Volume 2, Issue 8, September – 2022
www.ijsar.net
CrossRef DOI: 10.54756/IJSAR.2022.V2.i8.1

Study of Air Quality Detection using Machine Learning Techniques

Dr. D.J. Samatha Naidu[1] and R. Aruna[2]

[1]Professor, [2] P.G Scholar Master of Computer Application

Annamacharya P.G. College of Computer Studies

Rajampet

India

ABSTRACT

Over the past few decades, due to human activities, industrialization, and urbanization, air pollution has become a life-threatening factor in many countries around the world. Air, an important natural resource, has been compromised in terms of quality by economic activities. Pollution is a severe problem in areas where population density is high such as metropolitan cities. Various sorts of emissions caused by people's actions, like transportation, power, and fuel use, are affecting air quality. Considerable research has been dedicated to predicting instances of poor air quality, but most studies are limited by insufficient longitudinal data, making it difficult to account for seasonal and other factors. We forecast air quality by using machine learning to predict the air quality index of a given area. The air quality index is a dataset for a typical measure used to indicate the pollutant (SO2 NO2, RSPM, SPM, and more) levels over a period. The ML models like a Decision tree and Random Forest Classifier are implemented and compared to show better accuracy.

Key Words: Air Quality, Air Quality Index, Decision Tree, Machine Learning Models, Random Forest.

1. INTRODUCTION

Worldwide, air pollution is responsible for around 1.3 million deaths annually according to the World Health Organization (WHO). The depletion of air quality is just one of harmful effects due to pollutants released into the air. Other detrimental consequences, such as acid rain, global warming, aerosol formation, and photochemical smog, have also increased over the last several decades. The recent rapid spread of COVID-19 has prompted many researchers to investigate underlying pollution-related conditions contributing to COVID-19 pandemics in countries. Several shreds of evidence have shown that air pollution is linked to significantly higher COVID-19 death rates, and patterns in COVID-19 death rates mimic patterns in the high population density and high PM2.5 exposure areas. All the above mentioned raises an urgent need to anticipate and plan for pollution fluctuations to help communities and individuals better mitigate the negative impact of air pollution. To do so, air quality evaluation plays a significant role in monitoring and controlling air pollution. In the developing countries like India, the rapid increase in population and economic upswing in cities have led to environmental problems such as air pollution, water pollution, noise pollution and many more. Air pollution has direct impact on humans' health. There has been increased public awareness about the same in our country. Global warming, acid rains, increase in the number of asthma patients are some of the long-term consequences of air pollution. Précised air quality forecasting can reduce the effect of maximal pollution on the humans and biosphere as well. Hence, enhancing air quality forecasting is one of the prime targets for the society. Sulphur Dioxide is a gas. It is one of the major pollutants present in air. It is colorless and has a nasty, sharp smell. It combines easily with other chemicals to form harmful substances like sulphuric acid, sulphurous acid etc. Sulphur dioxide affects human health when it is breathed in. It irritates the nose, throat, and airways to cause coughing, wheezing, shortness of breath, or a tight feeling around the chest. The concentration of sulphur dioxide in the atmosphere can influence the habitat suitability for plant communities, as well as animal life. The proposed system is capable of predicting quality of air using ML models.

1.1 MOTIVATION

Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. With this increasing air pollution, we are in need of implementing models which will record information about concentrations of air pollutants (so2, no2, etc.). The deposition of this harmful gases in the air is affecting the quality of people's lives, especially in urban areas. Lately, many researchers began to use Big Data Analytics approach as there are environmental sensing networks and sensor data available.

1.2 OBJECTIVE

Air Quality Index (AQI), is used to measure the quality of air. Earlier classical methods such as probability, statistics were used to predict the quality of air, but those methods are very complex to predict the quality of air. Due to advancement of technology, now it is very easy to fetch the data about the pollutants of air using sensors and then store the data in files.

1.3 EXISTING WORK

Currently there are no such optimized solution for air quality prediction depending on climate condition. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature.

Supervised learning is built to make prediction, given an unforeseen input instance. A supervised learning algorithm takes a known set of input dataset and its known responses to the data (output) to learn the regression/classification model. An algorithm is used to learn the dataset and train it to generate the model for prediction of rainfall for the response to new data or test data. Supervised learning uses classification algorithms and regression techniques to develop predictive models.

1.3.1 Limitations of Existing Work

- Methods have performance limitations because of wide range of variations in data and amount of data is limited.
- Issue involved in rainfall classification is choosing the required sampling recess of Observation- Forecasting of rainfall, which is dependent upon the sampling interval of input data.

