Journal of Inventive Engineering and Technology (JIET)
ISSN: 2705-3865
Volume-2, Issue-3, pp-30-44
www.jiengtech.com

Research Paper Open Access

SMART CITY PM2.5 AIR POLLUTION MODELING TECHNIQUES: TRAIN-TEST DATA SPLIT VERSUS K-FOLD CROSS VALIDATION TECHNIQUES

[1](**Asogwa, Samuel Chibuzor,** Department of Computer Science, Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria) sasogwa@gmail.com

[2](**Obodoeze, Fidelis Chukwujekwu,** Department of Computer Engineering, Akanu Ibiam Federal Polytechnic Unwana, Ebonyi State, Nigeria) fcobodoeze@gmail.com

[3](**Etuk Enefiok A.,** Department of Computer Science, Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria)

[4](**Ugwuja N.E.,** Department of Computer Science, Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria)

Corresponding Author: fcobodoeze@gmail.com

ABSTRACT : *Due to the substantial risks it poses to both human health and the environment, air pollution is a major issue that urban residents and city managers must deal with. Environmental deterioration, respiratory ailments, cardiac health difficulties, and other challenges have all been brought on by air pollution, particularly in densely populated cities or metropolises. To assess the concentrations of air pollutants in the nearby ambient environment, a variety of research techniques have been used in the literature. One such strategy uses a variety of statistical data-driven methods such as machine learning modeling and prediction tools. This is due to the fact that data-driven approaches, as opposed to the so-called chemical models approach, which is rather quite complex and time-consuming, are simpler and more cost-effective for estimating the levels of air pollutants dispersion within a certain place. Artificial intelligence (AI) has several subfields, including machine learning and deep learning, which can be used to train prior historical datasets to detect patterns in an occurrence that can be used to predict or forecast future occurrences of air pollution in a particular location or city. In this paper, two different air pollution modeling and simulation techniques (Train-Test Data Split and K-Fold Cross Validation methods) were used to model/predict the particulate matter (PM2.5) emission in Awka Metropolis. Some historical datasets comprising past air pollutants and meteorological datasets from 2008 to 2013 from the city of Awka Anambra State of Nigeria was utilized carry our PM2.5 emission modeling using eight different machine learning algorithms such as Multi Linear Regression (MLR), Decision Trees, Multi Layer Perceptron (MLP) Artificial Neural Network (MLP-ANN), Support Vector Regressor (SVR), Random Forest, AdaBoost, Extreme Gradient Boosting (XGBoost),and Extra Trees. Performance metrics such as coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Square (RMSE) were used to compute the performances of the machine learning algorithms in terms of their modeling and prediction performances on the training and testing datasets. T*he results obtained from the experimental runs show that the models or algorithms such as - MLR, MLP ANN, Decision Tree, Random Forest (RF), AdaBoost, XGBoost and Extra Trees with the following R^2 scores (0.9856 versus 0.9802; 0.9815 versus 0.8825; 0.9782 versus 0.9742; 0.9886 versus 0.9722; 0.9854 versus 0.9503; 0.9870 versus 0.9696 and 0.9886 versus 0.9716 for the Train-Test Data Split Method and 10-Fold Cross-Validation Test method respectively. These results from the two different modeling methods show that some levels of similarities in terms of prediction accuracy and errors of prediction. Therefore, the two prediction modeling techniques are adequate and suitable for the prediction modeling and estimation of PM2.5 pollution levels within Awka Metropolis. The prediction results obtained in Train-Test Data Split method are validated by the results obtained from using a K-fold Cross Validation approach.

This shows that the two air pollution modeling and estimation techniques are suitable for modeling and prediction of air pollutant levels, since the results obtained from the two approaches show close correlation in terms of prediction accuracy and residual errors of prediction.

KEYWORDS: *PM2.5, air pollution, particulate matters, data-driven models, chemical models, Train Test Data Split, K-Fold Cross Validation, meteorological dataset, machine learning algorithms, modeling*

Date of Submission: 17-03-2023 Date of acceptance: 22-03-2023

I. INTRODUCTION

Air pollution is a very big challenge all over the world, especially in big cities where there is population explosion, so there is urgent need to provide smart-city air pollution monitoring and management solutions to forestall the challenges of air pollution. Population explosion is the main reason that leads to increased air pollution issues in big cities; this is due to increased human activities, use of industrial machines, motor vehicle exhausts, and also emissions from small-scale businesses and domestic activities. The main air pollutants include particulate matters such as PM10.0, PM2.5, PM1.0 and gaseous effluents such as SO_2, NO_2, NO, CO, CO_2, Ozone (O_3), Volatile Organic Compounds (VOCs), Formaldehydes (HCHO) as well as noise pollution; out of these air pollutants, the PM2.5 (i.e. smaller inhalable particles, with diameters that are generally 2.5 micrometers and smaller) are more dangerous because they can get into the deep parts of your lungs — or even into your blood. On long term basis, patients with respiratory diseases such as asthma and lung diseases can suffer more severe consequences such as death if affected by air pollutants such as PM2.5 (EPA, 2023). Bigger or coarser particles such as PM10.0 can cause eyes, nose, and throat irritations. Dust from roads, farms, dry riverbeds, construction sites, and mines are types of PM10.0 particles.

