

# **Prediction of Air Quality Index using Machine Learning Models**

## **PROJECT REPORT**

Submitted in partial fulfillment of the requirements for

The award of the degree of

## **BACHELOR OF TECHNOLOGY**

By

**KONDAM SUSHANTH REDDY** **181127**

**PEECHU SAICHARAN REDDY** **181237**

**KAMARAPU VISHNU VARDHAN** **181125**

Under the esteemed guidance of

**Dr.P. Venkateswara Rao**

(Associate Professor, Department of Civil Engineering)



**DEPARTMENT OF CIVIL ENGINEERING**

**NATIONAL INSTITUTE OF TECHNOLOGY**

**WARANGAL -506004**

**(2020-21)**

**NATIONAL INSTITUTE OF TECHNOLOGY  
WARANGAL**



**CERTIFICATE**

This is to certify that project work entitled “**PREDICTION OF AIR QUALITY INDEX USING MACHINE LEARNING MODELS**”, carried out by Kondam Sushanth Reddy (181127), Peechu Saicharan Reddy (181237), Kamarapu Vishnu Vardhan (181125 ) of final year B.Tech Civil Engineering during the year 2021-2022 is a bonafide work submitted to the National Institute of Technology, Warangal, in partial fulfilment of the requirements for the award of the degree of Bachelors of Technology.

---

**Project Supervisor**

**Dr. P. Venkateswara Rao**

Associate Professor,

Department of Civil Engineering,

NIT Warangal.

---

**Head of the Department**

**Dr. P. Rathish Kumar**

Professor,

Department of Civil Engineering,

NIT Warangal.

## Acknowledgement

We consider it a great privilege to express our deep gratitude to many respected personalities who guided, inspired and helped us in the successful completion of our project.

We would like to express our deepest gratitude to our guide, **Dr. P. Venkateswara Rao**, Associate Professor, Department of Civil Engineering, National Institute of Technology, Warangal, for his constant supervision, guidance, suggestions and invaluable encouragement during this project.

We are grateful to **Dr. P. Rathish Kumar**, Head of Department of Civil Engineering, National Institute of Technology, Warangal, for his moral support to carry out this project.

We are very thankful to **Dr. Ajey Patel**, Project Coordinator for his continuous support and guidance during our project throughout the year.

We are very thankful to the Project Evaluation Committee, for their strenuous efforts to evaluate our projects.

<b>KONDAM SUSHANTH REDDY</b>	<b>181127</b>
<b>PEECHU SAICHARAN REDDY</b>	<b>181237</b>
<b>KAMARAPU VISHNU VARDHAN</b>	<b>181125</b>

## **PROJECT WORK APPROVAL FOR B. TECH**

This Project entitled “**Prediction of Air Quality Index using Machine Learning Models**”, submitted by Kondam Sushanth Reddy (181127), Peechu Saicharan Reddy (181237), Kamarapu Vishnu Vardhan (181125 ) is approved for the degree of Bachelor of Technology, Department of Civil Engineering.

### **Examiners:**

---

---

---

### **Supervisor:**

---

**Dr. P. Venkateswara Rao**

(Associate Professor, Department of Civil Engineering)

### **Chairman**

---

**(Dr. P. Rathish Kumar )**

(Professor, Head of Department of Civil Engineering)

**Date:** \_\_\_\_\_

**Place:** \_\_\_\_\_

## **Declaration**

I declare that this written submission represents our ideas in our own words and where other ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all the principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea / data / fact / source in our submission. We understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom permission has not been taken when needed.

---

**KONDAM SUSHANTH REDDY (181127)**

---

**PEECHU SAICHARAN REDDY (181237)**

---

**KAMARAPU VISHNU VARDHAN (181125)**

Date:

Place:

# TABLE OF CONTENTS

## CHAPTER 1 – INTRODUCTION

1.1 General	1
1.2 Introduction to Air Pollution	2
1.3 Introduction to Air Quality Index (AQI)	3
1.4 Proposed Breakpoints for AQI	4
1.5 Health statements for AQI categories	4
1.6 Objectives	5

## CHAPTER 2 – LITERATURE REVIEW

2.1 Previous studies	6
2.2 Introduction to Machine Learning	8
2.3 Applications of Machine Learning	8
2.4 Advantages of Machine Learning	9
2.5 Case study regarding Delhi AQI	10

## CHAPTER 3 – METHODOLOGY

3.1 Requirement Analysis	12
3.2 Proposed Methodology	13
3.3 Implementation	14
3.3.1 Data Collection	14
3.3.2 Importing libraries	16
3.3.3 Data Pre Processing	18
3.3.4 Calculation and preparation of AQI Data Frame	20
3.3.5 Split the Data set	20
3.3.6 Applying ML Models	21
3.3.7 Model Evaluation	23
3.3.8 Time Series Forecasting	25

## CHAPTER 4 – DISCUSSION

4.1 Experimental Results and Analysis	26
4.2 Graphical representation of results by Time Series Forecasting	27
4.3 Conclusion	28
4.4 Project Work Plan	29
4.5 References	30

## **LIST OF FIGURES**

1. Formation of an Aggregated Air Quality Index
2. Applications of Machine Learning
3. Graphical representation of Table 3
4. Flow chart of proposed Methodology
5. Website of pollutants concentration data
6. Website of meteorological data
7. Raw excel data file of pollutant concentration
8. Raw excel data file for pollutant concentration and meteorological data
9. Importing libraries
10. Reading csv file for case 1 and case 2
11. Finding null values and data visualization of null values
12. Filling Null Values using KNN method
13. Preparation of AQI data frame
14. Splitting into training and testing data set
15. Flow chart of support vector machine(SVM)
16. Flow chart of Random Forests(RF)
17. Applying ML Models and Model Evaluation
18. Output of Time Series Forecasting
19. Graphical representation of results by Time series Forecasting of case-1 and case-2
20. Graphical representation of results by Time series Forecasting of case-1 vs case-2

## **LIST OF TABLES**

1. AQI Scale
2. Health Statements for AQI Categories
3. Sub indices of air pollutants for 7days
4. Results of case-1
5. Results of case-2

## **Abstract :**

Air pollution is increasing as a result of human activities such as industrialization and urbanization. As a result, air pollution levels have risen, causing serious worry among emerging countries. One of the most important natural resources for the existence and survival of all species on this planet is air. CO, PM<sub>2.5</sub>, PM<sub>10</sub>, O<sub>3</sub>, SO<sub>2</sub>, and NO<sub>2</sub> are the most common air pollutants. Meteorological variables such as atmospheric wind speed, wind direction, relative humidity, and temperature influence the concentration of air pollutants in ambient air. The older approaches such as probability, statistics, and others were tried to estimate air quality, but because those methods are difficult to predict, Machine Learning (ML) is a better approach. As a result, this project employs Machine Learning algorithms and methodologies to create an Air Quality Index (AQI) model. Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Linear Regression, Gradient boost regressor, and XG boost regressor are used in the work. We developed a time series forecasting model (FB prophet) to forecast the air quality index based on historical data from previous years and projecting the accuracy over a specific future year using the Root Mean Square Error as a parameter of evaluation. The meteorological department can use this tool to forecast air quality.

## **Key-words :**

Air pollution, Air quality index(AQI), Meteorological variables, Machine learning, Accuracy, Forecasting.



# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 General**

Many countries have witnessed increasing urbanization in recent years. It is one of the main causes of air pollution since increased transportation facilities send more pollutants into the atmosphere, and industrialization is another major driver of air pollution. CO, PM<sub>2.5</sub>, PM<sub>10</sub>, O<sub>3</sub>, SO<sub>2</sub>, and NO<sub>2</sub> are the primary pollutants. They have a significant impact on human health. Meteorological data also plays an important role in finding AQI. Temperature, wind and relative humidity are among the meteorological parameters taken into account by our model. Because of temperature inversion, when warm air above colder air serves as a lid, limiting vertical mixing and keeping the cooler air at the surface, temperature has an impact on air quality. Pollutants produced into the air by vehicles, fireplaces, and industry are trapped near the ground by the inversion. Pollution is diluted significantly by wind speed. Strong winds remove contaminants, but mild winds create stagnant conditions that allow pollutants to accumulate over a large region. The dispersion of contaminants may be affected by humidity.

The Air Quality Index is used to assess air quality. Individual air pollutant levels (e.g. SO<sub>2</sub>, CO, PM<sub>10</sub>) are converted into a single number, which provides a simple and lucid depiction of air quality for citizens. AQI also connects air quality levels to health warnings.

The quality of air we breathe largely impacts the quality of life of people. Therefore, demand for predicting air quality has become a necessity so that we take precautionary measures to control the air pollution levels. We predict using Machine learning models. Machine learning is a method of data analysis that automates analytical model Building. With minimum human intervention, they learn from data, see patterns, and make judgments. Supervised Learning, Unsupervised Learning, and Reinforcement Learning are three types of learning algorithms used in machine learning, which falls under artificial intelligence. We adopted a supervised learning strategy in the suggested research. Linear Regression, K Nearest Neighbor, Support vector machine (SVM), Decision tree, Gradient boost regressor, XG boost regressor, and Random Forest are examples of supervised learning techniques.

We gathered the information from an Indian government database that provides pollution concentrations at various locations around the country. We start by determining the AQI for the region by calculating the individual pollutant index for each accessible data point. We developed a model for predicting the air quality index. We can backtrack the principal pollution-causing pollutant by estimating the air quality index. We can anticipate future values using a time series forecasting model (FB prophet) with the help of existing data for a specified area.

This project was done because it is one specific contribution towards the issue of predicting air quality index. The data set we considered consists of major air pollutants as well as the meteorological data.

## **1.2 Introduction to Air pollution**

Due to rapid urbanization, many environmental hazards took place in the 20th century, including rise in air pollution levels.

**Air pollution:** The Phenomenon of change in concentration of variable gases in nature and exceeds the standard permissible limits which leads to an adverse effect on human, animal and plant life along with damage to properties is considered as air pollution.

Long-term and short-term health impacts can be caused by air pollution. Air pollution has been demonstrated to have a greater impact on the elderly and young children.

The proximity of activities that produce high quantities of pollution, such as:

1. Heavy traffic on the roads, automobiles that do not meet pollution standards.
2. Polluting smoke from thermal (coal-based) power plants and other factories.
3. Sites of uncontrolled construction or demolition.
4. Domestic energy demands, such as cooking, can be met with biomass fuel.
5. Firecrackers are going off.
6. Burning household waste, hospital waste, electronic waste, crop residues, and other materials.

