International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 07 | July 2022 www.irjet.net p-ISSN: 2395-0072

Air Pollution Prediction using Machine Learning

Deepu B P[1], Dr. Ravindra P Rajput[2]

[1]*Student, Dept. of Electronics and Communication Engineering, University BDT College of Engineering, Karnataka, India.*
[2]*Professor, Dept. of Electronics and Communication Engineering, University BDT College of Engineering, Karnataka, India.*

Abstract - *The air quality monitoring system collects data of pollutants from different location to maintain optimum air quality. In the current situation, it is the critical concern,. The introduction of hazardous gases into the atmosphere from industrial sources, vehicle emissions, etc. pollutes the air. Today, the amount of air pollution in large cities has surpassed the government-set air quality index value and reached dangerous levels. It has a significant effect on a human health. The prediction of air pollution can be done by the Machine Learning (ML) algorithms. Machine Learning (ML) combines statistics and computer science to maximize the prediction power. ML is used in order to calculate the Air Quality Index. Various sensors and an Arduino Uno microcontroller are utilized to collect the dataset. Then by using K- Nearest Neighbor (KNN) algorithm, the air quality is predicted.*

Keywords: Machine Learning, KNN, AQI, Arduino, sensors.

1.INTRODUCTION

Among the most crucial challenges faced in the world today is air pollution. Industrial activity is increasing more regularly due to the explosive growth of economy, which is causing air pollution to increase more rapidly. Environmental pollution is a serious issue that affects all living things, including humans, with pollution from industry accounting for a significant portion of it. Solid particles such as dust, pollen, and spores, and gases, contribute to air pollution. Carbon monoxide, Carbon dioxide, Nitrogen dioxide, Sulphur oxide, Chlorofluorocarbons, Particulate Matter, and other air pollutants that cause air pollution are released by the combustion of natural gas, coal, and wood, as well as factories, cars, and other sources. Prolonged exposure to air pollution leads to serious health problems, such as lung and respiratory illnesses

The annual death toll from household exposure to gasoline smoke is 3.8 million. Exposure to the outdoor air pollution will cause 4.2 million deaths annually. 9 out of 10 people on the earth reside in areas with air quality that is worse than recommended by the World Health Organization. As per the Greenpeace Southeast Asia Analysis of IQAir statistics, air pollution and associated problems caused over 120,000 deaths in India in 2020.According to the report, air pollution caused economic losses of ₹2 lakh crore in India. This demonstrates how crucial it is to pay attention on the air quality.

Primary pollutants and the secondary pollutants are the two major classifications of air pollutants. One that is directly emitted into the atmosphere from its source is referred to as a primary pollutant, whereas a secondary pollutant is one that is produced due to the interaction between two primary pollutants or with other elements of the atmosphere. One of the detrimental effects of pollutants emitted into the environment is the degradation of air quality. Also, other harmful effects, such as acid rain, global warming, aerosol production, and photochemical smog has increased in past years.

Predicting the air quality is crucial for preventing the problem of air pollution. The Machine Learning (ML) models can be used for this. With the use of training data, a computer can learn how to build models via a technique called as Machine Learning. It is a branch of Artificial Intelligence that gives computer program the ability to forecast outcomes with ever-increasing accuracy. ML can examine a variety of data and identify patterns and particular trends. Machine learning is the ability given to a computer program to do a task without any external programming and this is task is achieved by using some statistical and advanced mathematical algorithms.

As air pollution has been rising every day, monitoring has proven to be a significant task. The amount of pollution in a given area is determined through continuous air quality monitoring at that location. The information obtained by the sensors reveals the source and concentration of the pollutants in that area. Measures to minimise pollution levels can be taken using that knowledge and the ML model.

The hardware device consists of three different sensors like MQ-135 air quality sensor, MQ-5 sensor, Optical dust sensor connected to the Arduino uno board, which helps in collecting the pollutants information of the current place.

