Air Quality Prediction of Relative Humidity Using IoT Sensors

EGBOH DANIEL CHUKWUNONSO
University of Bradford

Abstract- Air quality prediction is critical to environmental monitoring and public health management. Relative humidity, an important atmospheric parameter, significantly influences air quality and human comfort. This technical report focuses on the prediction of air quality based on relative humidity using IoT sensors and machine learning techniques, highlighting its significance and potential applications. IoT sensors offer a cost-effective and scalable solution for real-time data collection, including relative humidity measurements, in various locations and environments. These sensors provide continuous monitoring and enable the acquisition of high-resolution data on relative humidity levels. The report discusses the deployment of IoT sensors in indoor and outdoor settings, considering factors such as sensor placement, network architecture, and data transmission protocols. Machine learning algorithms are employed to analyse the collected data and develop predictive models for air quality based on relative humidity. These algorithms utilize historical air quality data of an Italian city, meteorological parameters, and relative humidity measurements as input to train and validate the models. The technical report explores various machine learning techniques, including regression models, decision trees, neural networks, and support vector machines, highlighting their capabilities in capturing complex relationships and dependencies between relative humidity and air quality parameters.

I. INTRODUCTION

For the existence and survival of all species on this planet, air is the most crucial natural resource. All living things, including plants and animals, need air for survival. Therefore, clean air that is free of dangerous pollutants is necessary for the life of all living things. argon, carbon dioxide, nitrogen, oxygen, and other minor amounts of additional gases are all present in the air as a combination. Air pollution is defined as any change in the chemical composition of the air. It describes the release of toxins into the air that are dangerous to both human health and the environment. The average lifespan of people is shortened by chronic exposure to pollution in the air which also raises the likelihood of having a stroke, lung cancer, heart condition, or respiratory illnesses. The emissions of PM2.5, NO2 and CO are mostly caused by motor vehicle exhaust, industrial processes, and fuel combustion. Asthma, bronchitis, and other respiratory illnesses are brought on by NO2, a smothering irritant. Inhaling PM2.5, which has a diameter of roughly 2.5 m and is 30% smaller than the typical human hair, raises your chance of developing cardiovascular disease and respiratory issues. And CO hinders the body's ability to transfer oxygen. (J. Srishtishree et al, 2020).

The progress of technology can address these issues. The introduction and development of more compact, portable, and affordable sensor devices capable of measuring values and practically real-time reporting of air quality is the result of the new generation of sensors, IoT platforms, and other advances. To get insights into the causes and variations in air pollution levels, large data skills, including analytics and machine learning, may be utilised to evaluate the data of the pollutant and other associated data sets.

To evaluate the effectiveness of the predictive models, rigorous validation and performance assessment are conducted. The abstract discusses performance metrics R-squared value, mean absolute error, and root mean square error are examples used to measure the precision and dependability of the predictions. Comparative analyses against existing models and benchmarking against reference measurements are carried out to verify the superiority of the suggested IoT sensor-based approach. The results demonstrate that the integration of IoT sensors and machine learning techniques significantly improves air quality prediction based on relative humidity. The models exhibit higher accuracy and better temporal resolution compared to traditional approaches. The real-time nature of IoT sensor data enables timely adjustments to ventilation systems, pollutant control measures, and public health interventions to mitigate air quality issues.

1.1 PROBLEM STATEMENT

Risks to human health and ecological health are posed by air pollution, a major environmental problem. Predicting air quality accurately and promptly is essential for enabling preventive actions to lessen the negative consequences of pollution. Traditional monitoring systems may have limited real-time data and spatial coverage. Therefore, it is necessary to create a solid and expandable system that uses IoT sensors to gather and process data on air quality.

1.2 REPORT OBJECTIVE

The main goals of air quality prediction using IoT sensors are to provide precise, real-time, and spatially extensive monitoring, deliver prompt alerts, aid in decision-making, and promote public awareness and involvement. Together, these goals seek to solve the problems caused by air pollution and improve the quality of the air for the benefit of both the environment and humans.