**Air pollutants:** The physical chemical or Biological agent which is a solid liquid or gaseous matter that exists either naturally or man made to impact adverse effects on environment by changing composition of air is called as air pollutant.

Pollutants are divided into two categories: main and secondary.

**Primary air pollutants:** The air pollutant released from source and in the same original form it causes adverse effects to humans plants animals buildings without modification of their original form is primary pollutant.

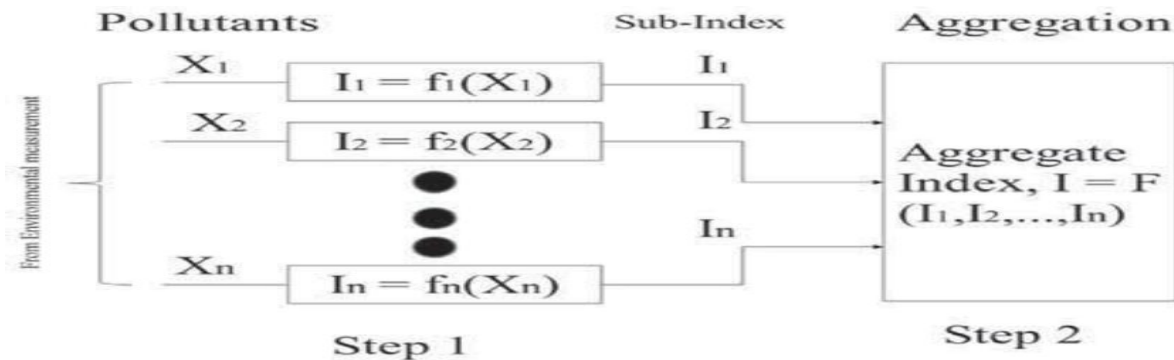
- CO, CO<sub>2</sub>, CH<sub>4</sub>, and VOCs are examples of carbon compounds.
- NO, N<sub>2</sub>O, and NH<sub>3</sub> are nitrogen compounds.
- Compounds of sulphur, such as H<sub>2</sub>S and SO<sub>2</sub>.
- Chlorides, fluorides, and bromides are examples of halogen compounds.
- Particulate Matter (PM or "aerosols") is a type of solid or liquid particle that is classified into groups based on their aerodynamic diameter (PM<sub>2.5</sub>, PM<sub>10</sub>).

**Secondary air pollutants:** The air pollutant which is formed due to chemical reaction between two primary pollutants or a primary air pollutant and a natural component in the atmosphere and then reached to ground level to cause air pollution is secondary air pollutant.

- NO produces  $\text{NO}_2$  and  $\text{HNO}_3$ .
- Ozone ( $\text{O}_3$ ) is produced by photochemical reactions between nitrogen oxides and volatile organic compounds (VOCs).
- Sulfuric acid droplets produced by  $\text{SO}_2$  and nitric acid droplets produced by  $\text{NO}_2$
- Sulfates and nitrates aerosols (e.g., ammonium (bi)sulfate and ammonium nitrate) are generated when sulfuric acid droplets react with  $\text{NH}_3$  and nitric acid droplets react with  $\text{NH}_3$ .
- In gas-to-particle processes, organic aerosols are generated from VOCs.

### **1.3 Introduction to Air Quality Index(AQI)**

A mathematical method that turns the weighed values of individual air pollution-related elements (such as pollutant concentrations) into a single number or collection of numbers is known as an air quality index. The ultimate result is a set of rules (i.e., most equations) for converting parameter values into a more basic form using numerical manipulation.



**Figure1: Formation of an Aggregated Air Quality Index**

In a nutshell, an AQI is useful for:

- (i) The general people to understand air quality in a simple manner.
- (ii) Politicians to take prompt action.
- (iii) Decision maker who can assess the current state of affairs and devise pollution control plans.
- (iv) A government officer to investigate the impact of regulatory actions.
- (v) A scientist who conducts scientific studies on the basis of data on air quality

## **1.4 Proposed Breakpoints for AQI**

**Table 1: AQI Scale**

AQI Category	PM 10 (24 Hours)	PM 2.5 (24 Hours)	NO2 (24 Hours)	O3 (8 Hours)	CO (8 Hours)	SO2 (24 Hours)	NH3 (24 Hours)	Pb (24 Hours)
Good (0-50)	0-50	0-30	0-40	0-50	0-1	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2	41-80	201-400	0.5-1
Moderately polluted (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2
Poor (201-300)	251-350	91-120	181-280	169-208	10.0-17	381-800	801-1200	2.1-3
Very Poor (301-400)	351-430	121-250	281-400	209-748	17-34	801-1600	1200-1800	3.1-3.5
Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

## **1.5 Health Statements for AQI Categories**

**Table 2: Health Statements for AQI Categories**

AQI	Associated Health Impacts
<b>Good(0–50))</b>	Minimal Impact
<b>Satisfactory (51–100)</b>	May cause minor breathing discomfort to sensitive people
<b>Moderately polluted (101–200)</b>	May cause breathing discomfort to the people with lung disease such as asthma and discomfort to people with heart disease, children and older adults
<b>Poor (201–300)</b>	May cause breathing discomfort to people on prolonged exposure and discomfort to people with heart disease
<b>Very Poor (301–400)</b>	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases
<b>Severe (401-500)</b>	May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity

## **1.6 Objectives**

### **Aim of the project:**

The goal is to create a machine learning model for predicting air quality. To anticipate the AQI value, we will use historical AQI and weather data. Multivariate analysis, missing value treatments, data validation, and data cleaning/preparation are all performed on the air quality data set. Then, using supervised machine learning techniques such as Linear Regression, Random Forest, Decision Tree, and Support Vector Machines, air quality is predicted. The smaller the gap between the predicted and actual AQI score, the better the model. We shall compare the above-mentioned models in order to determine which is the most efficient. Evaluation criteria such as Precision, Recall, and F1 Score were used to compare the performance of various machine learning algorithms.

### **Scope of the project:**

The project can be used to track and anticipate air quality in a certain location, allowing us to try to manage pollutants by taking the required precautions to avoid poor air quality. Inform the public about the current state of air quality and the health consequences of exposure to air pollution. Evaluate the AQI prediction model using data from a few major cities, and rank cities/towns for action prioritization based on the AQI measure.

## **CHAPTER 2**

### **LITERATURE REVIEW**

This chapter covers the following topics: Introduction to Machine Learning, Types of Machine Learning, Machine Learning Applications, and the working principles of the various Machine Learning algorithms employed in this study. In addition, the article focused on several Machine Learning research conducted in a civil engineering context.

#### **2.1 Previous studies :**

##### **Madhuri VM, Samyama Gunjal GH and Savitha Kamalapurkar (2020) :**

This Research paper titled “Air Pollution Prediction Using Machine Learning Supervised Learning Approach” was developed with an intent to predict air quality index using machine learning models. He discussed meteorological parameters affecting air quality index such as atmospheric wind speed, wind direction, relative humidity, and temperature. He talked about how meteorological factors including atmospheric wind speed, wind direction, relative humidity, and temperature affect the air quality index. He claims that older approaches such as probability, statistics, and others were tried to estimate air quality, but because those methods are difficult to predict, Machine Learning (ML) is a better approach. “The approaches he used are Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest Method (RF)” to predict the air Quality index and used Root Mean Square Error to predict the accuracy”.

##### **Mrs. A. GnanaSoundariMtech ,Mrs. J. GnanaJeslin M.E, and Akshaya (2019) :**

This Research paper titled “Indian Air Quality Prediction And Analysis Using Machine Learning” have predicted India's air quality using machine learning algorithms to estimate an area's air quality index (AQI). The Air Quality Index is a common metric for determining air quality. The agencies keep track of gas concentrations like SO<sub>2</sub>, NO<sub>2</sub>, and CO<sub>2</sub>. They developed a model to predict the air quality index based on previous year's historical data and forecasting for a certain year in the future as a gradient descent boosted multi variable regression issue. Cost Estimation for Predictive Problem was used to improve the model's efficiency. They claim that by given historical pollutant concentration data, this model can accurately estimate the air quality index of a whole country, state, or confined region.

##### **Kostandina Veljanovska and Angel Dimoski (2018) :**

This Research paper titled “Air quality index prediction using simple machine learning algorithms” was discussed about “three supervised learning techniques, k-nearest neighbour (KNN), Support Vector Machines (SVM), and Decision Tree (DT), and one unsupervised approach, Neural Network, were used to predict air quality index (NN). The data set was created using a model from the Ministry of Environment and Physical Planning of the Republic of Macedonia's official website. SO<sub>2</sub> (sulphur dioxide), NO<sub>2</sub> (nitrogen dioxide), O<sub>3</sub> (ozone), CO (carbon monoxide), suspended particulates PM<sub>2.5</sub> (fine particles),

and PM<sub>10</sub> (particulate matter) were all regarded major pollutants (large particles). The data set includes 365 samples (one for each day of 2017), 51 of which had a High Air Pollution Index, 94 had a Medium Air Pollution Index, and the remaining 220 had a Low Air Pollution Index". In this experiment, the samples were categorized as "low," "medium," and "high" levels of air pollution using a neural network. The six contaminants were utilized as inputs, with the air quality index (output) being divided into three categories: low, medium, and high. The neural net's hidden layer was made out of 10 neurons since it produced the least optimal error. Different values of k in the range of 1-21 were used for the K-NN classifier.

### **Nidhi Sharma , ShwetaTaneja , Vaishali Sagar and Arshita Bhatt (2018) :**

This Research paper titled "Forecasting air pollution load in Delhi using data analysis tools" had gone over the complete data analysis of air pollutants from 2009 to 2017, as well as suggested a critical observation of the air pollution trend in Delhi, India, from 2016 to 2017". "Sulfur Dioxide (SO<sub>2</sub>), Nitrogen Dioxide (NO<sub>2</sub>), Suspended Particulate Matter (PM), Ozone (O<sub>3</sub>), Carbon Monoxide (CO), and Benzene have all been anticipated to have future tendencies. They estimated the future levels of the pollutants listed previously on the basis of historical records utilizing data analytics Time series Regression forecasting. The monitoring stations of AnandVihar and Shadipur in Delhi are being investigated based on the findings of this study. The results demonstrate a significant increase in PM<sub>10</sub> concentrations, as well as increases in NO<sub>2</sub> and PM<sub>2.5</sub>, indicating increased pollution in Delhi. CO levels are expected to drop by 0.169 mg/m<sup>3</sup>, while NO<sub>2</sub> levels are expected to rise by 16.77 g/m<sup>3</sup>, Ozone levels are expected to rise by 6.11 mg/m<sup>3</sup>, Benzene levels are expected to drop by 1.33 mg/m<sup>3</sup>, and SO<sub>2</sub> levels are expected to rise by 1.24 g/m<sup>3</sup>".