1.4 PROPOSED WORK

Earlier techniques such as Probability, Statistics etc. were used to predict the quality of air, but those methods are very complex to predict the Machine Learning (ML) is the better approach to predict the air quality. With the need to predict air relative humidity by considering various parameters such as CO, Tin oxide, nonmetallic hydrocarbons, Benzene, Titanium, NO, Tungsten, Indium oxide, Temperature etc.

In this system the air quality of an environment can be predicted with the help of machine learning algorithm like RF and decision tree based on previous weather details.

Contributional Work

- Better accuracy.
- Can implement in real time process.
- Less time consumption process.
- Feature extracted for ML model is more accurate.

2 RELATED WORKS

The autoregressive integrated moving average model (ARIMA) is one of the most important and widely used models to forecast time series, the ability of ARIMA to forecast the monthly values for the air pollution index was studied in [10], demonstrating that it could produce forecasts that fall under the 95% confidence level. However, this method requires extensive manual intervention in terms of selecting the data fed into the system. The features to be considered must also be selected.

ML models are able to automatically look at large amounts of data and select important features, thus reducing the need for human intervention. ML models are able to achieve higher accuracies with large datasets, than classic statistical methods. Such models have long been used for AQI forecasting tasks. ML models are nonlinear, nonparametric in nature and hence are better able to handle the complexity of nonlinear elements like pollutant levels in the air [13].

Hence, they outperform statistical methods like ARIMA, which work well only with linear systems.

ML models like random forest and decision tree are able to find hidden patterns in vast quantities of data.

2.1 SYSTEM ARCHITECTURE

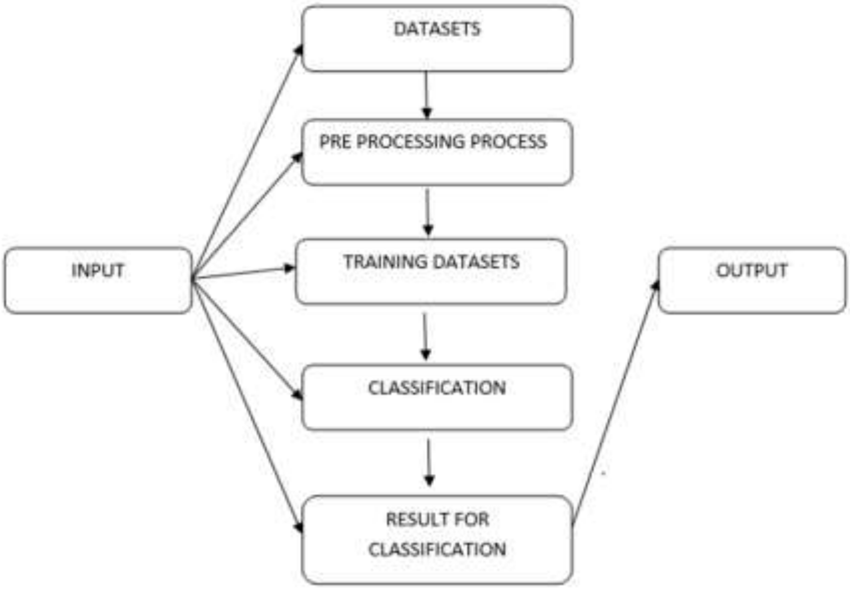


Fig 1 System Architecture

In this Figure 1 illustrate the System Architecture of Proposed work. In this we can collect the raw data as input. Once data is collected the data will be pre-processed. Then we can transform the data to Machine Learning Model. In between Feature selection is used to reduce input variable to the model by using relevant data. Here Feature Extraction includes choice of best input parameters of the chosen input dataset. After the data is transform to the Machine Learning Models. Then we can split the data into two sets. Training set and Testing set. In training set, we can train the machine learning models that means machine models will try to understand the correlations. Here validate the data and then predict the models. After that test those prediction models in testing set how accurately predict model from that we can evaluate the final model.

3.MODULES

1.Data Collection

To prognosticate the air quality of The NCR area, we want the pollutant concentration of all the elements available in the air. Which will be available in the cpcb.nic.in the website, which holds all the data that contaminates the area every year. We use data from several stations which measures many elements present in the atmosphere. Data is taken from 10 different stations in NCR. These data are stored in the form of a table which consists of a total of 3469 rows and having 8 columns in each row. The AQI formulae will be applied in order to calculate the AQI by using the various regression algorithm for a particular year.