II. LITERATURE REVIEW

Recently, machine learning has been widely used to estimate the value of the air quality index (AQI) using IoT sensors, which has piqued academics' curiosity. Mauro Castelli's study (2020) used meteorological information and air contaminants to estimate the air quality index (AQI) with a 0.75 accuracy in California. Although the aforementioned studies produced quite acceptable. Consequently, the models built for the trials still have certain drawbacks, like overfitting issues and inconsistent performance over a range of datasets. Huixiang Liu's other 2019 study (Liu et al. 2019) was primarily concerned with examining the correlation coefficients between characteristics. Next, the difference between an Italian city's NOx rating and Beijing's AQI value was predicted using the algorithms Regression using Support and Random Forest, respectively. With an accuracy for Support Vector Regression of 0.8923 and 0.9180 for Regression with a Random Forest, the prediction produced fairly positive results. In Samir Lemes and his team used the algorithm to assess the air quality index (AQI). colleagues demonstrated the variations in terms of the findings and analysis of air quality contamination while using various methods of AQI computation in 2018. Additionally, factors including accuracy, practicality, intricacy, and if the technique is simple enough for the general public to grasp were considered. (Bosnia 2018). The authors of the work proposed three criteria for calculating the Index of Air Quality that corresponded to the methods used in the USA, the EU, and the Western Balkan countries. In order to calculate the AQI value, there is a typical range of air pollution concentrations for each criterion. values. Additionally, Samir Lemes contrasted the classifications and upper limits of air pollution based on specific air contaminants that meet each standard. The amount in Bosnia and Herzegovina, a country with high levels of air pollution was then assessed using these methodologies using the same dataset, which provided clearer instructions on Using the information above, to determine how to calculate and classify the AQI values.

2.1 CASE STUDY (CITY OF VENICE IN ITALY)

This technical report is a case study of a city in Italy called Venice, 9358 samples of hourly averaged responses from a collection of five metal oxide chemical sensors that are included in an Air Quality Chemical Sensor dataset for Multisensor Devices. The gadget was located in a badly polluted area of an Italian city on a field at street level. most extended publicly accessible From March 2004 to February 2005, records of responses from on-field deployed air quality chemical sensor devices were made. (one year). a localised, verified reference source gave Every hour, Ground Truth averaged concentrations for carbon monoxide (CO), non-metallic hydrocarbons (NMH), Total nitrogen oxides (NOx), nitrogen dioxide (NO2), and benzene analyser.

2.2 DATA COLLECTION

The dataset is the air quality prediction of Relative Humidity using IOT sensors data obtained from the UCI dataset repository as can be seen on www.uci.com. As stated in the case study above, the 9358 answers from a group of five chemical metal oxide sensors integrated into a Multisensor chemical air quality device are included in the dataset on an hour-average basis.

III. RESEARCH QUESTION

When researching air quality prediction using IoT sensors and Support vector machines, random forests, and decision tree regression machines are a few examples of machine learning techniques. potential research questions can be explored which include:

1. How can IoT sensor data be effectively collected, pre-processed, and integrated for air quality prediction?
2. Which features extracted from IoT sensor data are most relevant for accurate air quality prediction?
3. Can machine learning models effectively capture the non-linear relationships and interactions between air quality parameters using IoT sensor data?
4. How does the performance of decision tree regression compare to the random forest and support vector machines in air quality prediction?
5. Can ensemble methods, such as combining Random forest, support vector machines, and decision tree regression machines, improve the accuracy of air quality prediction?

3.1 PRE-PROCESSING OF DATA AND ANALYSIS

The dataset used for this project is open-source data and hence contains some missing values and requires cleaning. The pre-processing involves the definition of the dataset attributes in other to determine the number of columns in the dataset and other variables as shown below.

S/N	ATTRIBUTES	DESCRIPTION
0	Date	Date (DD/MM/YYYY)
1	Time	Time (HH.MM.SS)
2	CO(GT)	CO true averaged concentration in milligrams per cubic metre (reference analyser)
3	PT08.S1(CO)	Tin oxide, PT08.S1 response of a sensor on an hourly average (usually focused at CO)
4	NMHC(GT)	According to reference analysers, the true hourly averaged total non-metallic hydrocarbon content is determined in micro g/m3.
5	C6H6(GT)	Benzene concentration (reference analysers) true hourly averaged in micro/m3.
6	PT08.S2(NMHC)	Hourly averaged sensor response for PT08.S2 (titania; ostensibly NMHC targeted)
7	NOx (GT)	Averaged hourly sensor response for Titania (PT08.S2; apparently NMHC targeted)
8	PT08.S3(NOx)	(Reference Analyzer) True hourly averaged NOx concentration in ppb
9	NO2(GT)	True hourly averaged NO2 concentration in microg/m^3 (reference analyser)
10	PT08.S4(NO2)	(Tungsten oxide) PT08.S4 Averaged over the last hour, NO2 targeted sensor response
11	PT08.S5(O3)	Hourly averaged sensor response for PT08.S5 (indium oxide; ostensibly O3 targeted)
12	T	Amount of heat in °C1
13	RH	Humidity Relative (%)1 AH Total Humidity
14	AH	(Tungsten oxide) PT08.S4 Averaged over the last hour, NO2 targeted sensor response

