IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

Air Pollution Modeling for Awka Metropolis using Ensemble Algorithms

Chris A. Nwabueze[1], Silas A. Akaneme[2], Fidelis C. Obodoeze[3]

Department of Electrical/Electronic Engineering, Chukwuemeka Odumegwu Ojukwu University,

Uli, Anambra State, Nigeria[1,2,3]

Abstract: Air pollution is a very serious problem facing urban the dwellers where various types of dangerous and poisonous air pollutants are discharged directly into the atmosphere on daily basis as a result of increased industrial and human activities due to increase in population and urbanization. These pollutants have serious and adverse impact on health and well-being of human beings and the environment. Air pollution prediction or forecasting can be adopted to predict or forecast the air quality index (AQI) of a city or area in advance before pollution occurs. This is helpful where air pollution monitors or stations are not installed or deployed. Awka Metropolis, the focus of this research, is a rapidly growing city due to the rising influx of people into it within the last ten years. The rapid population growth in Awka is as a result of it several important factors such as infrastructural, industrial and economic developments. Awka as a growing city has its own fair share of urbanization and environmental challenges. In this paper, ensemble technique of machine learning was used to develop a prediction model for air pollution one hour before time for PM2.5 (particulate matters) pollutant emissions within Awka Metropolis. A historical dataset consisting of about 12,958 one-minute of sensor readings for several air and noise pollutants such as PM1, PM2.5, PM10, TVOC (volatile organic compound), carbon dioxide, noise as well as historical weather or meteorological data comprising air temperature, humidity, pressure, light intensity were also used as input predictors to the model. Seven machine learning algorithms comprising about three traditional machine learning algorithms such as Linear Regression, Multi Layer Perceptron (MLP) Artificial Neural Network (ANN), Decision Tree, and four ensemble learning algorithms - Random Forest, XGBoost, AdaBoost, Extra Tree were used in the simulation modeling. Experimental results showed that the ensemble algorithms performed best in prediction accuracy having highest R^2 values and lower RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) scores. Random Forest and Extra Trees ensemble algorithms came first with the highest accuracy score (R^2=0.9886), followed by XGBoost R^2=0.9870, AdaBoost came fourth with R^2=0.9854. Equally the ensemble learning algorithms have the lowest prediction residual errors when compared to the traditional machine learning algorithms. The experimental test-bed and programming was carried out in Anaconda, Python 3 and Python machine learning module Scikit-learn. Jupyter Notebook IDE was used as programming development and simulation environment.

Keywords: Air Pollution, Regression, PM2.5, Ensemble, Ensemble Algorithm, Machine Learning.

I. INTRODUCTION

Air pollution is a very big threat to human existence because air pollutants are dangerous to certain categories of humans with varying types of health challenges such as lung and respiratory diseases, heart diseases, cancer, especially in children and the elderly. Statistics in literature have shown that these air pollutants such as particulate matters (PM1.0, PM2.5, and PM10) and gaseous effluents such as SO2, NO2, NO, CO, CO2, Ozone (O3), Volatile Organic Compounds (VOCs), have led to deterioration of human health in some patients and even mortality in some cases. Particulate matters such as PM2.5 in particular is of utmost interest to research because it is considered to be the most dangerous to human health [1][2]. Out of all the air pollutants because of its small size (less than 2.5µm, about 3% of the size of the human hair) which can easily be breathed in by human beings and can cause several chronic respiratory and cardiovascular diseases such as asthma, respiratory inflammations, jeopardisation of lung functions, COVID, laryngitis and even cancer [3]. The World Health Organization (WHO) in 2018 reported that about 2.4 million deaths occurred in 2005 as a result of air pollution [8]. The air pollution is majorly as a result of increased industrial activities and globalization due to the increase in population of big cities or metropolis. Emission of fossil fuels from vehicle combustion, industrial machines, and human activities and so on can lead to emission of these poisonous and dangerous air pollutants to the environment.

Greenhouse gases such as carbon dioxide and ozone gas can affect the environment negatively through the process of global warming. In big cities all over the world such as the rapidly-growing city of Awka, the administrative capital of Anambra State of Nigeria, controlling and managing air pollution menace has become a herculean task due to the lack of understanding of how to control it and partly because of lack of real-time air pollution monitoring stations or sub-

DOI: 10.17148/IJARCCE.2022.11107

stations installed in the cities. Situations where there is absence of real-time monitoring stations or sub-stations to detect various dangerous air pollutants in the atmosphere as they are emitted in real time, the only alternative is to deploy Air Quality Index (AQI) or air pollution prediction models to detect and forecast beforehand the emission of the dangerous air pollutants in a particular area or city. Ensembling is a technique of combining two or more similar or dissimilar machine learning algorithms to create a model that delivers superior and more accurate predictive power. Ensemble models or algorithms work independently and poll their prediction results or outcome together via some certain techniques such as voting or averaging to produce a better result. Ensemble learning techniques have a record of showing better performance in a variety of machine learning applications such as classification and regression problems. Ensemble algorithms such as Random Forest, Gradient Boosting models like XGBoost, AdaBoost, Light GBM and Extra Trees (ET) use a combination of weak learners to build up an ensemble. These traditional ensemble algorithms use homogeneous weak learners such as several Decision Trees to make predictions; the weak learners are grouped together to show their combined strengths.

It is also possible to combine the predictive power of the traditional weak learner machine learning algorithms such as Support Vector Machine (SVM), Decision Tree, Linear Regression, Multi Layer Perceptron Artificial Neural Network (MLP ANN), Naves Bayes, K Nearest Neighbour (kNN), Lasso etc. to produce a hybrid model; it is equally possible to combine the traditional ensemble learning algorithms such as Random Forest, Gradient Boosting Machine (GBM), Adaptive Boosting (AdaBoost), Extreme Boosting (XGBoost), Extra Trees etc together to form a more powerful and accurate hybrid ensemble model. An ensemble algorithm or model can be further categorized as either ensemble Classifier or ensemble Regressor depending on the type of machine learning problem that is intended to be solved.