2.Data Pre-Processing

After the input dataset is given, the data will be preprocessed by removing Null values from a data frame and replace NaN values with default values. Sometimes our data will be qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data. Then it fit the model to the data, then transform the data according to the fitted model. After the preprocessing, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers. Then using s_to_super function the first column of row(t) is shifted to last column of row (t-1) and concatenated. This act transforms a normal preprocessed dataset to recurrent dataset.

3.Feature Selection

DOI: 10.54756/IJSAR.2022.V2.i8.1

Feature selection is the method of choosing a subset from primary features that include important information to prognosticating output data. In the case of unnecessary data, feature extraction implies used. Feature extraction includes the choice of best input parameters of the chosen input dataset. The unified dataset hence gathered is used for further study. The maximum number of inputs available for review is seven, hence all the inputs are selected for the computations.

4.Splitting of Data

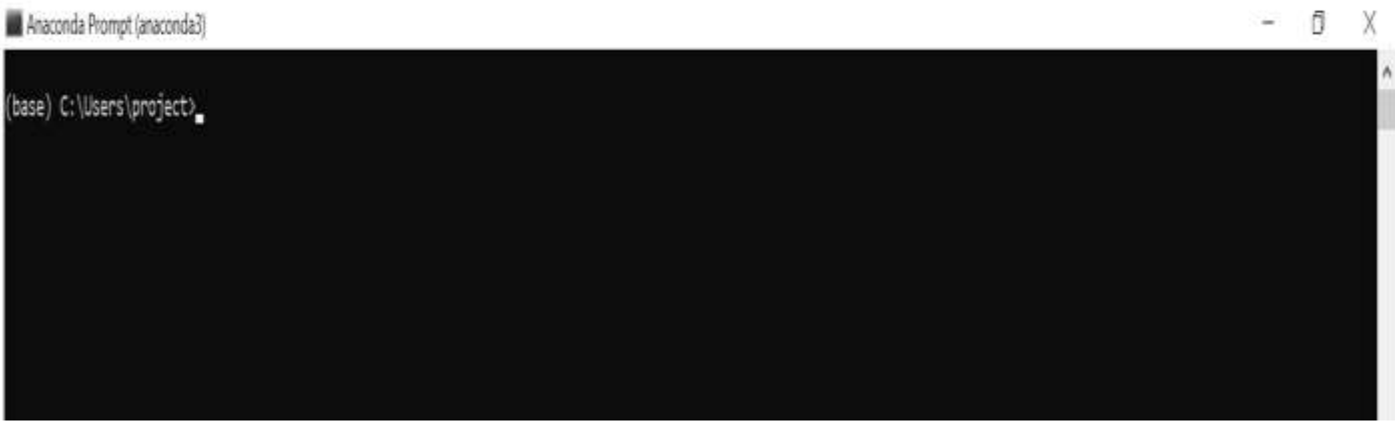
Now we need to split our dataset into two sets - a Training set and a Test set. We will train our machine learning models on our training set, i.e., our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80%of the dataset to training set and the remaining 20% to test set.

5 Training and Testing

Now to build our training and test sets, we will create 4 sets— X_train (training part of the matrix of features), x-Test (test part of the matrix of features), Y_train (training part of the dependent variables associated with the X train sets, and therefore also the same indices), Y_test (test part of the dependent variables associated with the X test sets, and therefore also the same indices). We will assign to them the test_train_split, which takes the parameters — arrays (X and Y), test size. Now, we need to build a model to train the data. Here the model used is Decision tree and random forest.

