JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Air Quality index prediction using machine learning and deep learning

Z. syed ijaz ahamed, Abdus shakoor. H. H, Mohammed thahir. A

Student

Aalim Muhammed salegh college of engineering

ABSTRACT

Air pollution is contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere. The foremost cause of air pollution is of two types, one is natural pollution by means of volcanoes, wildfire and wind-blown dust. The other one is of artificial pollution by humans such as burning woods, trashes, excess usages of fossil fuels and industrial emissions which can be reduce by taking certain measures. In order to maintain and reduce the causes, one should measure and monitor the gases and make the people aware of the cause and explain the changes in climate which increase global warming. Air pollution can also cause of acid rain with exchange of sulfuric or nitric acid that fall to the ground from the atmosphere in wet or dry forms. It is important to first be able to identify whether a site is contaminated before determining a solution. This project explores the classification of Air samples at a particular site (Bangalore) to investigate the natural or unnatural contamination of Air. The samples are addressed using Decision tree learning, Random forests, Support vector machines (SVMS), Auto-regressive integrated moving average (ARIMA), Long short-term memory (LSTM), Artificial Neural Networks (ANN) Algorithm. From these evaluations, contamination at the site of interest can be considered. The final prediction is viewed in application developed under Django using Python.

CHAPTER 1

1. INTRODUCTION

Pollutants in the air cannot be seen with our naked eyes, we don't realize the sources of the increasing pollution level. In order to understand the sources of air pollution, we need to first go through the basic causes of air pollution. PM2.5 is hazardous for the environment around the globe. Long term exposure by individuals to these tiny particles which can easily pass through deep into respiratory organs can cause a lot of disease. PM2.5 is 2.5 microns in size which is far smaller than the size of the human hair. Hence the cause and exposure on human body is extremely high that results in various endemic and epidemic diseases. Children and elderly citizens, sick people who are already victims of ailments like COPD, asthma, or other lung and heart disorders, pregnant ladies are easily at risk when the level of this concentration is high and predominates for a longer time. This poor air quality not only threatens the health and lives of individuals but the economies as well. An effective system for monitoring and predicting air pollution in advance has great importance for human health and government decision-making. However, the mechanism and process of PM2.5 formation are very complex due to the complexity of its properties, such as non-linear properties in time and space, which have a significant impact on the accuracy of prediction.

It thus requires an examining consideration. Furthermore, the air quality data is closely related to time, which means that it belongs to time series and has an apparent periodicity. Due to the data's timeliness, time predictions have become essential topics that undoubtedly require meticulous attention by academics and scholars. So doing showcases that Time series analysis plays a paramount role in many different applications, including economics, medicine, astronomy, geology, and others. Several PM2.5 prediction methods are developed by researchers based on statistical models and machine learning techniques. Recently, the academic community has begun using deep neural networks for pollutant concentration prediction. Deep learning may solve problems by using more layers and more extensive data sets and processing all layers simultaneously to obtain more accurate results. These favourable properties of deep learning make it suitable for modelling and predicting air pollution.

1.2MACHINE LEARNING:

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information.** Currently, it is being used for various tasks such as **image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system**, and many more.

Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by **Arthur Samuel** in **1959**. We can define it in a summarized way as: "Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed".

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:

