Air Quality Prediction By Given Attribute Based On Supervised Machine Learning Approach

R SANTHOSHI, *Department of Information Technology, Panimalar Engineering college, Chennai*
V SANGEETHA, *Department of Information Technology, Panimalar Engineering college, Chennai*
A PRIYADHARSHINI, *Department of Information Technology, Panimalar Engineering college, Chennai*
TAMIZHSELVI A, *Department of Information Technology, Panimalar Engineering college, Chennai*
Corresponding Author: Mrs. K. MUTHU LAKSHMI, M.Tech, Associate Professor, Panimalar Engineering College

Abstract
Generally, Air pollution refers to the release of pollutants into the air that are detrimental to human health and the planet as a whole. It can be described as one of the most dangerous threats that the humanity ever faced. It causes damage to animals, crops, forests etc. To prevent this problem in transport sectors have to predict air quality from pollutants using machine learning techniques. Hence, air quality evaluation and prediction has become an important research area. The aim is to investigate machine learning based techniques for air quality forecasting by prediction results in best accuracy. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi- variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in prediction of air quality pollution by accuracy calculation. To propose a machine learning-based method to accurately predict the Air Quality Index value by prediction results in the form of best accuracy from comparing supervise classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department dataset with evaluation of GUI based user interface air quality prediction by attributes

Keywords: *Graphical User Interface, Artificial Intelligence, Machine Learning, Air Quality Index.*

Date of Submission: 13-07-2021 Date of acceptance: 29-07-2021

I. INTRODUCTION

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

1.1.1 Need of the Project

The goal is to develop a machine learning model for real-time air quality forecasting, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm. Problem Description/ Problem Statements: Monitoring and preserving air quality has become one of the most essential activities in many industrial and urban areas today. The quality of air is adversely affected due to various forms of pollution caused by transportation, electricity, fuel uses etc. The deposition of harmful gases is creating a serious threat for the quality of life in smart cities. With increasing air pollution, we need to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area.

1.1.2 Scope of the Project

The scope of this project is to investigate a dataset of air pollutants records for India meteorological sector using machine learning technique. To identifying air quality is more difficult. We try to reduce this risk factor behind predicting from Air Quality Index (AQI) of India to safe human so as to save lots of meteorological efforts and assets and to predict whether assigning the air quality is bad or good.

1.2 EXISTING SYSTEM

To estimate the PM2.5 concentration by designing a photograph-based method. By observation, it is found that the saturation map is sensitive to air quality, exhibiting entirely different appearances under high and low PM2.5 concentrations. Specifically, it loses structures and most pixel values tend to be 0 under a high PM2.5 concentration. To compute the gradient similarity between the saturation and gray-scale maps to quantify the structural information loss. Then, utilizing the Weibull distribution to fit the saturation map and able to derive a value to estimate the color information. Finally, the PM2.5 concentration of an image can be estimated via the combination of the aforementioned two features followed by a nonlinear mapping procedure. Both numerical and visualized results on real captured data validate the effectiveness and superiority of the proposed method in comparison with the relevant stateof- the-art methods. Air pollution has become a worldwide concerned issue and automatical estimation of air quality can provide a positive guidance to both individual and industrial behaviors. Given that the traditional instrument-based method requires high economic, labor costs on instrument purchase and maintenance, proposes an effective, efficient, and cheap photo-based method for the air quality estimation in the case of particulate matter (PM2.5). This method lies in extracting two categories of features (including the gradient similarity and distribution shape of pixel values in the saturation map) by observing the photo appearances captured under different PM2.5 concentrations. Specifically the gradient similarity is extracted to measure the structural information loss with the consideration that PM2.5 attenuates the light rays emitted from the objects and accordingly distorts the structures of the formed photo. Meanwhile, the saturation map is fit by the Weibull distribution to quantify the color information loss. By combining two features, a primary PM2.5 concentration estimator is obtained. Next, a nonlinear function is adopted to map the primary one to the real PM2.5 concentration. Sufficient experiments on real data captured by professional PM2.5 instrument demonstrate the effectiveness and efficiency of the proposed method. Specifically, it is highly consistent with real sensor’s measures and requires low implementation time.

1.2.1 Drawbacks

- ➢ Photo graphic method is not sufficient to calculate PM2.5 and it taken only one pollutants of concentration.
- ➢ the work is given based on only one PM2.5 and to collect more monitoring data from other cities to verify the generalization of our work and more factors, e.g., geomorphic conditions.
- ➢ It can’t thereby better determine the regularity of air pollutant data and achieve more accurate prediction results.

