IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

APPLICATION OF MACHINE LEARNING FOR MODELLING CONCENTRATION AND DISPERSAL OF AIR POLLUTANTS IN ALESA-ELEME, RIVERS STATE NIGERIA

[1]Ahmad K, [2]Leton T. G & [2]Ugbebor J. N.
[1,2 &3]Center for Occupational Health and Safety, Institute of Petroleum Studies, University of Port Harcourt, Nigeria.

Abstract: This research was conducted to explore the use of machine learning approach in modeling the relationship between air pollutants (CO_2, VOC, PM_{10},) and meteorological parameters (Wind Speed, Air temperature, Solar Radiation) in Alesa-Eleme, River State. The data was gathered using AQM 65 at three (3) sites spread over the study area for a period of fourteen (14) Months. Statistical analysis of the data reveled the relationship between air pollutants concentrations and meteorological parameters. The correlated parameters were subjected to Machine Learning (ML) techniques; RF, NB, ANN, SVM and LR to predict concentration and dispersal of air pollutants in relation to meteorological dynamics. The five ML technique were evaluated and validated, and the result showed that RF was more accurate than the other considered ML techniques, and therefore was used in the prediction of pollutants concentration and dispersal using Orange Canvas and WEKA software. Applying the RF, pollutants concentrations were estimated with CA of 0.874 and Precision of 0.881. This implies that the application of ML concept using high quality and accurate data can bring more advances in Nigeria not only for air quality prediction, but any type of environmental monitoring to help preparedness, raise awareness and build resilient Environmental Management System, especially in areas more prone to industrial pollution.

INTRODUCTION

There have been several investigations on the use of machine learning algorithms for classifying air quality and assessments. According to Muhammed et al. (2015), machine learning methods are ideal for forecasting air quality.

With the advent of AI in recent years, classical machine learning and deep learning have also been effectively applied to the field of air quality prediction with positive results (Chen et al., 2016). Popular decision tree algorithms like ID3, C4.5, and CART are just a few examples of how the decision tree approach has been put in to use (Punia et al., 2011; Wang and Kong, 2019). Furthermore, the k-nearest neighbor algorithm has also been employed. Random forests were also used in this area with the advent of ensemble learning (Breiman, 2001). According to Graves, (2012), random forests employed a method in which many decision trees were constructed using subsets of data, with the aggregated forecasts serving as the final prediction. The use of big data, neural networks, and deep learning has been expanding in this area in recent years. A recursive neural network (RNN) was trained to predict future air quality changes by analyzing past concentration changes in NO_2, CO_2, SO_2, $PM_{2.5}$, and other air pollutants. The earliest artificial neural network (ANN) was utilized to analyse time-series data (Graves, 2013; Lipton et al., 2015). Later, the issue of partial gradient disappearance was fixed by creating a model with both long- and short-term memory (LSTM). Good promise was shown in the model's used for predicting air quality (Geer, 2001). Almost all previous research in this area has concentrated on increasing accuracy at the expense of algorithm complexity.

Evaluation of air quality is also critical in the fight against air pollution. As the state of the atmosphere worsens, there has been an uptick in the number of studies that use categorization as a means of assessing air quality. Constant progress is being made in the effectiveness and precision of relevant algorithms, and new approaches are continually being created. Air quality evaluation is a field with a wide variety of prediction techniques. The concentration value of specified air pollutants may be used to derive the air pollution index (API), which is the more direct and efficient way. The evaluation of the pollution index provides immediate information on the state of the air. This technique works well for the assessment of current conditions and short-term changes in air quality (Bin, 2008). Traditional methods of analysis have been used to determine the state of the air for quite some time. Mathematical and statistical methods are the foundation of the standard methods for predicting air quality (Niharika and Rao, 2014; Kujaroentavon et al., 2014). Such methods begin with the development of a physical model and the use of mathematical equations to encode data. There is a lot of math involved here. Furthermore, the precision provided by these techniques is rather low. Alternatives to the conventional methods that make use of big data and machine learning have been developed more recently (Kang et al., 2018).

Air quality around industrial facilities or an area prone to air pollution from industrial and other human activities needs to be monitored on a continuous long-term basis (Al-Salem and Khan, 2008; Simpson et al., 2013; Sanchez et al., 2019). Current air quality reports suggest further study is needed to better understand the complex relationship between ambient air quality before any pollutant may be forecasted using information about another pollutant. Novel approaches that incorporate state-of-the art technology deployment to study air pollutants concentrations and the level of association/relationship among pollutants and with meteorological variables will accelerate air pollution studies and management.

Concerns over air pollution in the Niger Delta Region of Nigeria have prompted many studies to assess criteria air pollutants concentrations in Rivers State, most notably Port Harcourt, and the possible association with airborne diseases. However, most of these studies could not account for temporal changes between samples or the human errors in measurement as the data were collected using handheld devises that could not continuously take measurements in real-time. To have a better understanding of air pollution problem at minimal cost and time, a Machine Learning approach and mathematical models can be deployed, and this fully rely on access to high-quality, accurate, and continuous air quality and meteorological data from specialized air quality and meteorological monitoring stations. In this research, high quality air pollutants and meteorological data (generated using AQM 65 - Continuous Monitoring Station) was used for Machine Learning Technic deployment to predict concentration of air pollutants, understand the relationship among air pollutants and their dispersion/conversion as affected by meteorological variables and seasonal variations. This was achieved by collecting and analysing continuous, real-time air quality and metrological data aimed at improving our understanding and management of air pollution issues in the study area and other locations with comparable characteristics.