### **Aditya C R, Chandana R Deshmukh, Nayana D K and Praveen Gandhi Vidyavastu (2018) :**

This Research paper titled "Detection and Prediction of Air Pollution using Machine Learning Models" was on the basis of a data set including atmospheric conditions in a particular city, machine learning methods were used to detect and forecast the PM<sub>2.5</sub> concentration level. They also forecasted PM<sub>2.5</sub> concentrations for a specific day. They employed the Logistic Regression technique to classify the air as contaminated or not polluted, and then the Auto Regression algorithm to forecast the future value of PM<sub>2.5</sub> based on historical records.

## **2.2 Introduction to machine learning:**

Machine Learning is a subset of AI that is based on providing data to computers and allowing them to learn and explore on their own. It is involved with finding patterns in large amounts of data.

**Supervised learning :** Learning through examples of which we know the desired output. They generate two kind of results:

**Regression:** Output is continuous Ex: price, temperature

**Classification:** Output is a discrete variable Ex: cat/dog.

**Unsupervised Learning :** When an algorithm learns from example data of which we don't know the desired output, leaving to the algorithm to determine the data patterns on its own.

**Reinforcement learning :** In Reinforcement Learning, our machines work as an agent in a virtual environment and perform possible actions in the environment.

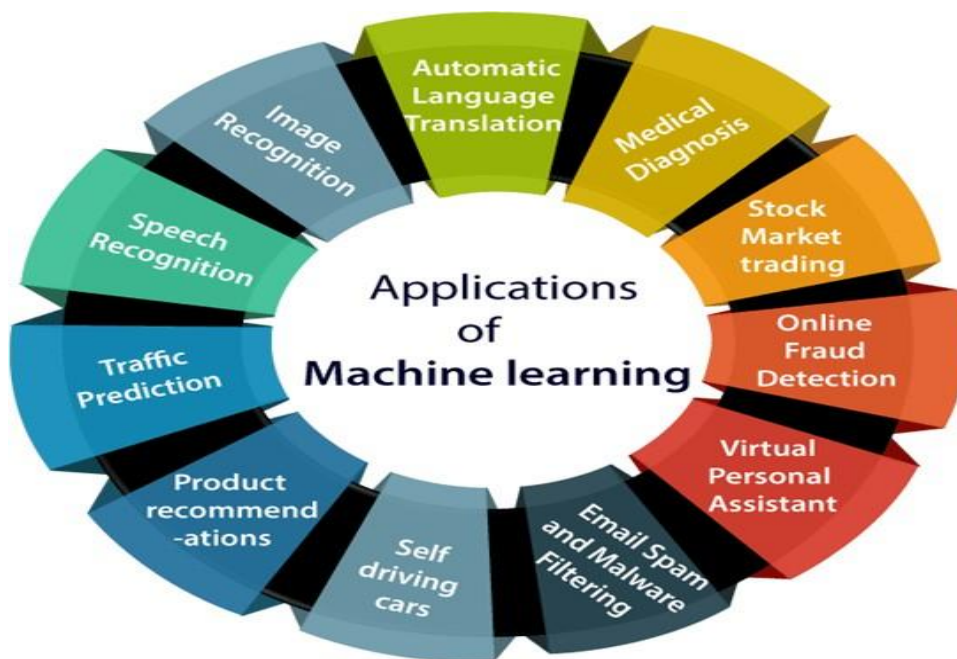
## **2.3 Applications of Machine Learning**

Machine learning is a hot topic in today's technology, and it's just getting hotter. We use machine learning in our daily lives without even realizing it, such as Google Maps, Google Assistant, Alexa, and so on.

**1.Civil Engineering Applications** : In civil engineering, they will be applied in colorful ways to finish planning, construction, and analogous tasks. For illustration, if you want the machine learning or AI result to use available data and design a plan, you need to tell it where to look for the applicable specs, how to find it, and what to do with it. You also need to define the parameters for that design, so the system doesn't produce outside the needed boundaries and tolerances.

**2.Non Civil Engineering Applications** : There are various applications of Machine learning in our day to day life. Some of them are Image recognition , speech recognition , self driving cars , Product recommendations and Email spam detection.





**Figure 2: Applications of Machine Learning**

## **2.4 Advantages of Machine Learning**

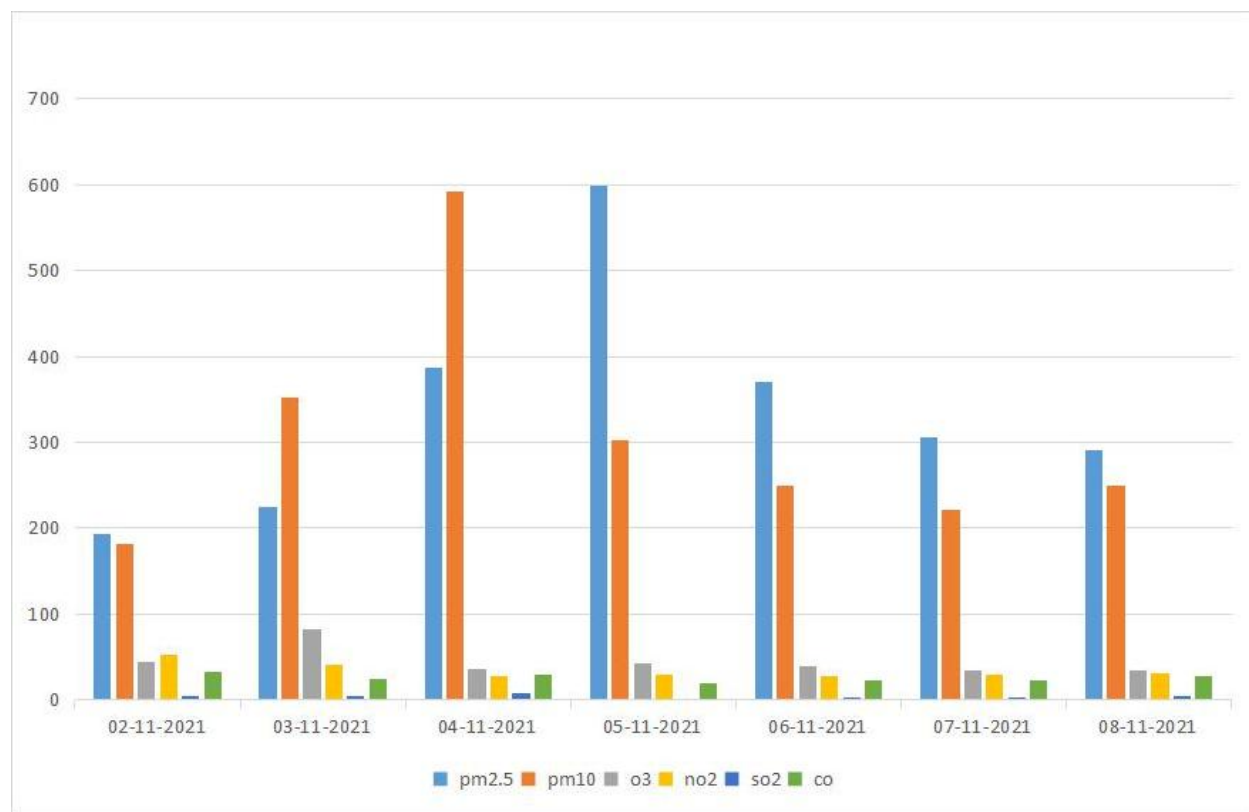
- Machine Learning easily identifies trends and patterns by reviewing large volumes of data that would not be apparent to human beings.
- Machine learning improves the algorithms on its own so we don't need to babysit the project every step of the way. It allows machines to make predictions because it gives them the ability to learn.
- Machine Learning algorithms gain experience, they keep improving in accuracy and efficiency which lets them make better decisions and algorithms learn to make more accurate predictions faster.
- Machine Learning algorithms are good at handling data that are multi dimensional and variety, they can do this in dynamic or uncertain environment.

## 2.5 Case study regarding delhi AQI

### AQI comparision of Delhi before and after Diwali

date	pm2.5	pm10	o3	no2	so2	co
02-11-2021	193	182	45	53	5	32
03-11-2021	224	352	82	41	4	24
04-11-2021	387	592	36	27	8	30
05-11-2021	599	303	42	29	2	19
06-11-2021	370	250	40	27	3	23
07-11-2021	305	221	35	30	3	22
08-11-2021	291	249	35	31	5	27

**Table 3: Subindices of air pollutants for 7days**



**Figure 3: Graphical representation of table 3**

The most talked topic is about the rise in the air quality index of Delhi. Now let us dig into the problem: what are the reasons for the rise in pollutant concentration and the pollutant that is the cause for the rise in AQI.

Now let us consider the AQI of Delhi before and after Diwali. Considering the pollutant concentrations from November 2nd to 5th we can infer that there is a gradual increase in the PM concentrations (PM<sub>10</sub>, PM<sub>2.5</sub>). This is Because of burning crackers amidst the festival and after Diwali it has seen a slight decrease in the pollutant concentration.

### **The main reasons for the rise in the AQI of Delhi is:**

1. Diwali fireworks
2. urban emissions from vehicles
3. waste burning, stubble burning
4. Emissions from industries, power plants
5. construction activities.

### **The steps taken by the government of Delhi in this crisis :**

1. Trucks are not allowed to enter Delhi unless they are carrying critical goods.
2. To stop construction and demolition activities in NCR
3. State Governments shall allow Work from Home (WFH) for at least 50% of their staff in offices in NCR
4. State Governments shall encourage at least 50% of staff working in private establishments in NCR are also allowed to work from home.
5. Until further orders, all public and private schools, colleges, and education institutions in the NCR will be shuttered, allowing only online education..

During the last ten years, Delhi has taken many steps to reduce air pollution in the city. However, there is still work to be done to reduce air pollution even further. Existing things must be strengthened and scaled up. Government actions alone are insufficient. Participation of people is important in order to have a tangible impact on pollution reduction. . The use of public transportation should be encouraged. Providing a sufficient number of feeder buses that run at the necessary frequency at Metro stations helps boost the use of Metro rail.