In this research paper, different experimental methodologies were used to show the prediction power of traditional ensemble algorithms such as Random Forest, XGBoost, AdaBoost, Extra Trees over traditional weaker learner algorithms such as Decision Tree, Linear Regression, MLP Artificial Neural Network. Performance metrics such as coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) were used to determine the effectiveness and accuracy of the models. About 12,958 air and noise pollutant datasets generated from a real-time air and noise pollution monitoring station for Awka Metropolis comprising of one-minute air pollutants concentrations such as PM1.0, PM2.5, PM10, TVOC, CO2, noise and weather parameters such as air temperature, light intensity, pressure were used to predict PM2.5 one hour ahead for Awka Metropolis using ensemble models.

II. REVIEW OF RELATED LITERATURE

This section deals with the review of past works on ensemble learning algorithms and hybrid ensemble machine learning problems and their prediction performances.

Authors in [4] presented a research paper on the use of supervised machine learning approach to predict Air Quality Index (AQI) of a city in India. The paper used approaches such as statistical analysis and probability and supervised machine learning prediction algorithms such as Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest Method (RF). The input datasets include air pollutant concentrations of Carbon Monoxide (CO), Tin oxide, nonmetallic hydrocarbons, Benzene, Titanium, NO, Tungsten, Indium oxide and meteorological parameters such as Temperature and relative humidity (RH). Experimental results with the datasets showed that Random Forest (RF) an ensemble algorithm performed the best among all the prediction models in terms of prediction accuracy having the lowest RMSE.

Authors [5] presented a research paper on the performance evaluation of hybrid ensemble machine learning algorithms using adaptive boosting technique. The authors conducted comparative performance experiments using python 3.6 open source programming with machine learning modules Keras, Tensorflow and Scikit-learn library for ten different traditional-based and ensemble-based algorithms such as Decision Tree (DT), Random Forest (RF), Elastic Net (EN), MLP Neural Network (MLP ANN), Linear Regression, Extra Trees (ET), DT+AdaBoost (using adaptive boosting), RF+AdaBoost (using adaptive boosting), and XGBoost. They equally conducted another experiment with the proposed model (a hybrid ensemble model of Extra Trees (ET) +AdaBoost, to forecast PM2.5 concentrations in a smart city of Delhi, India. Experimental results showed that the hybrid ensemble model of Extra Trees+AdaBoost outperformed all the ten algorithms in terms of speed of prediction and accuracy with the highest R2 score and the lowest error values of MAE and RMSE.

The author [6] applied Extreme Gradient Boosting algorithm (XGBoost) to predict PM2.5 concentrations on hourly basis in the city of Tianjin, China. The author made use of historical dataset of PM2.5 concentrations (December 1, 2016 – December 30, 2016) of about Nineteen air pollution monitoring stations. The set of input variables or

DOI: 10.17148/IJARCCE.2022.11107

parameters in the modeling include hourly concentrations of PM2.5, SO2, NO2, CO and Ozone (O3). The dataset was about 6845 samples of historical dataset in total. XGBoost algorithm was compared to other machine learning (ML) algorithms in the experiments. The experimental results showed that XGBoost another ensemble algorithm outperformed other ML algorithms such as Random Forest (RF), Multiple Linear Regression (MLR), Decision Trees (DT), and Support Vector Regression (SVR) in terms of prediction accuracy(i.e. highest R^2 value) and lowest error value (i.e. MAE and RMSE).

Authors [7] presented a research paper on Air quality prediction of the city of Bjelave, Sarajevo using different machine learning algorithms – Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Multi-Layer Regression (MLR) and Multi-Layer Perceptron (MLP) ANN using three years of historical dataset (2016-2018) comprising of five air pollutant concentrations of PM10, NO2, SO2, O3 and CO together with eight (8) meteorological parameters – minimum temperature, maximum temperature, average temperature, wind speed, wind direction, humidity, pressure and precipitation. Experimental results showed that Random Forest (RF) regression algorithm outperformed other machine learning algorithms in terms of the highest R^2 and lowest RMSE, closely followed by XGBoost algorithm.

III. MATERIALS AND METHODS

The section describes all the materials and method used in this research paper.

A. Experimental Methodology

This research papers adopted machine learning techniques known as ensemble learning using five ensemble machine learning algorithms to predict air pollution concentrations of PM2.5 within Awka Metropolis. Fig.1 shows the methodology adopted in this research paper. Here the historical air and noise pollution dataset as well as the meteorological or weather dataset was obtained from a real-time air and noise pollution monitoring system for Awka Metropolis. The datasets was split into 80% for training machine learning algorithms while the remaining 20% was used for testing and validation.

B. Methods

Traditional and ensemble machine learning algorithms were employed in the modeling and prediction of one-hour ahead PM2.5 minutely concentrations for Awka Metropolis. Three traditional machine learning algorithms were used to determine their accuracy in predicting PM2.5 concentrations within Awka Metropolis. They include Multi Linear Regression (MLR), Multi-Layer Perceptron Feed Forward Artificial Neural Network (MLP-ANN) and Decision Tree algorithms to serve as control for the remaining four ensemble algorithm.. Several experiments were carried out using four ensemble learning algorithms such as AdaBoost, Random Forest (RF), Extreme Gradient Boosting algorithm (XGBoost) and Extra Trees algorithms and their prediction performances in terms of accuracy and residual errors were evaluated and compared.