AIM OF THE STUDY

The aim of this study is to understand the relationship between air pollutants and meteorological parameter and explore the application of Machine Learning to predict concentration and dispersal of air pollutants in the study area.

2. MATERIAL AND METHODS / METHODOLOGY

2.1 Study Area

This study was performed in Port Harcourt Refining Company Ltd, complex located at Alesa-Eleme between long. 4º45'33"N and lat. 7º06'15"E of the Greenwich meridian, Port Harcourt Rivers State, Nigeria. Major activity in the study area is petroleum refining. The focus of the study was on the wastewater treatment plant which was designed to treat 16 m3 of sanitary wastewater and 495.6 m3 of combined process wastewater. Only sanitary wastewater (3.13% of the wastewater treatment capacity) is being treated during the study period due to process plants being on shut down mode. These created an environment that support the conduct of base line air quality studies in the area. Figure 2.1 shows the google snapshot of the study area and location of air quality/weather monitoring stations.

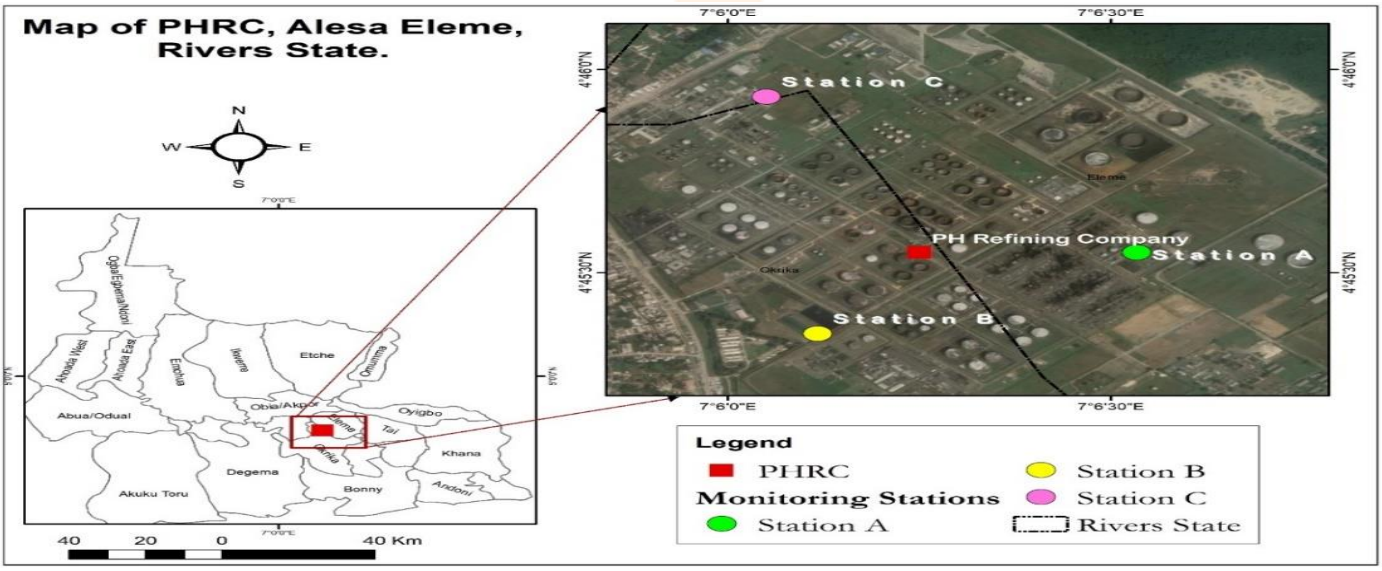

Figure 2.1: Map of the Study Area (Source: Google earth)

2.2 Research Design

In this research, three (3) air quality and weather monitoring stations (AQM 65) were sited at different locations within the study area, as Stations (A – C); STATION A (4º45'33"N 7º06'32"E), STATION B (4º45'21"N 7º06'07"E) and STATION C (4º45'56"N 7º06'03"E) in accordance with Specio-temporal consideration. Air pollutants and meteorological data were collected using AQM 65 in accordance with manufacturer's recommendations for a period of fourteen (14) months (November 2019 – January 2021). The air sample collected via gas and particulate matter inlets of the AQM 65 were analyzed by the different modules located in the equipment cabin. To ensure consistency of data, descriptive statistics was applied for air pollutants and meteorological data as well as correlation matrix between the pollutants considered.

2.2.1 Computation of Air Quality Index

The purpose of AQI is for the machine learning computation. Calculating the Air Quality Index included arithmetically averaging the concentrations of PM10, PM2.5, NO2, and SO2 relative to their reference values (AQI). Taking the mean and multiplying by 100 yields the AQI index. After that, the AQI was compared to other metrics using a scale (Fitz-Simon, 1999). With this formula, we were able to calculate the equivalent AQI for each individual pollutant:

$$\text{AQI} = \frac{\text{C}}{\text{C}_\text{s}} \times 100 \qquad (2.1)$$

where

AQI is Air Quality Index, C is the observed value of the air quality parameters based on CPCB standard (CPCB, 2009)

2.2.2 Machine Learning

In this research, we focused on supervised ML to discover data clusters, learn from past occurrences, and create a classification model for the future. Using this ML technique in conjunction with previously collected data yields the best results. Multiple data mining applications, including XLSTAT, were used for this goal. The XLSTAT was used for this purpose so that more ML methods may be tested on a wider range of training and testing data sizes. The program may be accessed with minimal effort and has a nice interface.