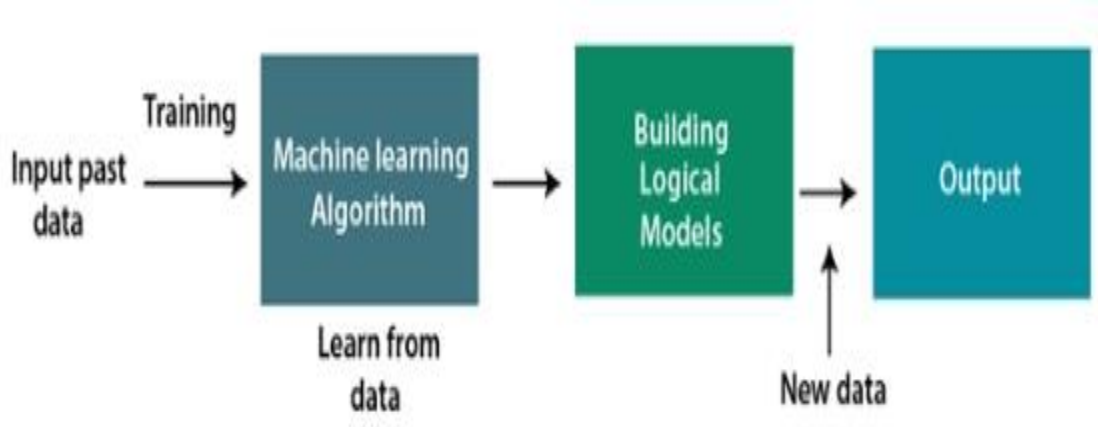


Fig:1.1 Working of Machine Learning

Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

3.1 REQUIREMENT ANALYSIS:

Requirements are a feature of a system or description of something that the system is capable of doing in order to fulfil the system's purpose. It provides the appropriate mechanism for understanding what the customer wants, analyzing the needs assessing feasibility, negotiating a reasonable solution, specifying the solution unambiguously, validating the specification and managing the requirements as they are translated into an operational system.

3.1.1 PYTHON:

Python is a dynamic, high level, free open source and interpreted programming language. It supports object-oriented programming as well as procedural oriented programming. In Python, we don't need to declare the type of variable because it is a dynamically typed language.

For example, x=10. Here, x can be anything such as String, int, etc.

Python is an interpreted, object-oriented programming language similar to PERL, that has gained popularity because of its clear syntax and readability. Python is said to be relatively easy to learn and portable, meaning its statements can be interpreted in a number of operating systems, including UNIX-based systems, Mac OS, MS-DOS, OS/2, and various versions of Microsoft Windows 98. Python was created by Guido van Rossum, a former resident of the Netherlands, whose favourite comedy group at the time was Monty Python's Flying Circus. The source code is freely available and open for modification and reuse. Python has a significant number of users.

Features in Python

There are many features in Python, some of which are discussed below

- Easy to code
- Free and Open Source
- Object-Oriented Language
- GUI Programming Support
- High-Level Language
- Extensible feature
- Python is Portable language
- Python is Integrated language
- Interpreted Language

3.2 ANACONDA

Anaconda distribution comes with over 250 packages automatically installed, and over 7,500 additional open-source packages can be installed from PyPI as well as the conda package and virtual environment manager. It also includes a GUI, Anaconda Navigator,[12] as a graphical alternative to the command line interface (CLI).

The big difference between conda and the pip package manager is in how package dependencies are managed, which is a significant challenge for Python data science and the reason conda exists.

When pip installs a package, it automatically installs any dependent Python packages without checking if these conflict with previously installed packages. It will install a package and any of its dependencies regardless of the state of the existing installation. Because of this, a user with a working installation of, for example, Google Tensorflow, can find that it stops working having used pip to install a different package that requires a different version of the dependent numpy library than the one used by Tensorflow. In some cases, the package may appear to work but produce different results in detail.

In contrast, conda analyses the current environment including everything currently installed, and, together with any version limitations specified (e.g. the user may wish to have.

Tensorflow version 2,0 or higher), works out how to install a compatible set of dependencies, and shows a warning if this cannot be done.

Open source packages can be individually installed from the Anaconda repository, Anaconda Cloud (anaconda.org), or the user's own private repository or mirror, using the conda install command. Anaconda, Inc. compiles and builds the packages available in the Anaconda repository itself, and provides binaries for Windows 32/64 bit, Linux 64 bit and MacOS 64-bit. Anything available on PyPI may be installed into a conda environment using pip, and conda will keep track of what it has installed itself and what pip has installed.

Custom packages can be made using the conda build command, and can be shared with others by uploading them to Anaconda Cloud, PyPI or other repositories.

The default installation of Anaconda2 includes Python 2.7 and Anaconda3 includes Python 3.7. However, it is possible to create new environments that include any version of Python packaged with conda.

1 EXISTING SYSTEM:

- In the existing method, the project demonstrated in Air pollution prediction by deep learning model is Long short term memory[LSTM].
- In Long short tern memory[LSTM] classifier, there is loss of performance and PM2.5 Data used is minimum compare to the proposed method.
- This LSTM classifier uses only two parameter as input such as RMSE (Root of the Mean of the Square of Errors) and MAE (Mean of Absolute value of Errors).

4.1.1 DISADVANTAGE:

- Air quality control system in the existing model are using single
- classifier which in compared to proposed method gives less accuracy.
- Dataset used in the existing model is fewer than the proposed model.
- Less performance.