Figure 3.1 Attributes of the Dataset

After determining the attributes of the dataset, the datatypes were also defined to ensure they are aligned, also the number of columns and rows was confirmed and afterwards how many NAN values there are in each column was calculated which led to some missing values in columns such as CO_GT, NOX_GT, NO2_GT being filled by the average of that particular hour while the NHHC_GT column was dropped as it contains numerous amount of missing values.

3.2 FEATURE SELECTION

Feature selection was also carried to promotes interpretability, decrease computational complexity, eliminate overfitting, boost model performance, improve data quality, and make model maintenance easier. Feature selection allows models to concentrate on the crucial data needed for precise predictions and improves their applicability to real-world settings by choosing the most important characteristics. Below is the heat map which gives a better understanding of the linearity between relative humidity (RH) and other input features.

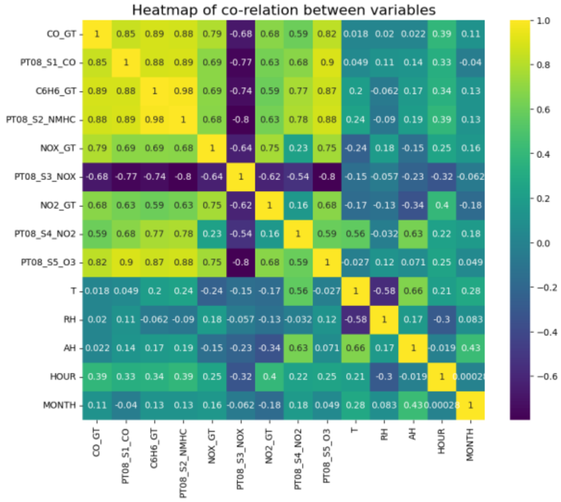
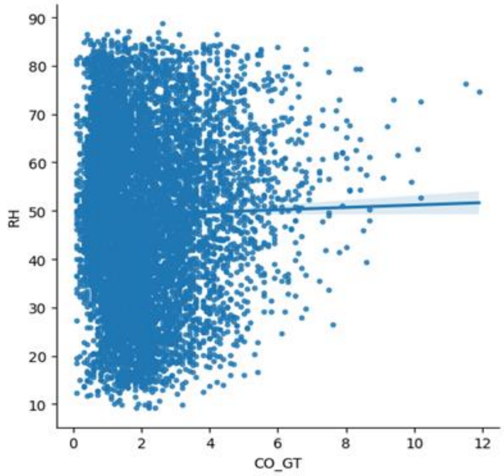
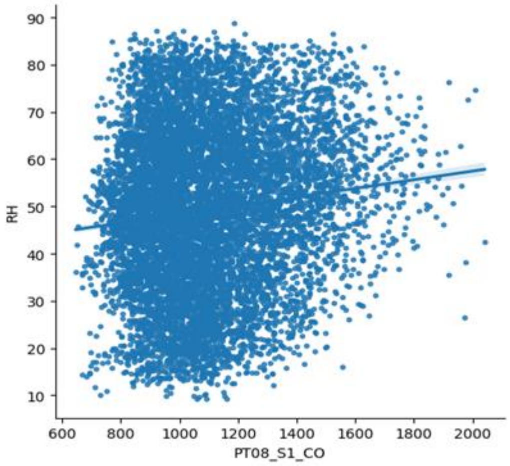


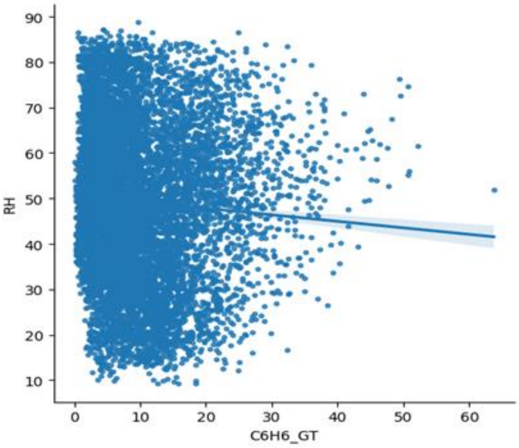
Figure 3.2. Heatmap of correlation between the dataset variables.