### **Methodology**

**3.1 Requirement Analysis:** In order to complete the project the requirements are :

#### **Software Requirements:**

##### **Libraries used:**

- 1.Numpy
- 2.Pandas
- 3.Matplot
- 4.Sklearn
- 5.Seaborn

##### **Other Requirement:**

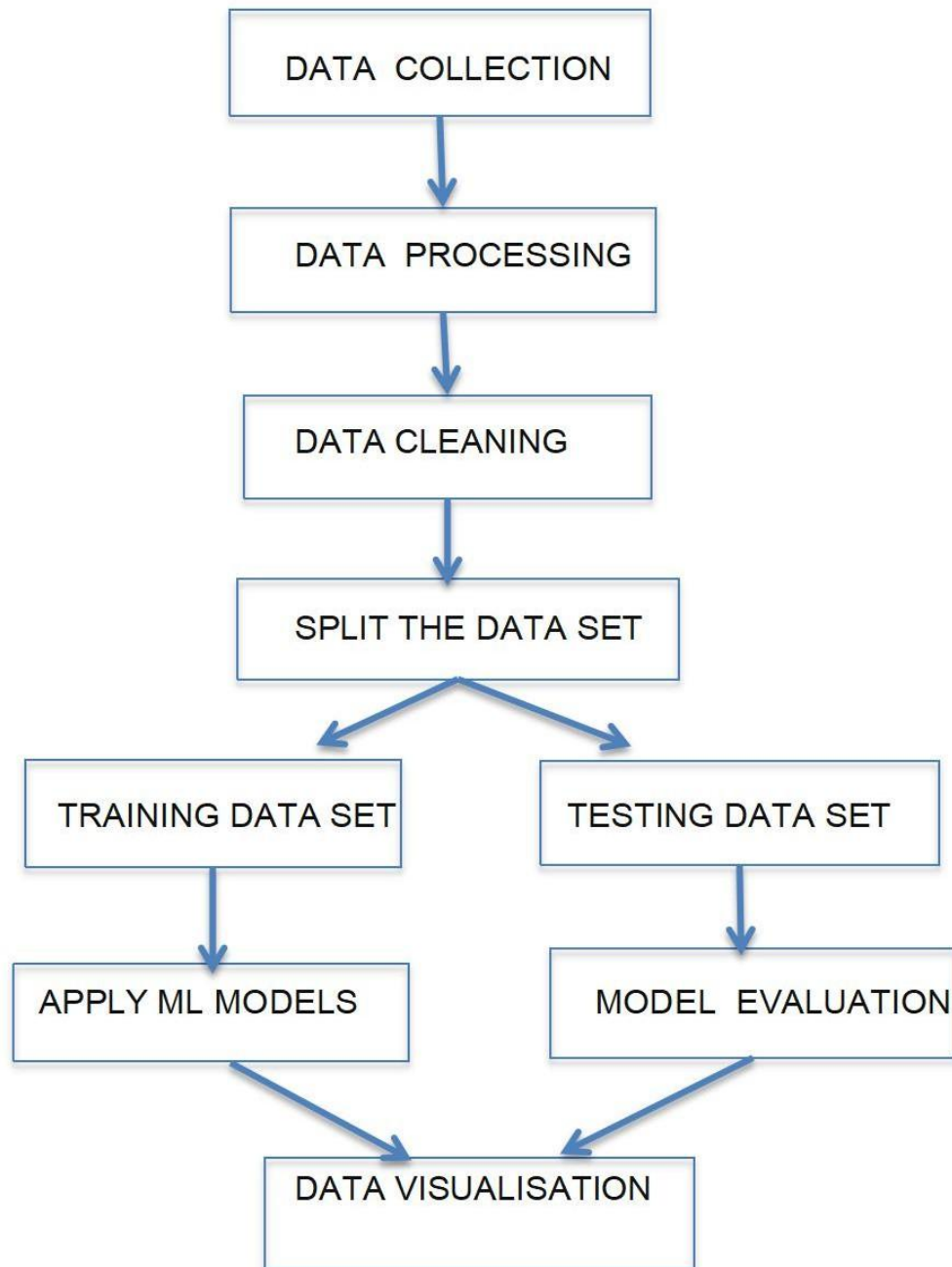
- 1.Anaconda Platform ( Jupyter Notebook)
- 2.Python 3.6.0

#### **Hardware Requirements:**

- 1.Microsoft Windows 10
- 2.Processor: Intel ® Core (TM) i5 -6200U CPU @2.30GHz
- 3.2.40GHz3.Ram : 4 GB and above
- 4.Disk Space : 1 TB

### **3.2 Proposed methodology :**

Project is followed in this order and carried out in Google Colab(Language :python)



**Figure 4 : Flow chart of proposed methodology**

### **3.3 Implementation:**

#### **3.3.1 Data collection :**

To estimate the air quality index of a certain place, we require the pollutant concentrations of all gases present in that region, as well as meteorological data, which has a significant impact on AQI value changes.

The data set we used was obtained from two sources they are:

1. <https://aqicn.org/data-platform/register/> (For pollutant concentration data)

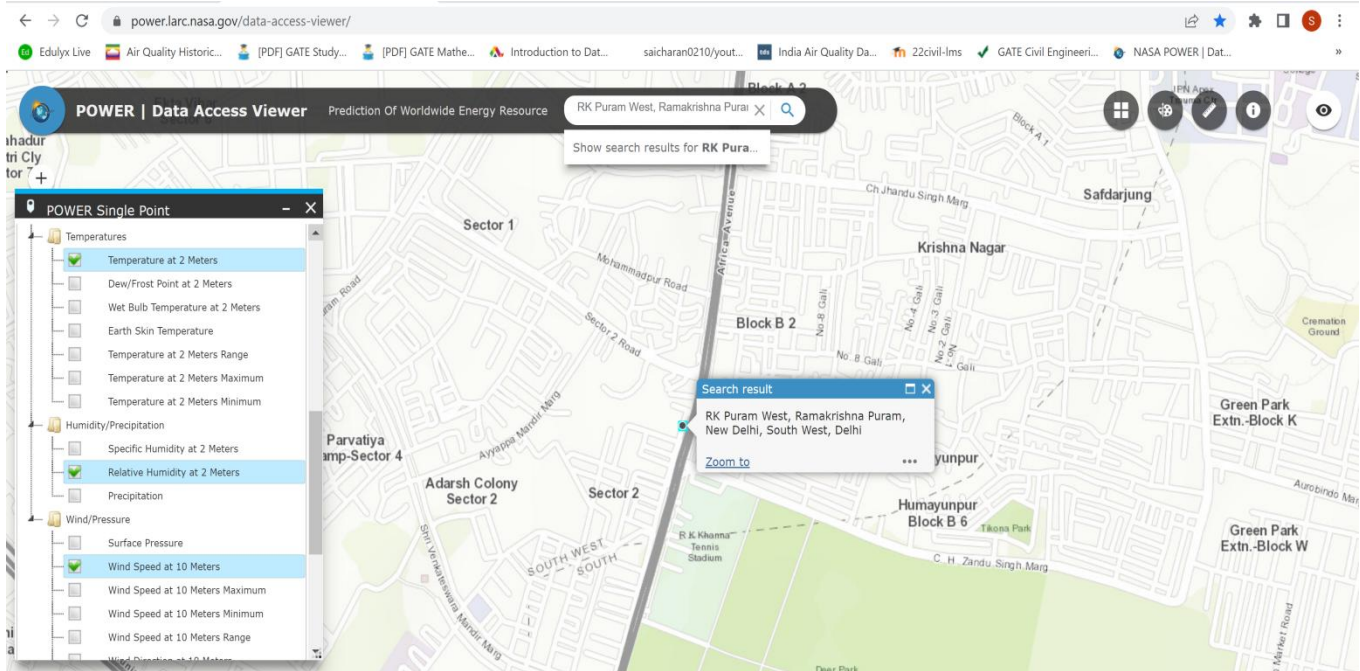
This is the first site picture where we collect pollutants concentration data of a particular place which is required.

The screenshot shows a web browser window with the URL [aqicn.org/data-platform/register/](https://aqicn.org/data-platform/register/). The page has a search bar with the text "Enter the name of a station, eg Beijing". Below the search bar, the text "R.K. Puram, Delhi, Delhi, India" is entered. To the right of the search bar is a magnifying glass icon. Below the search bar, there are five colored dots (green, yellow, orange, red, purple). Below the dots, the text "Or select on the station your previously visited or nearest stations:" is displayed. Below this text, there are five buttons: "Sanathnagar, Hyderabad", "Hyderabad US Consulate", "Bollaram Industrial Area, Hyderabad", "Zoo Park, Bahadurpura West, Hyderabad", and "Central University, Hyderabad". Below the buttons, there is a horizontal bar with ten colored segments representing AQI ranges: 0~25 (green), 25~50 (green), 50~75 (green), 75~100 (yellow), 100~125 (yellow), 125~150 (orange), 150~175 (red), 175~200 (red), 200~300 (purple), 300~400 (purple), and >400 (dark purple). Below the bar, the text "R.K. Puram, Delhi, Delhi, India past 99 months daily average AQI" is displayed. Below the text, there is a blue button with the text "Download the CSV data for R.K. Puram, Delhi, Delhi, India". Below the button, the text "Data Sources" is displayed. Below the text, there is a link: "Delhi Pollution Control Committee (Government of NCT of Delhi) - CPCB - India Central Pollution Control Board".

**Figure 5 : Website of pollutants concentration data**

2. <https://power.larc.nasa.gov/data-access-viewer/> (For meteorological data)

This is the Second site picture where we collect Meteorological data of a particular place from NASA website.