The program for collecting the information is written in the Arduino IDE according to the AQI level specified by the Central Pollution Control Board of India, in the report National Air Quality Index. The information collected from the sensors is recorded in the excel sheet, then it is stored in the required file path, which makes the dataset. Further the excel sheet in the .csv file is directly read in the ML program.

2. LITERATURE SURVEY

The authors of [1] proposed that Machine Learning algorithms plays important role in measuring air quality index accurately. Logistic regression and auto regression, ANN help in determining the level of PM2.5. ANN comes out with best results in the paper.

In [2] authors gives the prediction of the air quality index by using different machine learning algorithms like Decision Tree and Random Forest. From the results, concluded that the Random Forest algorithm gives better prediction of air quality index.

In [3] authors proposed model by using BILSTM which is the Deep Learning model to predicted the PM2.5 with improved performance comparing the existing model and produced exceptional MAE, RMSE.

In [4] authors used the prediction model results were based on Big Data Analytics and Machine Learning, which have helped to evaluate and contrast current assessments of air quality. The Decision Tree algorithm gave the best results among all the algorithms.

The authors of [5] used SVR, and LSTM Machine Learning models. The Machine Learning algorithms used for estimating the atmospheric pollutants (PM10 and PM2.5), it was demonstrated that SVR algorithms are the most suitable in forecasting the air pollutants concentrations.

3. METHODOLOGY

Information about air pollutants is obtained from the sensors, analysed, and then saved as a dataset. This dataset has been pre-processed with a variety of features, which includes attribute selection and normalisation. Once it is available, the dataset is divided into a training set and a test dataset. The training dataset is then used to apply a Machine Learning algorithm. The obtained results are matched with the testing dataset and results are analysed.

3.1 Machine Learning model

Machine Learning algorithm is implemented to predict the air pollution. Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that enables the software applications to be accurate in predicting the outcomes without being explicitly programmed to do so.

To predict the new outcomes, Machine Learning algorithms make use of existing past data as the input. With the help of Machine Learning, a user can provide a computer program huge amount of data, and the computer will only examine that data and draw conclusions from it.

KNN is the Machine Learning algorithm used for the prediction of air pollution. The K-Nearest Neighbors (KNN) algorithm is one of the types of Supervised Machine Learning algorithms. KNN is incredibly simple to design but performs quite difficult classification jobs. KNN is called the lazy learning algorithm as it lacks the training phase. Instead, it classifies a fresh data point while training on the entire dataset. It does not make any assumptions, hence it is called non-parametric learning method.

Steps in KNN:

- Determine the distance between each sample of the training data and the test data.
- To determine distance, we can utilise the Euclidian or Minkowski or Manhattan distance formula.
- Sort the estimated distances in ascending order.
- Vote for the classes.
- Output will be determined based on class having most votes.
- Calculate the Accuracy of the model, if required rebuild model.

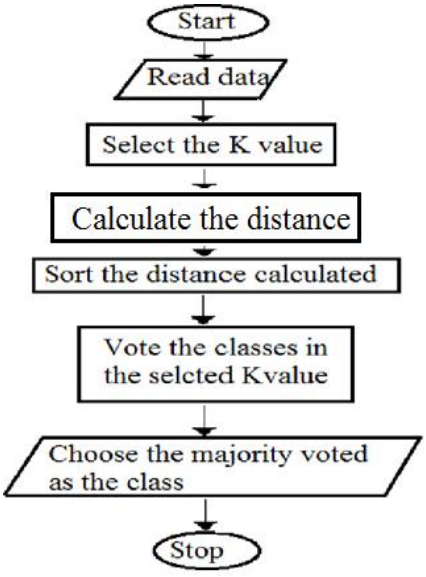

Fig-1: Flow chart of KNN

3.2 Sensors used

MQ-135 air quality sensor can detect gases like Ammonia (NH3), sulfur (S), Benzene (C6H6), CO2, and other harmful gases and smoke. MQ5 is a sensitive gas sensor that can detect or sense liquefied gas, propane, butane, natural, other combustible gases in the environment and smoke. Optical dust sensor that means it senses dust by using an optical sensing system such as a light source. It is used to detect dust particles in the air.

