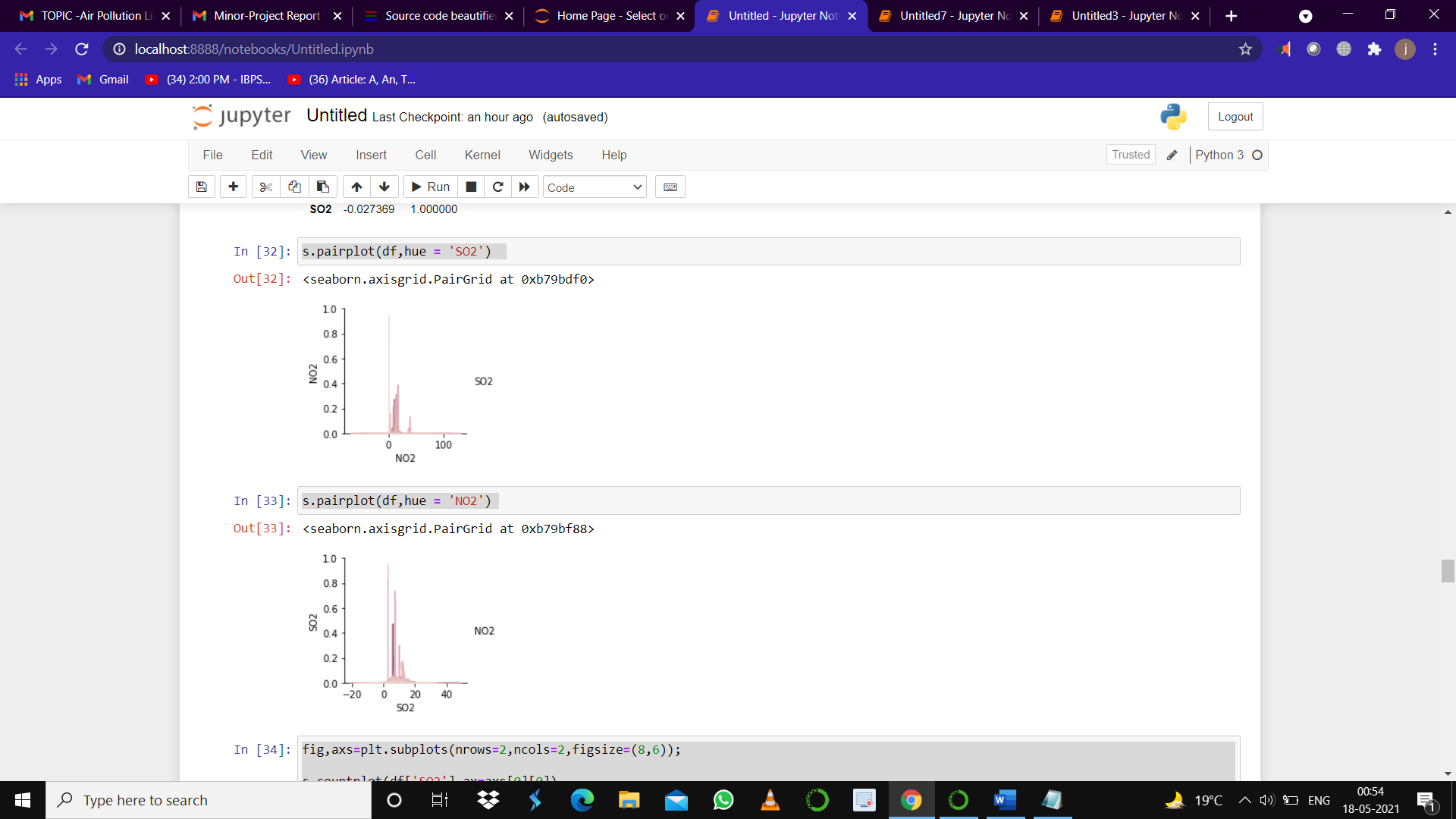


Fig 13: **PAIR PLOT** Fig 14: **LINE PLOT**

Page 21

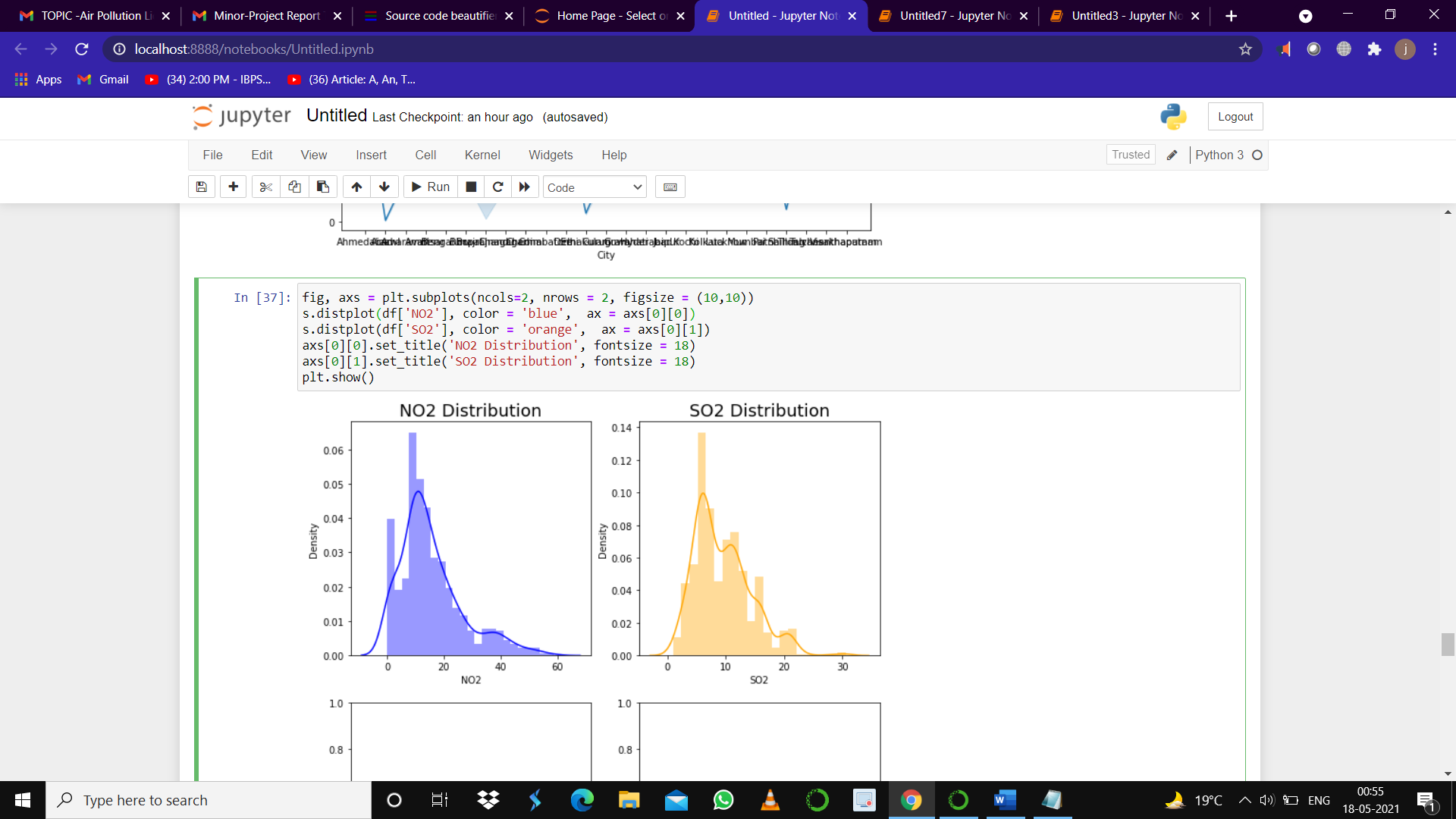