B.1 Data Collection

The historical dataset of PM2.5 air pollutants and other input predictor variables for forty (40) days (from October 25th to December 4th 2021) containing about 12,958 dataset samples of the city of Awka Nigeria (obtained from experiential runs outdoor of newly designed Awka Pollution Monitoring Station. CSV file can be downloaded at https://myrasoft.ng/awka-pollution-monitor/AWKA-POLLUTION-2022NEW.csv. The historical dataset in .CSV format was captured in an online cloud server Smart Citizen Station and was used to carry out the experiments. The input predictors to the models include the historical air pollutant concentrations including PM2.5, PM1.0, PM10.0 , equivalent carbon dioxide (ecarbondioxide), TVOC, and noise pollutant concentrations, UTC timestamp comprising of year, month, day, hour, minutes and seconds section of the time and meteorological parameters comprising the following- air temperature (temperature), relative humidity (humidity), light intensity (light) and barometric pressure (pressure). UTC is equivalent to GMT time zone and one hour ahead of Nigeria's local time. The experiments were carried out using Python 3 programming language with Sci-kit learn machine learning modules and Jupyter Notebook Integrated Development Environment running on Anaconda software environment.

Figs. 2 shows head data frame of the historical dataset containing about the first five rows of the dataset. Fig.3 shows the statistical analysis of the dataset during data pre-processing stage. Fig.4 shows the historical dataset frame in Pandas showing the data types of the dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

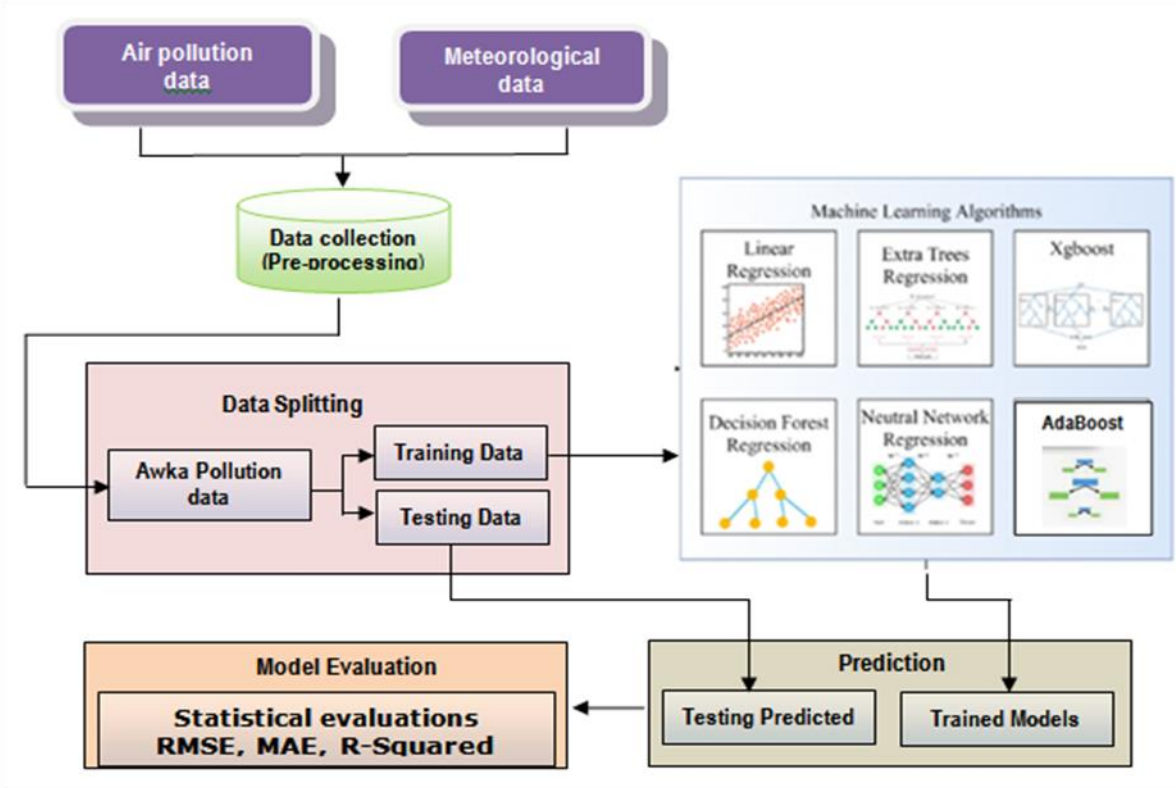

Fig. 1 The proposed ensemble machine learning methodology using regression analysis

Out[1]:

timestamp	humidity	temperature	pressure	ecarbondioxide	TVOC	light	PM_1	noise	PM10	PM2.5
2021-10-25 10:00:54 UTC	79.46	28.29	100.85	400.0	0.0	151	20.0	45.35	25.0	23.0
2021-10-25 10:01:54 UTC	79.68	28.22	100.84	400.0	0.0	178	20.0	47.63	25.0	23.0
2021-10-25 10:02:54 UTC	79.68	28.21	100.83	400.0	0.0	186	20.0	45.93	25.0	23.0
2021-10-25 10:03:54 UTC	79.78	28.22	100.84	420.0	3.0	186	20.0	46.40	25.0	23.0
2021-10-25 10:04:54 UTC	79.66	28.21	100.84	409.0	1.0	185	14.0	48.02	25.0	23.0