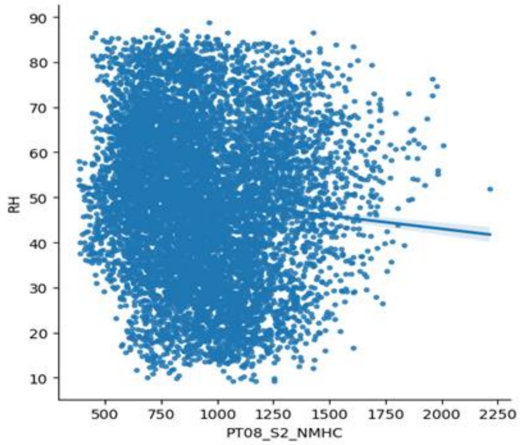
IV. EXPLORATORY DATA ANALYSIS

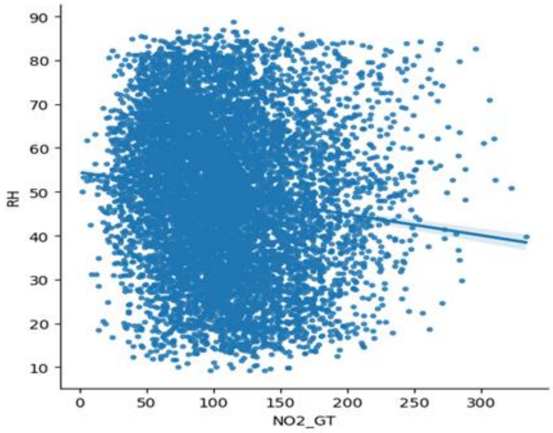
An exploratory data analysis was carried out on the dataset to determine the relationship between the main feature; relative humidity and the rest of the features and below is the plot showing the obtained result.