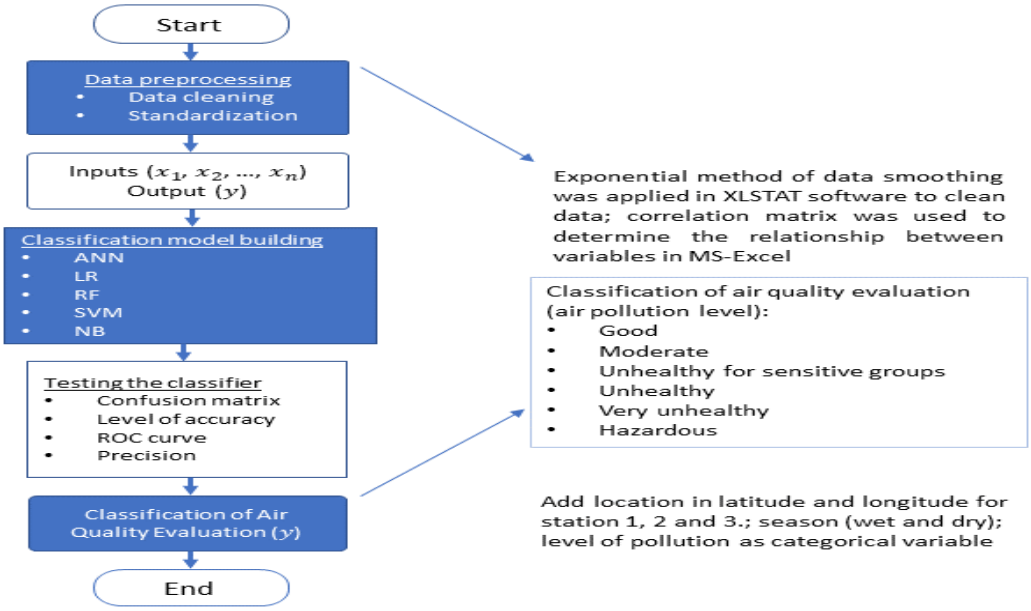

Figure 2.2: Algorithm for implementation of machine learning model

The information generated was separated into two. The first portion (75%) was used for the model's training and generation phases, while the second portion (25) was put in to use during validation and testing. To determine which ML method yields the best results, many models were built. However, a classification model for the future ambient air quality of the study area was developed using data from a time series of ambient air quality measurements and weather observations, considering the influence of both meteorological conditions and seasonal variations. This was accomplished with the help of machine learning (ML).

In this study, a data mining software known as orange canvas was selected to implement the machine learning protocols based on the time series air quality data and the meteorological characteristics. Orange canvas was chosen because it was designed specifically for data mining, and it is user friendly with very interactive visualizations.

2.2.3 Validation and Verification of Machine Learning

Accuracy of all ML models was checked using evaluation criteria such the confusion matrix (Liu et al., 2017), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) (Sammut and Webb, 2011). These metrics were used to summarize and assess the ML model's quality. The cells of this two-dimensional matrix are labelled as either true positives (TP), false positives (FP), true negatives (TN), or false negatives (FN), with the actual class of an item serving as the first dimension and the class assigned by the classifier serving as the second. Precision metrics such as specificity (SP), sensitivity (SS), positive predictive value (PPV), and negative predictive value (NPV) were all calculated using the confusion matrix.

2.3 Sample and Sampling Technique

Air Quality Monitor 65 (AQM 65) is an outdoor weather-proof integrated monitoring station that measures up to 20 gaseous air pollutants, particulate matter, and meteorological parameters simultaneously and continuously. Diagram describing AQM 65 is presented in Figure 3.2. Three (3) AQM 65 stations were strategically located at different monitoring locations as described in the study area above. To ensure consistency of sampling within the breathing zone and allow for data comparison among the three (3) monitoring stations, a height (above the ground level) of 1.8 m for gas inlet, 2.1 m for particulate matter inlet, 2.1 m for weather station and 1.7 m for solar meter was selected and adopted (AS/NZS 3580.1:2016)