Fig 15: **DISTRIBUTION PLOT**

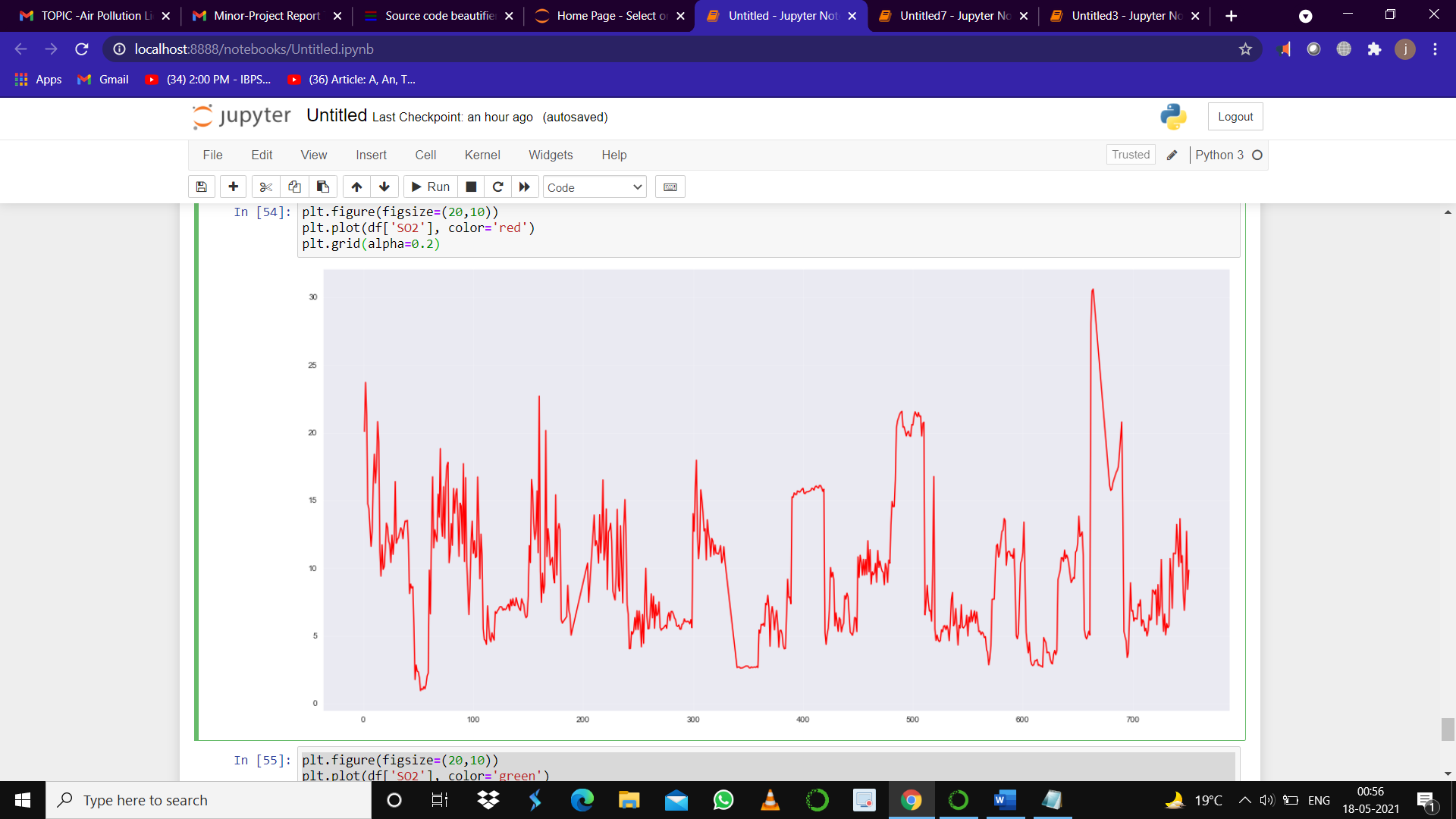


Fig 16: **XY FUNCTION PLOT**

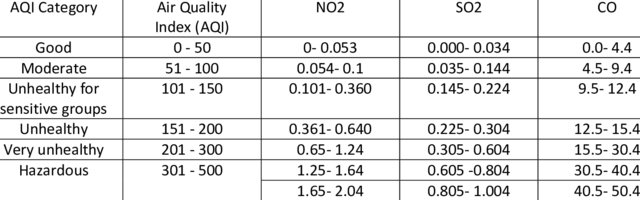


Fig 17: **PREDICTING AIR QUALITY**