Due to the observed harmful impacts air pollution has on people's health and the environment, it is becoming more and more crucial to be able to model, predict, and monitor the quality of the air in urban regions and developed cities. In order to monitor and manage air pollution in a city or metropolis using machine learning prediction algorithms or models, smart city solutions—which are the combination of Internet-of-Things (IoTs) ecosystems comprising sensors, actuators, communication networks, such as cloud server systems, and governance policies—are being used more and more.

However, few communities can afford to build expensive monitoring stations or sub-stations for air quality. As they can provide city administrators and the general public with precise and reliable air quality forecasts in advance, air quality prediction systems readily come to the rescue rather than relying on real-time air quality stations and monitors. Using historical weather or meteorological data as well as previous air pollutant concentrations of the specific air pollutant to measure, air quality prediction systems can use machine learning and deep learning algorithms and models to predict the likely outcomes of a city's air quality in advance. The local weather and meteorological parameters, such as atmospheric or air temperature, air pressure, relative humidity, wind speed, wind direction, rainfall or precipitation amount, snow level, and light intensity, etc. have a direct impact on the hourly and daily concentrations of PM2.5. To correctly and successfully conduct a predictive modeling of PM2.5 monitoring in a certain place or city, it is necessary to combine the historical data of these weather parameters together with the past air pollutant dataset.

The hourly, six-hourly, 12-hourly, and 24-hourly time series averages of the PM2.5 air pollution concentrations can be used to map pollution trends. Without having to spend money on additional air pollution monitoring stations, a much more expensive approach, it may be possible to gain actionable intelligence by understanding the relationship (i.e., developing a predictive model) between PM2.5 and weather. The management and control of an urban area's air quality index (AQI) has also shown this strategy to be successful. Air pollution modeling problem can be classified mainly as *classification prediction* or *regression prediction* time series problem.

Regressive prediction models provide the actual values of air pollutant concentration at a specific time period, while classification prediction or forecasting models provide AQI classes of the amount of air pollution in a city.

Models of the emissions that cause air pollution are representations of the actual emission-producing systems. They provide a consistent framework for describing the sizes, locations, and temporal variations of emissions from various sources.

PM2.5 pollutant concentrations and other air pollutants can be modeled using either *chemical models* or *data-driven models* but *chemical models* (i.e. classical deterministic models, sometimes referred to as *chemistry-transport models*) are based on the chemical laws to model all the relevant chemical processes that contribute to PM2.5 formation. Such models may describe up to hundreds of species such as troposphere, photochemistry and aerosols (Mallet & Spotisse, 2008).

Data-driven models also known as *stochastic or deterministic approaches* use historical data to make future predictions. They are based specifically on *statistical approaches*. Train-Test Data Split and K-Fold Cross Validation modeling techniques are typical popular approaches of data-driven modeling techniques.

Due to the complexity of chemical transformation models for air pollution and the difficulty in obtaining the current comprehensive list of emissions, there are several limits in the prediction of PM2.5 concentrations using chemical models. One of these restrictions is that the model inputs, such as emission inventories, are not very precise (for spatial and temporal distribution, for chemical speciation), and the meteorological fields are also imprecise (2). Too many parameters could result in heavy mathematical and computing cost (Mallet & Spotisse, 2008) .Several data-driven machine learning and statistical air pollutant modeling approaches have been adopted in literature.

In this paper, two different data-driven air pollution modeling and simulation techniques (Train Test Data Split and K-Fold Cross Validation methods) will be used to model/predict the particulate matter (PM2.5) emission in Awka Metropolis. Some historical datasets comprising past air pollutants and meteorological datasets from 2008 to 2013 from the city of Awka Anambra State of Nigeria was utilized carry our PM2.5 emission modeling using eight different machine learning algorithms such as Multi Linear Regression (MLR), Decision Trees, Multi Layer Perceptron (MLP) Artificial Neural Network (MLP-ANN), Support Vector Regressor (SVR), Random Forest, AdaBoost, Extreme Gradient Boosting (XGBoost), and Extra Trees. Performance metrics such as coefficient of determination (R2), Mean Absolute Error (MAE) and Root Mean Square (RMSE) were used to compute the performances of the machine learning algorithms in terms of their modeling and prediction performances on the training and testing datasets.

The aim of this research paper is to model PM2.5 emission in Awka Metropolitan City of Anambra State of Nigeria using two different modeling approaches viz: Train-Test Data Split method and K-Fold Cross Validation method and comparing the results obtained from these two modeling approaches

II. MATERIALS AND METHODS

This section presents the materials and methodology used in the experiments of this research paper.

A. THE MATERIALS

Awka City has no single Air Pollution Monitoring Station but a low-cost air pollution sub-station was constructed and implemented in the first phase of this research using real-time wireless air pollution sensors to collect historical records of air and noise pollution from October 25th 2021 to December 4th, 2021; 12,958 rows of datasets were collected to fit the models, covering a total of 43 days. Ground measured minutely pollutant concentrations including those of TVOC, PM10, PM2.5, PM1.0, noise, as well as the weather or meteorological dataset such as air temperature, pressure, relative humidity, light intensity from air and noise pollution monitoring station. The dataset for the experiments can be found at *http://www.myrasoft.ng/awka-pollution-monitor/AWKA-POLLUTION-2022NEW.csv*

B. METHODS

a. Experiment A (using Train-Test Data Split Method)

The train-test Data Split technique (Experiment A) was used to estimate the performance of each of the eight machine learning algorithms in predicting the PM2.5 air pollutant levels. This method is a fast and easy procedure to compute prediction accuracy (R^2) and prediction errors (RMSE and MAE) for each pollutant's concentration of each machine learning model results so as to compare them to one another.