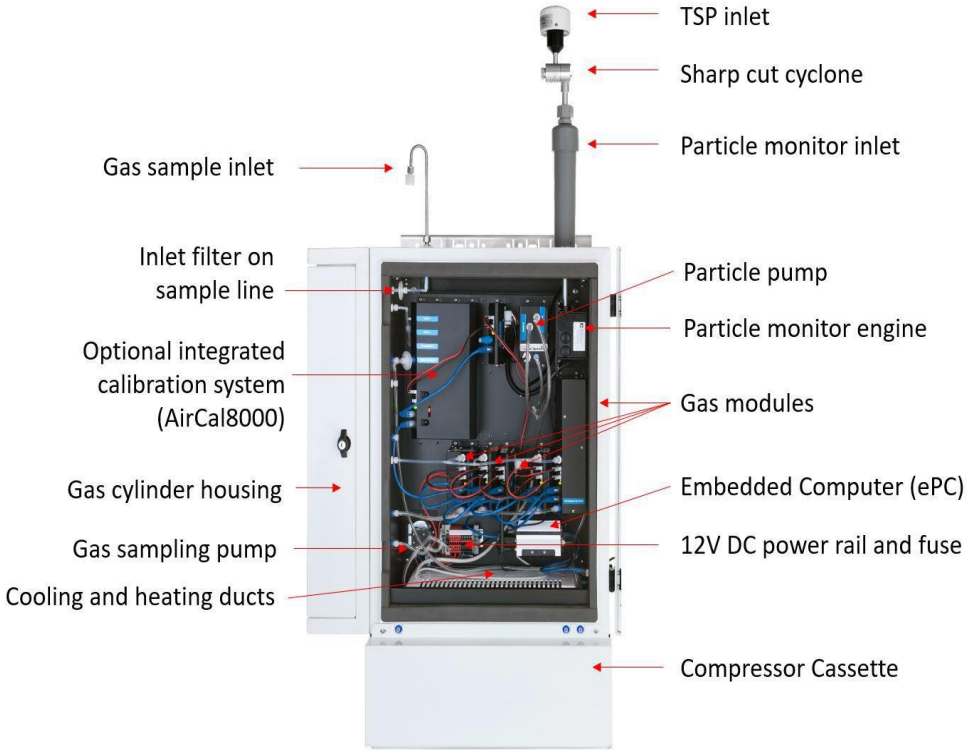

Figure2.3: Description of AQM 65

2.4 Method of Data Collection

The air sample collected from the AQM 65's gas and particulate matter inlets was analyzed by the modules housed in the equipment cabin. Modules determine CO_2, VOC, and PM_{10} concentrations and provide the findings in the proper units. The AQM 65 monitoring station had Vaisala weather meters for measuring wind speed (WS), temperature (T), and a pyranometer for measuring solar radiation (SR). Three (3) monitoring stations collected the data, which was subsequently sent to cloud plus and downloaded for analysis. All three (3) AQM 65 monitoring stations were calibrated on 07/08/2019 using NIST and ISO traceable test method 9722-1-6100 (ISO).

2.5 Method of Data Analysis

Initially, the air pollutant and meteorological data were described using descriptive statistics. The data were further analyzed using the Spearman correlation approach. additionally, correlation matrix was generated for all the parameters in question. Finally, the associated parameters were subjected to Machine Learning for prediction which was validated using Orange Canvas and Weka software respectively.

3. RESULTS AND DISCUSSIONS

3.1 Descriptive Statistics Results

Daily meteorological data including wind speed, air temperature and solar radiation for the same period was also collected. The data was pre-processed using exponential data smoothing (because of high correlation between the pre-processed and the raw experimental data) in time-series via XLSTAT add-in in MS Excel 2019 application. The purpose of the smoothing was to remove outliers and prepare data for machine learning classification. After smoothening, the data was organized, processed and presented as mean, standard deviation, skewness, minimum and maximum values in Table 3.1- 3.3 below. Machine leaning and Random Forest pictorial models were developed to aid in the visual description of pollutant dispersion at varying location within the study area.

Table 2.1: Descriptive statistics of daily air pollutants and meteorological parameters obtained from STATION A

Parameter	Unit	Mean	St. Dev.	Skewness	Range	Min.	Max.
CO_2	ppm	315.8279	82.9515	-0.1791	244.6044	187.1623	431.7667
VOC	ppm	0.0798	0.0378	1.3941	0.2472	0.0235	0.2707
PM_{10}	µg/m³	26.5671	28.5990	2.0014	133.3338	4.8623	138.1962
WS	m/s	0.5723	0.1501	0.6371	0.7266	0.2948	1.0213
Air Temp	°C	27.4927	1.2482	0.0282	5.4920	24.8922	30.3843
Solar Rad.	W/m²	106.8467	32.0178	0.6098	118.1471	63.9829	182.1300

WS – wind speed, WD, Solar Rad. – solar radiation

Table 3.2: Descriptive statistics of daily air pollutants and meteorological parameters obtained from STATION B

Parameter	Unit	Mean	St. Dev.	Skewness	Range	Min.	Max.
CO_2	Ppm	377.4674	26.3246	0.8411	180.6301	316.3699	497.0000
VOC	Ppm	0.4050	1.2385	7.1280	12.8366	0.0000	12.8366
PM_{10}	µg/m³	18.8740	22.9918	2.3243	117.6541	0.4400	118.0941
WS	m/s	0.8270	0.2170	2.2439	2.0262	0.4738	2.5000
Air Temp	°C	27.0324	1.2699	0.0074	5.3421	24.5206	29.8628
Solar Rad.	W/m²	104.6396	27.5410	0.5225	165.3684	11.2000	176.5684

Table 3.3: Descriptive statistics of daily air pollutants and meteorological parameters obtained from STATION C

Parameter	Unit	Mean	St. Dev.	Skewness	Range	Min.	Max.
CO_2	Ppm	158.9154	128.7514	0.3937	482.9976	0.0024	483.0000
VOC	Ppm	0.1081	0.0346	0.2854	0.2204	0.0000	0.2204
PM_{10}	µg/m³	489.2392	3129.4836	8.8811	38573.1634	0.4000	38573.5634
WS	m/s	0.6380	0.2672	5.5591	3.1386	0.3114	3.4500
Air Temp	°C	27.0655	1.2271	0.1035	5.4599	24.5288	29.9887
Solar Rad.	W/m²	96.4245	29.3290	0.6285	164.4920	6.8000	171.2920

WS – wind speed, and Solar Rad. – solar radiation

3.2 Correlation Matrix between Air Pollutants and Meteorological Data

MS Excel 2019 version was used to compute the correlation matrix via the data analysis tool (Tables 3.4 – 3.6). The purpose of this computation was to find the strength of relationship between the different pollutants as well as pollutants versus meteorological data to generate mathematical model for predicting pollutants concentration.

