Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

MACHINE LEARNING ALGORITHM FOR PREDICTING AIR POLLUTION

[#1]SANGEM AJAY,

[#2]Dr.V.BAPUJI, ***Associate Professor& HOD,***

Department of Master of Computer Applications,

VAAGESWARI COLLEGE OF ENGINEERING, KARIMNAGAR, TELANGANA

Abstract: The air quality monitoring system collects information on contaminants from multiple sites to maintain optimal air quality. It is currently the most pressing issue. The release of hazardous chemicals from industrial sources, as well as car emissions, damage the atmosphere. Today's main cities have dangerously high levels of air pollution that exceed the government-mandated air quality index standard. It has a tremendous impact on a person's health. Air pollution predictions can be made using Machine Learning (ML) techniques. Machine learning (ML) combines statistics and computer science to improve prediction accuracy. Machine learning (ML) is used to calculate the Air Quality Index. A variety of sensors and a microcontroller known as an Arduino Uno are used to collect the data. The K-Nearest Neighbor (KNN) approach is then used to anticipate air quality.

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

1. INTRODUCTION

One of the most pressing problems humanity faces today is air pollution. Industry activity is picking up speed as a result of the brisk economic expansion. Because of this, levels of air pollution are rapidly increasing. contamination from industry is a major contributor to environmental contamination, which is bad for humans and all other forms of life. Solids and gases, such as dust, pollen, and bacteria, contribute to air pollution. Burning natural gas, coal, or wood creates carbon monoxide, carbon dioxide, nitrogen dioxide, sulfur oxide, chlorofluorocarbons, particulate matter, and other air pollutants. Major health issues, including lung and breathing disorders, have been linked to prolonged exposure to dirty air. About 3.8 million people each year are killed by exposure to gasoline fumes in their houses. Air pollution is responsible for the premature deaths of 4,2 million people worldwide every year. The air quality in which 90% of the world's population resides is below the standards set by the World Health Organization. According to the Southeast Asia Analysis of IQAir conducted by Greenpeace, approximately 120,000 people in India will be killed by air pollution and related ailments in 2020.The study found that air pollution cost India's GDP 2 trillion rupees.

This highlights the significance of maintaining a vigilant vigilance on air quality. Primary pollutants and secondary pollutants are the two most common forms of air pollution. Primary pollutants are those that enter the environment unfiltered. When two main pollutants combine or react with one another or with other components of their environment, a secondary pollutant is produced. Air pollution is just one effect that pollutants have on their environments. Other issues that have worsened in recent years include acid rain, global warming, aerosol generation, and photochemical smog.

Predicting the weather is a crucial step in reducing air pollution. Machine Learning (ML) models can be used for this purpose. In order to educate a computer to create models, machine learning is used. It is a subfield of AI that trains computers to anticipate future events with increasing accuracy. In order to identify patterns and tendencies, ML may examine a wide variety of data. Statistics and advanced mathematics are employed for this purpose.

Air quality has been difficult to monitor due to its steady decline. The frequency with which air quality is measured can be used to estimate the level of pollution present. According to the data gathered by the sensors, we can see exactly where and how much pollution is there. The ML model and this information can be used to develop strategies for reducing pollution. A MQ-135 air quality sensor, a MQ-5 sensor, and an optical dust sensor make up the hardware device. These three sensors, which are linked to an Arduino uno

board, are used to assess pollution levels in the area.

The AQI from the Central Pollution Control Board of India report was used to calibrate the Arduino IDE's data-gathering application. The sensor readings are imported into an Excel spreadsheet and then saved in the appropriate directory. Dataset describes this. The ML software may import an Excel sheet directly from a.csv file.

2 LITERATURE SURVEY

The advantages of forecasting the severity of air pollution using the Bidirectional Long-Short Memory [BiLSTM] approach. The proposed method improved forecasts by simulating PM2.5's long-term, short-term, and crucial effects. The provided procedure allows for estimations to be made at 6, 12, and 24 hour intervals. After 12 hours, results are consistent, but after 6 hours and 24 hours, they vary widely. According to Chao Zhang, air quality could be predicted with the help of web services. They provided a way for mobile phone users to send in evidence of polluted environments. The proposed method consists of two stages. They were able to obtain reports from adjacent air quality stations using their GPS coordinates. b) They employed language learning and a convolution neural network to analyze user-submitted images and make predictions about the air quality. The proposed approach is more error-prone than state-of-the-art algorithms such as PAPLE, DL, and PCALL, but it does not learn as well, hence it produces less precise results. Using data from the city of Shanghai, researcher Ruijun Yang constructed a DAG to show how polluted the air was. The dataset is split into training and evaluation subsets. This approach fails to account for contextual elements like location and society. Therefore, depending on these variables, the outcomes could be different. TemeseganWalelignAyelee demonstrated how to monitor air quality remotely using the IoT. Using a technique called Long Short-Term Memory (LSTM), we were able to foretell how the air will be. By reducing the training time, the proposed strategy improved the model's accuracy. However, by contrasting several approaches, such as the Random Forest method, accuracy can be increased. The most prevalent forms of air pollution, carbon monoxide and nitrogen oxides, were predicted using a nonlinear, fictitious Regressive model developed by NadjetDjebbriet. Things including wind velocity, wind direction, temperature, humidity, and the presence of potentially harmful substances were examined in industrial regions like Skikda. They evaluated performance using RMSE and MAE, although this technique ignored all contaminants except for NO and CO. Other significant toxins, such as sulfur dioxide, PM2.5, and PM10, are ignored.