By default Test set is split into 30% of actual data and the Training set is split into 70% of the actual data as shown in Fig. 1.

The dataset is split into training and testing sets to evaluate how well the machine learning models are performing. The *train set* is used to train or *fit the model*, the statistics of the train set are known. The second set is called the *test or testing dataset*, this set is solely used to *perform predictions* only and evaluate the model.

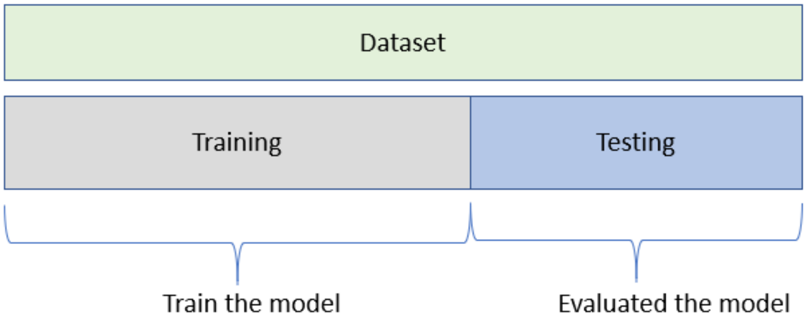
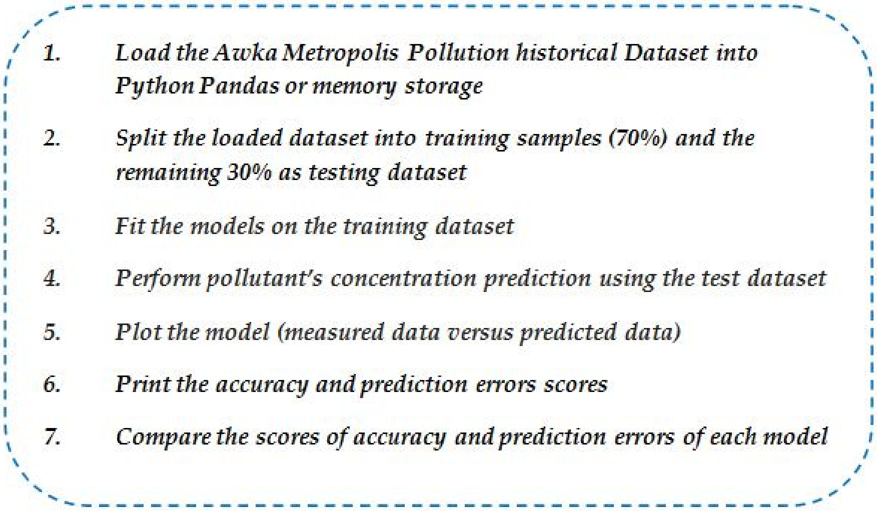


Fig. 1: Dataset splitting into training and testing datasets for Train-Test Data Split method

Fig.2 shows the following steps taken to actualize Experiment 4 using Train-Test Split Method:



1. *Load the Awka Metropolis Pollution historical Dataset into Python Pandas or memory storage*
2. *Split the loaded dataset into training samples (70%) and the remaining 30% as testing dataset*
3. *Fit the models on the training dataset*
4. *Perform pollutant's concentration prediction using the test dataset*
5. *Plot the model (measured data versus predicted data)*
6. *Print the accuracy and prediction errors scores*
7. *Compare the scores of accuracy and prediction errors of each model*

Fig.2 : Train-Test Data Split method used in the experiments

The experimental runs or simulations for each of the eight machine learning models were executed in Scikit-learn machine learning module in Python 3. The next section in the second experiment (Experiment B) uses the k-Fold Cross Validation technique to model and estimate the concentrations of PM2.5 in Awka Metropolis.

b. Experiment B (using K-Fold Cross-Validation technique)

K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called **folds**. For each learning set (training), the prediction function uses k-1 folds, and the rest of the folds are used for the test set. This approach is a very popular cross validation (CV) approach because it is easy to understand, and the output is less biased and more accurate than other methods.

The steps for k-fold cross-validation are:

Split the input dataset into K groups

- ***For each group:***
    - ***Take one (1) group as the reserve or test data set.***
    - ***Use remaining groups (k-1) as the training dataset***
    - ***Fit the models on the training dataset and perform prediction and validation***
    - ***Compute the validation accuracy scores and find the average for all the iterations or rounds***
    - ***Use the test data to perform the final test on the models using MAE, RMSE and R^2 performance metrics.***

Fig. 3 depicts the K-Fold cross validation test methodology used in this experiment and its sub-experiments. *E* represents the results of the regression performance evaluation criteria such as MAE, RMSE, and R$^{2.}$