Table 34: Correlation matrix for STATION A using the daily data

	CO_2 (ppm)	VOC (ppm)	PM_{10} (µg/m³)	WS (m/s)	AIR TEMP (°C)	Solar Rad. (W/m²)
CO_2 (ppm)	1.0000					
VOC (ppm)	0.4207	1.0000				
PM_{10} (µg/m³)	0.3597	0.0384	1.0000			
WS (m/s)	-0.2086	-0.3660	0.0397	1.0000		
AIR TEMP (°C)	0.3557	0.2377	0.4922	-0.2342	1.0000	
Solar Rad. (W/m²)	**0.7185**	0.2459	**0.6979**	0.0491	**0.6621**	1.0000

Table 3.5: Correlation matrix for STATION B using the daily data

	CO_2 (ppm)	VOC (ppm)	PM10 (µg/m³)	WS (m/s)	AIR TEMP (°C)	Solar Rad. (W/m²)
CO_2 (ppm)	1.0000					
VOC (ppm)	0.2449	1.0000				
PM_{10} (µg/m³)	-0.0497	-0.0887	1.0000			
WS (m/s)	0.2913	-0.1150	0.0706	1.0000		
AIR TEMP (°C)	-0.3866	-0.2401	0.4496	-0.1464	1.0000	
Solar Rad. (W/m²)	-0.1374	-0.1741	**0.6493**	0.0125	**0.6630**	1.0000

Table 3.6: Correlation matrix for STATION C using the daily data

	CO_2 (ppm)	VOC (ppm)	PM_{10} (µg/m³)	WS (m/s)	AIR TEMP (°C)	Solar Rad. (W/m²)
CO_2 (ppm)	1.0000					
VOC (ppm)	-0.5242	1.0000				
PM_{10} (µg/m³)	0.1303	-0.0841	1.0000			
WS (m/s)	0.3503	-0.3768	-0.0451	1.0000		
AIR TEMP (°C)	0.4129	-0.5561	0.2446	-0.1323	1.0000	
Solar.Rad. (W/m²)	**0.7061**	-0.4641	0.1042	-0.0715	**0.6858**	1.0000

From the Table 3.4 it was observed that solar radiation strongly correlates with Co2 and Air temperature, In Table 3.5 solar radiation strongly correlate with Co2, PM10 and air temperature while in Table 3.6 solar radiation strongly correlate with CO2 and air temperature.

3.3 Machine Learning Model Training and Testing Results

Meteorological and pollutants concentration data from all three locations were utilized as inputs. Using equation 2.1 from section 2, AQI was calculated to account for the Machine Learning (ML) classifications. To acquire qualitative data as the target variable for ML classifications such Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM), it was categorized as good, hazardous, moderate, unhealthy, harmful for sensitive groups, and extremely unhealthy. Therefore, AQI ranges were used in the ML categorization of the pollution. We filtered the data and eliminated the outliers to get down to 1237 occurrences that would be utilized for training. There was a 75% success rate in training, and a 25% success rate in testing using the data.

3.3.1 Prediction of CO2

In prediction of CO_2 using ML algorithm, RF model showed very high level of classification accuracy (0.977) than the other models (Table 3.7). Therefore, the RF model was utilized to predict the concentration and dispersion of CO_2 in the atmosphere. Figure 4.14 shows a very strong and positive correlation between predicted RF model and experimental values based on AQI level.

Table 3.7 Evaluation Results for CO_2 Machine Learning Classification

Model	AUC	CA	F1	Precision
Neural Network	0.500	0.534	0.372	0.285
Random Forest	0.998	0.977	0.978	0.980
SVM	0.613	0.534	0.372	0.285
Logistic Regression	0.310	0.534	0.372	0.285
Naïve Bayes	0.954	0.748	0.771	0.823

CA=classification accuracy; AUC=area under the curve; F1=balanced weighted average of precision and recall