1.3 Proposed System

- ➢ While applying photo graphic based method is critical to evaluate parameters and it’s taken data size is high
- ➢ To overcome this method to implement machine learning approach by user interface of GUI application
- ➢ Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

1.3.1 Advantages:

- ➢ These reports are to the investigation of applicability of machine learning techniques for air quality forecasting in operational conditions.
- ➢ Finally, it highlights some observations on future research issues, challenges, and needs.

2.1 Project Goals

➢ Exploration data analysis of variable identification

- Loading the given dataset
- Import required libraries packages
- Analyze the general properties
- Find duplicate and missing values
- Checking unique and count values

- Uni-variate data analysis
- Rename, add data and drop the data
- To specify data type
- Exploration data analysis of bi-variate and multi-variate
- Plot diagram of pairplot, heatmap, bar chart and Histogram
- Method of Outlier detection with feature engineering
- Pre-processing the given dataset
- Splitting the test and training dataset

2.2 Data flow diagram

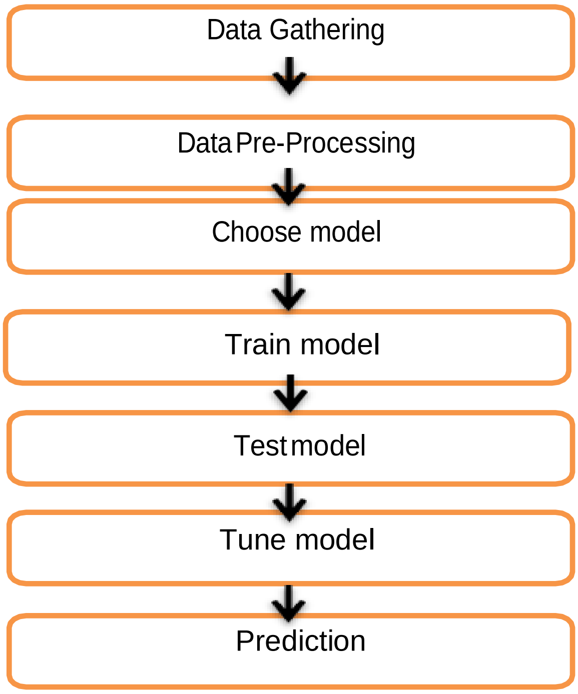


Figure 1: Process of dataflow diagram

2.4 Use case diagram

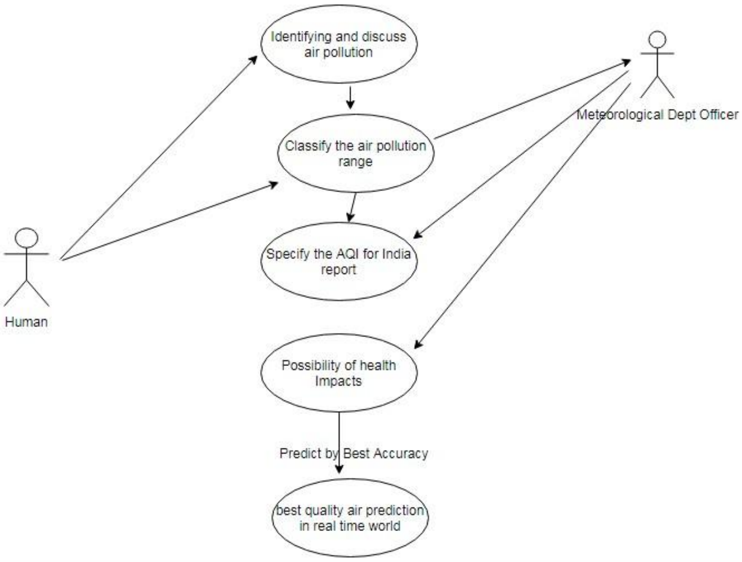

Identifying and discuss air pollution
Classify the air pollution range
Specify the AQI for India report
Possibility of health Impacts
Predict by Best Accuracy
best quality air prediction in real time world
Human
Meteorological Dept Officer

Figure 2: UML Use case diagram

3.1 Functional requirements

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

3.2 Non-Functional Requirements

Process of functional steps,

1. Problem defines
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

3.3 Hardware Requirements

Hard disk : minimum 80 GB
RAM : minimum 2 GB

3.4 Software requirements

Operating System : Windows
Tool: Anaconda with Jupyter Notebook
Language: Python

3.4.1 Software Description

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system "Conda". The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS. So, Anaconda distribution comes with more than 1,400 packages as well as the Conda package and virtual environment manager called Anaconda Navigator and it eliminates the need to learn to install each library independently.

(i) The Jupyter Notebook:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

(ii) Notebook document:

Notebook documents (or "notebooks", all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc…). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc.) as well as executable documents which can be run to perform data analysis.