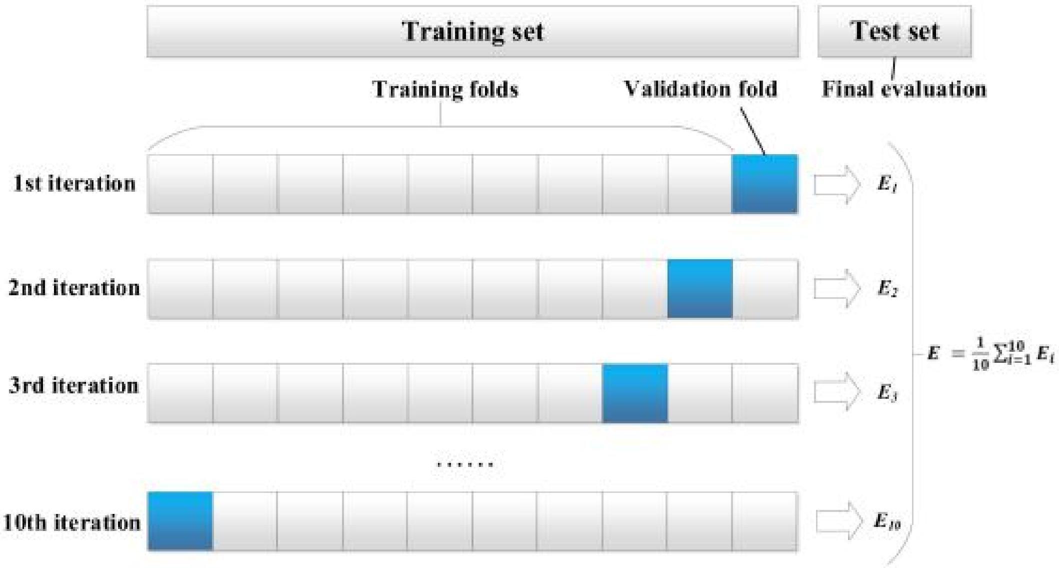


Fig.3: A 10-Fold Cross Validation method used to train, test and validate the selected machine learning modes for the prediction of PM2.5 Air pollutant levels within Awka Metropolis.

We implemented the k-Fold Cross Validation tests (with k=10) in Python simulation environment in all the sub-experimental runs to validate or confirm the results obtain in Experiment 4 using Test-Train data-split method. This K-fold cross validation method reduces the chances of the models of being overfitted or underfitted and increases the accuracy and effectiveness of the model on unseen future dataset.

The Figs. 4 and 5 summarized the steps taken to actualize Experiment 2 using k-Fold Cross Validation Test method:

1. ***Load the Awka Metropolis Pollution historical Dataset into the model stack***

2. ***Shuffle the dataset randomly***

3. ***Split the dataset into k- groups or folds***

4. ***For each unique group in k***

    1. Take the group as a hold out or test data set
    2. Take the remaining groups (k-1) as a training and validation data set
    3. Fit a model on the training set and evaluate it on the test set
    4. Retain the validation score and discard the model
    5. Compute the final accuracy score for all the rounds or iterations

Fig.4: Steps taken in the experiments to implement a 10-Fold Cross Validation technique

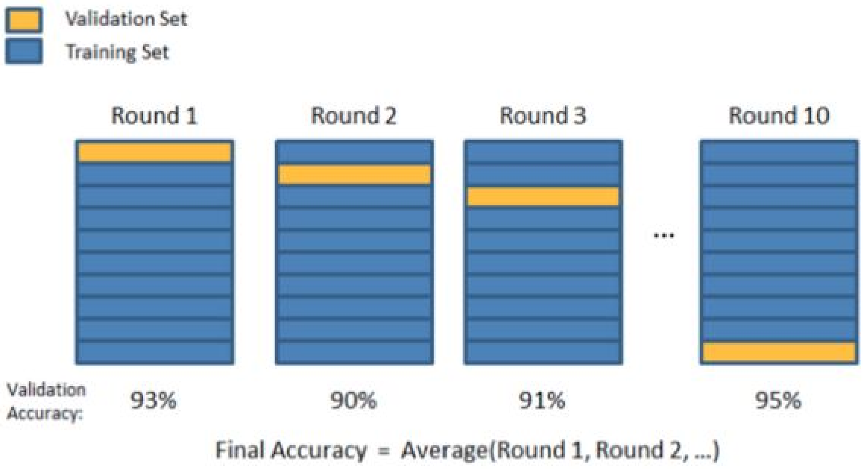


Fig. 5: Iteration steps for 10-Fold Cross validation method to computer the average final accuracy scores

The experimental runs or simulations for each of the eight machine learning models were executed in Scikit-learn machine learning module in Python 3.6.

The results of 10-Fold Cross Validation tests carried out in Experiment 5 are presented in section

C. Data Normalization

The dataset used for the experiments are *normalized or scaled* to remove any form of irregularities in the values or weights of the models. The range normalization function used is mathematically given as follows:

$$X_{normalisation} = \frac{(X_i - X_{min})}{(X_{max} - X_{min})} \qquad (1)$$

where $X_{normalisation}$ **i**s the normalized value, $\mathbf{X_i}$is the i_{th} value passed, and $\mathbf{X_i}$ and $\mathbf{X_i}$ are the minimum and maximum value for $\mathbf{X_i}$ value respectively.

The data normalization was performed in Python using MinMax Scalar() inbuilt function in Scikit-learn machine learning module.

D. PERFORMANCE EVALUATION METRICS

In order to determine or evaluate the best machine learning Air pollution prediction models quantitatively in terms of error bands or the prediction accuracy, the following statistical performance metrics- Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination or Variance (R^2) were employed and calculated as shown in Eqs. (4)-(6).

A. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - M_i| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

B. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|P_i - M_i|^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(3)$$

C. Coefficient of Determination (R^2)

$$\boldsymbol{R^2 = 1 - \frac{\sum(M_i - P_i)^2}{\sum(M_i - \overline{M_i})^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

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**.**

c. RESULTS AND DISCUSSION

This section describes the experimental results and discussion of the results obtained from the experiments as carried out in the research paper.

A. Experiment A Result (PM2.5 prediction using Train-Test Data Split method)

Figs. 6-13 show the regression scatterplots of the various machine learning algorithms as they performed during experimental runs (in Experiment 4) on the dataset for PM2.5 pollution concentrations prediction using Train-Test Data Split method. Table 1 also presents the performance evaluation result obtained for PM2.5 Pollution prediction results using the various eight (8) machine learning algorithms using Train-test Split data method.

Table 1: PM2.5 Pollution prediction results using various machine learning algorithms and Train-Test Data Split method

ML model or algorithm	RMSE	MAE	R^2
MLR	1.0319	0.4470	0.9856
SVR	1.6745	0.5218	0.9622
MLP ANN	1.1711	0.6018	0.9815
Decision Tree	1.2709	0.3812	0.9782
Random Forest	0.9207	0.2802	0.9886
AdaBoost	1.0386	0.3076	0.9854
XGBoost	0.9814	0.3133	0.9870
Extra Tree	0.9175	0.2847	0.9886

From Fig. 6, the prediction of PM2.5 Concentrations was performed using MLP Neural Network algorithm and the following results were obtained: - RMSE=1.1711µg/m^3, MAE= 0.6018 µg/m^3 and R^2= 0.9815 or 98.15% prediction accuracy.

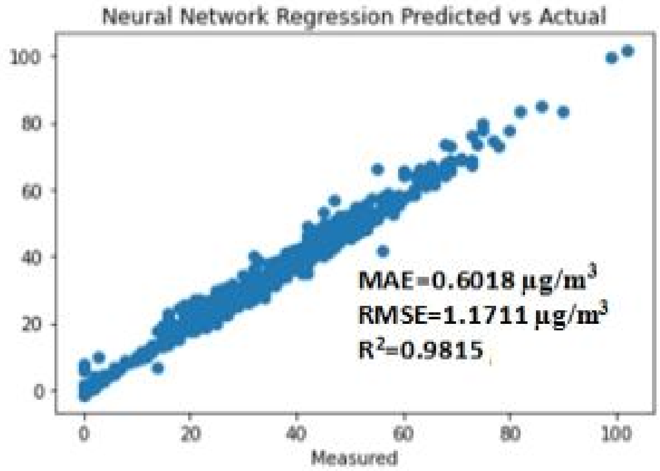


Fig.6: Regression scatterplot of PM2.5 Pollution using MLP Artificial Neural Network (ANN) algorithm

Fig.7 shows the regression plot of PM2.5 concentrations prediction using XGBoost algorithm and the following results were obtained:- RMSE=0.9814µg/m^3 , MAE= 0.3133µg/m^3 and R^2= 0.9814 or 98.14% prediction accuracy.

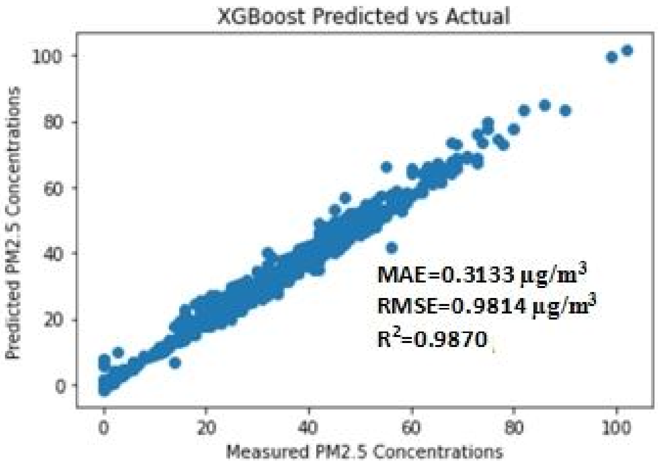


Fig.7: Regression scatterplot of PM2.5 Pollution using XGBoost algorithm

Fig. 8 shows the regression plot of PM2.5 concentrations prediction using Multiple Linear Regression (MLR) algorithm and the following results were obtained:- RMSE= 1.0319µg/m^3 , MAE= 0.4470µg/m^3 and R^2= 0.9856 or 98.56% prediction accuracy.

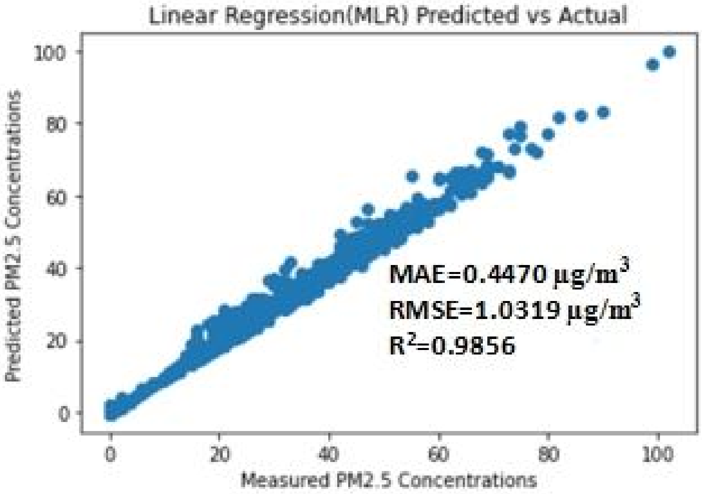


Fig.8: Regression scatterplot of PM2.5 Pollution using MLR algorithm

Fig. 9 shows the regression plot of PM2.5 concentrations prediction using Decision Tree algorithm and the following results were obtained:- RMSE= 1.2709µg/m^3 , MAE= 0.3812µg/m^3 and R^2= 0.9782 or 97.82% prediction accuracy.