3.3 Air Quality Index

AQI	Associated Health Impacts
Good (0–50)	Minimal Impact
Satisfactory (51–100)	May cause minor breathing discomfort to sensitive people
Moderate (101–200)	May cause breathing discomfort to the 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 with short exposure
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 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

Fig-2: AQI

The Central Pollution Control Board of India provided the AQI in the report National Air Quality Index, which is shown in the Fig 2 above.

According to this AQI the program is written in the Arduino IDE to collect the dataset from the current place. Further this dataset is recorded in the excel sheet and saved in the particular file path as required.

3.4 Implementation

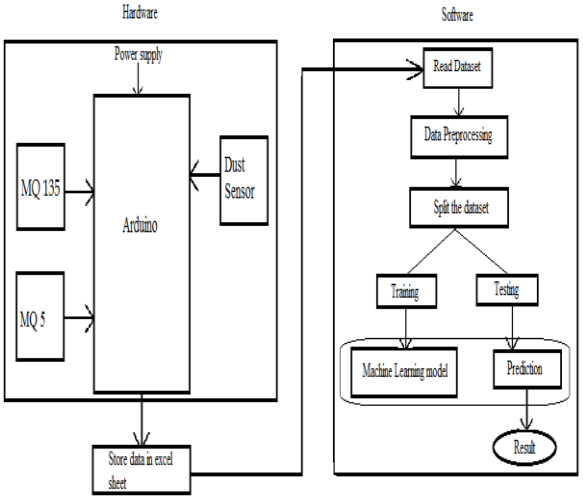

Fig-3: Block diagram

Hardware Connections:

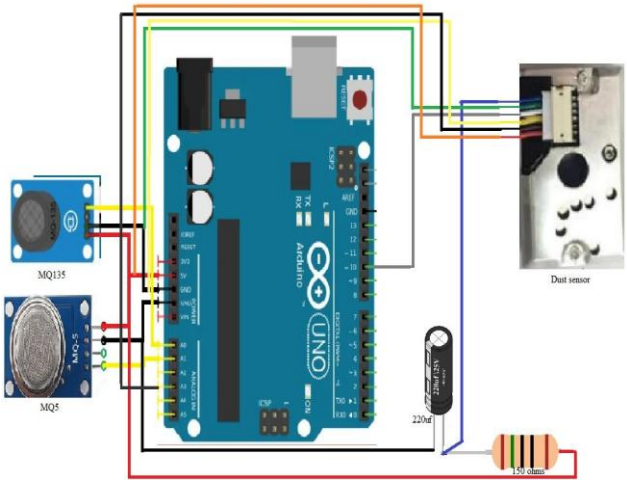

Fig- 4: Hardware connections

MQ135 sensor connections: The MQ135 sensor Vcc is wired to the Arduino 5V pin. The GND of MQ5 and GND of Arduino are wired together, AO of MQ5 and A0 of the Arduino are connected.

MQ5 sensor connections: The MQ5 sensor Vcc is wired to the Arduino 5V pin. The GND of MQ5 and GND of Arduino are wired together, AO of MQ5 and A1 of the Arduino are connected.

Dust sensor connections: Connect the V-LED (blue) pin of sensor to Arduino 5V pin with a capacitor of 220uf and a resistor of 150 ohms in between. Now connect LED-GND (green) and S-GND (yellow) to the GND pin of Arduino. Connect the Vcc (red) of sensor to Vcc of Arduino. Next connect VOUT (black) to A3 of Arduino and connect LED (white) of sensor to digital pin 10 of Arduino.