4.2 PROPOSED SYSTEM:

- The proposed method is to provide an alternative solution for quality analysis which minimizes the required time and cost.

- The proposed method takes pre-existing dataset from kaggle website.
- The proposed method consists of three different classifiers in both ML and DL.
- The one successful classifier with high prediction in both ML & DL undergoes with Django for user interface.
- The Django is used for development of websites and application, from where the application gives the final predicted results.

4.2.1 ADVANTAGES

- High accuracy and performance are made when using two models.
- By monitoring the air quality in industries, can reduce the effect on workers in plant and also maintain the release of gases properly.
- The emission of PM2.5 can cause major respiratory disease and cancer which can be reduced in advance by use of proposed method.

4.3 SYSTEM ARCHITECTURE

4.3.1 SYSTEM ARCHITECTURE DIAGRAM

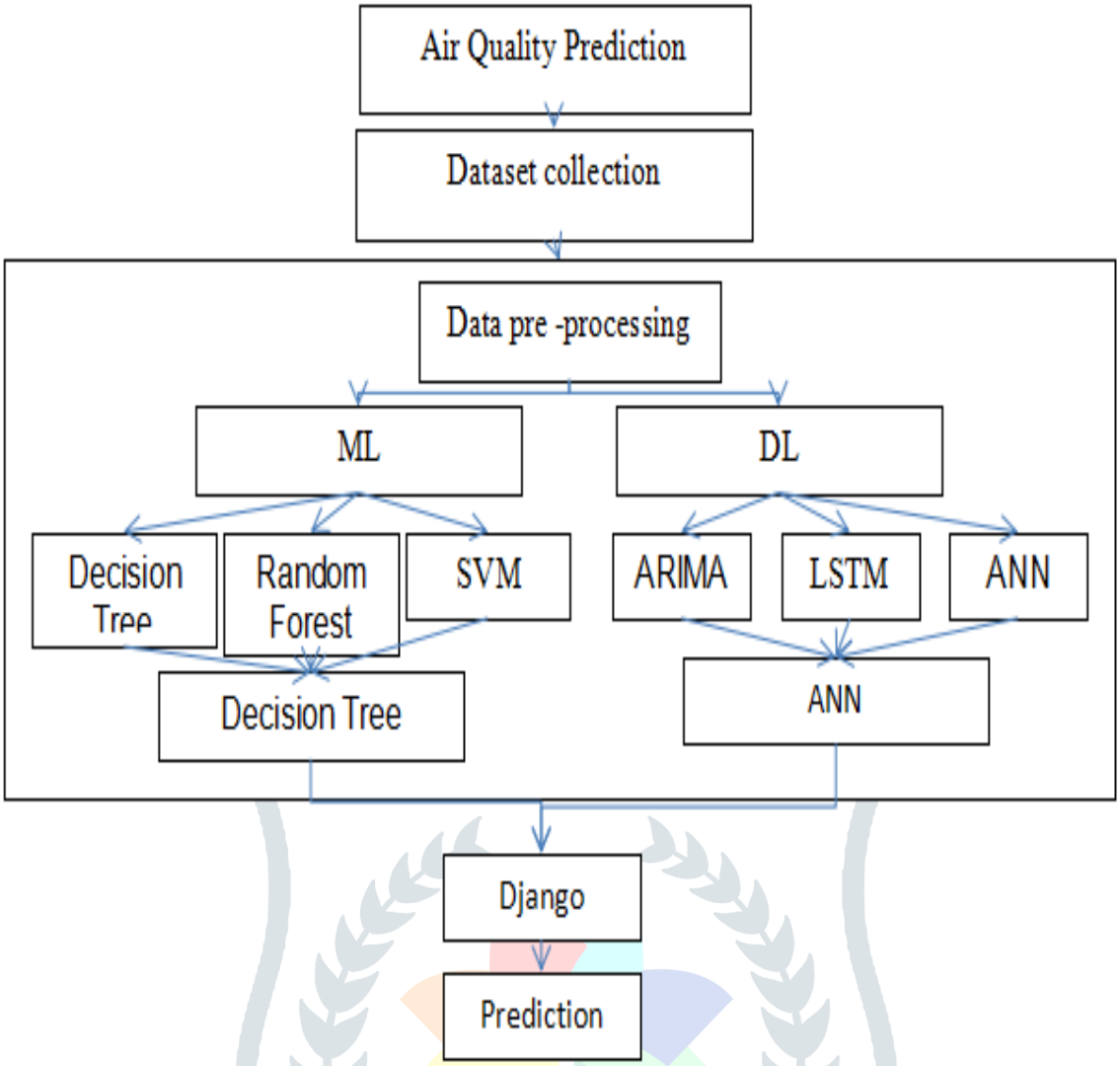
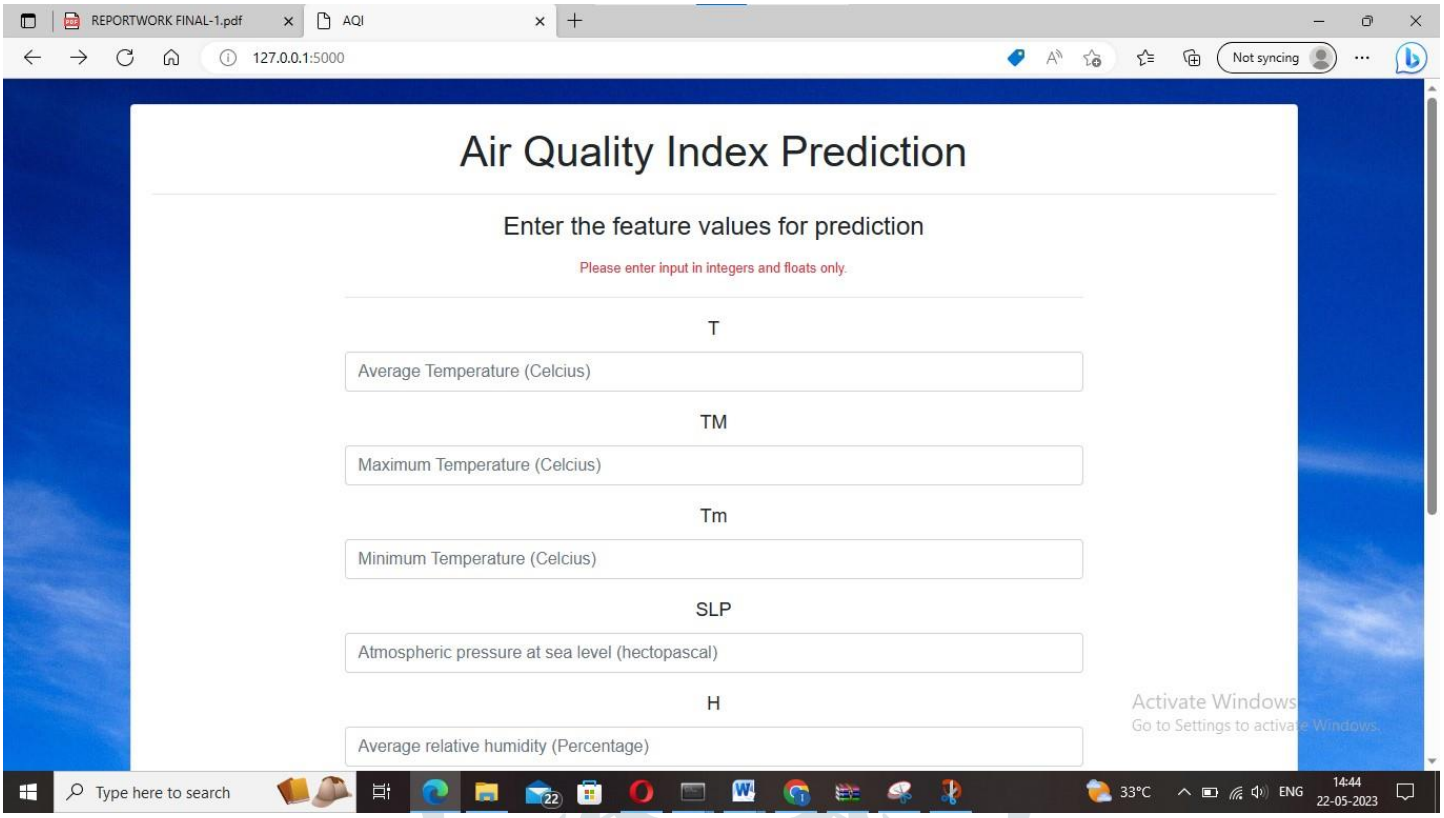
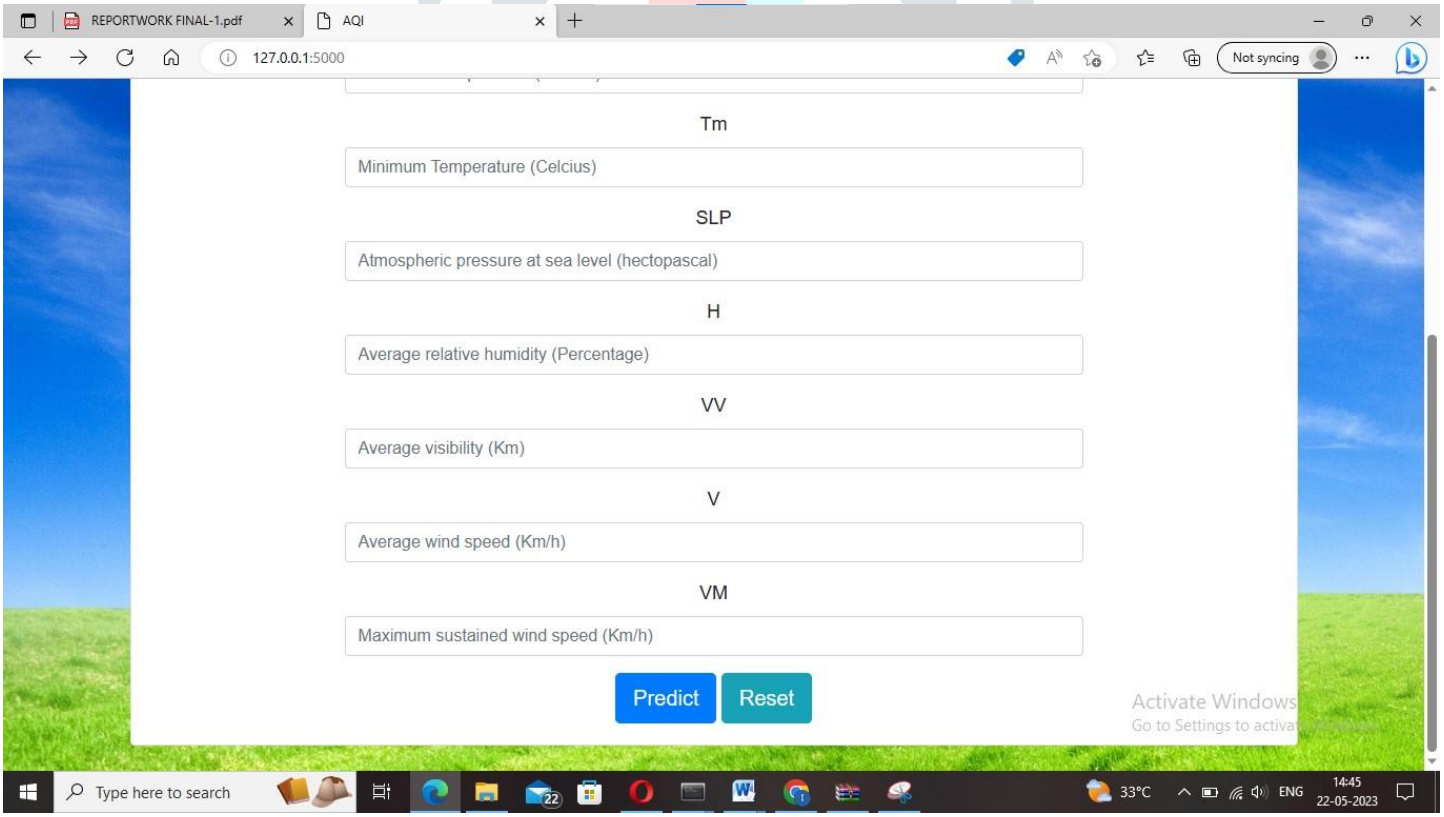
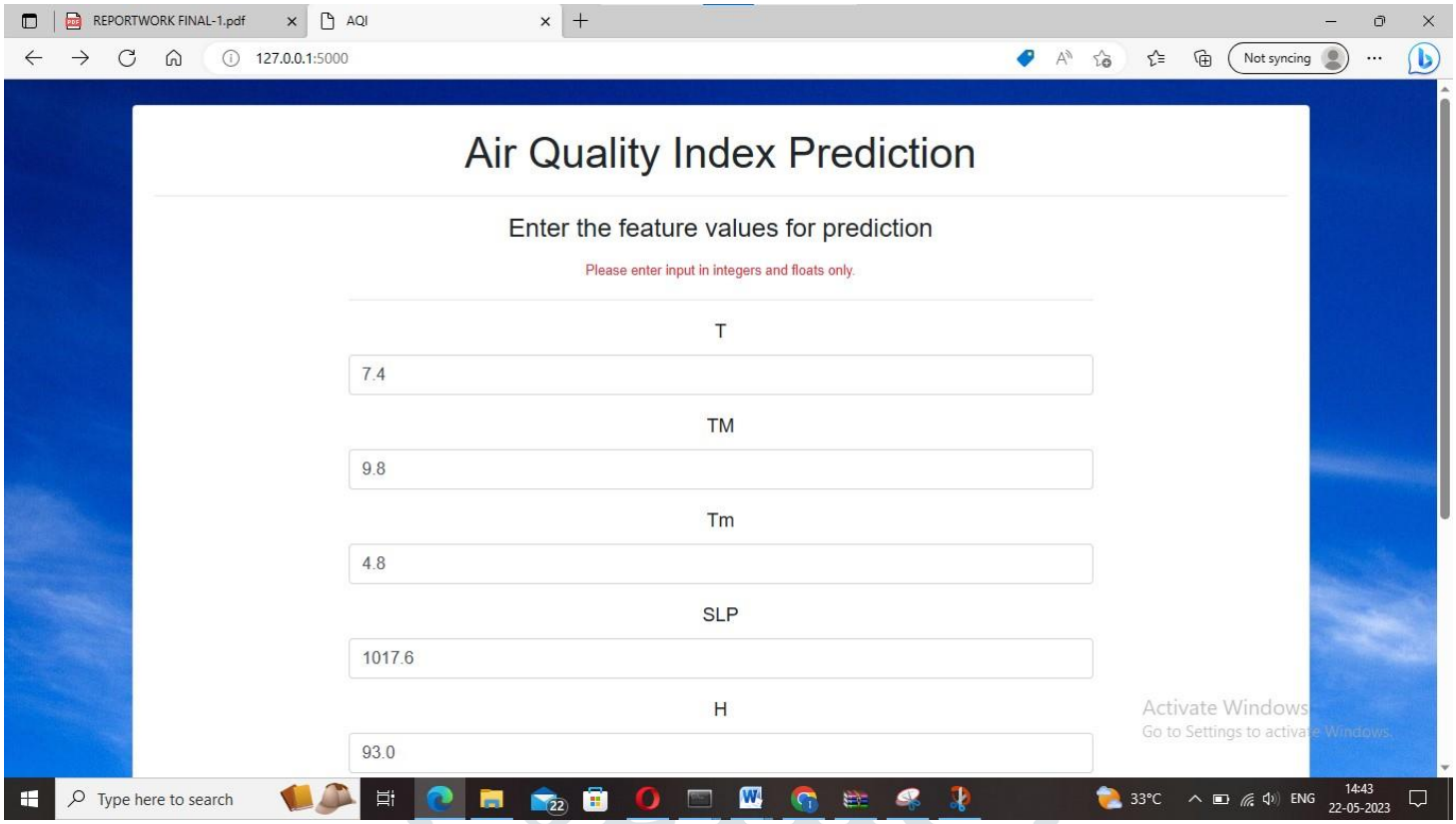