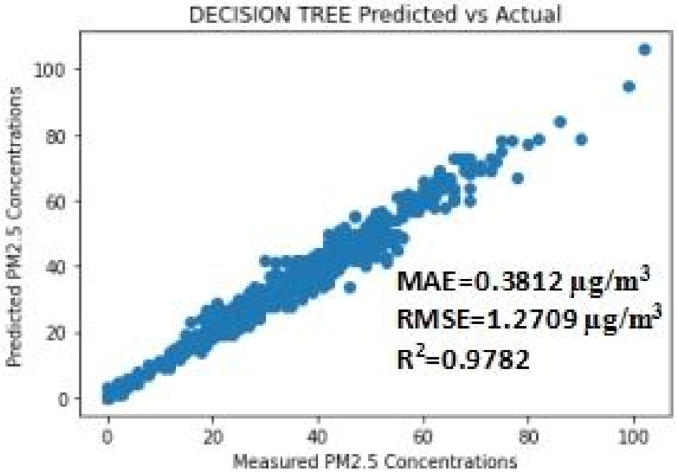


Fig.9: Regression scatterplot of PM2.5 Pollution using Decision Tree algorithm

Fig. 10 shows the regression plot of PM2.5 concentrations prediction using Adaptive Boosting (AdaBoost) algorithm and the following results were obtained:- RMSE= 1.0386 µg/m^3 , MAE= 0.3076 µg/m^3 and R^2= 0.9854 or 98.54% prediction accuracy. The hyperparameters used for AdaBoost during the experiment include Learning rate=1, number of estimators=200 and random state=1234.

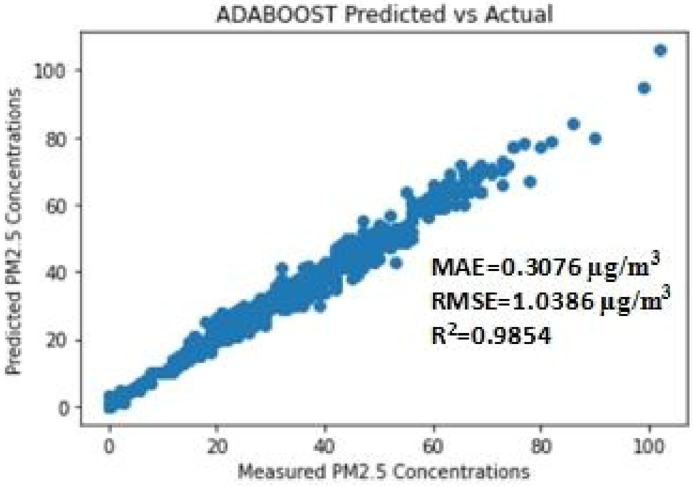


Fig.10: Regression scatterplot of PM2.5 Pollution using AdaBoost algorithm

Fig. 11 shows the regression plot of PM2.5 concentrations prediction using Extra Tree algorithm and the following results were obtained:- RMSE= 0.9175 µg/m^3 , MAE= 0.2847 µg/m^3 and R^2= 0.9886 or 98.86% prediction accuracy. The hyperparameters used for Extra Trees during the experiment include Learning rate=1, number of estimators=200 and random state=1234.

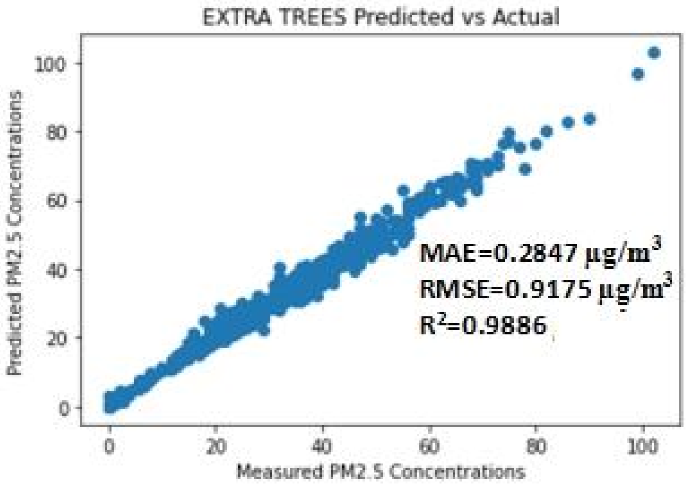


Fig.11: Regression scatterplot of PM2.5 Pollution using Extra Trees algorithm

Fig. 12 shows the regression plot result of PM2.5 concentrations prediction using Random Forest algorithm and the following results were obtained:- RMSE= 0.9207 µg/m^3 , MAE= 0.2802 µg/m^3 and R^2= 0.9886 or 98.86% prediction accuracy. The hyperparameters used for Extra Trees during the experiment include Learning rate=1, number of estimators=200 and random state=1234.