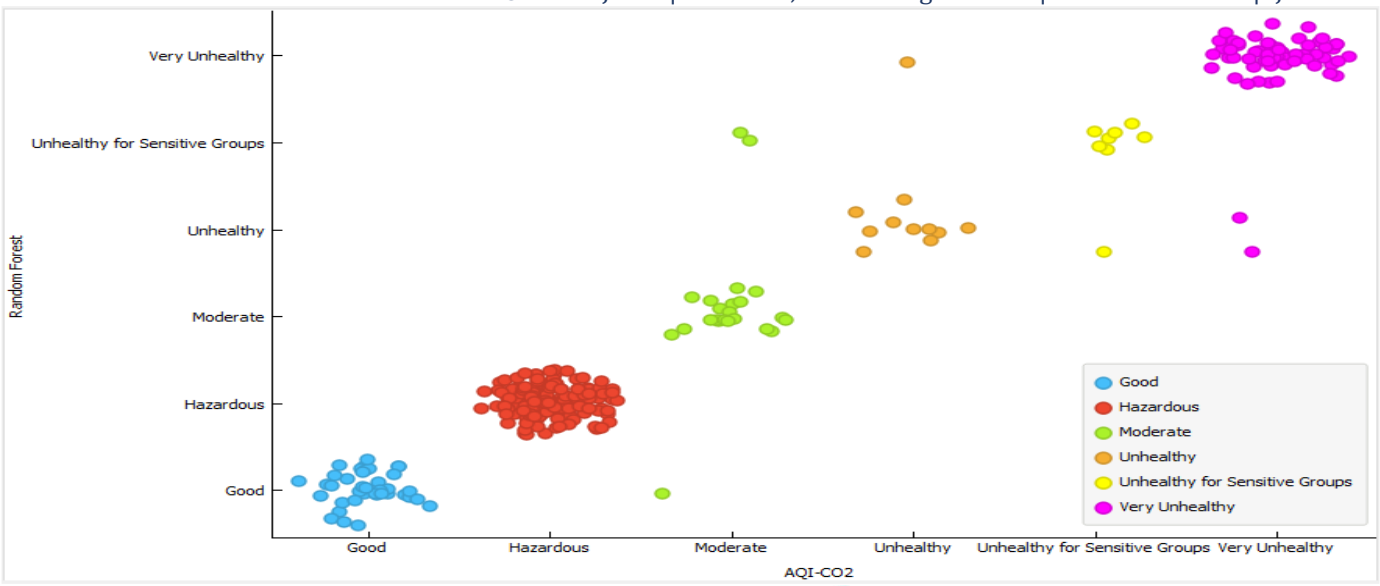

Figure 3.1RF predicted versus experimental CO_2 using Orange Canvas

In Figure 3.2 we can infer that increasing solar radiation increases the dispersion of CO_2 in the atmosphere.

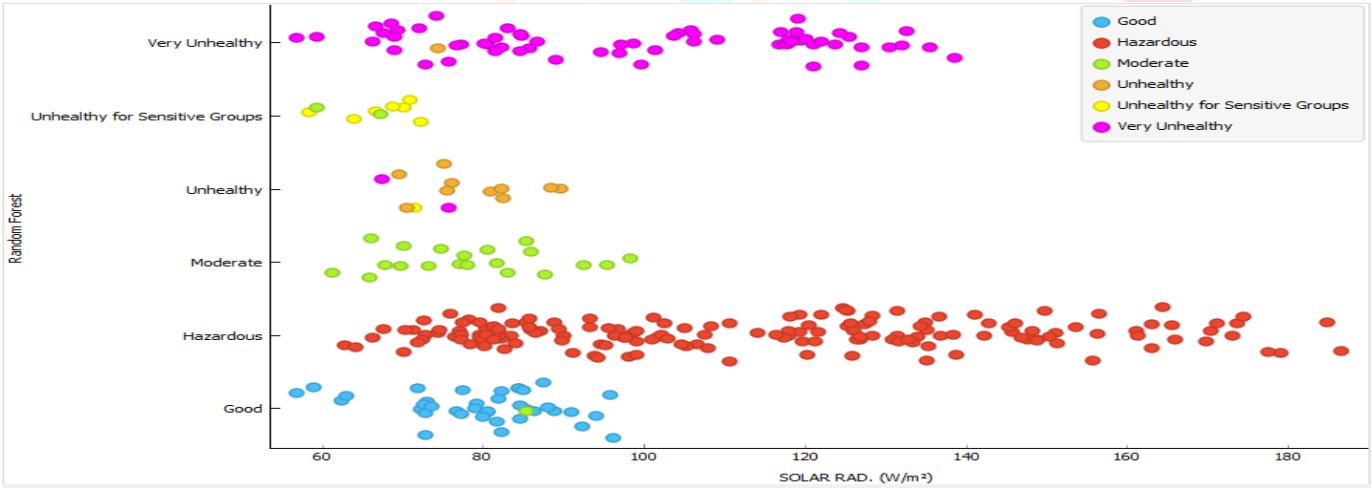

Figure 3.2: Effect of solar radiation on CO_2 distribution across study location

Based on RF model prediction, it was observed that STATION A is hazardous, STATION C looks very unhealthy and STATION B is mixed with unhealthy, and good condition based on AQI level (Figure 3.3). Since, CO_2 is not a pollutant of health concern, however, the implication of this outcome is greenhouse effect due to high levels.