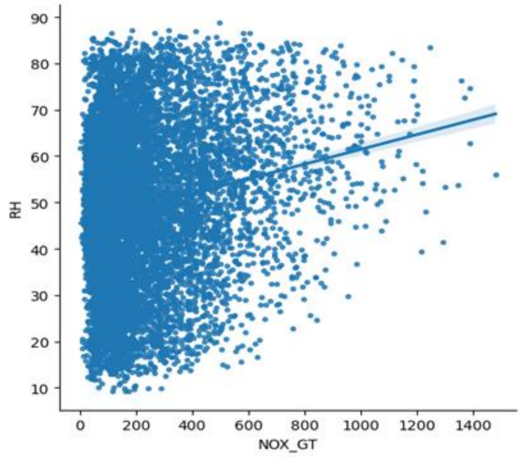


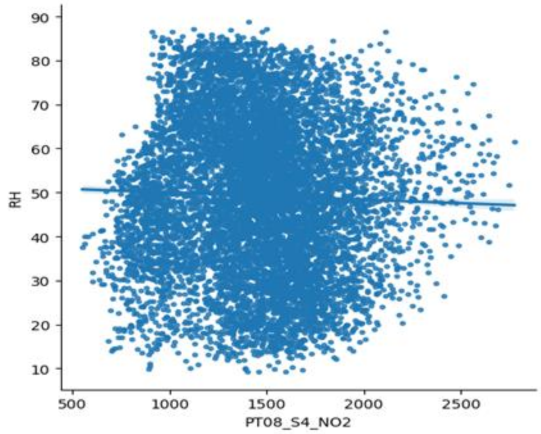


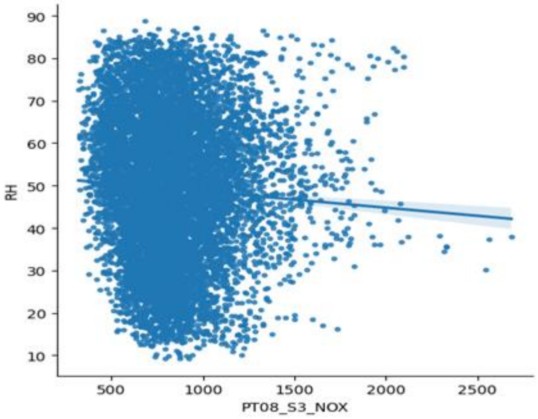


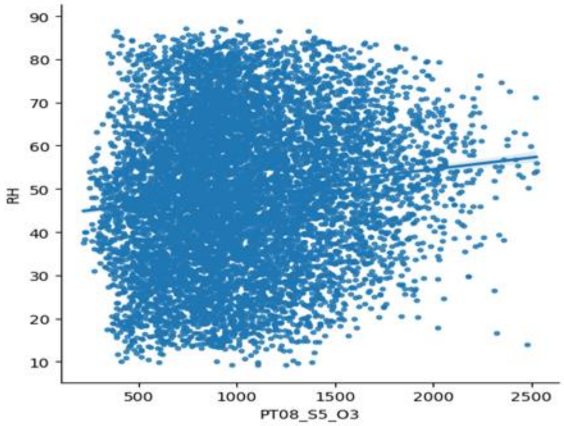

RH
PT08_S2_NMHC
10 20 30 40 50 60 70 80 90
500 750 1000 1250 1500 1750 2000 2250

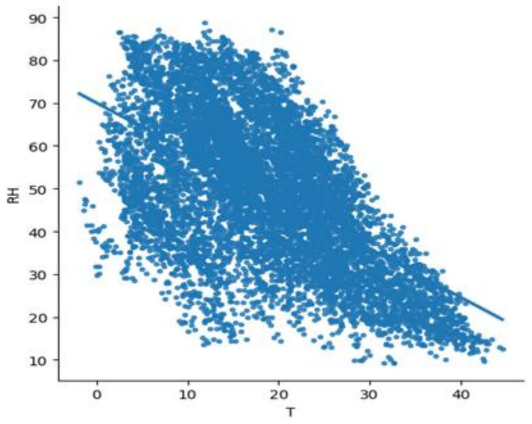

RH
NO2_GT
10 20 30 40 50 60 70 80 90
0 50 100 150 200 250 300

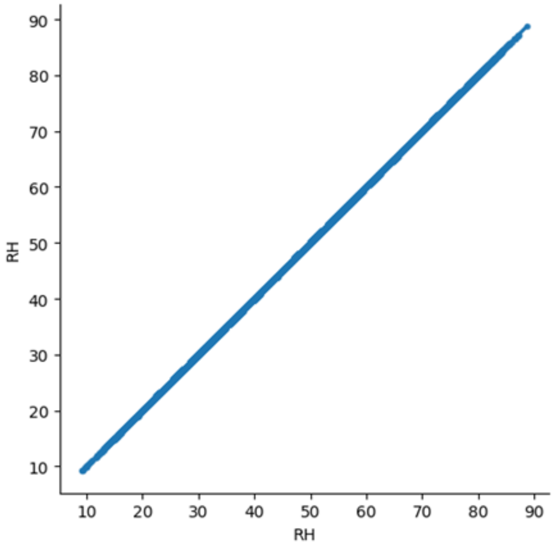

RH
NOX_GT
10 20 30 40 50 60 70 80 90
0 200 400 600 800 1000 1200 1400

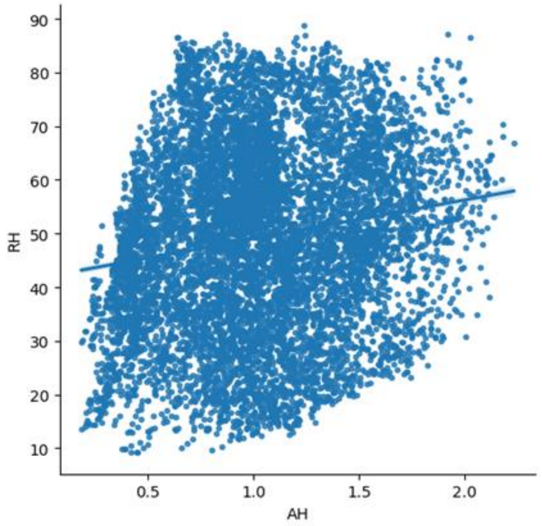

RH
PT08_S4_NO2
10 20 30 40 50 60 70 80 90
500 1000 1500 2000 2500

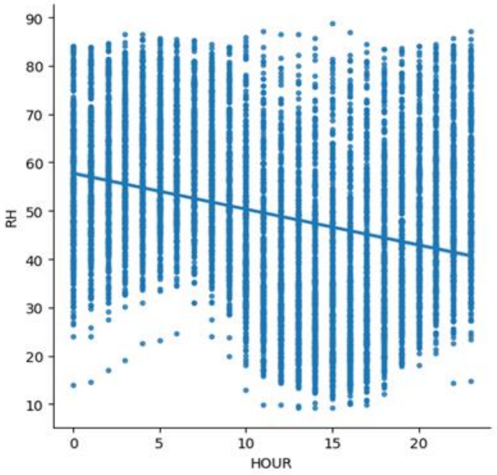

RH
PT08_S3_NOX
10 20 30 40 50 60 70 80 90
500 1000 1500 2000 2500

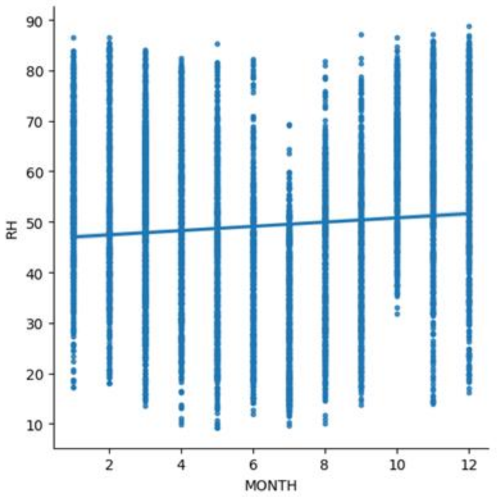

RH
PT08_S5_O3
10 20 30 40 50 60 70 80 90
500 1000 1500 2000 2500









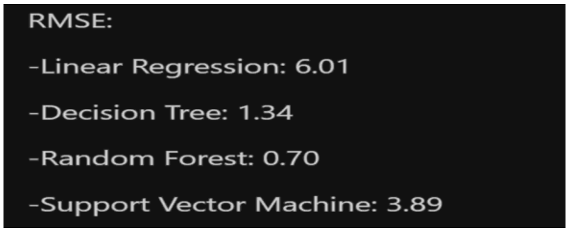


V. RESULT AND DISCUSSION

The dataset was split into testing and training data, this is done to compare models fairly and evaluate performance accurately by dividing the dataset into testing and training sets. It ensures that the generated model can handle unforeseen data and produce accurate predictions in practical situations. Also, the training and test data size was determined which was followed by the root mean square value for all columns of the features. Furthermore, a few machine learning algorithms were applied to the sensor-generated data to determine which algorithm is best fit for its prediction, this includes:

1. Linear Regression
2. Decision Tree regression