Fig. 2: The data frame of the historical dataset showing the first five rows

In [2]: `dataset.describe()`

Out[2]:

	humidity	temperature	pressure	ecarbondioxide	TVOC	light	PM_1	noise	PM10	PM2.5
count	12958.000000	12958.000000	12958.000000	12958.000000	12958.000000	12958.000000	12958.000000	12958.000000	12958.000000	12958.000000
mean	72.183491	31.916019	100.511330	636.302439	38.647091	1222.338169	19.638216	53.223142	26.461568	24.199722
std	14.541698	4.958644	0.158447	269.692861	59.768807	4359.001327	4.962155	6.712010	9.192102	8.119768
min	29.130000	24.930000	99.710000	0.000000	0.000000	0.000000	0.000000	10.000000	0.000000	0.000000
25%	61.745000	28.110000	100.400000	450.000000	9.000000	0.000000	20.000000	48.320000	25.000000	23.000000
50%	75.560000	30.525000	100.530000	549.000000	22.000000	11.000000	20.000000	53.815000	25.000000	23.000000
75%	84.390000	34.830000	100.630000	732.000000	50.000000	481.750000	20.000000	58.100000	25.000000	23.000000
max	94.250000	56.860000	100.910000	2506.000000	1794.000000	50700.000000	77.000000	90.470000	110.000000	106.000000

Fig. 3: Statistical analysis of Awka Pollution Dataframe after data pre-processing

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

```
In [3]: dataset.info()

<class 'pandas.core.frame.DataFrame'>
Index: 12958 entries, 2021-10-25 10:00:54 UTC to 2021-12-04 14:48:25 UTC
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   humidity        12958 non-null  float64
 1   temperature     12958 non-null  float64
 2   pressure        12958 non-null  float64
 3   ecarbondioxide  12958 non-null  float64
 4   TVOC            12958 non-null  float64
 5   light           12958 non-null  int64
 6   PM_1            12958 non-null  float64
 7   noise           12958 non-null  float64
 8   PM10            12958 non-null  float64
 9   PM2.5           12958 non-null  float64
dtypes: float64(9), int64(1)
memory usage: 1.1+ MB
```

Fig. 4: Awka pollution monitor data frame in pandas showing the data types

B.2 Data pre-processing

The raw historical datasets collected from the real-time Awka Pollution monitoring station was cleaned up using different data pre-processing techniques. All the null values in PM2.5, PM1.0, PM10.0, TVOC, ecarbondioxide and noise columns were successfully removed using the mode values of each respective columns. Data normalization or scaling was introduced in the columns and rows to make the sums of the standard deviation to be 1 and sum of mean to zero. This was achieved in Python using Scikit-learn standard scaler() function.

B.3 Data splitting for training and validation testing

The dataset was split into 80% for training and 20% for testing and validation tests.

C. Performance Evaluation Methods

In order to determine or evaluate the best machine learning Air pollution prediction models quantitatively in terms of the prediction accuracy, the following statistical performance metrics- Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of determination or RSquared (R^2) were employed and calculated as shown in Eqs. (1)-(3). Eq.4, the Pearson correlation coefficient was used in data analysis to determine the correlation effects of the various input predictors or features to the output variable the PM2.5 concentration.

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - M_i| \qquad (1)$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|P_i - M_i|^2} \qquad (2)$$

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum(M_i - P_i)^2}{\sum(M_i - \overline{M_i})^2} \qquad (3)$$

where n is the number of data in the test dataset, P_i and M_i are the predicted and measure value for the i^{th} hour and $\overline{M_i}$ is the mean of all the measured values for the i^{th} hour. The higher the value of R^2, the more accurate and better the prediction result while the lower the values of RMSE and MAE, the higher the accuracy of the prediction model or algorithm.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

Pearson Correlation coefficient (r):

$$r = Cor(X,Y)\frac{Cov(X,Y)}{std(X)std(Y)} \quad (4)$$

IV. Results and Discussion

This section describes the experimental results and discussion of the results obtained from the experiments.

A. Results

This section describes the results obtained after several experimental runs were carried out on the datasets.

Fig.5 shows the Pearson correlation matrix plot obtained to determine the correlation between the target variable $PM_{2.5}$ concentrations and other input predictors or independent variables such as the other air pollutant concentrations, noise and meteorological variables such as air temperature, relative humidity, pressure and light intensity.

Fig.6 depicts the distribution of $PM_{2.5}$ pollution in Awka Metropolis from October 28th 2021 to December 4th 2021 using Histogram Plot.

Fig.7 shows how a $PM_{2.5}$ concentration was distributed in Awka Metropolis from October 28th 2021 to December 4th 2021 using density plot. Fig.8 captures the correlation matrices run in Python 3 showing how PM2.5polution was correlated with other pollution and meteorological variables.

Figs.10-16 show the regression scatterplots of the various traditional machine learning algorithms and ensemble learning algorithms as they perform during experimental runs on the dataset for PM2.5 minutely prediction.