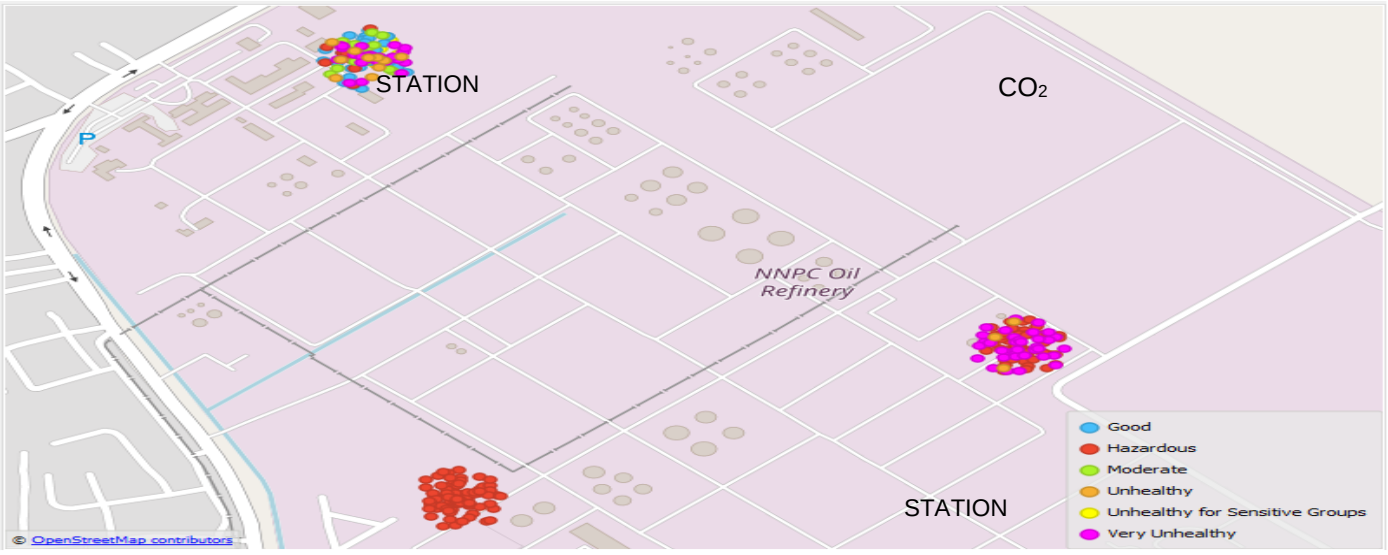

Figure 3.3: Random Forest prediction of CO_2 distribution across study locations

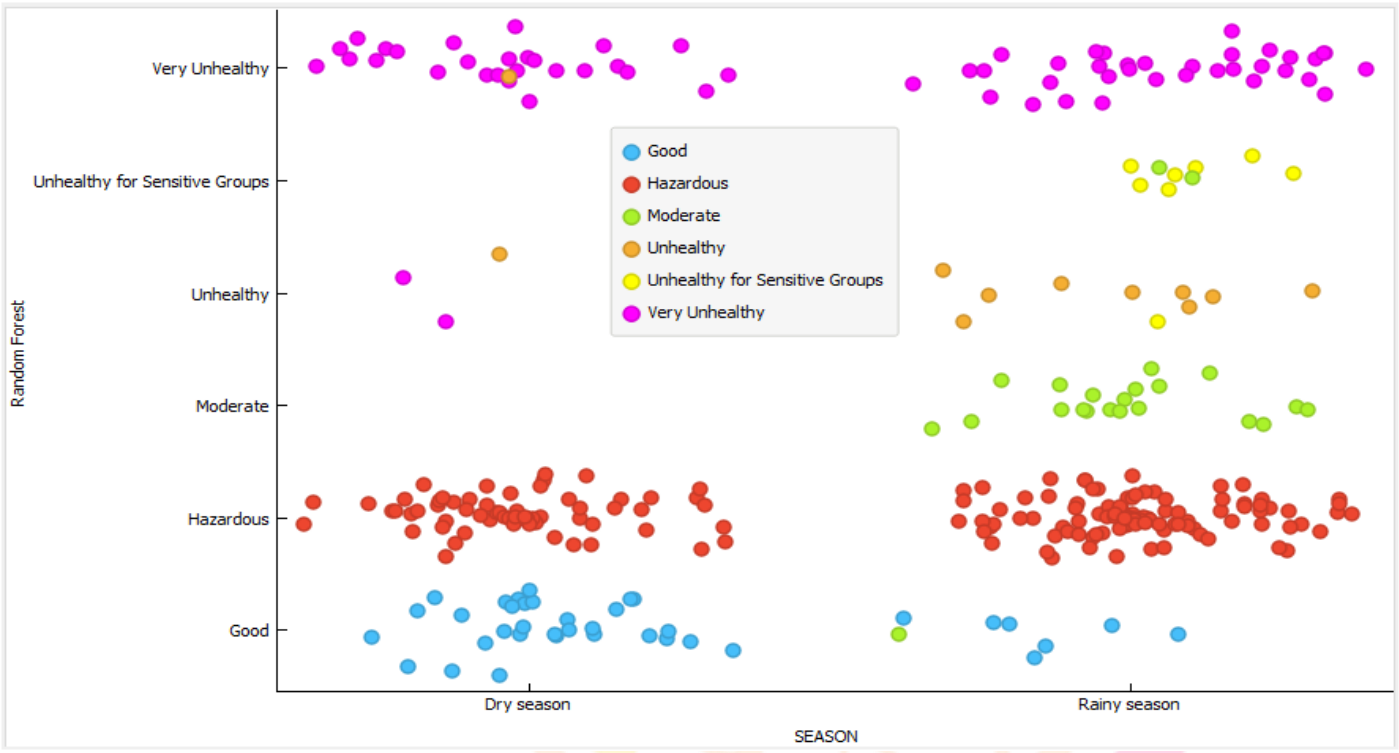

Figure 3.4: Seasonal variation in CO_2 distribution across study location

In Figure 3.5, the effect of wind speed and temperature in the distribution of CO_2 is shown. it was observed that at wind above 0.6m/s pollutant dispersed faster.

