Journal of Applied Information Science
10 (1) 2022, 10-16
http://www.publishingindia.com/jais

Prediction of Air Pollution using Supervised Machine Learning

Sudha Rajesh[1*], P. Amudhavalli[2] and Dhanapriya A. P.[3]
[1]Assistant Professor, Department of Computer Application, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India. Email: sudharajesh@crescent.education
[2]Assistant Professor, Department of Computer Application, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India. Email: amudhavalli@crescent.education
[3]Master of Computer Application, Department of Computer Application, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India. Email: dhanapriya385@gmail.com
*Corresponding Author

Abstract: The prediction and estimation of the air pollution stands a necessary analysis area. The main abstract of this is to explore machine learning which is used for checking of air quality forecasting to predict the outcome in best accuracy. In supervised machine learning this analysis is done to measure information like identification of variable and analysis of univariate, bivariate and multivariate analysis and identify the values which are missed and used to analyze validation of data and cleaning of data and visualization of data is done by entire by dataset. This analysis is to provide complete guide to sensitive analysis of parameters in addition to forecasting of air pollution by accurate prediction. This analysis is to suggest framework to expect the index of air excellence by predicting the results to form good accuracy and used to compare supervise classification machine learning algorithms. In addition it is to discuss the algorithm from the dataset taken from traffic department with graphic user interface air quality prediction.

Keywords: Air pollution, Air quality, Classification method, GUI results, Prediction, Python, Quality.

I. Introduction

Air is very important for each and every person in this world to live their life. Due to growth of industries and uses of unhealthy gases in the air, it becomes poisonous to breathe. These gases will cause issues in human body like lung disease and other breathing issues.

II. Domain Introduction

Machine learning algorithms are used to predict the upcoming with the help of past datasets. Machine learning comes under the part of artificial intelligence (AI). Machine learning is of three types. They are supervised, unsupervised, reinforcement supervised is defined as use of labeled data which is to train different algorithms that is used to predict outcomes. Unsupervised has no labels. Reinforcement learning is dynamically interacts with the environment.

III. Existing System

Guanghui Yue, Ke, Gu, and Junfei Qiao [1] Existing system focuses on photograph method. It is originated to be soaking map which is said in the direction of sensitive quality air, and it demonstrates the entire appearance in many different ways under high and low PM2.5 concentrations. This one also misses the configurations and pixel principles of PM2.5 and it is predictable via amalgamation of the double features which is followed through the non-linear mapping methods.

Rahul Ahuja, Ishan Verma, and Hardik Meisheri [2] Recurrent Neural Network (RNN) is proved as very efficient for processing temporal data. Hypothetically, an enormous delay can be recycled, but in training the prediction results may decrease if interruption is too big.

Temesegam Walelign Ayele, Rutvik Mehta [3] It is more risks for the people who worked in factory and plants too affected due to the breathing of chemicals substances and other gasses, smokes, fumes and vapors which exposure their harmful gases. Air contamination is considered as wide-reaching issue including international organizations etc. By using this concept, it can associate uncountable low power-driven objects to the internet.

Abdullah Kadri, Khaled Bashir Shaban, and Eman Rezk [4] It is widely believed that urban pollution incorporates a direct impact on human health, especially in developing industrial countries, wherever air quality measures aren't obtainable or enforced or implemented. Recent studies have proven that substantial evidences that are in introduction to region toxins has sturdy links towards deadly illnesses like bronchial asthma in addition to respiratory organ infection. The modules area unit is answerable for receiving and storing the info, pre-processing

and changing the information into helpful information, prediction the pollutants supported historical info, and finally presenting the non-inheritable information or data through completely different channels, like mobile application, Web application portals, and short message service. The main target the watching system and its prediction module.

Xia Xi, Zhao Wei and Rui Xiaoguang [5] Urban pollution prediction is the important one in all the foremost necessary tasks within the treatment of WRF-Chem that may be a numeric prototype, the prediction consequences are also not accurate and it takes full projection on weather, pollution and chemical element starting from WRF-Chem prototype as contribution sources, style a wide-ranging analysis framework on the way to boost the prediction performance. Experiment's area unit enforced with completely different options and teams. Classification algorithms in machine learning are used in seventy-four cities of China, in order to come out the most effective methods for every town. From experimental purposes, from various town, the most effective result will be attained by dissimilar cluster of features choice and model choice. Experimental outcomes and conclusions show that a lot of features used, a lot of risk to boost the accurateness. For technical side, the result from their proposed work is superior than distinctive idea. In modern times, by way of fast development of China's economy, the concentration of pollutants present in the atmosphere has decreased considerably. Problems in conservation air quality valuation and toxic waste management dragged community attention and also associated with (AQI) Air Quality Index, may be a range employed by government assistances to speak the general communal. However, impure the air currently is or however impure it's forecast will become.