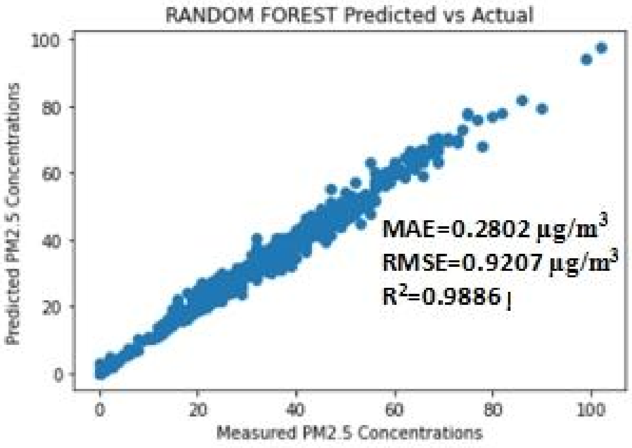


Fig.12: Regression scatterplot of PM2.5 Pollution using Random Forest algorithm

Fig. 13 shows the regression plot result of PM2.5 concentrations prediction using Support Vector Regression (SVR) algorithm and the following results were obtained: RMSE= 1.6745 µg/m^3 , MAE= 0.5218 µg/m^3 and R^2= 0.9622 or 96.22% prediction accuracy.

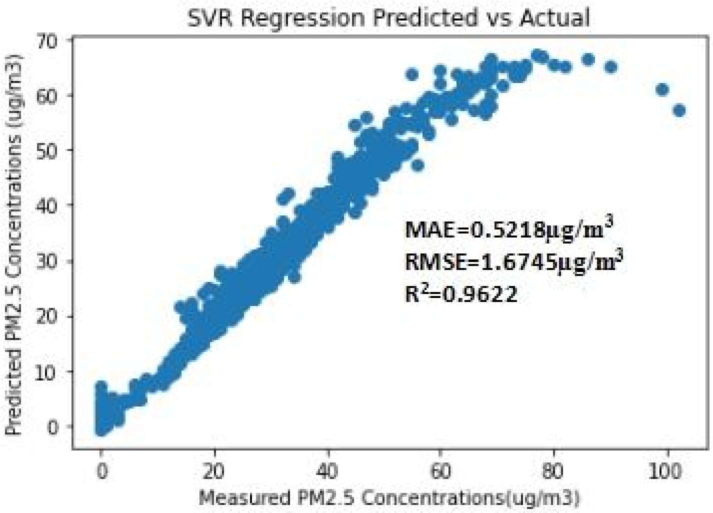


Fig.13: Regression scatterplot of PM2.5 Pollution using SVR algorithm

From Table 1, the results of PM2.5 prediction simulation using different ML algorithms, Extra Trees and Random Forest (RF) algorithms outperformed other six ML algorithms with the highest accuracy score (R^2=0.9886 or 98.86% both tied) and the lowest prediction errors (MAE=0.2847 µg/m^3 and RMSE=0.9175 µg/$m^3$for Extra Trees and (MAE=0.2802µg/m^3 and RMSE=0.9207µg/m^3 for Random Forest. Also, from Fig. 11 and 12, the regression scatterplots of PM2.5 using Extra Trees and Random Forest algorithms respectively, the regression plots are very smooth and the line-of-best fit are very smooth; the variance between the predicted values from the observed or measured values from the sensors are well explained by the regression lines. Therefore any one of Extra Trees algorithm and Random Forest is selected for the future prediction of PM2.5 concentrations in Awka Metropolis.

We are going to compare the prediction modeling results obtain in the first part of the experiments using Train-Test Data Split method and K-Fold Cross Validation method in the subsequent tables below.

B. Experiment B Result (PM2.5 prediction using K-Fold Cross Validation method)

Experimental results from prediction of Air and noise pollutant levels in Awka Metropolis using k-Fold Cross Validation (k=10) Test method:

Tables 2 presents the results from experimental runs for PM2.5 air pollution levels within Awka Metropolis using k-Fold Cross Validation method. The number of k used is 10, i.e. k=10. The objective of this experiment is to validate the results obtained in Experiment A using machine learning algorithms. Table 3 shows the comparison between the results obtained using Train-Test Data Split method (in Experiment A) to 10-Fold Cross-Validation Test (in Experiment B. These results will help to compare and determine the machine learning algorithm to select for future prediction and deployment within Awka Metropolis.

Table 2: PM2.5 Pollution prediction results using various machine learning algorithms and 10-Fold Cross Validation technique

ML model or algorithm	**RMSE**	**MAE**	**R^2**
MLR	1.0244	0.4350	0.98018
SVR	5.2103	0.5102	0.8752
MLP ANN	4.3047	3.5525	0.8825
Decision Tree	1.1382	0.3366	0.9742
Random Forest	1.3415	0.4134	0.9722
AdaBoost	1.8090	1.0139	0.9503
XGBoost	1.3849	0.4355	0.9696
Extra Tree	1.3663	0.4201	0.9716

Table 3: Comparison of results from Train-Test Data Split method and 10-Fold Cross Validation technique for PM2.5 levels prediction