Page 22

**CHAPTER 04**

**RESULTS**

**PREDICTIONS ACCORDING TO OUR DATA**

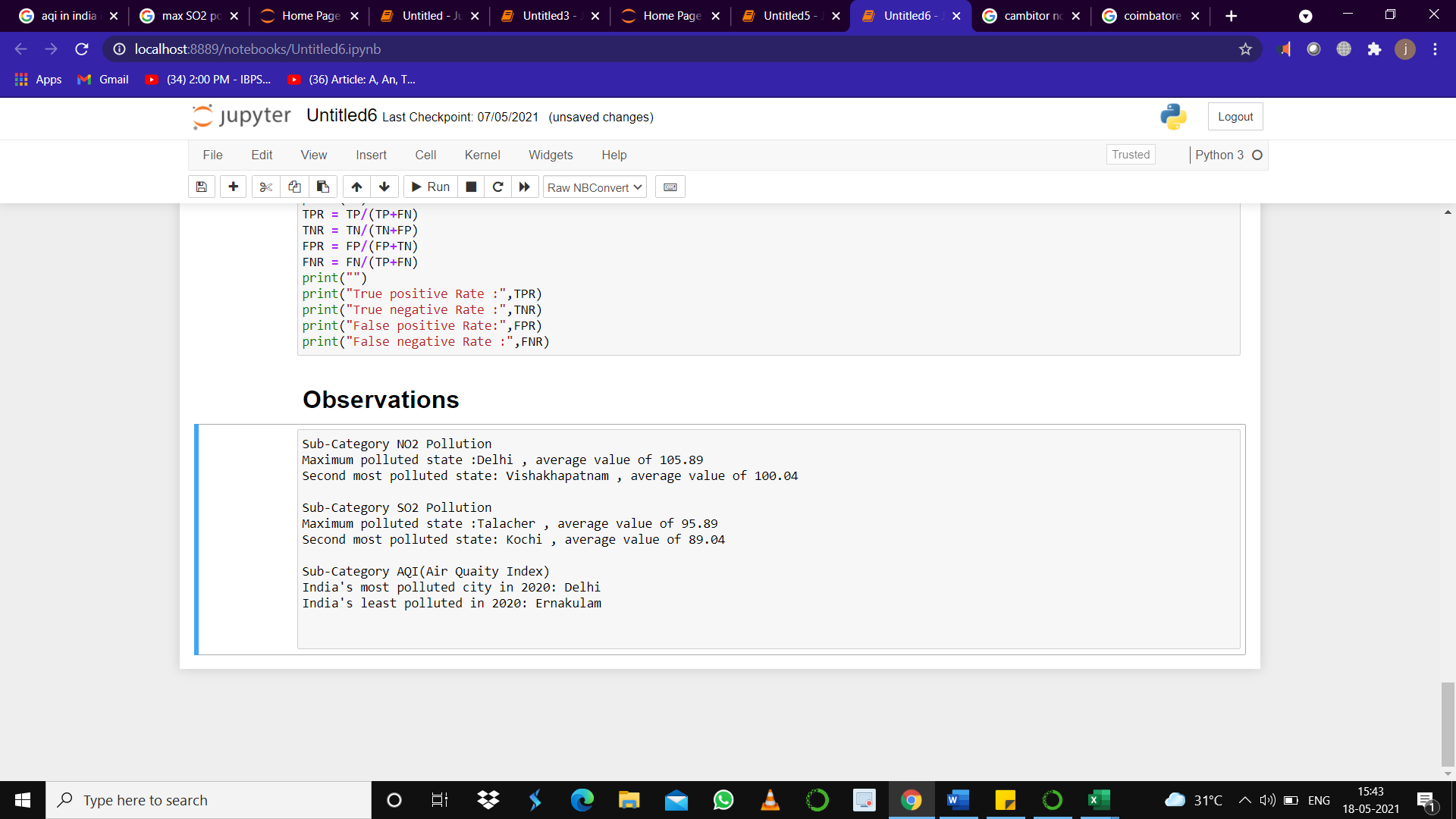


Fig 17 **OBSERVATIONS**