Luke Curtis, William Rea, Patricia Smith- Willis [6] The outside air every so often encompasses of organically expressive levels of many pollutants including carbon monoxide, volatile organics, oxides of nitrogen and sulfur, ozone, particulates such as PM10 or PM2.5, metals, bio aerosols and pesticides. These pollutants type are created as a result of anthropogenic activities. Many peoples are spending their time in outdoor and indoor. So people are mostly affected the air because of air pollution. Most of the patients has asthmatics, COPD patients, chemical sensitivities, heart and stroke patients, pregnant women, diabetics, the elder people and children are particularly affected by out-of-doors and indoor air quality. Nowadays air is more polluted by many industrial air pollution and traffic pollution and many other air pollutions.

J. Guo, Q. Liu, H. Xu, and L. Ye, [7] Nowadays air are affected by many things like industrial air pollution, fuel air pollution, traffic air pollution, and many more things. Sand is also one pollution because sand-dust is also affecting your breathing part.

Sand-dust are mostly accrued when a strong wind is appeared in the atmosphere. Sand is a tiny particle we can't see with your naked eye the particle is PM10 it is so tiny so it can easy affect your berthing part. It accrued when a high wind is created. It mostly affects who working in sand place. The particle matter is PM10 or PM2.5. The sand polluted areas are Odisha, Delhi and New Delhi. These types of pollutants affect your eyes, nose and respiration track, and affecting especially children. Short term diseases such as chest pain, nausea and vomiting, headache, eye irritations, skin rash, tiredness, cough, and lung problems. Long term disease such as Kidney and liver damage, depression of the central nervous system, Nervous system damage and Neuromuscular blockage.

Table I: Air Pollutant Range Estimated by the Government

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

The above table shows the AQI level and Associated Health Impacts Particulate Matter by means of a thickness not more than 10 micrometers (PM10), Particulate Matter by a thickness of not more than 2.5 micrometers (PM2.5), Nitrogen Dioxide (NO2), Ozone (O3) and Carbon Monoxide (CO). These are the pollutions which affect humans and animal. The AQI Level is divide by decent (0-50) is minimum impact satisfactory (51-100) is may be the basis of major breathing problem to sensitive people, Moderately polluted (101-200) is may be the reason of breathing problem people who affected by lungs disease etc., Poor (201-300) is may cause problem people who have heart disease, Very Poor (301-400) is may be the source of problems to the people who affected by lungs and heart disease, Severe (401-500) is may cause people has major problem in lungs and heart disease.

IV. Proposed System

In proposed system we are going to analyze and predict whether the air quality pollutants are high or low using supervised machine learning algorithms.

The steps are given below:

Step 1: Collect the data set from difference places that are preprocessed.

Step 2: After preprocessed divided the dataset into training and testing.

Step 3: Train the dataset by applying ML algorithm.

Step 4: In final process compare the accuracy of the algorithms to get high accuracy.

The below Fig. 1 shows the working flow of proposed work. First we need to collect the dataset and we need to pre-processed the dataset. Data transformation into splitting data example testing and training. Then we need to compare with algorithm in this proposed work we used two algorithm - Decision tree and K-Nearest Neighbor by comparing with these two algorithm we will get one best accuracy rate by using that algorithm we will create your GUI based output.