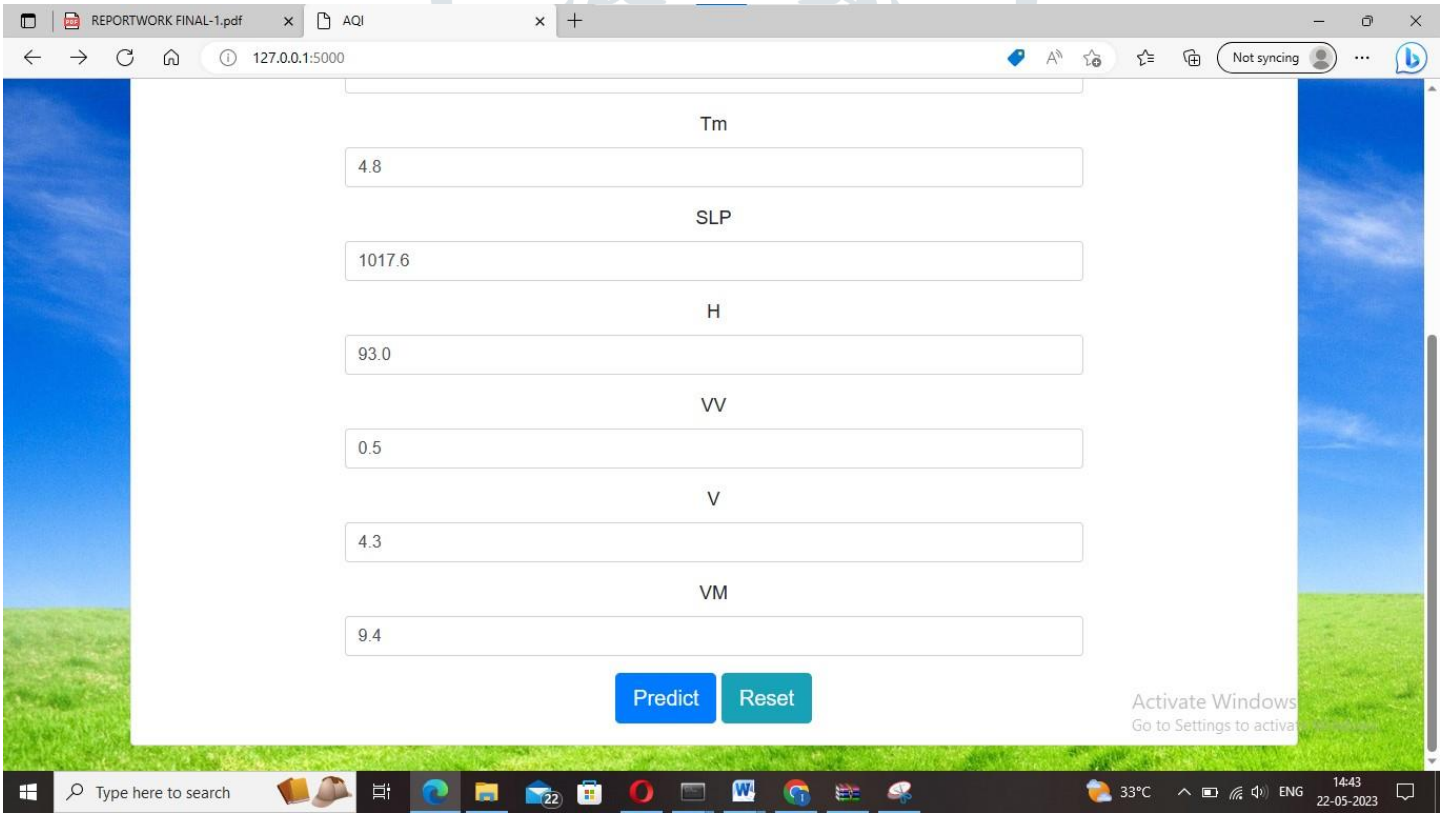
Fig 4.1 : System Architecture

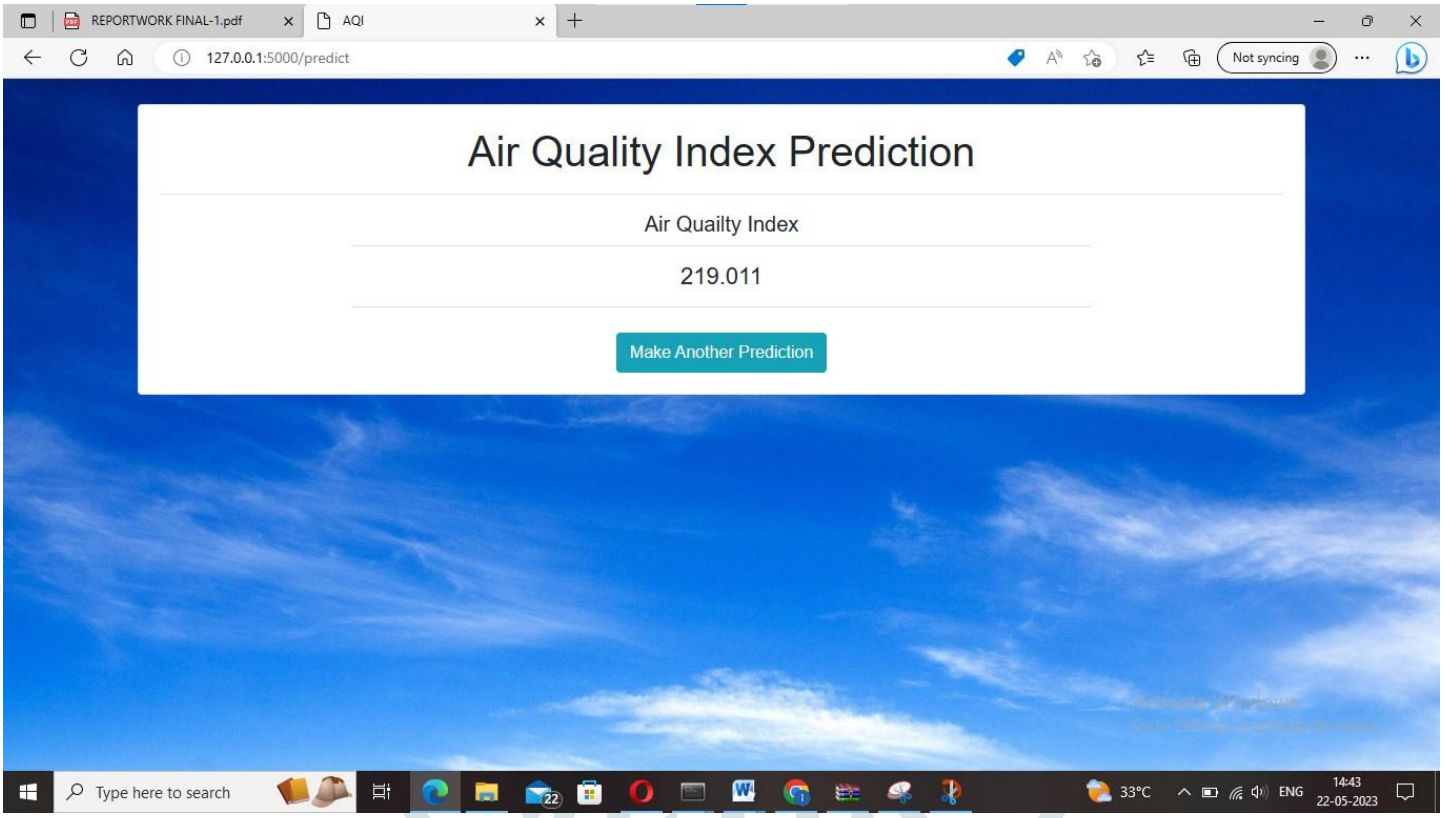
OUTPUT SCREENSHOT:






REPORTWORK FINAL-1.pdf
AQI
127.0.0.1:5000
Not syncing
Air Quality Index Prediction
Enter the feature values for prediction
Please enter input in integers and floats only.
T
7.4
TM
9.8
Tm
4.8
SLP
1017.6
H
93.0
Activate Windows
Go to Settings to activate Windows.
Type here to search
33°C
ENG
14:43
22-05-2023


REPORTWORK FINAL-1.pdf
AQI
127.0.0.1:5000
Not syncing
Tm
4.8
SLP
1017.6
H
93.0
VV
0.5
V
4.3
VM
9.4
Predict
Reset
Activate Windows
Type here to search
33°C
ENG
14:43
22-05-2023



CONCLUSION

From the analyses of the data with respect to classification, it can be stated with high confidence that the air contamination. This contamination is characterized by especially high concentrations of PM2.5 particles, Nitrogen Dioxide, Sulphur Dioxide, and Ozone. The source of the contamination is yet unknown, given the provided data, but may be related to depth of soil sample or specific site operations. Classification of air sample contamination is one that is constantly undergoing change. Most available data uses hierarchical classification to determine clusters of samples, along with principal component analysis. In future work, I would like to investigate the accuracy of hierarchical classifications. Exposure to PM2.5 when its level exceeds the limit is hazardous to humans and predicting pollutants with deep learning algorithms accurately helps the government body to signal its citizens as well as cut or regularize the sources of pollutants to a great extent. In future work, I would like to increase accuracy by giving more dataset and use upcoming model for better future.