4.IMPLEMENTION SCREENS

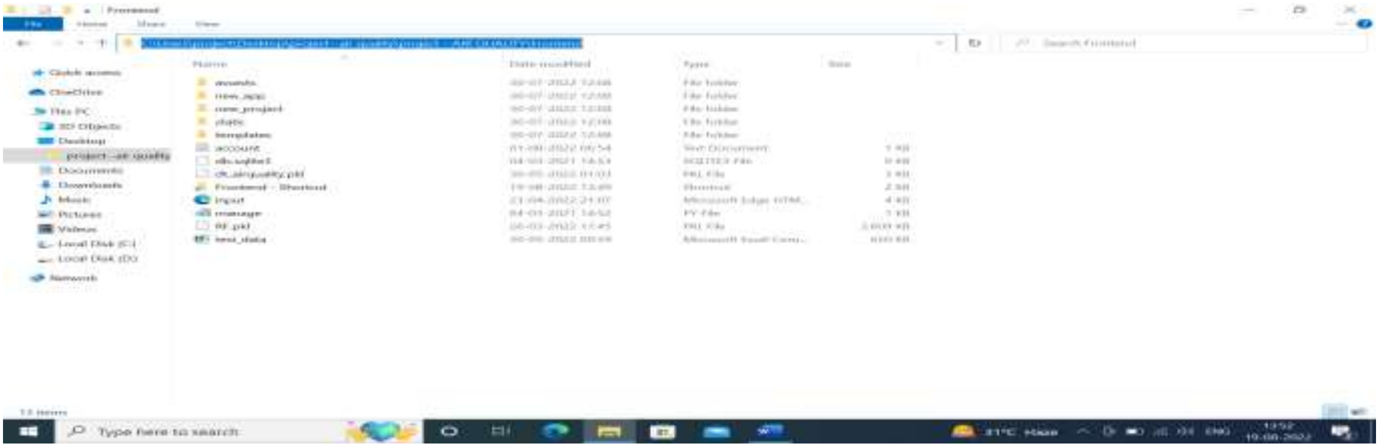
Open Anaconda Prompt is shown in figure 2.



```
(base) C:\Users\project>
```

Fig 2 Open Anaconda Prompt

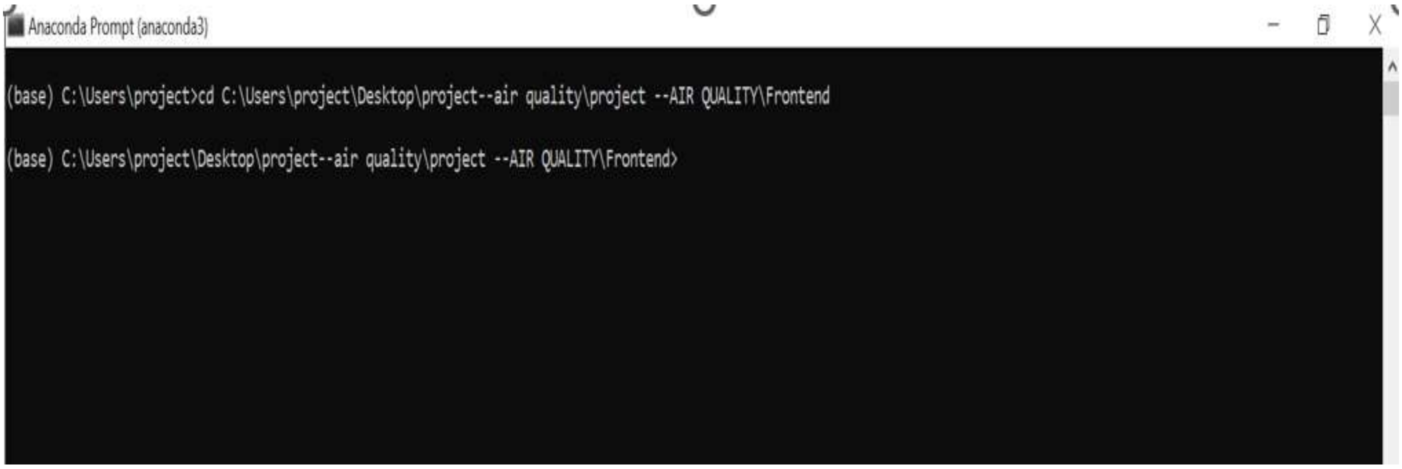
Figure 3 shows the copy of the path.