V. Architecture Diagram

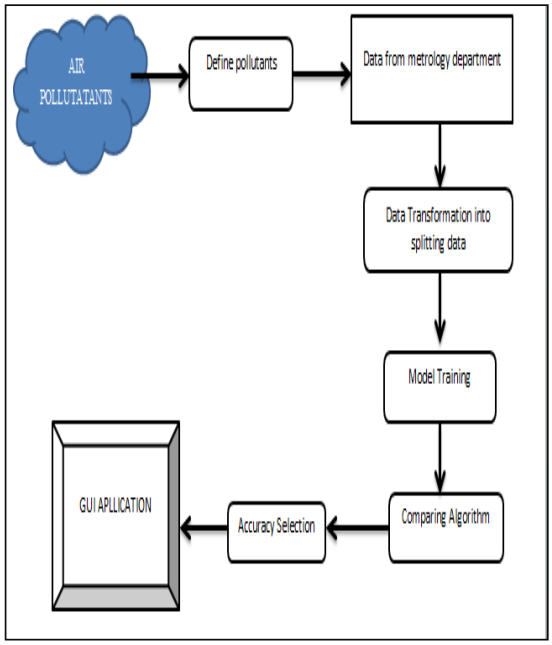

Fig. 1: Architecture Diagram

Project Goals

- Examination of data analysis for variable identification:
  - Load the particular dataset
  - Import vital libraries packages
  - Examine the over-all possessions
  - Find identical and misplaced values
  - Check the unique and total values
- Univariate data analysis
  - Change the name, enhance and drop the data
  - To specify data type
- Exploration data analysis of bi-variate and multi-variate
  - Bar chart, pie chart, chart, Line chart, scatter plot, box plot, Dot plot etc.
- Method of Outlier detection with feature engineering
  - Preprocess the given data
  - Divide the data for testing and training
  - Compare the Decision tree and Logistic regression model and random forest.
- Compare the algorithms to get the predicted outcome based on the best accuracy.

Advantages

- These intelligences are highly used for the exploration of data with appropriate machine learning approach to forecast the quality of air in all operational conditions.
- At last, it highlights the best few observations arranged in the near future examination issues, experiments, and requests.

Scope

The main scope of the proposed work is to examine a dataset of air pollutants proceedings for India meteorological subdivisions with the help of machine learning approach. The identification of air quality is more difficult and try to reduce this risk factor behind predicting from Air Quality Index (AQI) of India to safe human so as to save many meteorological determinations and belongings and to predict whether assigning the air quality is bad or good based on given attributes.

Objectives

The goal is to cultivate a machine learning prototype for real-time air quality forecasting, to potentially replace the updatable supervised machine learning classification models by guessing outcomes in the form of best precision by comparing supervised algorithm.

VI. Working Process

The dataset we are taken from different places which is needs to be generalized format to resolve the missing and null values.

Module Description

- Data Pre-Processing
- Data Visualization
- Comparing with algorithm GUI based output

Dataset

It has 824 tuples and 9 attributes. They are name, city, country, state, average, minimum and maximum etc.

The below Fig. 2 shows the data set your proposed project. The data set contain Country, Stage, City, Average, Minimum, Maximum and pollutants (like SO_2, NO_2, CO, O_3, PM10 and PM2.5 etc.)

	A	B	C	
1	Country	State	city	place
2	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - A
3	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - A
4	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - A
5	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - A
6	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - A
7	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - A
8	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - A
9	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Ra
10	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Ra
11	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Ra
12	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Ra
13	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Ra
14	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Ra
15	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Ra
16	India	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPC
17	India	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPC
18	India	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPC
19	India	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPC
20	India	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPC

Fig. 2: Dataset

Data Pre-Processing

The first part of the project is data pre-processing. In this process we will preprocess the data to remove missing data, duplicate data. This is the first stage for this project to gain good result for pre-processing show the efficiency of the algorithm will be developed. Then we need to train the dataset and we need to test with new data set for best accuracy.

The below Fig. 3 shows your data pre-processing. In this process we will clean or remove the duplicate data null data.

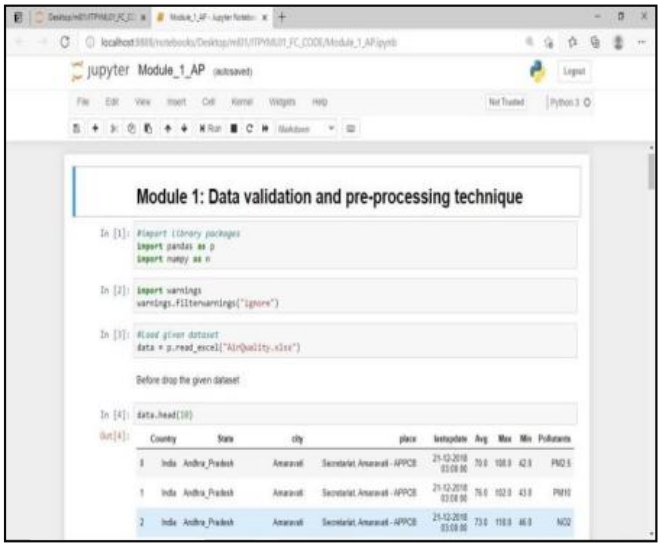

Fig. 3: Data Pre-Processing

Data Visualization

It is the most important skill in applied statistics and device learning. To know a dataset, we can explore data visualization. It also helps to identify patterns, corrupt data and outliers etc.