**# Step 3 Performance Measurements of Linear regression**

# 

Fig 18: **LINEAR REGRESSION RESULTS**

Page 23

**# Step 4 Performance Measurements of Logistic regression**

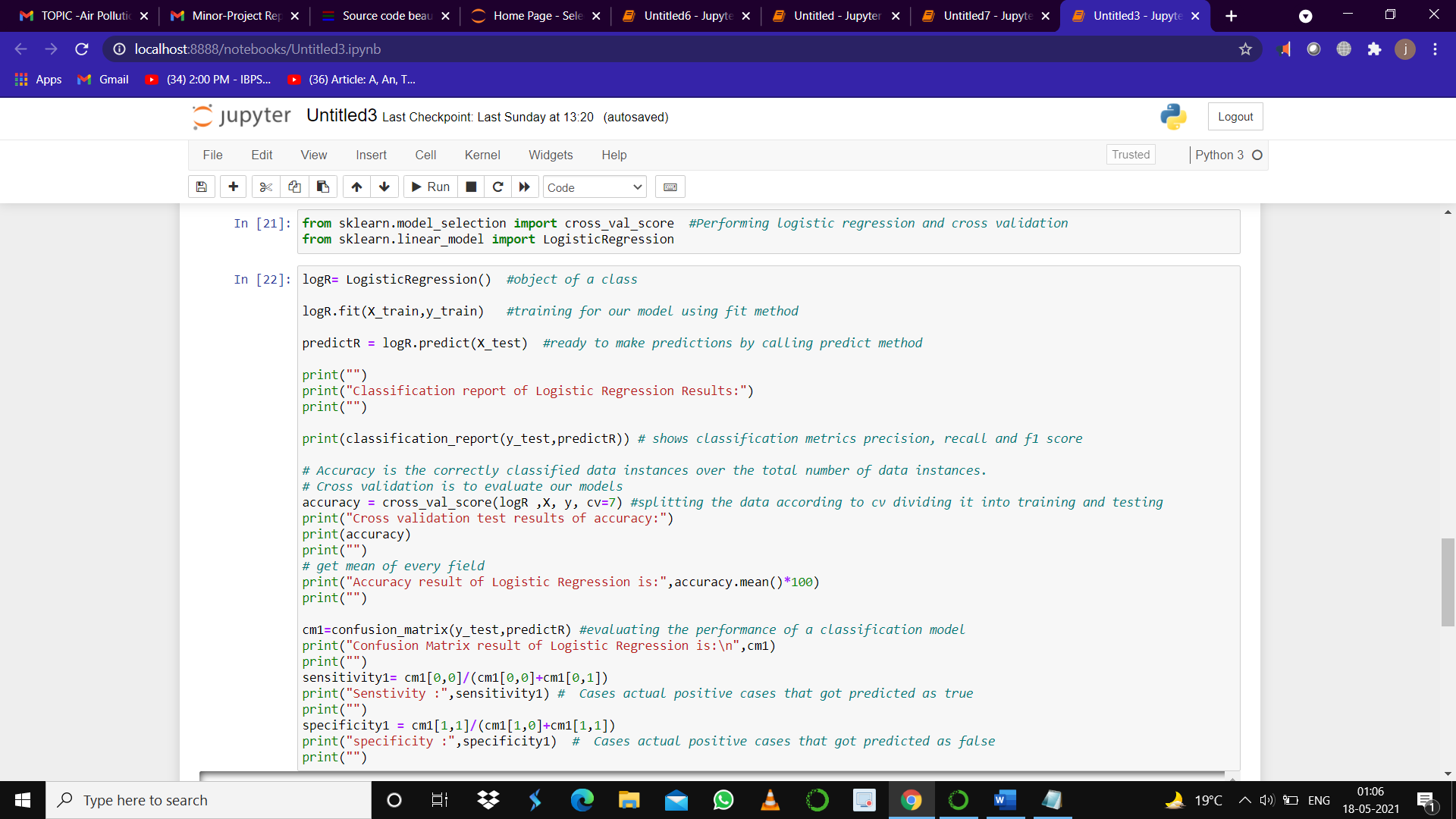
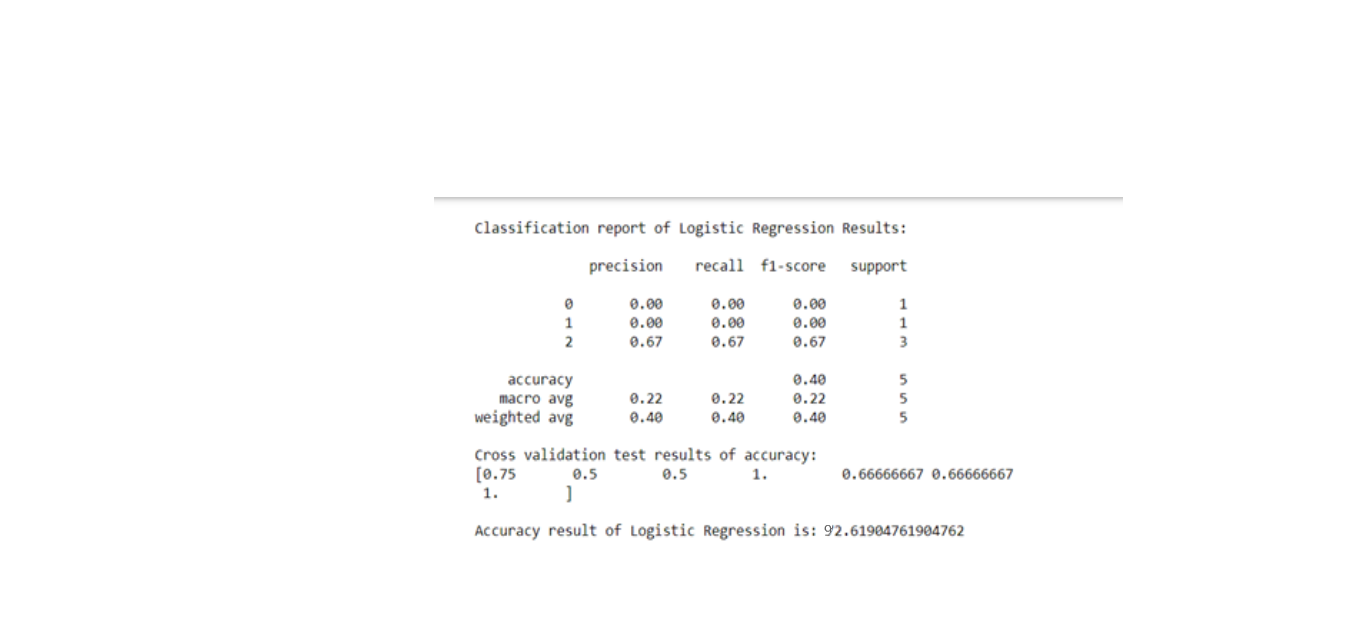
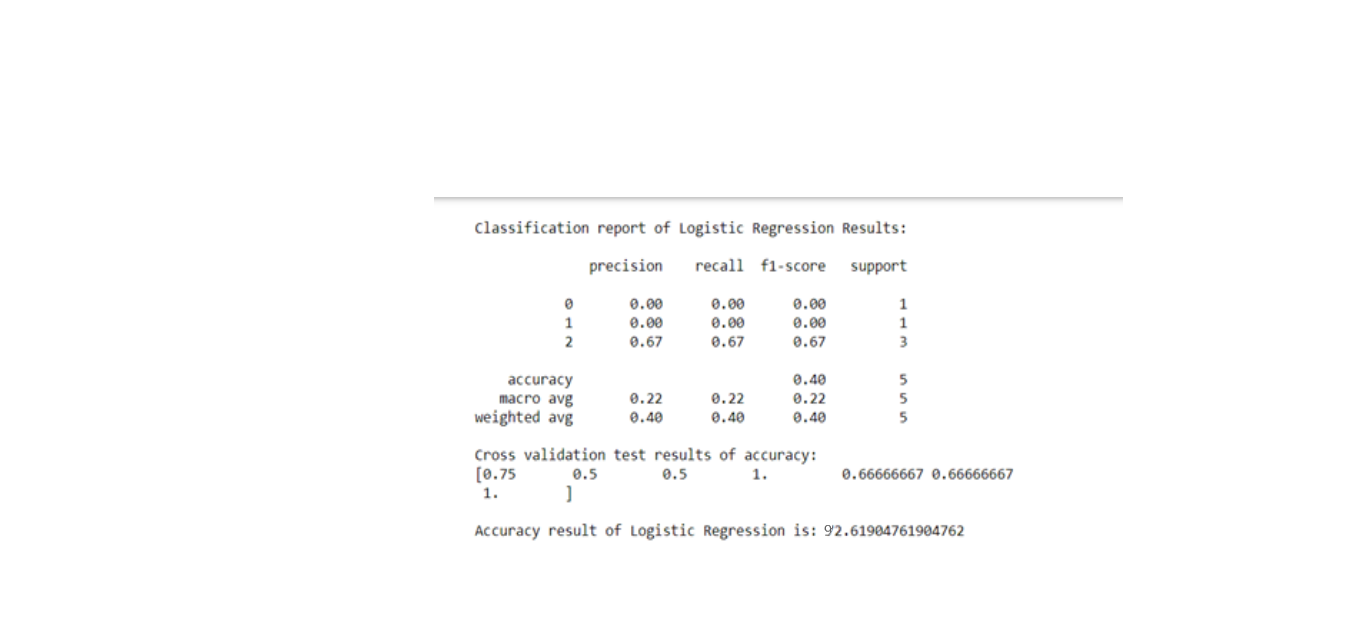