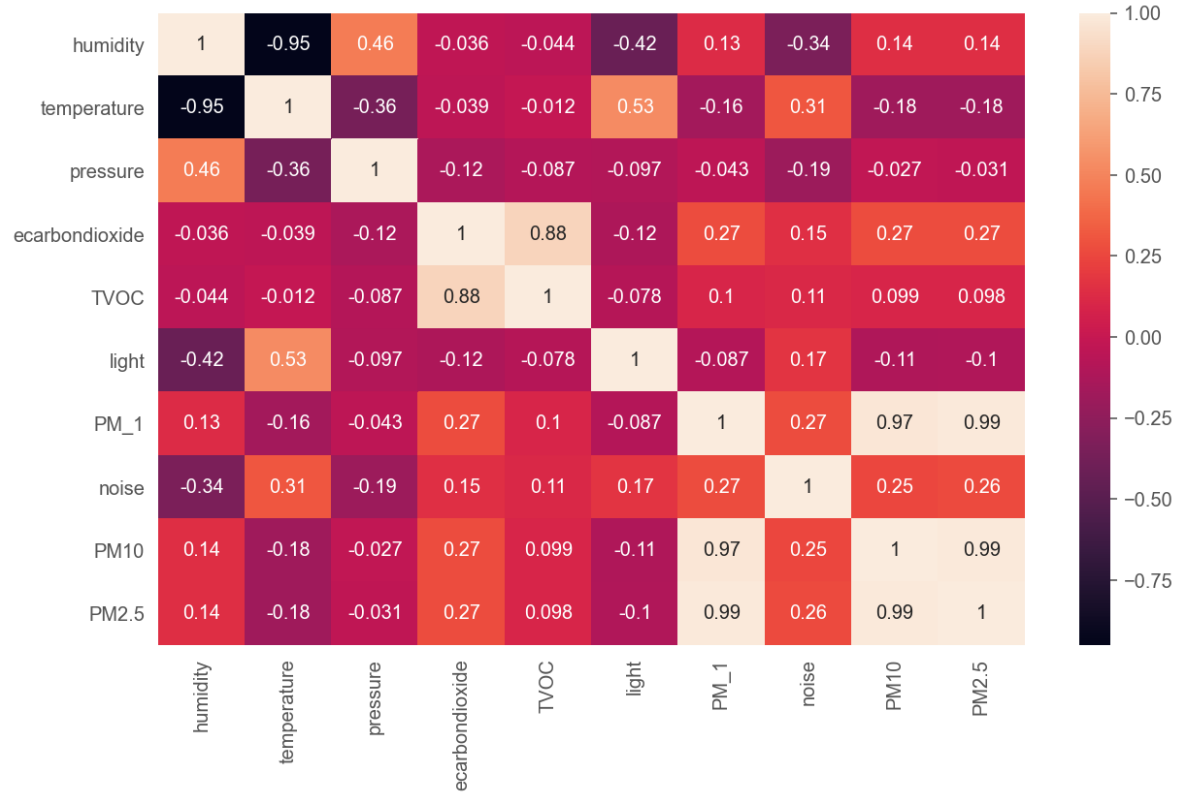

Fig. 5: Awka pollution monitor Correlation matrices plot using Pearson Correlation coefficient, r..

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

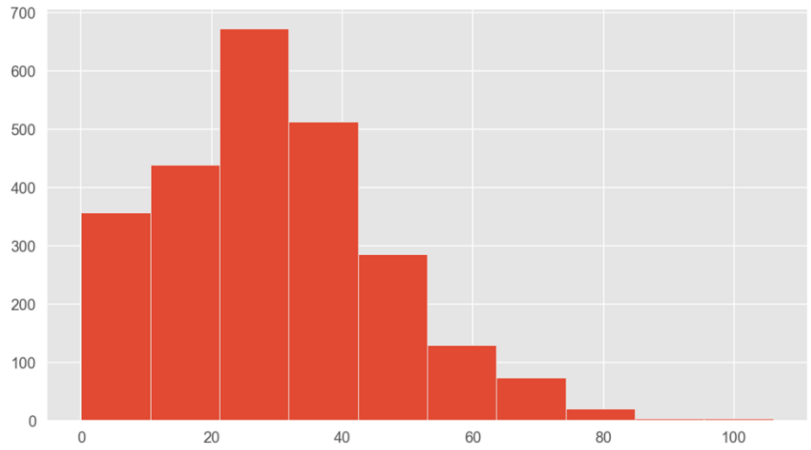

Fig.6: Histogram Plot of PM2.5 Pollution levels in Awka metropolis by minute distribution from October 28th 2021 to December 4th 2021.

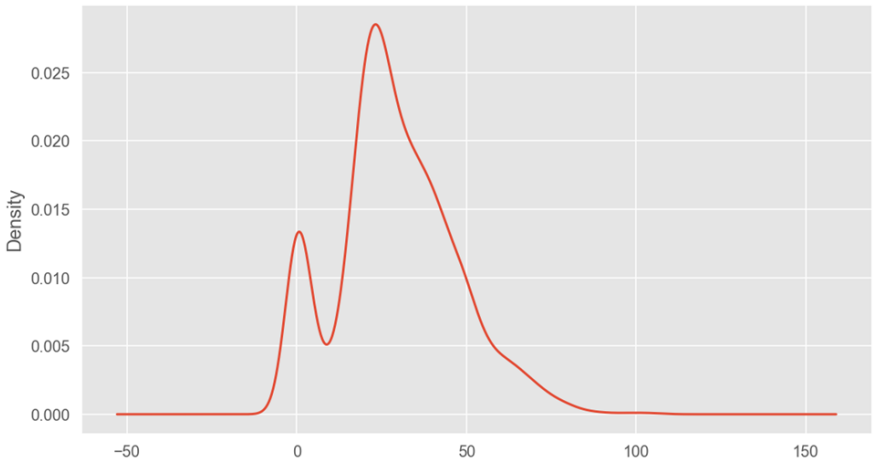

Fig. 7: Density plot for $PM_{2.5}$ Pollution distribution for Awka Metropolis from October 28th 2021 to December 4th 2021.

```
In [90]: dataset.corr()['PM2.5'].sort_values()

Out[90]: temperature   -0.176958
         light         -0.099904
         pressure      -0.031180
         TVOC           0.098164
         humidity       0.142697
         noise          0.261746
         eCO2           0.269801
         PM_1           0.985809
         PM10           0.989569
         PM2.5          1.000000
         Name: PM2.5, dtype: float64
```