3. SYSTEM DESIGN

Air pollution data is gathered via sensors, analyzed, and documented in a database. Many operations have already been performed on this dataset, including attribute selection and standardization. Upon completion of the data collection process, the data is partitioned into a training set and a test set. A Machine Learning technique is then applied to the sample data used for training. The received data is compared to the test data, and the differences are examined.

Machine Learning model

Predictions of air pollution levels are made using a machine-learning algorithm. Machine learning (ML) is a subfield of AI that enables autonomous, accurate prediction of future events within software systems. Predictions made by Machine Learning systems are grounded in historical data. With machine learning, one can provide a large amount of data to a computer program, and the program will analyze the data and form conclusions. K-Nearest Neighbor (KNN) is a machine learning technique used to forecast the air quality.

K-Nearest Neighbors (KNN) is a machine learning technique that takes advantage of observational learning. Although KNN is a very straightforward classification method, it is capable of performing complex classification tasks. Since KNN doesn't require any training, it's sometimes referred to as the "lazy learning" algorithm. Instead, it uses the existing data to learn how to identify new data as it accumulates. An "non-parametric learning approach" is one that does not use parameters or parameters to guide its analysis.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

Steps in KNN:

- Learn how widely off your training data and your test data actually are.
- Distance can be calculated using the Euclidean, Minkowski, or Manhattan approaches.
- Arrange the lengths in alphabetical order.
- The voting system will determine the groupings.
- The winning group will be determined by the group with the most votes cast.
- Determine the precision of the model and recreate it if necessary.

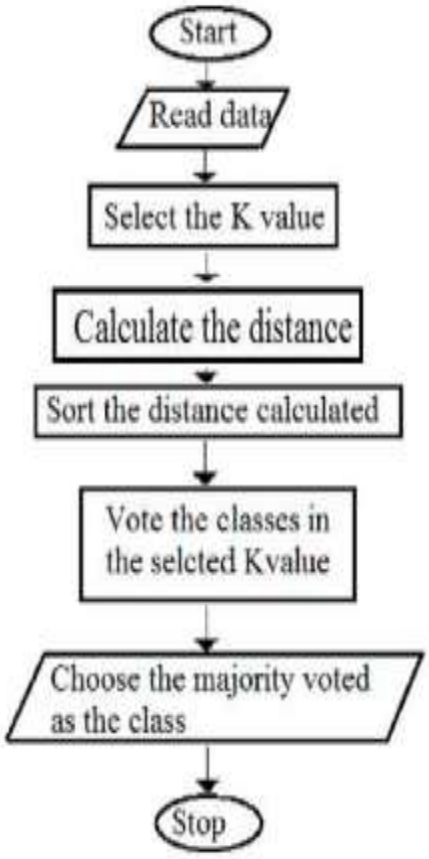

Fig-1: Flow chart of KNN

Sensors used

The MQ-135 air quality tester is capable of detecting dangerous gases and particles in the air, including ammonia (NH3), sulfur (S), benzene (C6H6O), and carbon dioxide (CO2). The MQ5 gas detector is sensitive enough to detect combustible gases such as liquefied gas, propane, butane, natural gas, and smoke. A lens or other optical detector is used in an optical dust monitor to detect the presence of dust. It is a tool for determining the level of dust in the atmosphere.

Air Quality Index

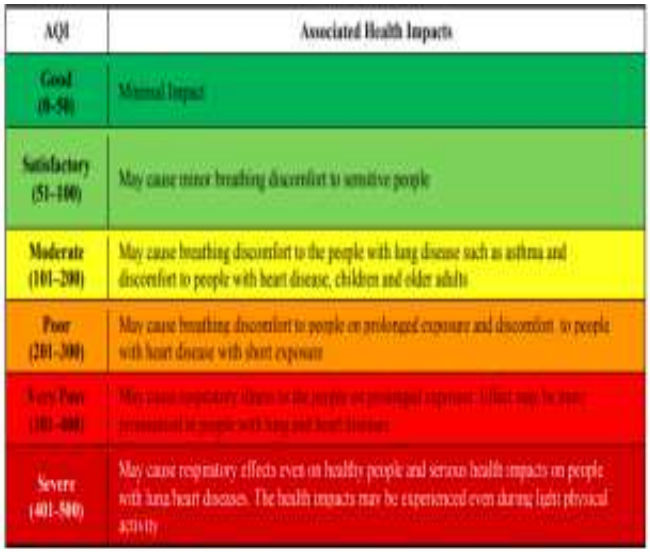

Fig-2: AQI

As can be seen in Figure 2, the AQI was included in the National Air Quality Index published by the Central Pollution Control Board of India.