Fig 19: **LOGISTIC REGRESSION ALGORITHM**



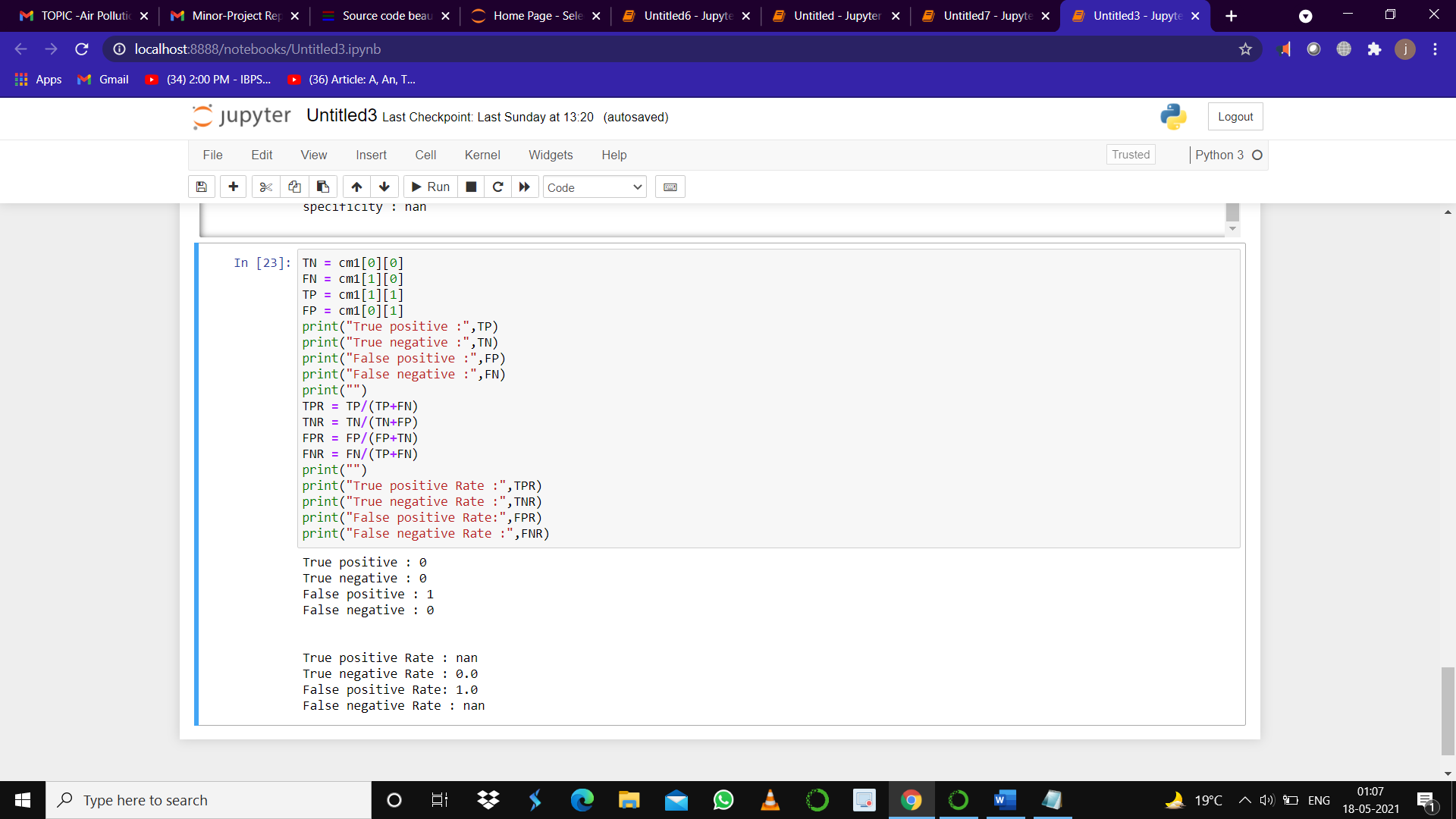
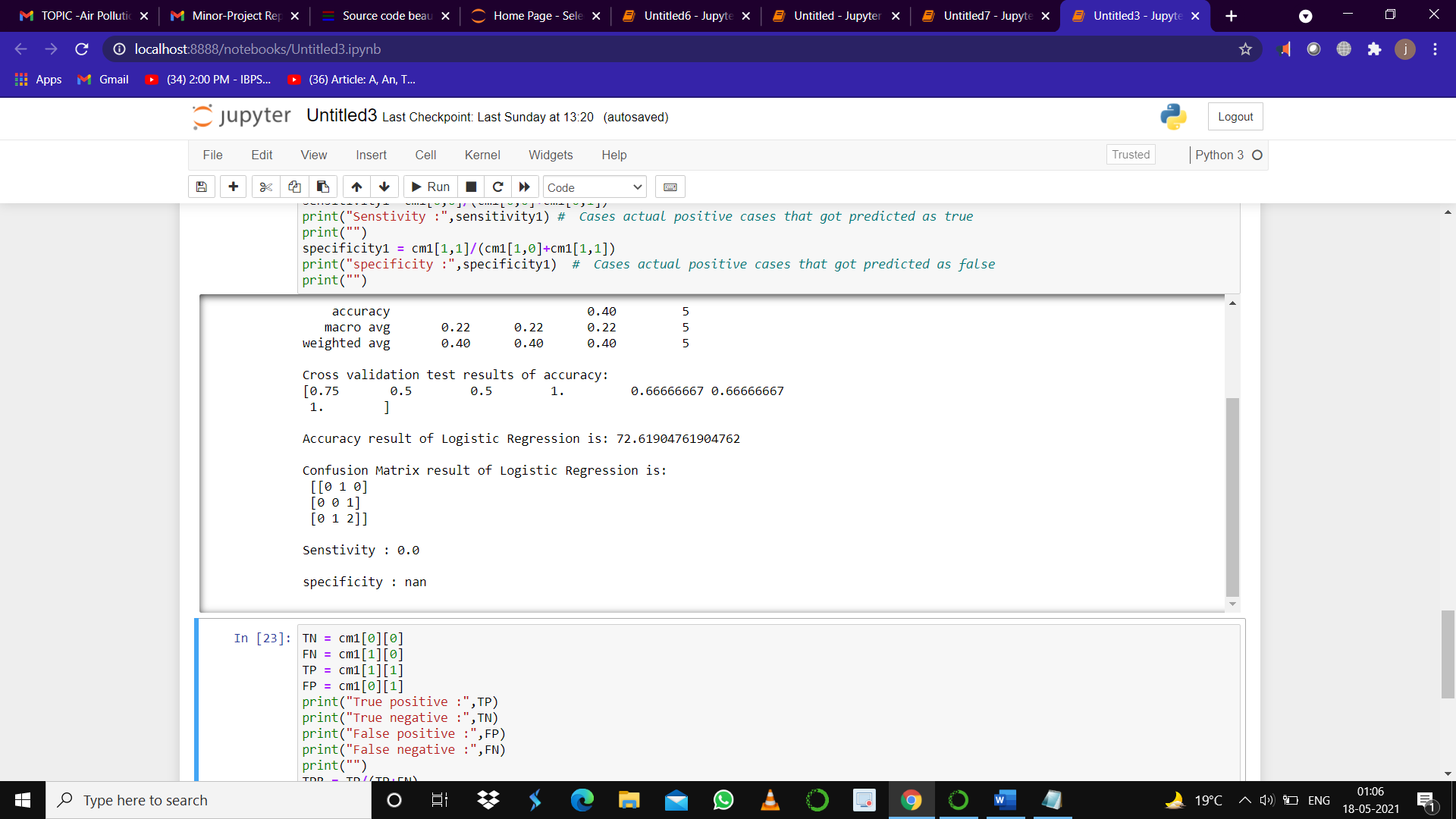