DOI: 10.54756/IJSAR.2022.V2.i8.1

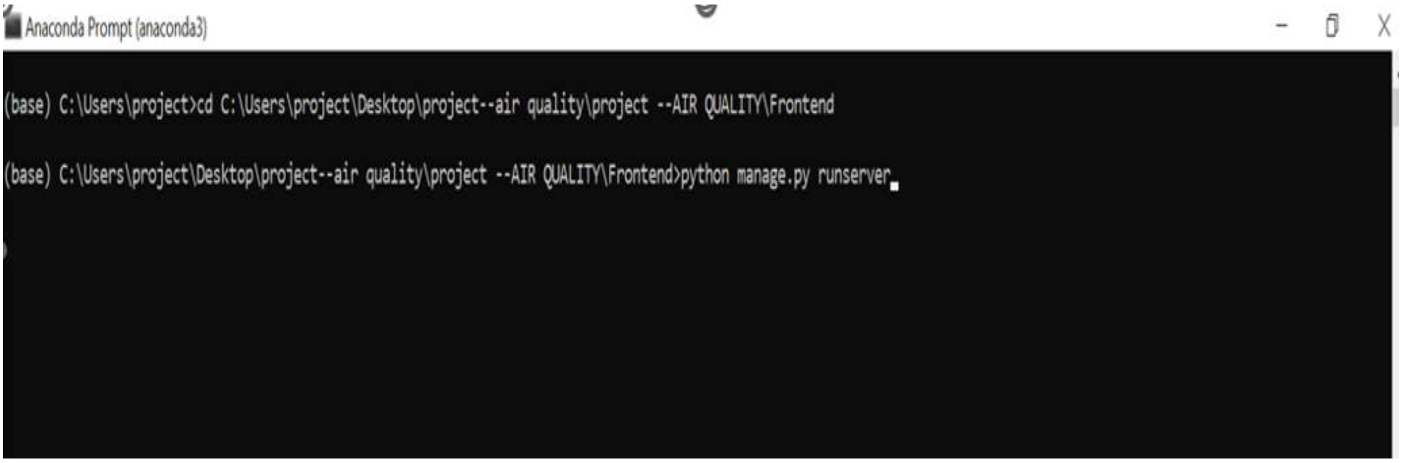
Fig 3 Copy the Path

Figure 4 shows the paste path and figure 5 depicts the python manage.py runserve.



```
(base) C:\Users\project>cd C:\Users\project\Desktop\project--air quality\project --AIR QUALITY\Frontend

(base) C:\Users\project\Desktop\project--air quality\project --AIR QUALITY\Frontend>
```

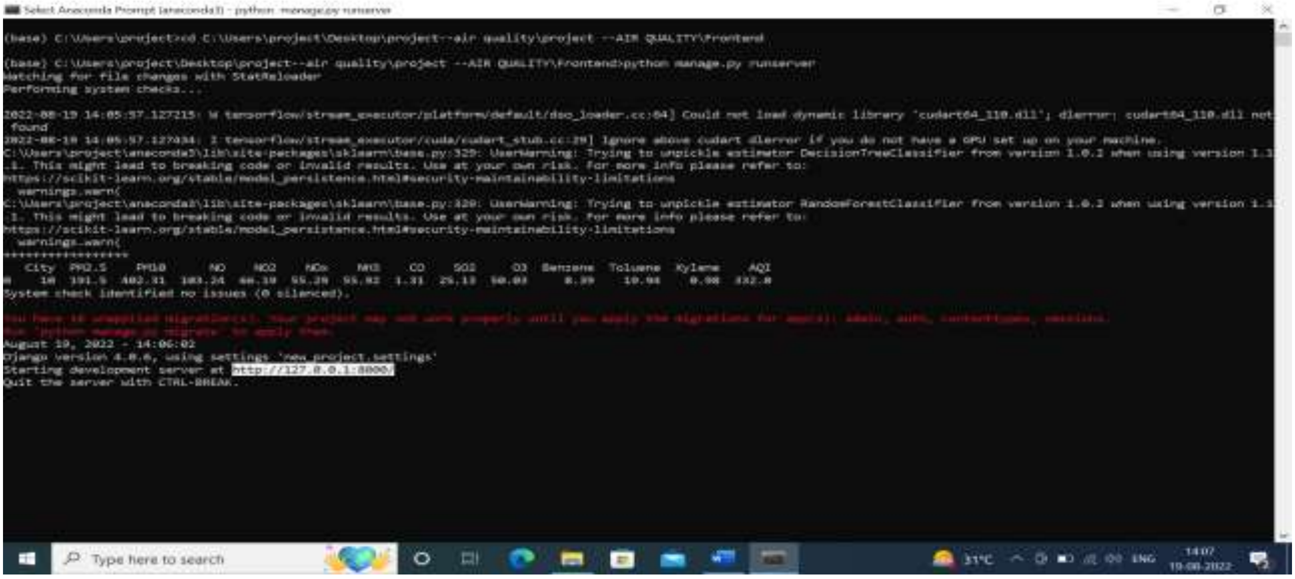
Fig 4 Paste path



```
(base) C:\Users\project>cd C:\Users\project\Desktop\project--air quality\project --AIR QUALITY\Frontend

(base) C:\Users\project\Desktop\project--air quality\project --AIR QUALITY\Frontend>python manage.py runserver
```