This AQI indicates that the Arduino IDE was developed to collect data in the immediate vicinity. This data collection is also described in an Excel file that has been properly filed away.

4. SYSTEM ANALYSIS

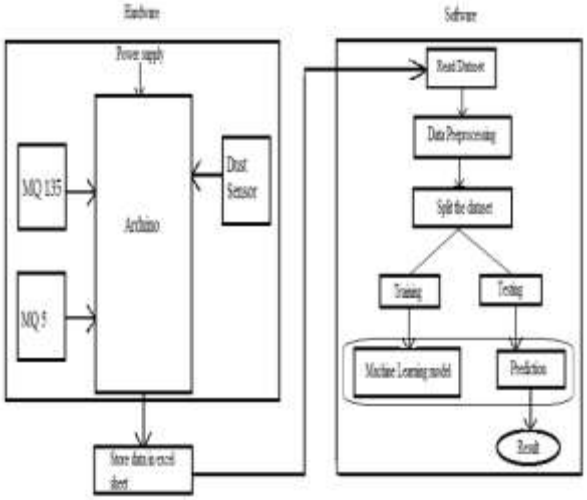

Fig-3: Block diagram

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

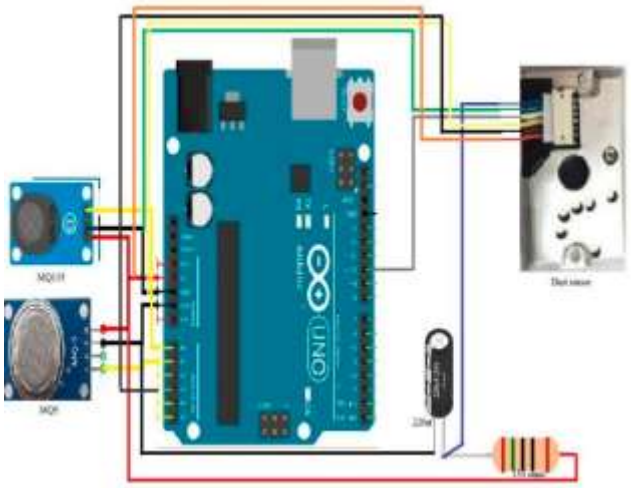

Fig- 4: Hardware connections

The MQ135 sensor's Vcc is connected to the Arduino's 5V supply. The MQ5's AOs are connected to the Arduino's A0s, and both boards' GNDs are connected together.

Connectors for MQ5 sensors: The MQ5 sensor's Vcc is connected to the Arduino's 5V supply. Both the AO and A1 pins on the MQ5 and the GND pins on the Arduino are connected.

Dust sensor connections:

Connect the sensor's V-LED (blue) pin to the Arduino's 5V pin using a 220uF capacitor and a 150 ohm resistor. The GND pin on the Arduino should be connected to the green LED-GND and yellow S-GND cables. The red Vcc wire on the sensor must be connected to the red Vcc wire on the Arduino. Plug the sensor's white LED into digital pin 10 and the black VOUT into A3 on the Arduino.

Steps to Collect Data:

Load the code into the Arduino IDE once you've gotten everything set up. To access the data source, open the corresponding Excel file.

Select Data Streamer from the menu.

Select "Connect the Device" from the menu that appears.

It's important to select "COM PORT."

Start by selecting "Begin Data" from the toolbar.

To begin gathering information, select the Record Data button.• When you're done recording or collecting data, select Stop Data or Stop Recording, respectively. In order to save this Excel document, you must specify a save location.