**Figure 6 : Website of meteorological data**

### **Case1 : AQI prediction by considering only pollutants concentration :**

	A1								
	A	B	C	D	E	F	G	H	
1	date	pm25	pm10	o3	no2	so2	co		
2	01-01-2018	351	440	20	40	12	23		
3	02-01-2018	386	282	22	32	9	18		
4	03-01-2018	289	211	22	42	8	17		
5	04-01-2018	244	276	13	34	8	19		
6	05-01-2018	282	260	32	37	9	22		
7	06-01-2018	277	236	36	35	14	11		
8	07-01-2018	248	195	35	37	19	7		
9	08-01-2018	233	239	40	40	23	9		
10	09-01-2018	266	197	39	36	25	9		
11	10-01-2018	233	165	37	36	28	9		
12	11-01-2018	223	240	34	46	33	20		
13	12-01-2018	243	278	35	44	19	22		
14	13-01-2018	272	231	38	41	22	11		
15	14-01-2018	244	256	33	39	29	7		
16	15-01-2018	257	233	43	33	24	19		
17	16-01-2018	230	293	21	37	11	18		
18	17-01-2018	300	350	110	37	17	21		
19	18-01-2018	331	329	78	40	20	18		
20	19-01-2018	311	252	8	32	22	8		
21	20-01-2018	262	225	60	43	23			
22	21-01-2018	248	305	94	43	18	9		

**Figure 7 : Raw excel data file of pollutants concentration**



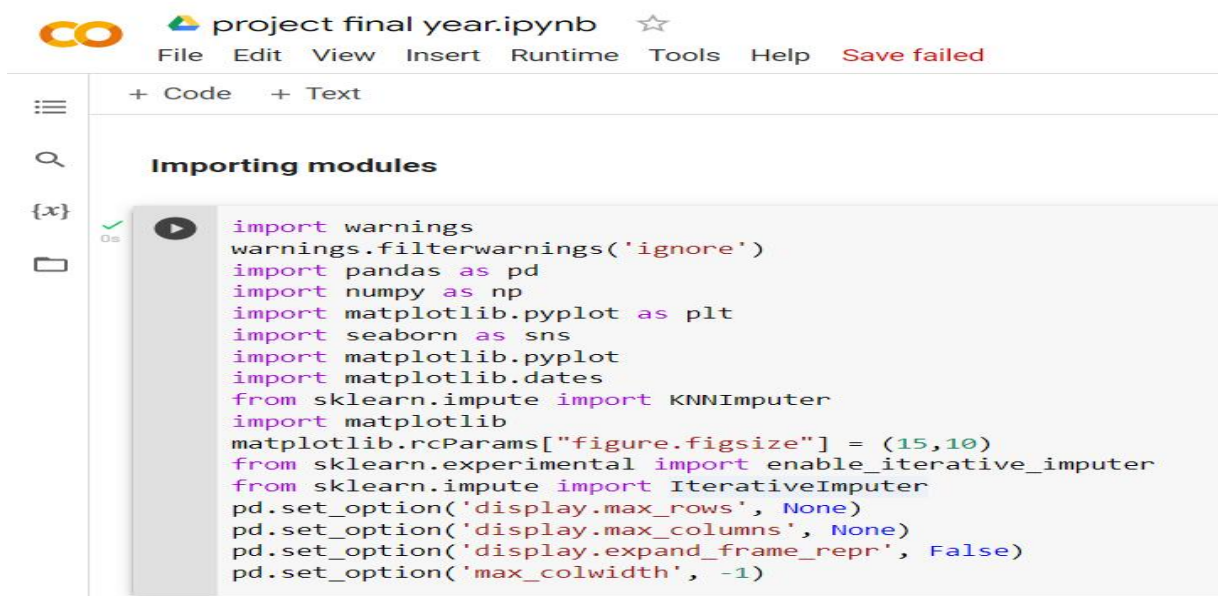
## Case-2 : AQI prediction by considering both pollutants concentration and meteorological data :

A1												fx	date
	A	B	C	D	E	F	G	H	I	J	K		
1	date	pm25	pm10	o3	no2	so2	co	T2M	RH2M	WS10M			
2	01-01-2018	351	440	20	40	12	23	12.07	34.56	2.41			
3	02-01-2018	386	282	22	32	9	18	10.98	38.5	2.58			
4	03-01-2018	289	211	22	42	8	17	11.01	39.62	3.02			
5	04-01-2018	244	276	13	34	8	19	12.51	33.56	2.09			
6	05-01-2018	282	260	32	37	9	22	13.02	30.94	1.95			
7	06-01-2018	277	236	36	35	14	11	11.71	40.5	3.15			
8	07-01-2018	248	195	35	37	19	7	11.65	41.19	3.8			
9	08-01-2018	233	239	40	40	23	9	11.98	38.25	3.07			
10	09-01-2018	266	197	39	36	25	9	11.94	38.06	3.41			
11	10-01-2018	233	165	37	36	28	9	12.55	32.69	4.23			
12	11-01-2018	223	240	34	46	33	20	13.25	29.88	4.14			
13	12-01-2018	243	278	35	44	19	22	15.13	29	3.35			
14	13-01-2018	272	231	38	41	22	11	14.88	32.38	2.88			
15	14-01-2018	244	256	33	39	29	7	14.12	31.38	2.86			
16	15-01-2018	257	233	43	33	24	19	13.23	31.31	4.02			
17	16-01-2018	230	293	21	37	11	18	14.54	24.56	3.18			
18	17-01-2018	300	350	110	37	17	21	15.02	27	2.11			
19	18-01-2018	331	329	78	40	20	18	16.6	26.75	2.53			
20	19-01-2018	311	252	8	32	22	8	15.33	33.81	3.34			
21	20-01-2018	262	225	60	43	23		13.98	26.31	3.8			
22	21-01-2018	248	305	94	43	18	9	14.24	24.44	2.39			
23	22-01-2018	296	198	16	45	6	22	15.6	21.62	1.7			
24	23-01-2018	225	136	31	28	5	22	12.98	41.81	3.03			

**Figure 8 : Raw excel data file of pollutants concentration and meteorological data**

### **3.3.2 Importing libraries :**

First we are importing required libraries to get start with the code (python language).



```
import warnings
warnings.filterwarnings('ignore')
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.pyplot
import matplotlib.dates
from sklearn.impute import KNNImputer
import matplotlib
matplotlib.rcParams["figure.figsize"] = (15,10)
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
```

**Figure 9 : Importing libraries**



## Reading raw data file :

Now we will be having two raw data excel files. We should arrange both excel files in a structured way and merge them into one csv file which consists date, pollutant concentrations data and meteorological data columns. Now we have to upload this merged csv file in jupyter notebook and read that file by using Pandas read\_csv() function to import a CSV file to Data Frame format.

### Case 1:

```
READING RAW DATA FILE

RAW DATA FILE CONSISTS OF POLLUTANT CONCENTRATIONS ONLY

[3] df = pd.read_csv('/content/drive/MyDrive/project/rk_puram_pollutant.csv', encoding= 'unicode_escape')

[4] df.head()
```

	date	pm25	pm10	o3	no2	so2	co
0	01-01-2018	351	440	20	40	12	23
1	02-01-2018	386	282	22	32	9	18
2	03-01-2018	289	211	22	42	8	17
3	04-01-2018	244	276	13	34	8	19
4	05-01-2018	282	260	32	37	9	22

```
[5] len(df), len(df.columns)

(1486, 7)
```

### Case 2:

```
{x}

CASE 2:INCLUDING METEROLOGICAL PARAMETERS

READING RAW DATA FILE

[ ] dfs = pd.read_csv('/content/drive/MyDrive/project/r.k.-puram, delhi, delhi, india-air-quality.csv', encoding= 'unicode_escape')

[ ] len(dfs), len(dfs.columns)

(1486, 10)

[ ] dfs.head()
```

	date	pm25	pm10	o3	no2	so2	co	T2M	RH2M	WS10M
0	01-01-2018	351	440	20	40	12	23	12.07	34.56	2.41
1	02-01-2018	386	282	22	32	9	18	10.98	38.50	2.58
2	03-01-2018	289	211	22	42	8	17	11.01	39.62	3.02
3	04-01-2018	244	276	13	34	8	19	12.51	33.56	2.09
4	05-01-2018	282	260	32	37	9	22	13.02	30.94	1.95

**Figure 10 : Reading csv file of case-1 and case-2**

### **3.3.3 Data pre processing :**

Inconsistent data, missing values, and repeated data may be found in the data we obtain from various sources. To acquire a good prediction result, the data set must be cleaned, and missing values must be dealt with either by deleting them or replacing them in with mean values or another way. In order to avoid biasing the results, redundant data must also be deleted or discarded. Some datasets may contain outliers or extreme values that must be eliminated in order to achieve acceptable prediction accuracy. Only if all of this pre processing is done on the data then classification and clustering algorithms, as well as other data mining approaches, perform properly.

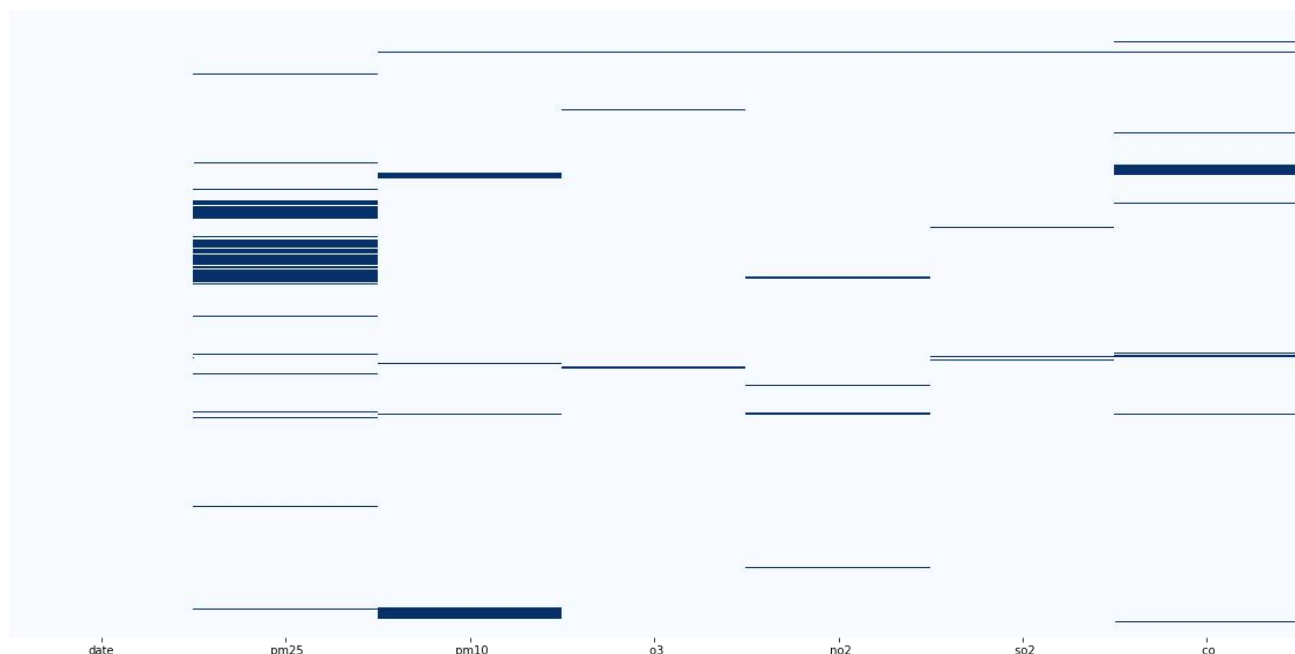
**Data Cleaning :** Data cleaning is the process of determining whether parts of data are erroneous, incomplete, inaccurate, irrelevant, or missing, and then changing, replacing, or deleting them as needed.