Steps to Collect Data:

- After the hardware setup, upload the code in the arduino IDE.
- Open the excel sheet with data streamer downloaded.
- Click on the Data Streamer on the menubar.
- Click on the Connect the Device.
- Select the COM PORT.
- Click on Start Data on the toolbar.
- To start recording of data click, Record Data.
- After collecting the data click on Stop Recording to stop recording, and then click on Stop Data to end collecting data.
- This excel file can be saved in the required file path.

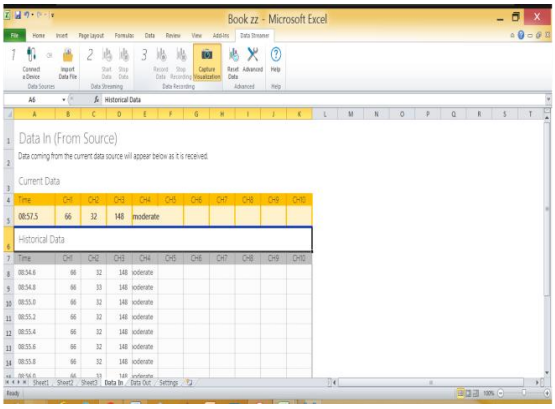

Fig-5: Data collection

The software used is Anaconda Navigator for Python, which features the web-based Interactive Development Environment for data, code and notebooks called Jupyter Notebook. The users could build and arrange the workflows in data sciences, machine learning and scientific computing, using its interface. The Jupyter Notebook is the original web tool for producing computational documents.

Steps in Software Implementation

1) Read dataset:

The required libraries are imported and then the dataset is read in the python code.

Parameters considered in the dataset are the sensor values of air quality, smoke and dust and the respective quality of air for the values from the sensors of the current place. Hence, there are four columns in the dataset and the number of rows depends on the time that the data is recorded. This dataset is saved as the .csv file in excel.

```
data=pd.read_csv(r'C:\Users\USER\Desktop\Air Pollution Prediction\air pollution.csv')
print(data)
```

```
      air  smoke  dust        quality
0      61     37    50   satisfactory
1      61     37    50   satisfactory
2      61     37    50   satisfactory
3      61     37    50   satisfactory
4      61     37    50   satisfactory
...   ...    ...   ...            ...
1114   63     36   498         severe
1115   63     36   498         severe
1116   63     36   498         severe
1117   63     37   498         severe
1118   63     36   498         severe

[1119 rows x 4 columns]
```

Fig-6: Reading dataset

2) Split the training and testing dataset:

The training set is used to train the model, and the testing set is used to determine whether the model generalises well to new and unexplored data.The better outcomes are attained when 20% to 30% of the data are used for testing and the rest 70% to 80% for training.

This is done by importing the train_test_split library from the Sci-kit, where training and testing ratio is taken 80% and 20% respectively.

3) Choosing the Machine Learning model

KNN is the Machine Learning model chose for the prediction of air pollution.

4) Prediction

After the ML model is fit, it gives the prediction of air quality based the AQI described, of the current place i.e., whether the air quality is satisfactory or moderate to breathe, or poor so that people can decide the impact of air pollution, or very poor and severe to survive in that place.

```
plt.scatter(y_test,KNN_predicted)
```

```
<matplotlib.collections.PathCollection at 0xf1a7f84670>
```

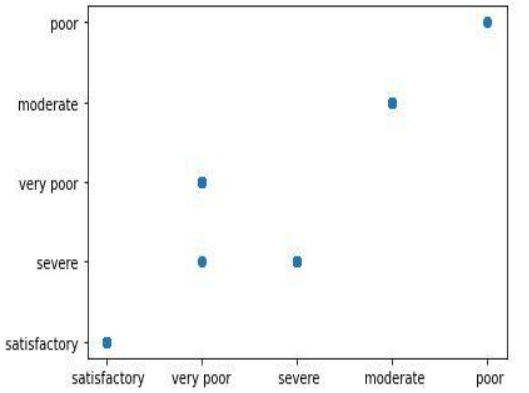

Fig-7: Scatterplot of y_test and predicted values

4. RESULT AND DISCUSSION

The confusion matrix of the particular dataset air pollution, which is read initially at the time of read dataset is as shown in the Fig 8 below.