(iii) Jupyter Notebook App:

The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet. In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a "Dashboard" (Notebook Dashboard), a "control panel" showing local files and allowing opening notebook documents or shutting down their kernels.

(iv) kernel:

A notebook *kernel* is a "computational engine" that executes the code contained in a Notebook document. The ipython kernel, referenced in this guide, executes python code. Kernels for many other languages exist (official kernels). When you open a Notebook document, the associated *kernel* is automatically launched. When the notebook is *executed* (either cell-by-cell or with menu *Cell -> Run All*), the *kernel* performs the computation and produces the results. Depending on the type of computations, the *kernel* may consume significant CPU and RAM. Note that the RAM is not released until the *kernel* is shut-down.

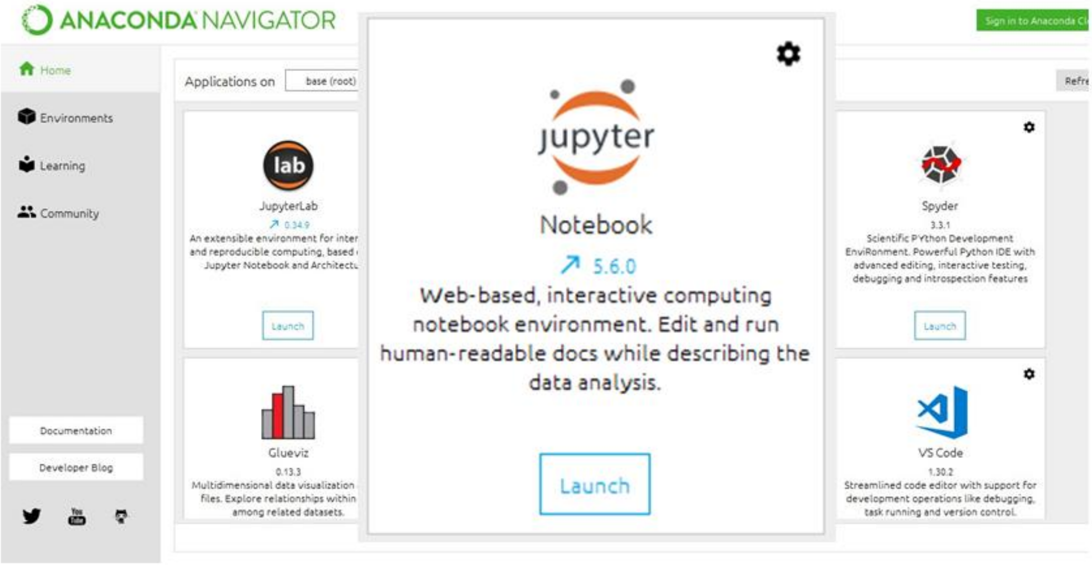
(v) Notebook Dashboard:
The *Notebook Dashboard* is the component which is shown first when you launch Jupyter Notebook App. The *Notebook Dashboard* is mainly used to open notebook documents, and to manage the running kernels (visualize and shutdown). The *Notebook Dashboard* has other features similar to a file manager, namely navigating folders and renaming/deleting files. Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems. There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming.

The best way to get started using Python for machine learning is to complete a project.

- It will force you to install and start the Python interpreter (at the very least).
- It will give you a bird's eye view of how to step through a small project.
- It will give you confidence, maybe to go on to your own small projects.

IV. RESULT AND DISCUSSION

4.1.1 Software involvement steps:



An extensible environment for inter and reproducible computing, based Jupyter Notebook and Architectu

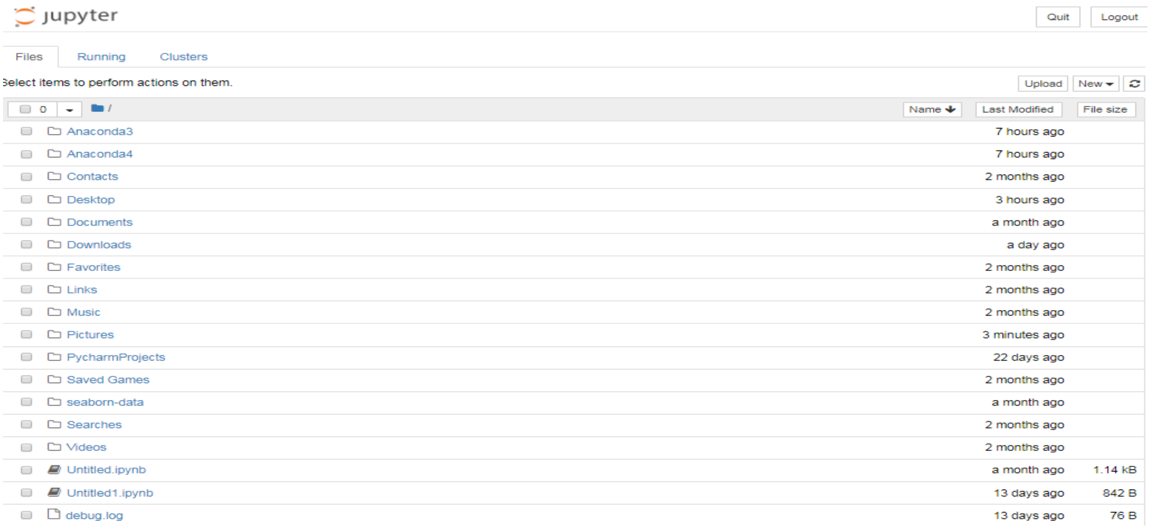
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Multidimensional data visualization files. Explore relationships within among related datasets.