Finding number of null values: `Df_name.isnull().sum()`

```
FINDING NULL VALUES IN DATA SET

✓ [8] df1.isnull().sum()
0s
date      0
pm25     158
pm10      51
o3        16
no2       22
so2       17
co        57
dtype: int64
```

**Data visualization :** Null values ( Heat map)



**Figure 11 : Finding null values and data visualization of null values**

We can fill null values by different methods in our case we have used KNN method for filling null values.

**KNN:** KNN for missing values is based on the concept that a point value may be approximated by the values of the points that are closest to it, based on other factors. It can handle continuous, discrete, ordinal, and categorical data, making it particularly effective for dealing with missing data of any kind.

#### FILLING NULL VALUES USING K NEAREST NEIGHBOURS

```
✓ [21] feature_cols = [col for col in df3.columns if col not in ['date']]  
0s knn_imputer = KNNImputer(n_neighbors=10)  
df3[feature_cols] = knn_imputer.fit_transform(df3[feature_cols])
```

```
✓ [22] df3.isna().sum()  
0s
```

```
date      0  
pm25      0  
pm10      0  
o3        0  
no2       0  
so2       0  
co        0  
dtype: int64
```

#### **Data visualization for checking null values :**



**Figure 12 : Filling null values with KNN method**

Same procedure is followed in case-2 for filling null values.

### 3.3.4 Calculation and preparation of AQI Data Frame :

We have data set with pollutants concentration and meteorological data. We need AQI column in our data frame to proceed with model training. So to find AQI for each row a MAXIMUM OPERATOR system has been used and AQI column is prepared in our data frame. Maximum concentration among pollutants concentration is AQI.

$$AQI = \text{MAX}(I_1, I_2, I_3, \dots, I_n)$$

$I_1, I_2, \dots$  are pollutants concentration

{x}

CREATING A COLUMN AQI USING MAX OPERATOR

```
[26] df5=df3[[' pm25', ' pm10', ' o3', ' no2', ' so2', ' co']]
```

```
[27] df3['AQI']=df5.max(axis=1)
```

```
[28] df3.head()
```

	date	pm25	pm10	o3	no2	so2	co	AQI
0	01-01-2018	351.0	440.0	20.0	40.0	12.0	23.0	440.0
1	02-01-2018	386.0	282.0	22.0	32.0	9.0	18.0	386.0
2	03-01-2018	289.0	211.0	22.0	42.0	8.0	17.0	289.0
3	04-01-2018	244.0	276.0	13.0	34.0	8.0	19.0	276.0
4	05-01-2018	282.0	260.0	32.0	37.0	9.0	22.0	282.0

**Figure 13 : Preparation of AQI Data Frame**

### 3.3.5 Split the data set:

To begin, we must divide the data set into two groups: training and testing. The training data set is used to train the prediction model. It will then be put to the test with the testing set to find accuracy of model.

#### Case-1:

ASSIGNING INDEPENDENT AND DEPENDENT VALUES TO X AND Y

```
[30] X = df3[[' pm25', ' pm10', ' o3', ' no2', ' so2', ' co',]]
```

```
Y = df3[['AQI']]
```

```
[127] from sklearn.model_selection import train_test_split
```

APPLYING TRAIN TEST SPLIT USING 80% AS TRAIN DATA AND 20% AS TEST DATA

```
[33] X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=5)
```

#### Case-2 :

ASSIGNING INDEPENDENT AND DEPENDENT VALUES TO X AND Y

```
[90] X = df3[[' pm25', ' pm10', ' o3', ' no2', ' so2', ' co', 'T2M', 'RH2M', 'WS10M']]
```

```
Y = df3[['AQI']]
```

```
from sklearn.model_selection import train_test_split
```

```
df3.dtypes
```

date	int64
pm25	float64
pm10	float64
o3	float64
no2	float64
so2	float64
co	float64
AQI	float64
dtype:	object

APPLYING TRAIN TEST SPLIT USING 80% AS TRAIN DATA AND 20% AS TEST DATA

```
[92] X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=5)
```

**Figure 14 : Splitting into training and testing data**

### **3.3.6Applying ML models:**

**Linear regression:** Linear regression analysis is a statistical technique for predicting the value of one variable based on the value of another. The dependent variable is the variable you want to forecast. The independent variable is the one you're using to forecast the value of the other variable.

The procedure is known as simple linear regression when there is only one input variable (x). The procedure is referred to as multiple linear regression when there are several input variables.

#### **Simple linear regression equation :**

$$Y = MX + C$$

Y - y coordinate

X - x coordinate

M - slope

C - y intercept

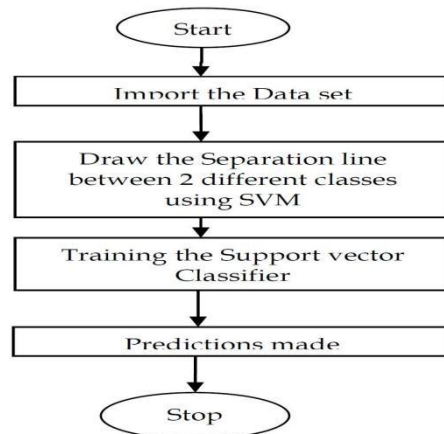
**Gradient boosting:** Gradient boosting is a machine learning technique for regression and classification that creates a prediction model from an ensemble of weak prediction models. This method builds a model step by step and then generalizes it by allowing any differentiable loss function to be optimized. Gradient boosting is an iterative procedure for combining multiple weak learners into a single strong one. For each weak learner, a new model is fitted to generate a more accurate estimate of the response variable. Gradient boosting is a method for merging a number of relatively weak prediction models to produce a more powerful prediction model.

$$Y = MX + C + E$$

E - Error term

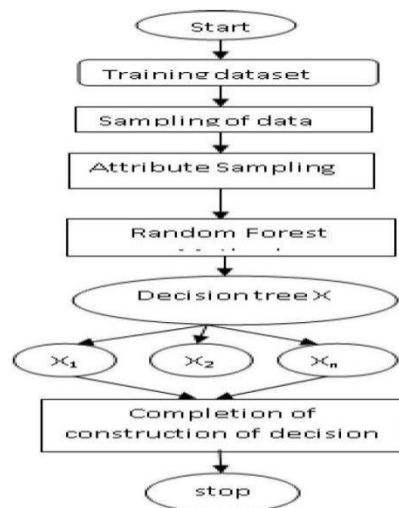
**Decision tree:** One of the administered learning procedures used to portray the choice made in light of the condition is the Decision Tree. It tends to be applied to both characterization and regression. The choice tree is generally built in a hierarchical way. The primary hub from the top is known as a root hub. The last hub in the chain is known as a leaf hub. Inner hubs exist between the root hub and the leaf hub. Inner hubs are isolated into bunches relying upon determined models, and afterward choices are made. As the quantity of factors expansions continuously, the tree develops bigger and the methodology turns out to be more modern. The two kinds of choice trees are arrangement and relapse trees. A grouping tree is utilized to order the information assortment, making it more straightforward to investigate. Notwithstanding, utilizing this technique, we can't make a forecast. The relapse tree is a type of tree used to figure consistent qualities.

**Support Vector Machine :** SVM is a supervised learning method that divides the plane into two sections by drawing a line between the two classes. The hyperplane is the line that divides the plane into multiple segments. It always returns a perpendicular distance between the data point and the separation line. It can classify in both linear and nonlinear ways. It's primarily utilized for classification and regression analysis.



**Figure 15 : Flow chart of Support Vector Machine(SVM)**

**Random Forest :** It's a collection of decision trees used for regression and classification. To determine the majority vote, classification is employed. The mean value is calculated using regression. This method is more accurate, resilient, and capable of handling a wide range of data types, including binary, category, and continuous data. Random Forest is nothing more than a collection of decision trees. The training takes into account 80% of the data set. The training data is sampled, and distinct decision trees are built using the Random Forest algorithm based on attribute sampling.



**Figure 16 : Flow chart of Random Forest(RF)**

**XGBoost:** Extreme gradient boosting is the name of a software library called XGBoost. The library is an implementation of gradient boosting machines. Gradient Boosting, Regularized Gradient Boosting are among the optimization features offered by XGBoost, which focuses on computational speed and model performance. To enhance the efficiency of compute time and memory resources, the XGBoost algorithm was employed. The approach uses Sparse Aware implementation, which handles missing values from data sets automatically. In addition, the approach allows an already fitted model to be trained on additional data. XGBoost is a prominent open source software library, primarily because of its speed.

**K nearest neighbour (KNN):** K Nearest Neighbor is an essential calculation that keeps up with every accessible model and arranges new information or cases utilizing a comparability measure. It's most regularly used to arrange an information point in light of the order of its neighbors.

### **Training models:**

Above explained Machine Learning models(Linear regression, SVM, DT etc) can be applied to train data set.

### **3.3.7 Model evaluation :**

To predict accuracy, we use the Mean Squared Error (MSE) and the Root Mean Squared Error (RSME).

**MSE:** Mean squared error is one of the most basic and straightforward regression metrics. It's the sum of the squares of the difference between the actual and predicted values, or the average squared errors of the prediction.

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n}$$

MSE stands for Mean Squared Error.

$y_i$  is the actual output


$\hat{y}_i$  average of observation of  $y_i$

$n$  is the total observations considered

The disadvantages of MSE include the fact that it is ineffective when data is more noisy. As a result, the article employs the RMSE method.

**RMSE:** Root mean Squared Error is defined as Square root of Mean square error. The square root minimize the errors. The RMSE value is used to verify the results. The lower the RMSE value, the better the accuracy.