Fig. 8: PM2.5 Correlation with other pollution and meteorological variables

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

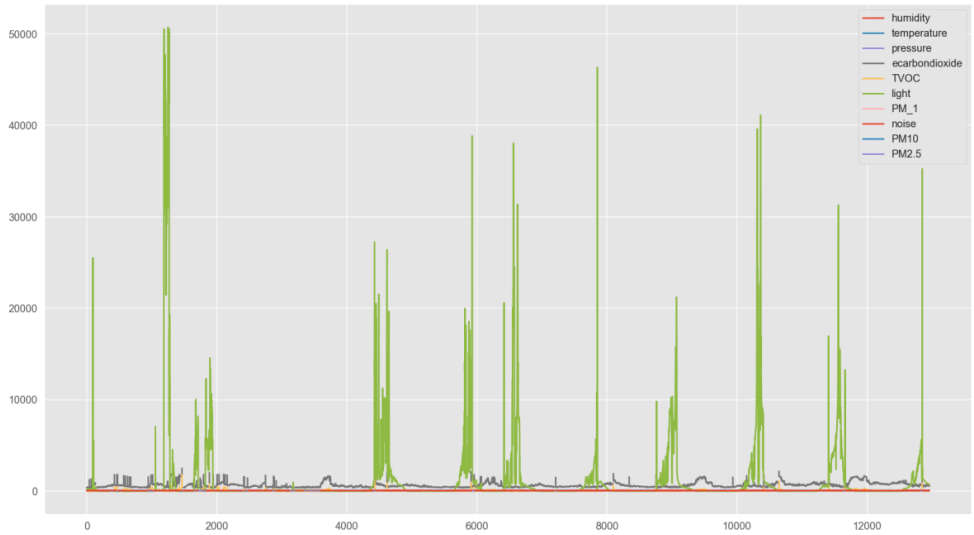

Fig.9: Time series distribution for the pollution variables from October 28th 2021 to December 4th 2021.

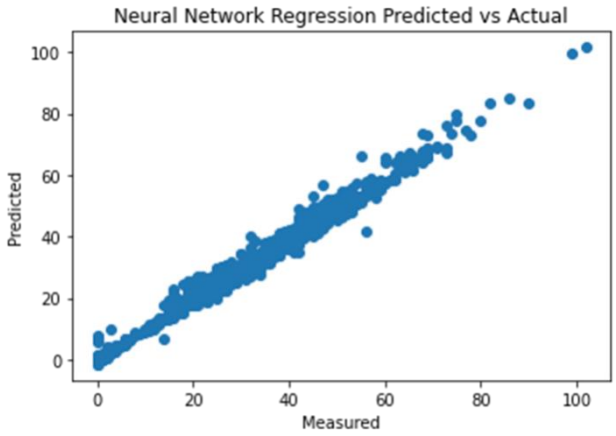

Fig. 10: Regression Scatterplot of PM2.5 Pollution of MLP Neural Network Algorithm

For MLP Neural Network algorithm the following output was obtained:- RMSE=1.1711,MAE= 0.6018 and R^2= 0.9815.

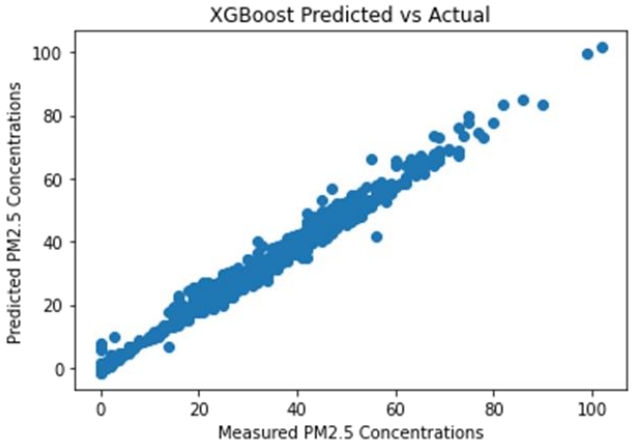

Fig. 11: Regression Scatterplot of PM2.5 Pollution of XGBoost Ensemble Algorithm

For Fig.11, RMSE=0.9814, MAE= 0.3133 and R^2= 0.9870.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

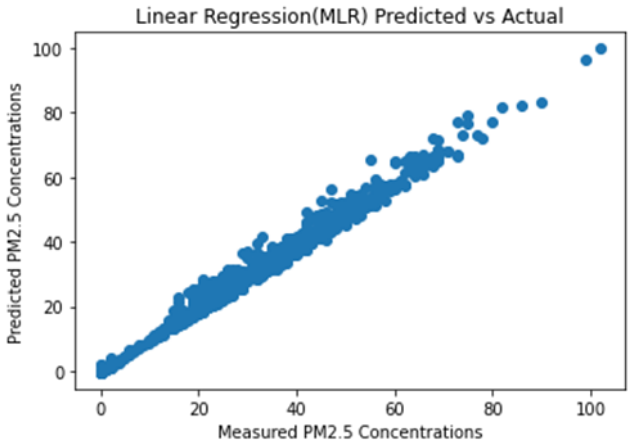

Fig. 12: Regression Scatterplot of $PM_{2.5}$ Pollution of Multi Linear Regression (MLR) algorithm

For Fig. 12 RMSE=1.0319 MAE=0.4470 R^2= 0.9856.

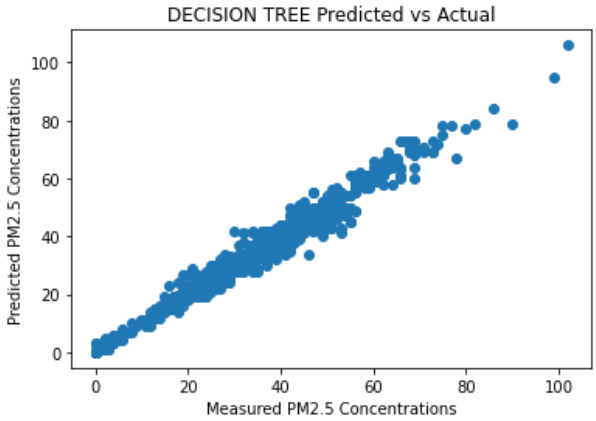

Fig. 13: Regression Scatterplot of $PM_{2.5}$ Pollution of Decision Trees algorithm

For Decision Trees' experimental result in Fig.13 RMSE=1.2709 MAE=0.3812 and R^2=0.9782.