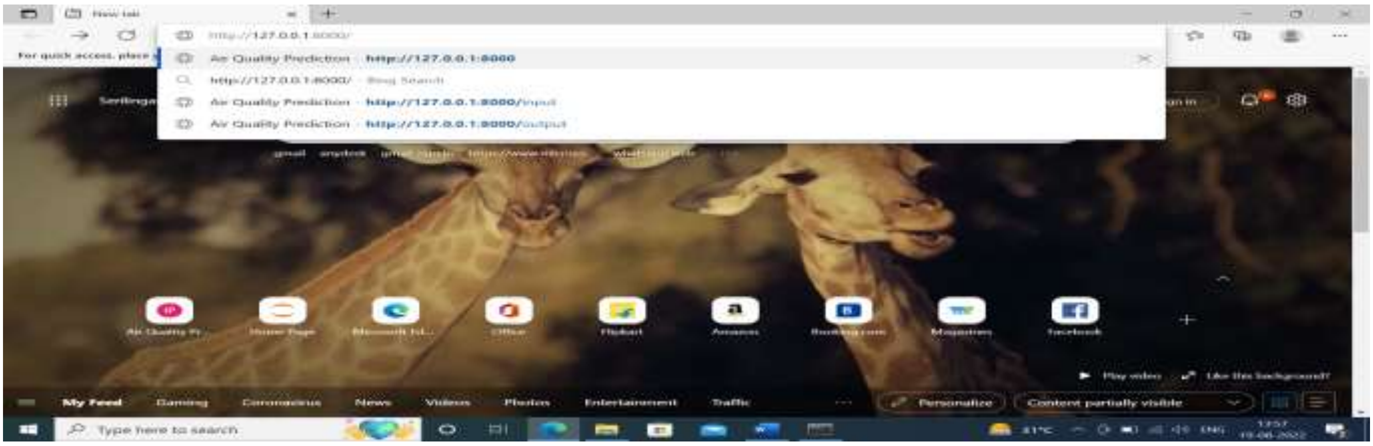
DOI: 10.54756/IJSAR.2022.V2.i8.1

Fig 5 Enter the python manage.py runserve



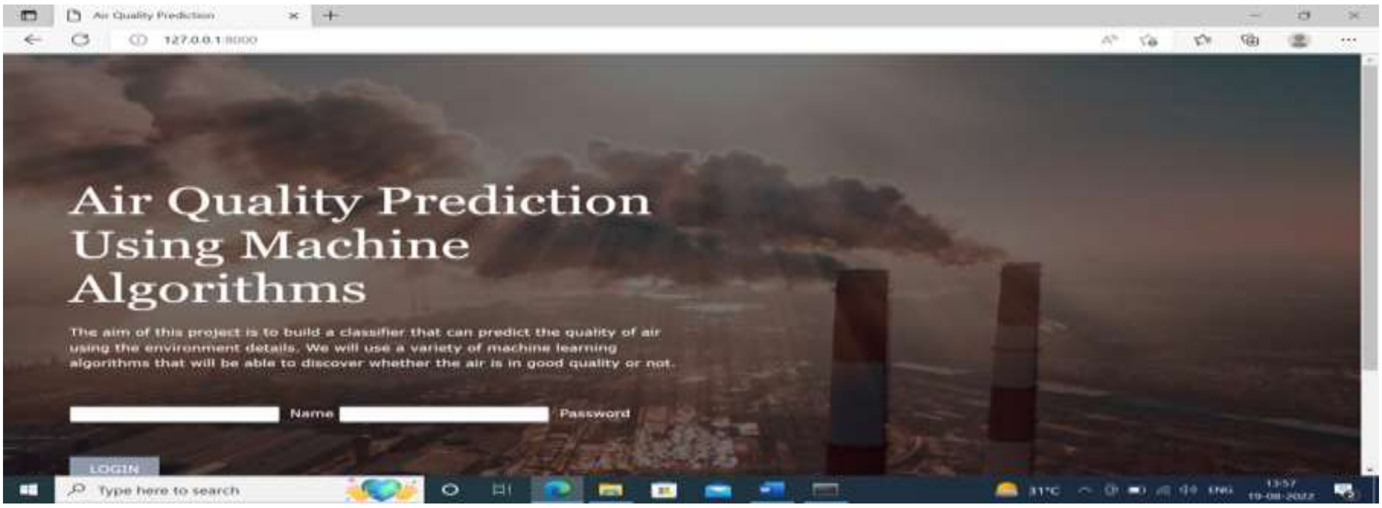
```
Select Anaconda Prompt (anaconda3) - python  manage.py runserver

(base) C:\Users\project>cd C:\Users\project\Desktop\project--air quality\project --AIR QUALITY\Frontend

(base) C:\Users\project\Desktop\project--air quality\project --AIR QUALITY\Frontend>python manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

2022-08-19 14:05:57.127215: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'cudart64_110.dll'; dlerror: cudart64_110.dll not
 found
2022-08-19 14:05:57.127434: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if you do not have a GPU set up on your machine.
C:\Users\project\anaconda3\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator DecisionTreeClassifier from version 1.0.2 when using version 1.1
.1. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
  warnings.warn(
C:\Users\project\anaconda3\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator RandomForestClassifier from version 1.0.2 when using version 1.1
.1. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
  warnings.warn(
+++++++++++++++++
   City  PM2.5    PM10      NO     NO2     NOx     NH3     CO     SO2      O3  Benzene  Toluene  Xylene    AQI
0    10  191.5  402.31  103.24  46.10   55.28   55.82  1.31  25.13  50.03     0.39    10.98    0.98  332.0
System check identified no issues (0 silenced).

You have 18 unapplied migration(s). Your project may not work properly until you apply the migrations for app(s): admin, auth, contenttypes, sessions.
Run 'python manage.py migrate' to apply them.
August 19, 2022 - 14:06:02
Django version 4.0.6, using settings 'new_project.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

Fig 6 Copy URL