ML model/Algorithm	**Training, Testing and Validation Technique**	**RMSE** **(µg/m^3)**	**MAE** **(µg/m^3)**	**R^2**
MLR	Train-Test Split	1.0319	0.4470`	0.9856
	10-Fold Cross Validation Test	1.0244	0.4350	0.98018
SVR	Train-Test Split	1.6745	0.5218	0.9622
	10-Fold Cross Validation Test	5.2103	0.5102	0.8752
MLP Neural Network Regressor	Train-Test Split	1.1711	0.6018	0.9815
	10-Fold Cross Validation Test	4.3047	3.5525	0.8825
Decision Tree	Train-Test Split	1.2709	0.3812	0.9782
	10-Fold Cross Validation Test	1.1382	0.3366	0.9742
Random Forest	Train-Test Split	0.9207	0.2802	0.9886
	10-Fold Cross Validation Test	1.3415	0.4134	0.9722
AdaBoost Regressor	Train-Test Split	1.0386	0.3076	0.9854
	10-Fold Cross Validation Test	1.8090	1.0139	0.9503
XGBoost Regressor	Train-Test Split	0.9814	0.3133	0.9870
	10-Fold Cross Validation Test	1.3849	0.4355	0.9696
Extra Tree Regressor	Train-Test Split	0.9175	0.2847	0.9886
	10-Fold Cross Validation Test	1.3663	0.4201	0.9716

From Table 2, PM2.5 prediction results, it can be seen that Multiple Linear Regression (MLR) algorithm outperformed other machine learning algorithms in terms of prediction accuracy for PM2.5 prediction with R^{2}score of 0.98018 (98% accuracy) and the lowest prediction residual errors (RMSE=1.0244µg/m^3 and MAE=0.4350 µg/m^3) compared to other algorithm, though other machine learning algorithms performed very well in this sub-Experiment and thus validated the results they posted in Experiment A using Train-Test Data Split method.

In this section of the analysis of result, efforts is made here to compare the results obtained in Experiment A using Train-Test Data Split method and Experiment B (10-Fold Cross-Validation Test method) respectively. This will help to determine the pollutant predictions that are correct or validated by the 10-Fold Cross-Validation Test since results obtained from the Train-Test Data Split method alone could be misleading.

Results obtained as presented in Tables 3 shows the comparison of the prediction results in terms of performance accuracy (R^2) and residual errors of prediction in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The essence of these comparisons is to determine the prediction results that were accurately validated by the experiments carried out in Experiment A using a 10-Fold Cross-Validation test method in Experiment B.

From Table 3 on the prediction of the particulate matter PM2.5, the results obtained from the experimental runs show that the following models or algorithms- MLR, MLP ANN, Decision Tree, Random Forest (RF), AdaBoost, XGBoost and Extra Trees with the following R^2 scores (0.9856 versus 0.9802; 0.9815 versus 0.8825; 0.9782 versus 0.9742; 0.9886 versus 0.9722; 0.9854 versus 0.9503; 0.9870 versus 0.9696 and 0.9886 versus 0.9716 for the Train-Test Data Split Method and 10-Fold Cross-Validation Test method respectively. These results from the two different modeling methods show that some levels of similarities in terms of prediction accuracy and errors of prediction. Therefore, the two prediction modeling techniques are adequate and suitable for the prediction modeling and estimation of PM2.5 pollution levels within Awka Metropolis. The prediction results obtained in Train-Test Data Split method are validated by the results obtained from using a K-fold Cross Validation approach.

IV. CONCLUSIONS

Smart city solutions can be deployed in the management of the environmental pollution and can be deployed to air quality prediction in order to forecast the concentrations of air pollutant parameters present in a city or metropolis in advance before it occurs. This is important so as to alert city administrators and the general public to know the implications of the environment in order to protect their health. Machine learning is a branch of Artificial Intelligence that can be used to predict or forecast the possible concentrations of air pollutants in the atmosphere before it occurs using past historical dataset of air pollutants and meteorological parameters of the same locality. This paper has demonstrated that it is possible to use machine learning algorithms to model and predict air quality of a city or metropolis.

In this work, the particulate matter $PM_{2.5}$ prediction modeling were carried out using nine (9) different machine learning models and two major modeling techniques known as Train-Test Data Split and K-Fold Cross Validation technique where K=10 was used in the experiments. The prediction performances of these machine learning models were evaluated using statistical performance metrics such as MAE, RMSE and R^2.

From the two modeling techniques, it is very clear that K-Fold Cross Validation technique gave a slightly closer results to the Train-Test Data Split technique in terms of R^2, MAE and RMSE but these results are considered far more accurate and reliable because K-Fold Cross Validation technique is used to detect and reduce prediction errors such as bias and variance as a result of over-fitting and under-fitting associated with other modeling techniques such as Train-Test Data method. K-fold Cross validation generalizes very well on unseen future dataset when compared to Train-Test Data Split which can take in noise from the input dataset as true dataset; this leads to errors from over-fitting but K-Fold Cross validation technique eliminates these errors. The only issue with K-fold Cross Validation technique is high CPU processing overhead as a result of repeated iteration of each k-group for training, validation and testing. K-Fold Cross Validation technique is also suitable when there are few or limited historical dataset for training and testing compared to the Train-Test Data Split technique that requires a large amount of historical dataset obtained from past air pollutants and meteorological parameters.

Conflict of interest

The authors declare that there is no conflicting interest in the publication of this paper.

REFERENCES

EPA (2023). "What are the Harmful Effects of PM?" Accessed online at https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

Mallet, Vivien and Spotisse, Bruno (2008). Air Quality modeling: From deterministic to stochastic approaches. Computers & Mathematics with Applications. Volume 55, Issue 10, May 2008. pp. 2329-2337