The below Fig. 4 shows the data visualization. With is very easy to understand for the user to view the rating and value of the data set in colorful view. This bar chat shows the Rate and percentage of pollutions like (like SO_2, NO_2, CO, O_3, PM10 and PM2.5 etc.).

Algorithm and Techniques

Algorithm: Decision Tree

Decision tree is considered as simple and popular Machine learning algorithm. It helps to build a classification or regression models in a tree structure. The main aim of this algorithm is to create a model which is used to predict the end result variable of the class by knowing the decision rules which is inferred from prior training data.

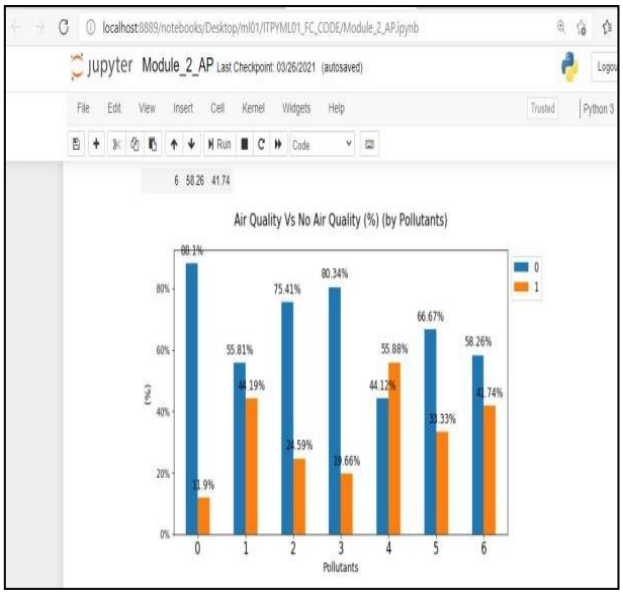

Fig. 4: Estimating the Dataset Value in Bar Chart

Step 1: We will say base node as S, it contains entire dataset.

Step 2: We can find the best attribute using (ASM).

Step 3: Then we need to divide the subset for best quality.

Step 4: We need to generate.

Step 5: It continues the process to reach the final stage called leaf node.

The below Fig. 5 shows the accuracy value of decision tree. The accuracy value is 99%.

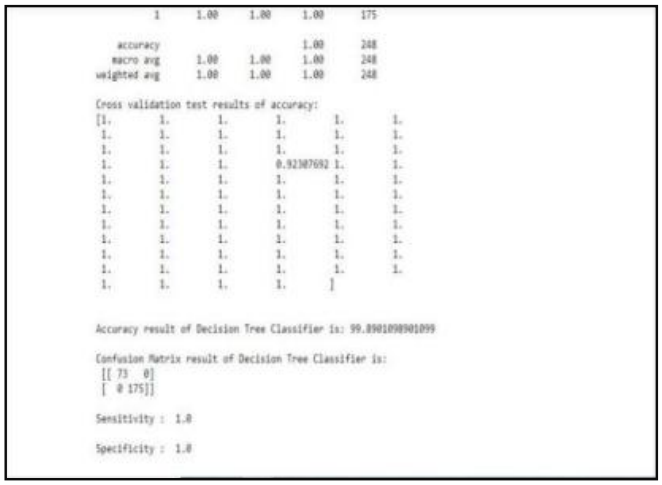

Fig. 5: Decision Tree Accuracy Value

K-Nearest Neighbor

This algorithm is very simple which stores the available data which is related to the given data points in n-value space and it separates the new data by parallel measure. If an unknown dataset is received it will analyze the k closest number of instance which proceeds the most communal class for real value data.

Step 1: Implementing any algorithmic program, we want dataset. Thus throughout the primary step of KNN, we tend to should load the coaching moreover as check information.

Step 2: We has to decide on the worth of K.

i.e., the closest information. K may be any number.