```
http://127.0.0.1:8000/
Air Quality Prediction - http://127.0.0.1:8000
http://127.0.0.1:8000/ - Bing Search
Air Quality Prediction - http://127.0.0.1:8000/input
Air Quality Prediction - http://127.0.0.1:8000/output
```

Fig 7 Paste the URL in the Browser



Air Quality Prediction

127.0.0.1:8000

Air Quality Prediction Using Machine Algorithms

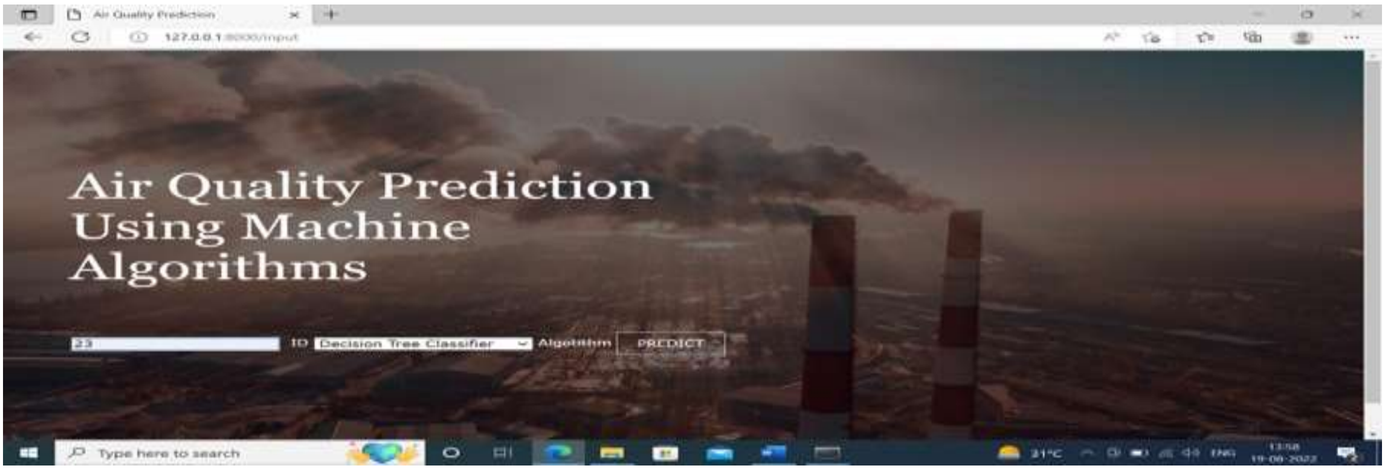
The aim of this project is to build a classifier that can predict the quality of air using the environment details. We will use a variety of machine learning algorithms that will be able to discover whether the air is in good quality or not.

Name Password

LOGIN

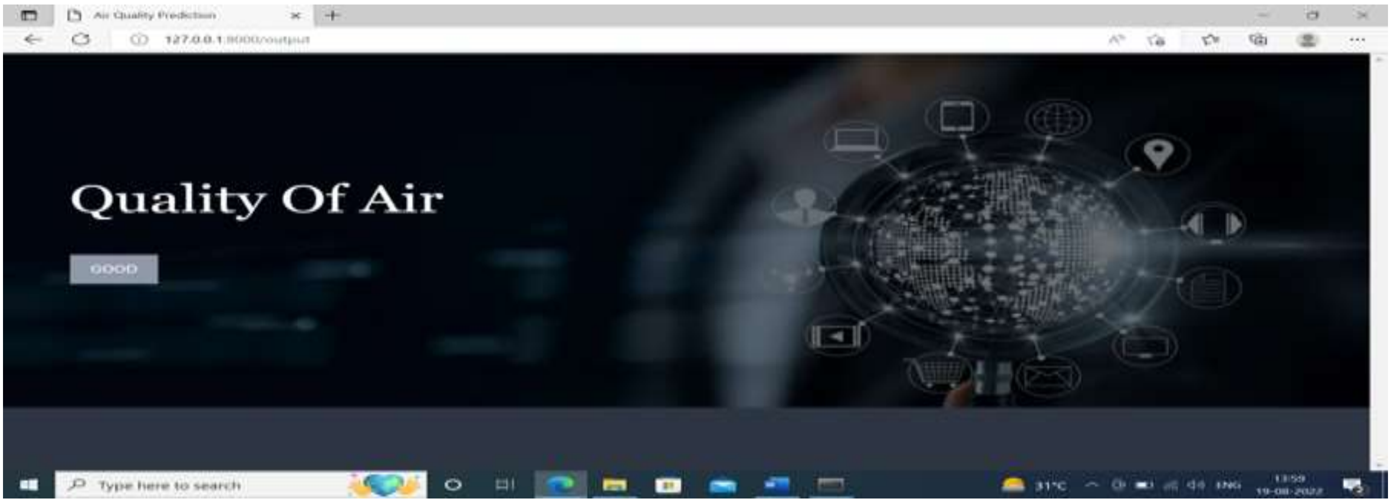
Fig 8 Login Page

DOI: 10.54756/IJSAR.2022.V2.i8.1



Air Quality Prediction Using Machine Algorithms

Fig 9 Air Quality Prediction



Quality Of Air

Fig 10 Results of the Air Quality

5.CONCLUSION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The results show that machine learning models (logistic regression and auto regression) can be efficiently used to detect the quality of air and predict the level of AQI in the future. The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities. The agenda of our work is not only to bring awareness but also to minimize pollution through proper measures and ensure that the vehicles are emitting the pollutants within the range of regular pollution check. This can lead to a pollution free region in the area.