Fig 20: **LOGISTIC REGRESSION RESULTS**

Page24

**# Step 5: Performance Measurements of Support Vector Machine**

# 

# FIG 21: SUPPORT VECTOR MACHINE ALGORITHM

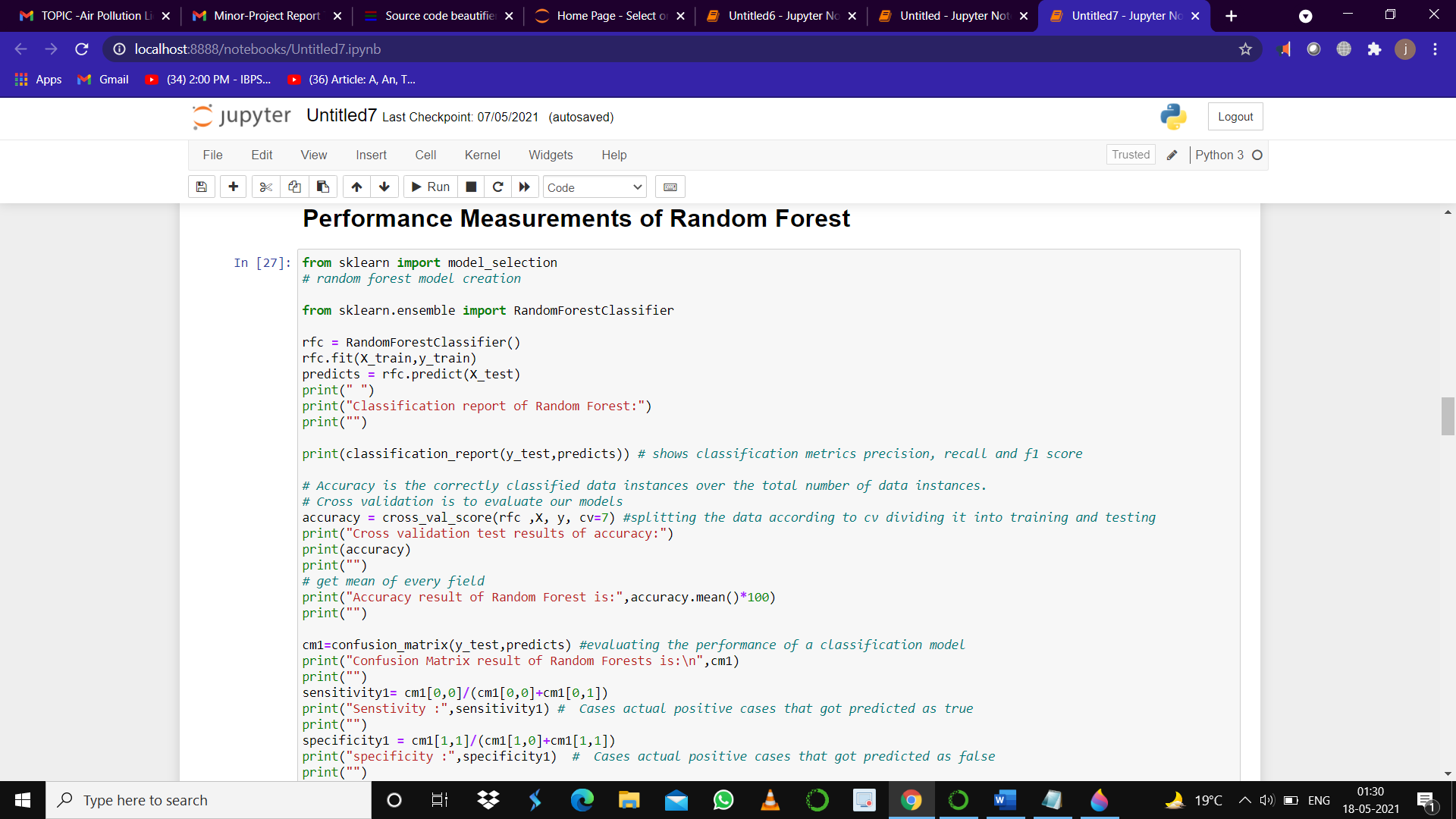
# 

# 

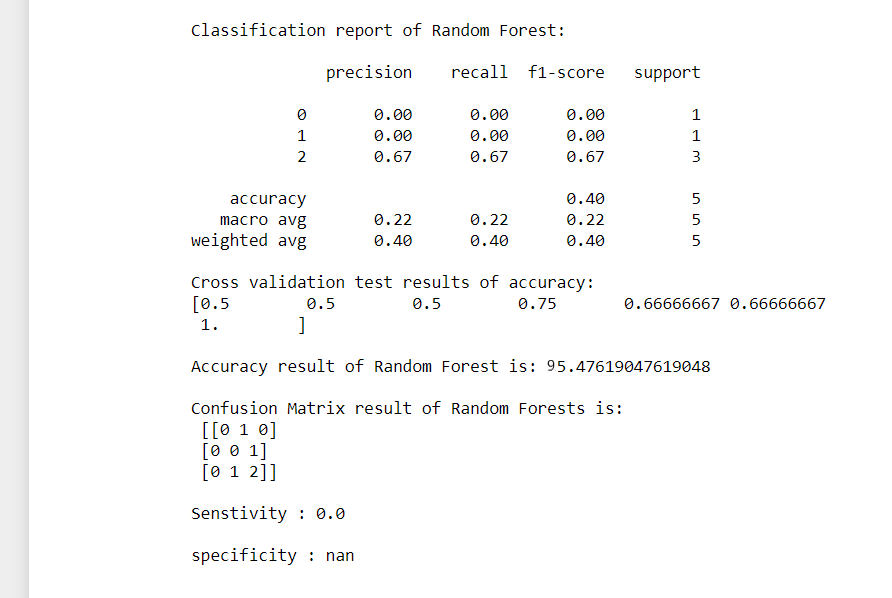
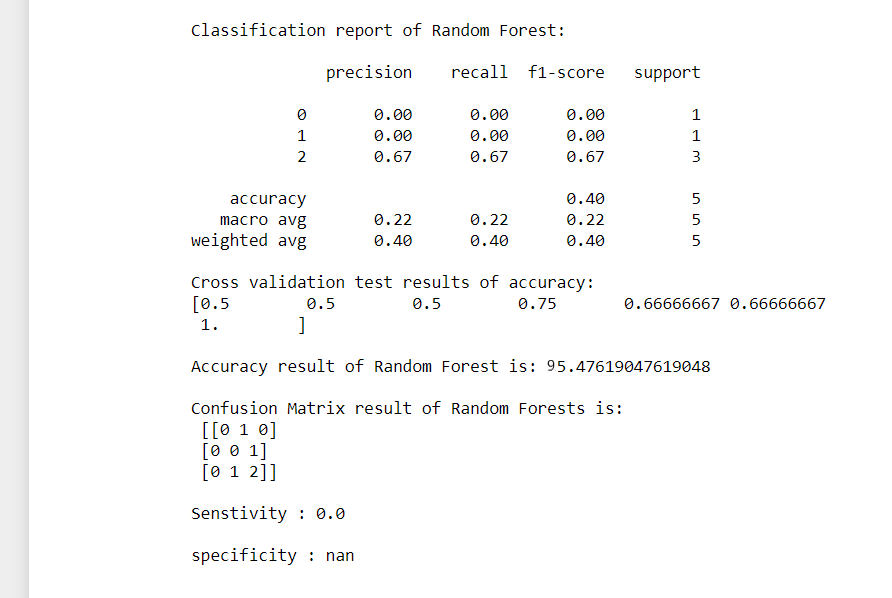
Fig 22: **SUPPORT VECTOR MACHINE** **RESULTS**