## Case-1:



The image shows a Jupyter Notebook interface with a sidebar on the left containing a search icon, a file icon, and a code icon. The main area is titled "K NEIGHBOURS REGRESSOR". It contains two code cells. The first cell imports KNeighborsRegressor, fits it to training data, and prints the score. The second cell predicts on test data and prints MAE, MSE, and RMSE metrics.

```
K NEIGHBOURS REGRESSOR
```

```
[ ] from sklearn.neighbors import KNeighborsRegressor
    regressor2=KNeighborsRegressor(n_neighbors=1)
    regressor2.fit(X_train,y_train)
    regressor2.score(X_test,y_test)

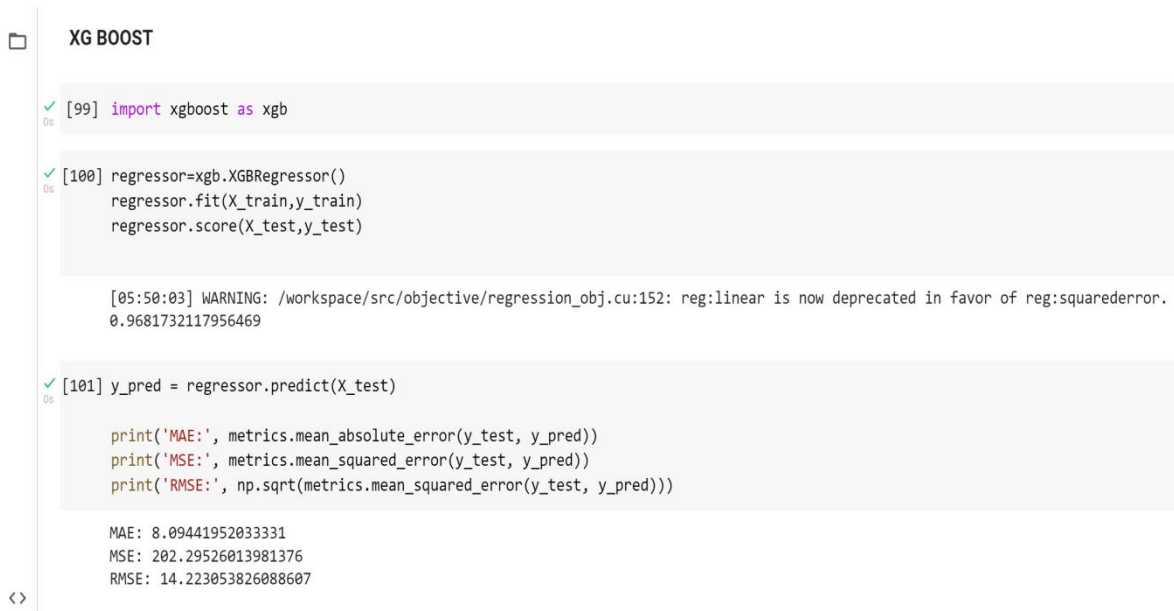
0.9885178465459052
```

```
[ ] y_pred = regressor2.predict(X_test)

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

MAE: 5.718791946308725
MSE: 72.98208053691275
RMSE: 8.542955023697171
```

## Case-2 :



The image shows a Jupyter Notebook interface with a sidebar on the left containing a file icon. The main area is titled "XG BOOST". It contains three code cells. The first cell imports xgboost as xgb. The second cell creates an XGBRegressor, fits it to training data, and prints the score. The third cell predicts on test data and prints MAE, MSE, and RMSE metrics. A warning message is also displayed.

```
XG BOOST
```

```
✓ [99] import xgboost as xgb
0s
```

```
✓ [100] regressor=xgb.XGBRegressor()
0s
       regressor.fit(X_train,y_train)
       regressor.score(X_test,y_test)
```

```
[05:50:03] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
0.9681732117956469
```

```
✓ [101] y_pred = regressor.predict(X_test)
0s

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

MAE: 8.09441952033331
MSE: 202.29526013981376
RMSE: 14.223053826088607
```

**Figure 17 : Applying ML Models and Model Evaluation**



### 3.3.8 TIME SERIES FORECASTING AND PREDICTION OF FUTURE VALUES

After model training we choose the best model among them and predicted the values for whole data set considering all parameters and took that data for time series forecasting and predicting the future values.

**Fb Prophet Model** Prophet is a time series data forecasting process that uses an additive model to accommodate non-linear trends with yearly, weekly, and daily seasonality, as well as holiday impacts. It works effectively with time series with significant seasonal influences and historical data from numerous seasons. Missing data and trend changes are tolerated by Prophet, and outliers are usually handled satisfactorily. Prophet produces a model by identifying the line, which is made up of the elements listed below:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- $g(t)$  = Overall growth trend.  $g(t)$
- $S(t)$  = Yearly seasonality/weekly seasonality
- $h(t)$  = Holidays effects
- $\epsilon_t$  = error term

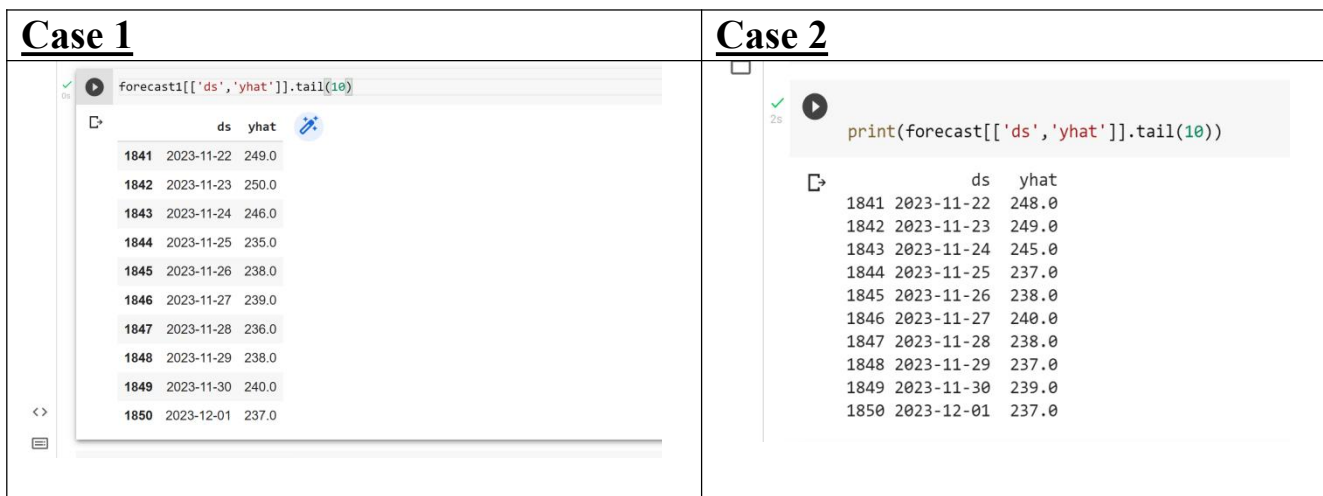
Prophet is one of the best model for time series forecasting because it is **Accurate and Fast**.

Prophet is utilized in a variety of Facebook apps to generate accurate forecasts for planning and goal-setting. In the vast majority of circumstances, we've found it to outperform any other strategy. We use Stan to fit models so that you may get forecasts in a matter of seconds.

**Fully Automatic** It does not require manual effort, we can get a fair forecast on sloppy data. Outliers, missing data, and significant changes in time series are not a problem for Prophet.

**Tunable forecasts** The Prophet method gives us a lot of options for tweaking and adjusting forecasts. By combining our subject expertise with human interpretable parameters, we can improve our forecast.

Prophet model is implemented for the best model after performing model training and time series forecasting is done to achieve the future values of the data.



**Figure 18 : Output of Time Series Forecasting**

### Discussion

#### 4.1 Experimental Results And Analysis :

After training and testing models of Machine Learning , we came up with a new model which uses highest accuracy model and prophet (time series analysis). Using Time Series Analysis, we are able to predict future data points.

##### Case-1 : AQI prediction by considering only pollutants concentration

	Variance ( $R^2$ value)	Root Mean Squared Error
LR	0.935	20.34
GB	0.984	9.97
DTR	0.952	17.55
XGB	0.982	10.69
RFR	0.953	17.34
KNN	0.989	8.54
SVM	0.868	28.96

**Table 4 : Results of case-1**

It is clear from the above table that most accurate model that fits our data is KNN regressor model.

##### Case-2 : AQI prediction by considering both pollutants concentration and meteorological data

	Variance ( $R^2$ value)	Root Mean Squared Error
LR	0.919	22.57
GB	0.967	14.43
DTR	0.932	20.71
XGB	0.968	14.22
RFR	0.943	19.10
KNR	0.957	16.49
SVM	0.850	30.87

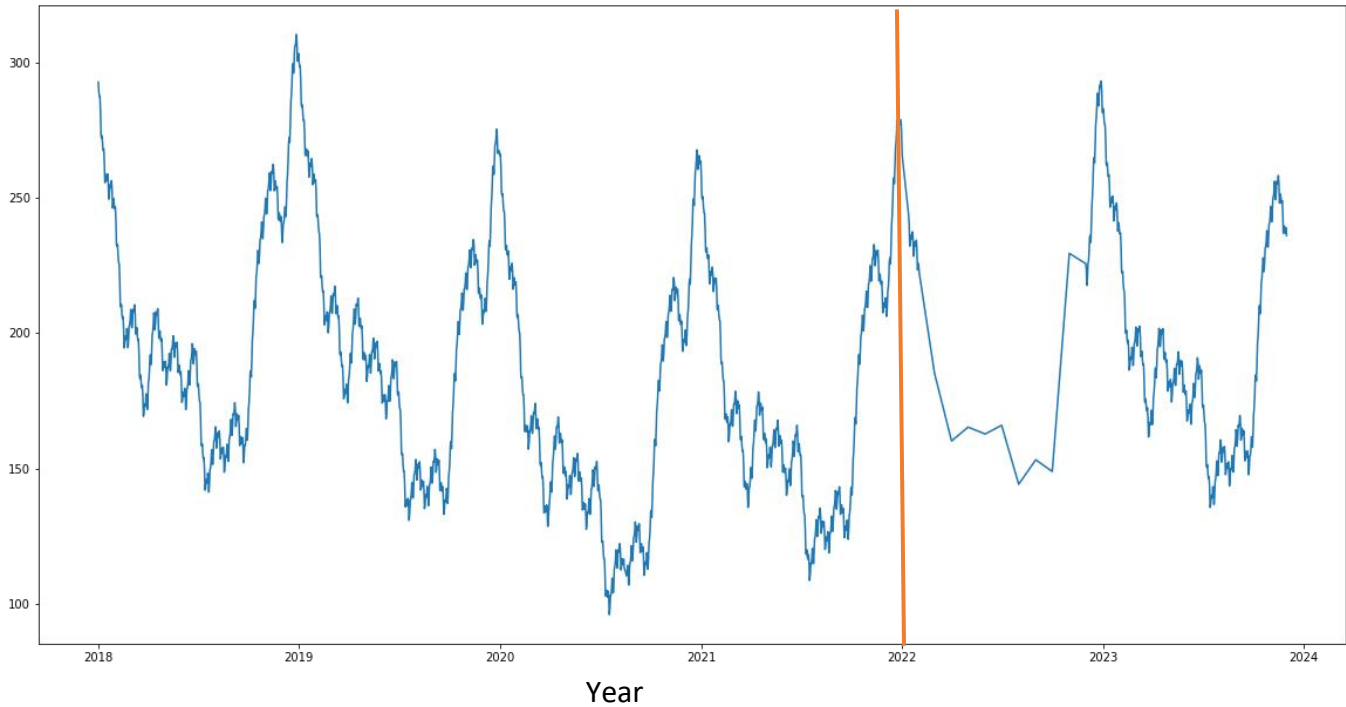
**Table 5 : Results of case-2**

It is clear from the above table that most accurate model that fits our data is XGBoost model.