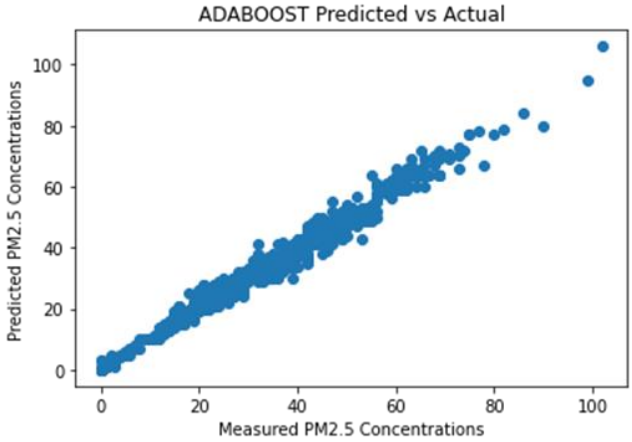

Fig.14: Regression Scatterplot of $PM_{2.5}$ Pollution of AdaBoost Ensemble algorithm

For AdaBoost Ensemble algorithm in Fig.14 Learning rate=1, number of estimators=200 random state=1234

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

RMSE= 1.0386 MAE= 0.3076 R^2=0.9854.

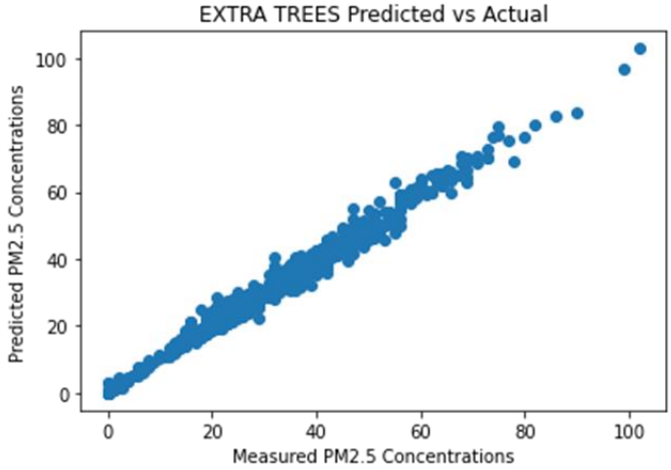

Fig.15: Regression Scatterplot of $PM_{2.5}$ Pollution of Extra Trees Ensemble algorithm

For Extra Trees ensemble algorithm, the following hypermarkets were used:- Learning rate=1, number of estimators=200 random state=1234. The output obtained were RMSE= 0.9175, MAE=0.2847 and R^2=0.9886.

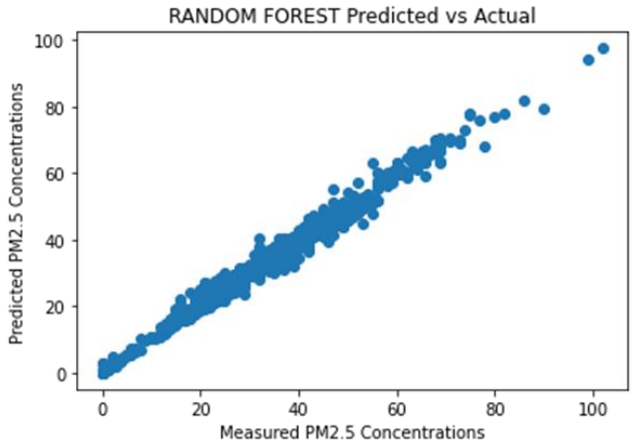

Fig.16: Regression Scatterplot of $PM_{2.5}$ Pollution of Random Forest (Decision Forest) Ensemble algorithm

For Random Forest (Decision Forest) ensemble algorithm, the following hypermarkets were used:- Learning rate=1, number of estimators=200 random state=1234. The output obtained were RMSE=0.9207, MAE=0.2802 and R^2=0.9886.

Tables 1 and 2 shows the printout of the Actual versus predicted values for PM2.5 concentrations using the different machine learning models from Python 3.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

Table 1. Multi Linear Regression PM2.5 Prediction

	Actual-Values	Predicted-Values
0	23.0	23.109703
1	23.0	23.169186
2	23.0	23.080872
3	23.0	23.173619
4	23.0	23.094174
5	24.0	23.100543
6	23.0	23.107283
7	23.0	23.151462
8	23.0	23.073775
9	23.0	23.102005
10	23.0	23.100982
11	23.0	23.163082
12	24.0	22.032364
13	20.0	18.874741
14	23.0	23.155343
15	23.0	23.129356

Table 2. Random Forest PM2.5 Prediction: Actual values versus Predicted Values

	Actual-Values	Predicted-Values
0	23.0	23.000
1	23.0	23.000
2	23.0	23.000
3	23.0	23.000
4	23.0	23.000
5	24.0	24.320
6	23.0	23.000
7	23.0	23.000
8	23.0	23.000
9	23.0	23.000
10	23.0	23.000
11	23.0	23.000
12	24.0	23.360
13	20.0	20.025
14	23.0	23.000
15	23.0	23.000

Table 3 shows the performance comparisons of the seven (7) machine learning algorithms- the three traditional machine learning algorithms and the four ensemble learning algorithms in terms of computed performance metrics – RMSE, MAE and coefficient of determination or R^2 value**.**

Table 3: The comparison of results of machine learning algorithms (traditional and ensemble machine learning algorithms obtained from experimental runs.