Step 3: In every purpose check information do the subsequent

- Check the space and information in every row of coaching information in tactic such has: geometer and playing distance. All most it is used to measure distance.
- Let, support the space worth, which arrange in ascending order.
- Sorted array of K row will be chosen.
- Then it will allocate category for checking purpose supporting most frequent class of those rows.

Step-4: Four finish.

The below Fig. 6 shows the K-Nearest Neighbor accurate value. The accuracy value stands as 97%.

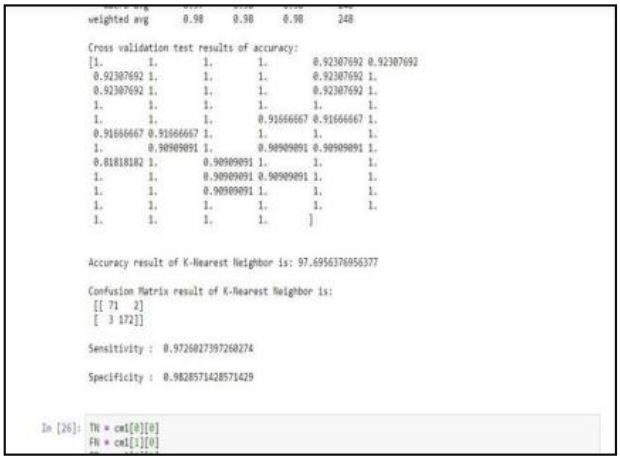

Fig. 6: K-Nearest Neighbor Tree Accuracy Value

Accuracy Calculation: The accuracy values are calculated for all algorithm through general formula:

Accuracy

$$= \frac{(\text{TruePositive}+\text{TrueNegative})}{(\text{TruePositive}+\text{TrueNegative}+\text{FalsePostive}+\text{FalseNegative})}$$

Accuracy is an important parameter and it is without problem. It has a proper ratio records.

Precision: The percentage of confident. It gives correct value:

$$\text{Precision} = \frac{\text{TruePositive}}{(\text{TruePositive}+\text{FalsePositive})}$$

$$\text{TruePositiveRate} = \frac{\text{TruePositive}}{(\text{TruePositive} + \text{FalseNegative})}$$

$$\text{FalsePositive(FalsePostiveRate)} = \frac{\text{FalsePostive}}{(\text{FalsePostive} + \text{TrueNegative})}$$

Recall: The positive values are correctly Predicted.

F-Measure: F1 is a measure of test accuracy.

General Formula:

$$\text{F-Measure} = \frac{2\text{TruePositive}}{2\text{TruePositive} + \text{FalsePositive} + \text{FalseNegative}}$$

F1-Score the Precision and Recall is used to compute the score of procedure.

If the F-score is higher, then the improved power of predictive classification procedure.

F1-Score Formula:

$$\text{F1Score} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

Python Package:

- *Sklearn:* This is a package which has a lot of ML algorithms.
- *Numpy:* Numpy is numeric module in python which is used for mathematical calculation.
- *Pandas:* It is used to read and write.
- *Matplotlib:* We can easily do the data manipulation by this.
- *TKinter:* TKinter is python package.
- It is standard GUI library for Python.
- It is easy way to create GUI application.

The Fig. 7 shows the output of GUI based Air Quality Prediction in this output we will give the data like Name, State Name and City after giving the value we will give AQI value (Above-10 etc.).

Then we need to click Check Prediction of pollutants button to see the pollution prediction values like SO_2, NO_2, CO, O_3, PM10 and PM2.5 etc. Then clicking the check AQI stages button to see the AQI Cause will show the condition of air like good, bad etc. with the people who will affect by this air.

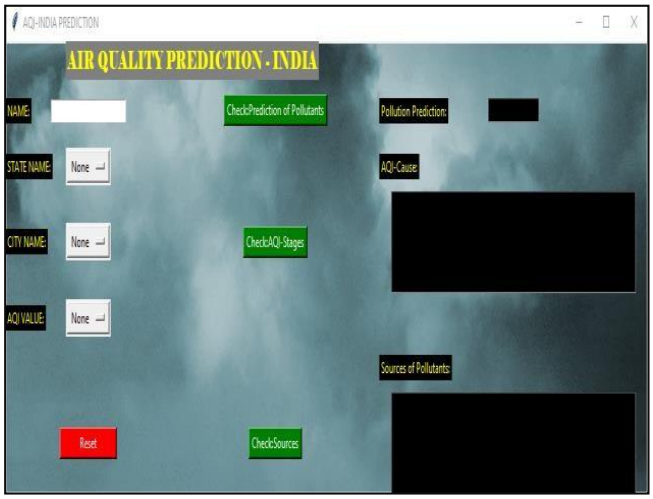

Fig. 7: GUI based Air Quality Prediction

Then we need to click the button check sources to see the sources of pollutants it will show the reason of the pollution like - the air is polluted by industrial boilers, Pollutants emitted by cars etc.