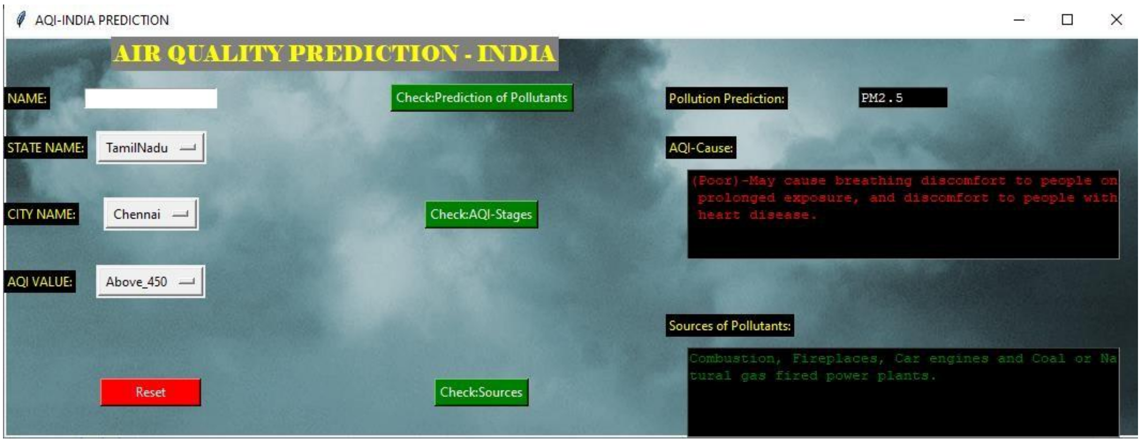
Streamlined code editor with support for development operations like debugging, task running and version control.

Figure 3: Launch the jupyter notebook platform



Select items to perform actions on them.

Name	Last Modified	File size
Anaconda3	7 hours ago	
Anaconda4	7 hours ago	
Contacts	2 months ago	
Desktop	3 hours ago	
Documents	a month ago	
Downloads	a day ago	
Favorites	2 months ago	
Links	2 months ago	
Music	2 months ago	
Pictures	3 minutes ago	
PycharmProjects	22 days ago	
Saved Games	2 months ago	
seaborn-data	a month ago	
Searches	2 months ago	
Videos	2 months ago	
Untitled.ipynb	a month ago	1.14 kB
Untitled1.ipynb	13 days ago	842 B
debug.log	13 days ago	76 B



AQI-INDIA PREDICTION
AIR QUALITY PREDICTION - INDIA
Check:Prediction of Pollutants
Check:AQI-Stages
Reset
Check:Sources

NAME:

STATE NAME: TamilNadu

CITY NAME: Chennai

AQI VALUE: Above_450

Pollution Prediction: PM2.5

AQI-Cause:

```
(Poor)-May cause breathing discomfort to people on
 prolonged exposure, and discomfort to people with
 heart disease.
```

Sources of Pollutants:

```
Combustion, Fireplaces, Car engines and Coal or Na
tural gas fired power plants.
```

Figure 4: Open the correspondent result folder

V. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score for predicting the air quality by given attributes. This application can help India meteorological department in predicting the future of air quality and its status and depends on that they can take action. India meteorological department wants to automate the detecting the air quality is good or not from eligibility process (real time). To automate this process by show the prediction result in web application or desktop application. To optimize the work to implement in Artificial Intelligence environment.

REFERENCES

[1]. Effective and Efficient Photo -Based PM2.5 Concentration Estimation, Guanghui Yue , Ke Gu , and Junfei Qiao, Member, IEEE, 2020.

[2]. A Machine Learning Approach to Predict Air Quality in California Mauro Castelli Fabiana Martins Clemente,1 Alesˇ Popovicˇ, Received 25 January 2020; Accepted 23 June 2020; Published 4 August 2020.