3. Random Forest regression and
4. Support Vector Machine.

The above ML techniques were used in the prediction however for the design of the model for the prediction of relative humidity, the RMSE for the various techniques are shown below.



```
RMSE:
-Linear Regression: 6.01
-Decision Tree: 1.34
-Random Forest: 0.70
-Support Vector Machine: 3.89
```

Figure 5.1: Prediction Results from the Individual ML Techniques.

5.1 CONCLUSION

It can be therefore concluded that in relative humidity forecast for air quality using IOT sensors, Random Forest is selected as the best fit predicting technique. however, other use of machine learning algorithms can be explored to make a comparative analysis.

REFERENCES

[1] J. Srishtishree, S. Mohana Kumar and Chetan Shetty, H. S. Saini et al. (2020), Air Quality Monitoring with IoT and Prediction Model using Data Analytics Innovations in Computer Science and Engineering, Lecture Notes in Networks and Systems 103.

[2] Van, N.H., Van Thanh, P., Tran, D.N. et al, (2023), A new model of air quality prediction using lightweight machine learning. Int. J. Environ. Sci. Technol. 20, 2983–2994.

[3] Castelli M, Clemente FM, Popovicˇ A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Hindawi 2020:23

[4] Bosnia H (2018) Air Quality Index (AQI) – Comparative study and assessment of an appropriate model for B&H," Academia

[5] Liu H, Li Q, Dongbing Y, Yu Gu (2019) Air quality index and air pollutant concentration prediction based on machine learning algorithms. Appl Sci 9(19):4069.