6.REFERENCES

[1] L. Y. Siew, L. Y. Chin, P. Mah, and J. Wee, “Arima and integrated arfima models for forecasting air pollution index” The Malaysian Journal of Analytical Science, vol. 12, no. 1, pp. 257–263, 2008.View at: Google Scholar

[2] G. E. Box and D. A. Pierce, “Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,” Journal of the American statistical Association, vol. 65, no. 332, pp. 1509–1526, 1970.View at: Google Scholar

DOI: 10.54756/IJSAR.2022.V2.i8.1

[3] Khaled Bashir Shaban, Senior Member, IEEE, Abdullah Kadri, Member, IEEE, and Eman Rezk," Air Pollution Monitoring System With Forecasting Models.", IEEE(2016)

[4] Shweta Taneja, Dr. Nidhi Sharma, Kettun Oberoi, Yash Navoria," Predicting Trends in Air Pollution in Delhi using Data Mining", IEEE (2016)

[5] Nidhi Sharmaa, Shweta Tanejab*, Vaishali Sagarc, Arshita Bhattd, "Forecasting air pollution load in Delhi using data analysis tools.", Elseviere (ICCIDS 2018)

[6] KRZYSZTOF SIWEK, STANISŁAW OSOWSKI," Data mining methods for prediction of Air Pollution", amcs (2016)

[7] Mansi Yadav, Suruchi Jain and K. R. Seeja," Prediction of Air Quality Using Time Series Data Mining", Springer (2019)

[8] Manisha Bisht and K.R. Seeja," Air Pollution Prediction Using Extreme Learning Machine: A Case Study on Delhi.", Springer (2018)

DOI: 10.54756/IJSAR.2022.V2.i8.1