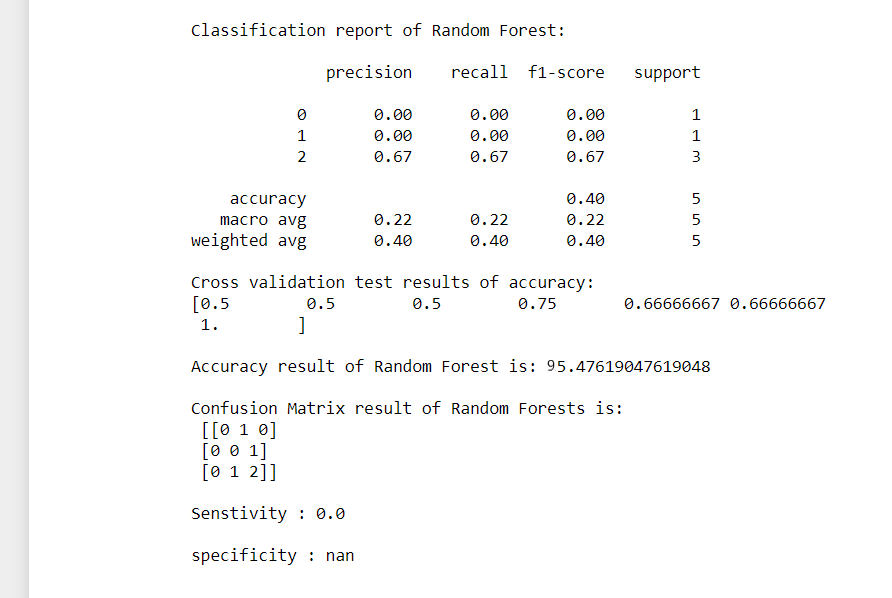
Page 25

**# Step 6: Performance Measurements of Random Forest**



# FIG 22: RANDOM FOREST ALGORITHM





**FIG 23: RANDOM FOREST RESULTS**

Page 26

**4.1 DISCUSSION ON THE RESULTS ACHIEVED**

Table 2: **PCA DATA SET**

|  |  |  |
| --- | --- | --- |
| **METHODS** | **TRAIN ACCURACY** | **TEST ACCURACY** |
| **SVM** | **0.80** | **0.84** |
| **RF** | **0.80** | **0.95** |
| **LR** | **0.80** | **0.92** |

**4.2 CONCLUSION**

In this work, we mainly solve the problem of fast classification of lidar signals for measuring air pollution.

By comparison, it can be concluded that the classification accuracy can be improved by introducing the

statistical features. By comparing the classification performances of SVM, LR, RF, Linear Regression, we

can find that RF performs best on the PCA dataset.

**4.3 APPLICATIONS OF PROJECT**

Since our project is based on how to detect air pollution using machine learning methods, thus it plays a vital role in detecting air pollution and to find out the concentrations of SO2 and NO2, we can detect the quality of air,

We can do pollution monitoring which is much helpful in metropolitan cities.

Reduction of population exposure to air pollution

Address air quality standards

**4.4 FUTURE WORK**

In future the project can be upgraded in the following ways:

Usage of interface sensors and can help us make an IOT based project with help of which we know detail content of all the gases present in air.

Design webpage and upload data on webpage with date and time.

**4.5 LIMITATIONS**

We cannot calculate the exact concentrations of the gases as there is humidity, in atmosphere which definitely effect in finding the exact concentrations of the SO2 and NO2 gases in the air.

The signal waveforms of NO2 are a bit similar, but the difference lies in value, which is also the difficulty of our classification.

Page 27

**REFERENCES**

**1.**Z Xu, Y Cao and Y Kang, "Deep spatiotemporal residual early-late fusion network for city region vehicle emission pollution prediction[J]", Neurocomputing, vol. 2019, pp. 183-199.

**2.**L Zhang, S Sun, B Xing et al., "Using psychophysiological measures to recognize personal music emotional experience[J]", Frontiers of Information Technology & Electronic Engineering, vol. 20, no. 7, pp. 964-974, 2019.

**3.**K Jamieson, M R Gupta, E Swanson et al., "Training a support vector machine to classify signals in a real environment given clean training data[C]", 2010 IEEE International Conference on Acoustics Speech and Signal Processing, pp. 2214-2217, 2010.

**4.**A Gharehbaghi and M Linden, "A deep machine learning method for classifying cyclic time series of biological signals using time-growing neural network[J]", IEEE transactions on neural networks and learning systems, vol. 29, no. 9, pp. 4102-4115, 2017.

**5.**L Jian and W Yan-Qing, "Research on Application of Decision Tree in Classifying Data[C]", 2011 Fourth International Conference on Intelligent Computation Technology and Automation, vol. 1, pp. 1098-1101, 2011.

**6.**Chao Wang, H. Gao and Xudong Zhang, "Radar signal classification based on auto-correlation function and directed graphical model", 2016 IEEE International Conference on Signal Processing Communications and Computing (ICSPCC), pp. 1-4, 2016.

**7.**C Ya-feng, W Xiao-bin, L Qiu-wu et al., "Mobile S02 Differential Absorption Lidar System[J]", ACTA PHOTONICA SINICA, vol. 2017, no. 7.

**8.**Z Xu, Y Kang, Y Cao et al., "Man-machine verification of mouse trajectory based on the random forest model[J]", Frontiers of Information Technology & Electronic Engineering, vol. 20, no. 7, pp. 925-929, 2019.

**9.**Z Xu, Y Kang and W Lv, "Analysis and prediction of vehicle exhaust emission using ANN[C]", 2017 36th Chinese Control Conference (CCC), pp. 4029-4033, 2017.

**10.**Z Xu, Y Kang, Y Cao et al., "Residual Autoencoder-LSTM for City Region Vehicle Emission Pollution Prediction[C]", 2018 IEEE 14th International Conference on Control and Automation (ICCA), pp. 811-816, 2018.

**11.**Z Xu, Y Cao and Y Kang, "Deep spatiotemporal residual early-late fusion network for city region vehicle emission pollution prediction [J]", Neurocomputing, vol. 355, pp. 183-199, 2019.

Page 28