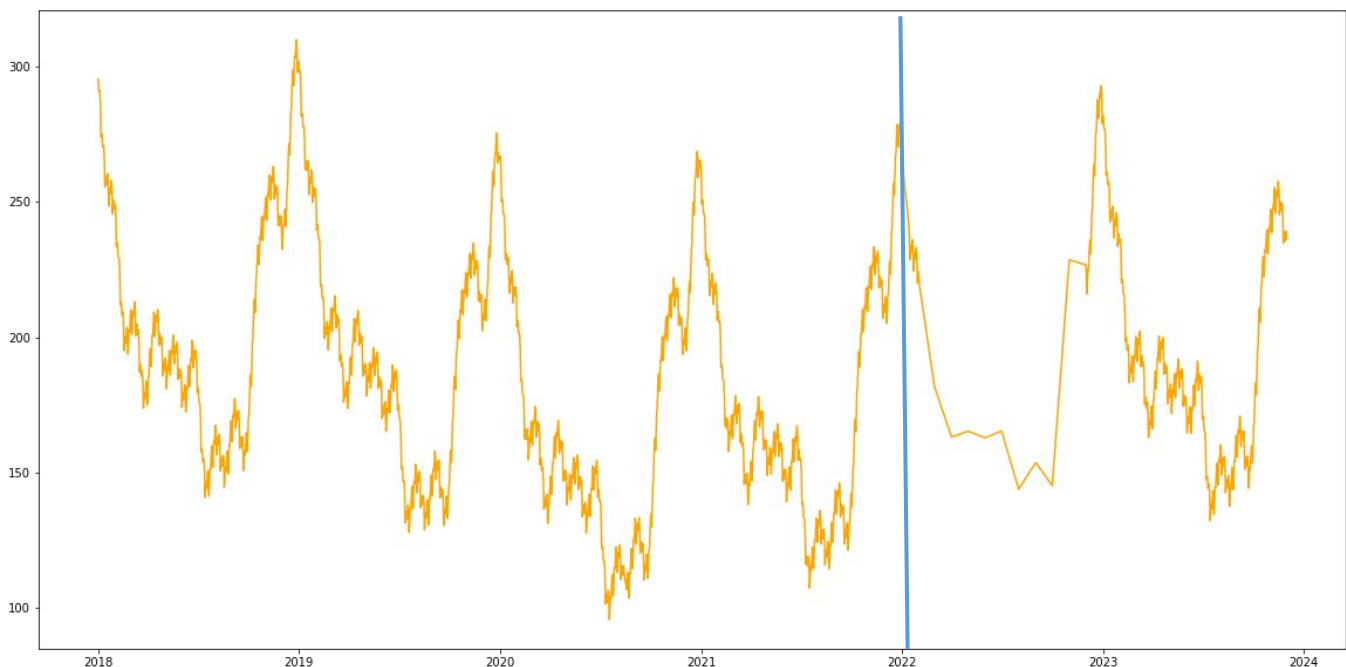
We have done time series analysis using FB Prophet model by taking the best accurate models into consideration.

## **4.2 Graphical representation of results by time series forecasting :**

### **Case-1 : Time series forecasting of data set(pollutants concentration) by using KNN regression model.**

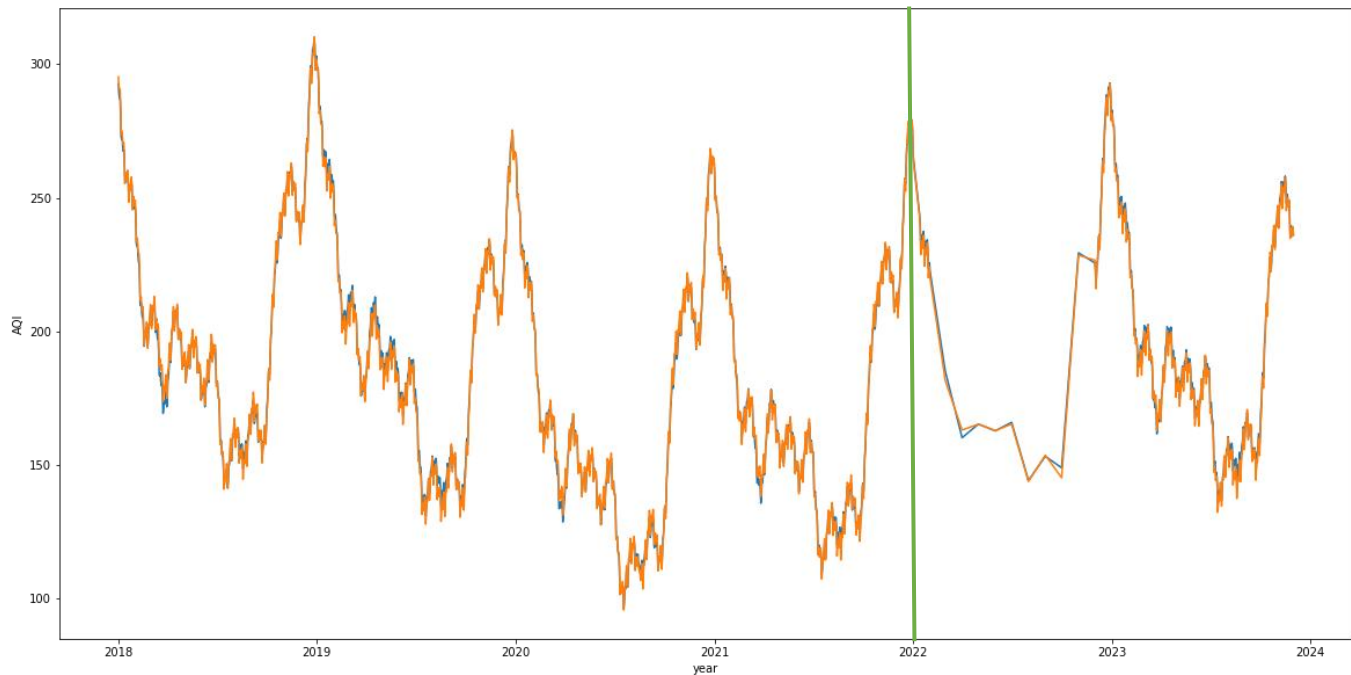


### **Case-2: Time series forecasting of data set(pollutants concentration and meteorological data) by using XGBoost regression model.**



**Figure 19 : Graphical representation of results by time series forecasting of Case -1 and Case -2**

## **Case-1 VS Case-2**



**Figure 20 : Graphical representation of results by time series forecasting of Case -1 Vs Case -2**

## **4.3 Conclusion**

It is estimated that air pollution kills around seven million people worldwide each year. This makes it very important for the air to be monitored continuously and taken care of . Air Quality is measured using AQI. AQI is the way of showing changes in the amount of pollution in the air.

The project performs the study of different machine learning models for the prediction of AQI using pollutant concentrations and meteorological parameters as well. The data we have used is of RK Puram Delhi for the study of patterns using Linear regressor, Decision tree regressor, Support vector machine, Random forests, K Nearest Neighbours, Gradient boost, XGboost algorithms and then performed time series analysis using prophet model for prediction of future values of AQI considering business as usual scenario. The models are evaluated using  $R^2$  Score, MAE, MSE, RMSE.

The need for predicting the AQI is necessary for safeguarding future and helping people know how the local air quality impacts their health. After performing the model training we can conclude that KNN Regressor and Xgboost works well for Data sets consisting of pollutant concentration and meteorological parameters.

This project helps the meteorological department in predicting the air quality values beforehand and taking necessary precautions to control the pollutant concentration. This would help in reducing the environmental hazards like global warming and many health issues like heart attacks and respiratory effects such as asthma attacks and bronchitis to which people are prone due to bad air quality.

#### 4.4 Project work plan

	September	October and November	December	January and February	March and April
<b>Selection of project</b>					
<b>Literature Study</b>					
<b>Data Collection</b>					
<b>Learning of Software ( Machine Learning)</b>					
<b>Report and Presentation Work</b>					
<b>Building of Machine Learning Models</b>					
<b>Results and Discussion</b>					
<b>Final Report Preparation</b>					

## **4.5 References**

- 1.Madhuri VM, Samyama Gunjal GH, Savitha Kamalapurkar, Air Pollution Prediction Using Machine Learning Supervised Learning Approach, International journal of scientific & technology research volume 9, issue 04, april 2020.
- 2.Report by Central Pollution Control Board (CPCB) on Air Quality Index.
- 3.K. Mahesh Babu, J. Rene Beulah,Air Quality Prediction based on Supervised Machine Learning Methods, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-9S4, July 2019.
- 4.Mrs. A. Gnana Soundari MTech, (PhD),Mrs. J. Gnana Jeslin M.E, (PhD),Akshaya Indian Air quality prediction and analysis using Machine learning International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019.
- 5.Datasets taken from following websites :
  - 1.<https://aqicn.org/data-platform/register/> (For pollutant concentration data)
  - 2.<https://power.larc.nasa.gov/data-access-viewer/> (For meteorological data)
- 6.Nidhi Sharma , ShwetaTaneja , VaishaliSagar , Arshita Bhatt, “Forecasting air pollution load in Delhi using data analysis tools”, ScienceDirect, 132 (2018) 1077– 1085.
- 7.Aditya C R, Chandana R Deshmukh, Nayana D K Praveen Gandhi Vidyavastu .” Detection and Prediction of Air Pollution using Machine Learning Models”. InternationalJournal of Engineering Trends and Technology (IJETT) – volume 59 Issue4 – May 2018.
- 8.Kostandina Veljanovska, Angel Dimoski , Air Quality Index Prediction using Simple Machine Learning Algorithms, International Journal Of Emerging Trends & Technology in Computer Science(IJETTCS), Volume 7 ,Issue 1,Jan-Feb 2018.