VII. Result Analysis

For this project we have used anaconda navigator which gives the next input immediately. It is very easy to use and user friendly.

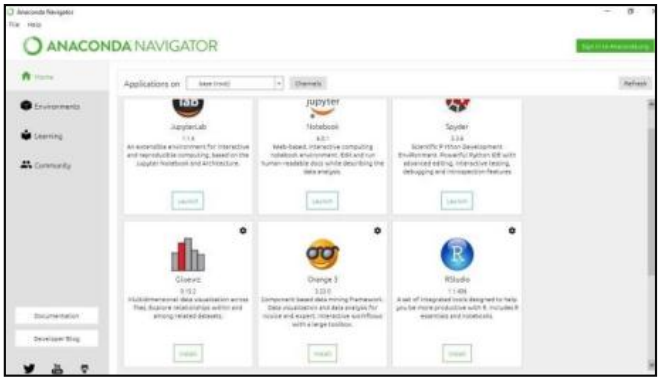

Fig. 8: Anaconda Navigator

The above Fig. 8 shows the user interface of anaconda navigator.

Jupyter Lab is a web-based interactive environment for jupyter notebook. It is flexible, extensible and modular. The above Fig. 9 shows the jupyter note book. After Launching the anaconda navigator then weneed to launch jupyter note book.

Fig. 9: Jupyter Notebook

Fig. 10: Folder

The above Fig. 10 shows the folder by clicking the folder the modules will open after that we can run the program. The Fig. 11 shows the final output of your project in GUI base format. It contains input Name, Country, State, City. The output is Pollution Prediction, AQI case, Sources of pollution.

Fig. 11: Final Output

VIII. Conclusion

It is started by analytical process and preprocessing of data, missing values and exploratory analysis is done. At finally model is built with evaluation. This application will be helpful for India metrological department for predicting the air quality and they will take action on air pollution.

IX. Future Enhancement

This application can be used by India metrological department in future for detecting whether the air quality is good or not from the real time process. We can automatically show the final results in web or desktop application. We can also implement this work in artificial intelligence.

References

[1] G. Yue, K. Gu, and J. Qiao, "Effective and efficient photo-based PM2.5 concentration estimation," *IEEE Transactions on Instrument and Measurement*, vol. 68, no. 10, Oct. 2019.

[2] I. Verma, R. Ahuja, H. Meisheri, and L. Dey, "Air pollutant severity prediction using bi-directional LSTM network," *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Santiago, Chile, 3-6 Dec. 2018.

[3] T. W. Ayele, and R. Mehta, "Air pollution monitoring and prediction using IoT," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT),* Coimbatore, India, 20-21 Apr. 2018.

[4] K. B. Shaban, A. Kadri, and E. Rezk, "Urban air pollution monitoring system with forecasting models," *IEEE Sensors Journal*, vol. 16, no. 8, Apr. 2016.

[5] X. Xi, Z. Wei, R. Xiaoguang, W. Yijie, B. Xinxin, Y. Wenjun, and D. Jin, "A comprehensive evaluation of air pollution prediction improvement by a machine learning method," *2015 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, Yasmine Hammamet, Tunisia, 15-17 Nov. 2015.

[6] L. Curtis, W. Rea, P. Smith-Willis, E. Fenyves, and Y. Pan, "Adverse health effects of outdoor air pollutants," *Environment International*, vol. 32, no. 6, pp. 815-830, May 2006, doi: 10.1016/j.envint.2006.03.012. P MID: 16730796.

[7] H. Xu, J. Guo, Q. Liu, and L. Ye, "Fast image dehazing using improved dark channel prior," in *Proc. IEEE Int. Conf. Inf. Sci. Technol.,* Mar. 2012, pp. 663-667.