S/N	Machine learning Regressor Algorithm	Type of ML Algorithm	R^2 or Accuracy Score	MAE Score	RMSE Score
1	Linear Regression	TRADITIONAL ML ALGORITHMS	0.9856	0.4470	1.0319
2	MLP-Neural Network		0.9815	0.6018	1.1711
3	Decision Tree		0.9782	0.3812	1.2709
4	Random Forest (Decision Forest)	ENSEMBLE ALGORITHM	0.9886	0.2802	0.9207
5	AdaBoost		0.9854	0.3076	1.0386
6	Extra Trees		0.9886	0.2847	0.9175
7	XGBoost		0.9870	0.3133	0.9814

B. DISCUSSION OF RESULTS

The results of the PM2.5 concentrations regression modeling and predictions are shown in Figs.10-16. As shown in the Fig.10, the regression scatterplot for MLP Neural Network is very close to the line of best fit with R^2= 0.9815, RMSE=1.1711 and MAE=0.6018 . This is high accurate prediction.

Fig.11 represents the result for the regression scatterplot for XGBoost algorithm. Here the measured and predicted values fall within the line of best fit with R^2= 0.9870, RMSE= 0.9814 and MAE= 0.3133. This is a result by an

DOI: 10.17148/IJARCCE.2022.11107

ensemble algorithm. It can be seen that this is a much more accurate result than that of a traditional machine learning algorithm, MLP Neural Network.

Fig.12 shows the regression result for Multi Linear Regression algorithm. Here, R^2=0.9856, RMSE=1.0319 and MAE=0.4470. The prediction accuracy of the Linear Regression is very high, better than prediction accuracy of the other two traditional machine learning algorithm MLP Neural Network and Decision Trees but lower than the most of the ensemble algorithms. Secondly, its residual errors RMSE and MAE are higher than that of the ensemble learning algorithms Random Forest, AdaBoost, XGBoost and Extra Trees. So in terms of prediction accuracy and residual errors of prediction, the ensemble learning algorithms are better.

Fig. 13 shows the regression scatterplot of Decision Trees algorithm. This algorithm is the principal weak learner for all the ensemble learning algorithms. It is expected that its prediction power can be boosted by adding trees from the ensemble algorithms. Therefore its prediction accuracy is expected to be lower than that of any ensemble algorithm. Its prediction accuracy R^2=0.9782, is the lowest among all the other algorithms under review.

Figs. 14, 15 and 16 shows the regression scatterplots of the ensemble algorithms AdaBoost, Extra Trees and Random Forest algorithms respectively. From the figures, it can be seen that the lines of best fit for their regression scatterplots are very smooth and the residual errors of prediction are very small in values having accuracy scores(R^2) of 0.9854, 0.9886 and 0.9886 respectively with Extra Trees and Random Forest having the best prediction in terms of accuracy rating.

Tables 1 and 2 equally confirmed that the majority of the regression models experimented upon has good prediction accuracy. There is very little variation or residual errors from the actual or measured values from the air pollution sensors when compared to the predicted values from simulation modeling.

So generally, from the results so far produced from the experimental runs, all the models produced good prediction results with very little or minimal error of less than 2%.

V. CONCLUSION

Since air quality is deteriorating worldwide, accurate air quality prediction has important theoretical and practical significance for protection of human lives and the environment. This study proposes the application of ensemble models that utilizes the efficient high-performing ensemble machine leaning algorithms such as AdaBoost, Random Forest, Extra Trees and XGBoost to model and predict one hour before time the PM2.5 pollutant concentrations within Awka Metropolis. The paper demonstrated through various experimental runs that ensemble algorithms such as Random Forest, AdaBoost, XGBoost and Extra Trees have higher accuracy in air pollution prediction than the traditional machine learning algorithms such as Decision Trees, MLP Neural Network and Multi Linear Regression (MLR). The results of the experiments show the relevance of using ensemble learning techniques in improving the accuracy of air quality forecasting in a smart city.

REFERENCES

[1]. Bingyne Pan .”Application of XGBoost algorithm in hourly PM2.5 Concentration prediction". IOP Conference Series: Earth and Environmental Science 113 (2018) 012127. Doi :10.1088/1755-1315/113/1/012127, 2017.

[2]. S. Kumar, S.Mishra, and S.K. Singh. “A machine learning-based model to estimate PM2.5 concentration levels in Delhi's atmosphere”. Heliyon. 2020 . Nov 6(11): e05618. doi: 10.1016/j.heliyon.2020.e05618

[3]. A. Valavanidis, K. Fiotakis, T.Vlachogianni. “ Airborne particulate matter and human health: toxicological assessment and importance of size and composition of particles for oxidative damage and carcinogenic mechanisms”,. J. Environ. Sci. Health, Part C. 2008;26(4):339–362.

[4]. V.M.Madhuri, G.Samyama, K.Savitha."Air Pollution Prediction Using Machine Learning Supervised Learning Approach". International Journal Of Scientific & Technology Research Volume 9, Issue 04, April 2020.pp.118-123.

[5]. S. Kumar, S.Mishra, and S.K. Singh. “A machine learning-based model to estimate PM2.5 concentration levels in Delhi's atmosphere”. Heliyon (6). 2020 .E0568, pp.1-10.

[6]. Bingyne Pan .”Application of XGBoost algorithm in hourly PM2.5 Concentration prediction". IOP Conference Series: Earth and Environmental Science 113 (2018) 012127. Doi :10.1088/1755-1315/113/1/012127, 2017.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11107

[7]. E.Džaferovic and K.K. Hadžiabdic. "Air Quality Prediction and Application using Machine Learning methods: A case study of Bjelave Neighborhood". Technological Systems and Applications. Lecture Notes in Network Systems 142, 2021.

[8] World Health Organisation .“Ambient air pollution”. Accessed online at https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/ambient-air-pollution on January 8, 2022.