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test,KNN_predicted)
```

```
array([[ 17,   0,   0,   0,   0],
       [  0,   2,   0,   0,   0],
       [  0,   0,  92,   0,   0],
       [  0,   0,   0, 102,   0],
       [  0,   0,   0,   2,   9]], dtype=int64)
```

Fig-8: Confusion matrix

The accuracy of confusion matrix shown in the figure is,

Accuracy = (17 + 2 + 92 + 102 + 9) /

(17 + 2 + 92 + 102 + 9 + 2)

= 222/ 224

= 99.1071 %

```
In [21]: inp = np.array([[75], [51],[48]])
         inp = inp.reshape(1, -1)
         output = KNN.predict(inp)
         output
Out[21]: array([' satisfactory'], dtype=object)

In [15]: inp = np.array([[75], [50],[148]])
         inp = inp.reshape(1, -1)
         output = KNN.predict(inp)
         output
Out[15]: array([' moderate'], dtype=object)

In [16]: inp = np.array([[68], [51],[248]])
         inp = inp.reshape(1, -1)
         output = KNN.predict(inp)
         output
Out[16]: array([' poor'], dtype=object)

In [17]: inp = np.array([[75], [51],[348]])
         inp = inp.reshape(1, -1)
         output = KNN.predict(inp)
         output
Out[17]: array(['very poor'], dtype=object)
```

Fig-9: Predicted outcomes

5. CONCLUSION

The quality of the air is determined by components like gases and particulate matter. These pollutants decrease the air quality, which can lead to serious illnesses when breathed in repeatedly. With air quality monitoring systems, it is possible to identify the presence of these toxics and monitor air quality in order to take sensible measures to enhance air quality. As a result, production rises and health problems caused by air pollution are reduced.

The prediction models built using machine learning have been shown to be more reliable and consistent. Data collecting is now simple and precise due to advanced technology and sensors. Only machine learning (ML) algorithms can effectively handle the rigorous analysis needed to make accurate and efficent predictions from such vast environmental data. In order to predict air pollution, the KNN algorithm is used, which is better suitable for prediction tasks.

The Machine Learning algorithm KNN, has given the accuracy of 99.1071% in the air pollution prediction.

REFERENCES

[1] Shreyas Simu,Varsha Turkar, Rohit Martires, “Air Pollution Prediction using Machine Learning”, 2020, IEEE

[2] Tanisha Madan, Shrddha Sagar, Deepali Virmani, “ Air Quality Prediction using Machine Learning Algorithms”, 2020, IEEE

[3] Venkat Rao Pasupuleti, Uhasri , Pavan Kalyan, “Air Quality Prediction Of Data Log By Machine Learning”, 2020 , IEEE

[4] S. Jeya, Dr. L. Sankari, “Air Pollution Prediction by Deep Learning Model”, 2020, IEEE

[5] SriramKrishna Yarragunta, Mohammed Abdul Nabi, Jeyanthi.P, “Prediction of Air Pollutants Using Supervised Machine Learning”, 2021, IEEE

[6] Marius, Andreea, Marina, “ Machine Learning algorithms for air pollutants forecasting”, 2020, IEEE

[7] Madhuri V.M, Samyama Gunjal G.H, Savitha Kamalapurkar, “Air Pollution Prediction Using Machine Learning Supervised Learning Approach”, 2020, International Journal Of Scientific & Technology Research, Volume 9, Issue 04.

[8] K. Rajakumari, V. Priyanka,“Air Pollution Prediction in Smart Cities by using Machine Learning Techniques”, 2020, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume 9, Issue 05.