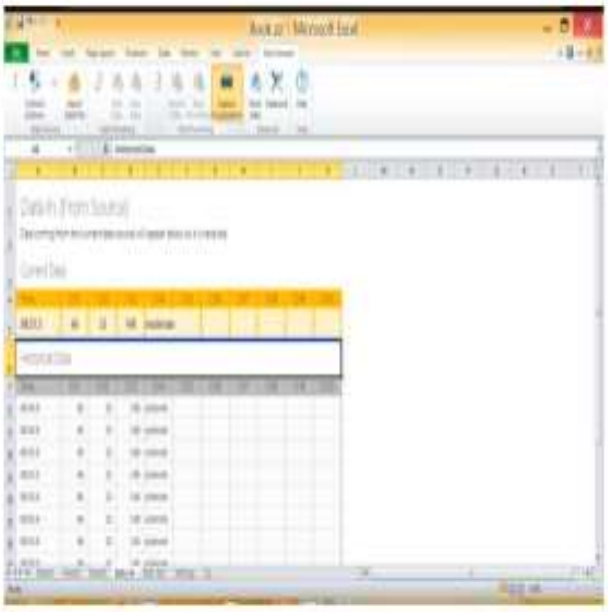

Fig-5: Data collection

The tool being utilized is Python's Anaconda Navigator. Included in the package is the web-based IDE (Interactive Development Environment) for notebooks, code, and data called Jupyter Notebook. Its user-friendly interface can be leveraged by those in the fields of scientific computing, machine learning, and data sciences to construct and manage workflows related to these activities. The first web-based tool for creating code papers was Jupyter Notebook.

Steps in Software Implementation

Read dataset:

The dataset is loaded into the Python code when the necessary libraries have been imported. The dataset contains data from air quality, smoke, and particulate matter sensors, as well as the corresponding data for the current location. Thus, there are four columns in the dataset, and the number of rows is determined at the time of data collection. This information has been exported from Excel as a.csv file.

```
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

Split the training and testing dataset:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

Validation on new, unanalyzed data is performed using the training set. Put the data in the training set to work for you.When just 20–30% of the data is used for evaluation, and the remaining 70–80% is used for training, the outcomes improve. To do this, import the Sci-kit train_test_split tool and change the ratio from 80:20 to 80:2 for training and testing, respectively.

Choosing the Machine Learning model KNN was selected as the machine learning model to measure air pollution levels.

Prediction After the ML model has been fitted, the AQI is used to make predictions about the air quality at the current location, indicating whether it is good enough for human habitation, slightly unacceptable for breathing, inadequate for estimating the consequences of air pollution, or extremely poor and severe.

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

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

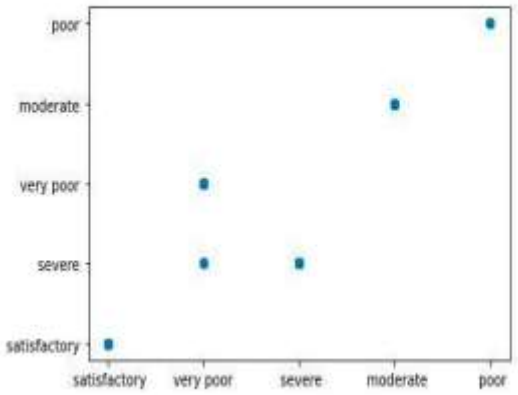

Fig-7: Scatterplot of y_test and predicted values

5. RESULTS

As can be seen in Figure 8, the uncertainty matrix for the air pollution dataset is also read during the initial data read.

```
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([[60], [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

Gases and tiny particles in the air determine its quality. These air contaminants reduce air quality and, when inhaled frequently, can cause serious health problems. In order to take the necessary measures to enhance air quality, it is necessary to detect these harmful compounds and monitor the air with air quality monitoring equipment.

Thus, output increases but air pollution-related health issues decrease. It has been discovered that machine learning-based prediction models are more accurate and consistent. New tools and sensors have made data collection simple and reliable. Only machine learning (ML) algorithms are capable of performing the kind of in-depth analysis required to reliably anticipate outcomes from this mountain of environmental data. Since the KNN approach is more accurate than other methods, it is used to foretell the presence of air pollution. The KNN machine learning technique has a 99.1071% success rate in predicting air pollution.

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.

[9] Biau, GA˜ Srard. "Analysis of a random forest model." ˇJournal of Machine Learning Research 13.Apr (2012): 1063- 1095.

[10] Biau, Gerard, and ErwanScornet. "A random forest ´ guided tour." Test 25.2 (2016): 197-227.

[11] Grimm, Rosina, et al. "Soil organic carbon concentrations and stocks on Barro Colorado Island— Digital soil mapping using Random Forests analysis." Geoderma 146.1- 2 (2008): 102-113.

[12] Strobl, Carolin, et al. "Conditional variable importance for random forests." BMC bioinformatics 9.1 (2008): 307.

[13] Svetnik, Vladimir, et al. "Random forest: a classification and regression tool for compound classification and QSAR modeling." Journal of chemical information and computer sciences 43.6 (2003): 1947-1958.

[14] Verikas, Antanas, AdasGelzinis, and MarijaBacauskiene. "Mining data with random forests: A survey and results of new tests." Pattern recognition 44.2 (2011): 330-349.

[15] Ramasamy Jayamurugan,1 B. Kumaravel,1 S. Palanivelraja,1 and M.P.Chockalingam2 International Journal of Atmospheric Sciences Volume 2013, Article ID 264046, 7 pages http://dx.doi.org/10.1155/2013/264046