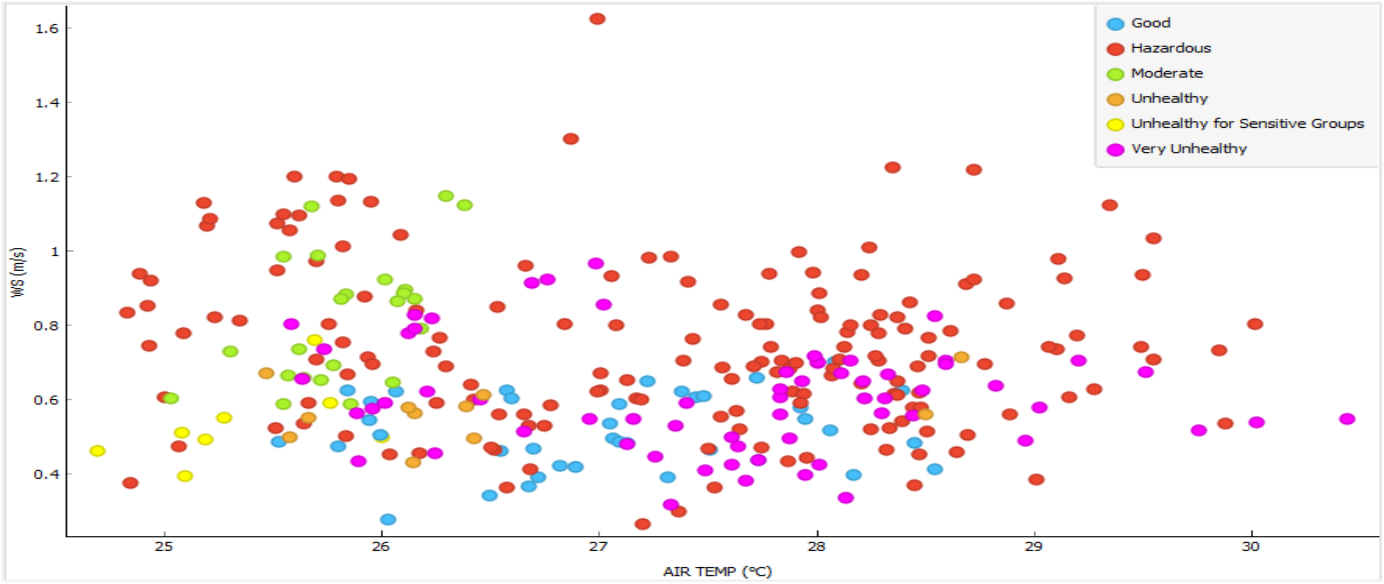

Figure 3.5: Effect of WS and Air Temp on CO_2 distribution across study location

Based on the confusion matrix, the RF model using WEKA software classified 696 out of 696 as hazardous, 286 out of 286 as very unhealthy, 42 out of 42 as unhealthy, 34 out of 34 as unhealthy for sensitive groups, 66 out of 66 as moderate, and 113 out of 113 as good correctly. The correctly classified instances are 100% (Figure 3.6). Similarly, in Figure 4.21 the NB model classified 551 out 696 as hazardous, 247 out of 286 as very unhealthy, 8 out 42 as unhealthy, 27 out of 34 as unhealthy for sensitive groups, 55 out 66 as moderate and 105 out 113 as good correctly. The correctly classified instances are 80.3% (Figure 3.7).

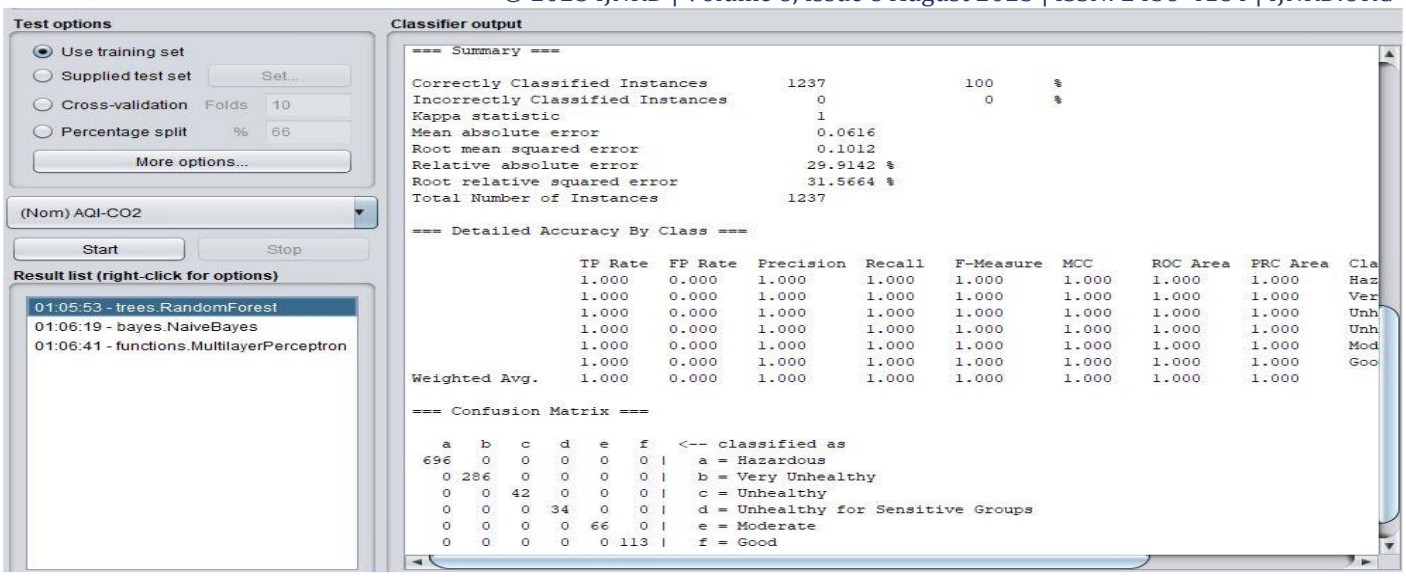

Figure 3.6: RF confusion matrix for CO_2

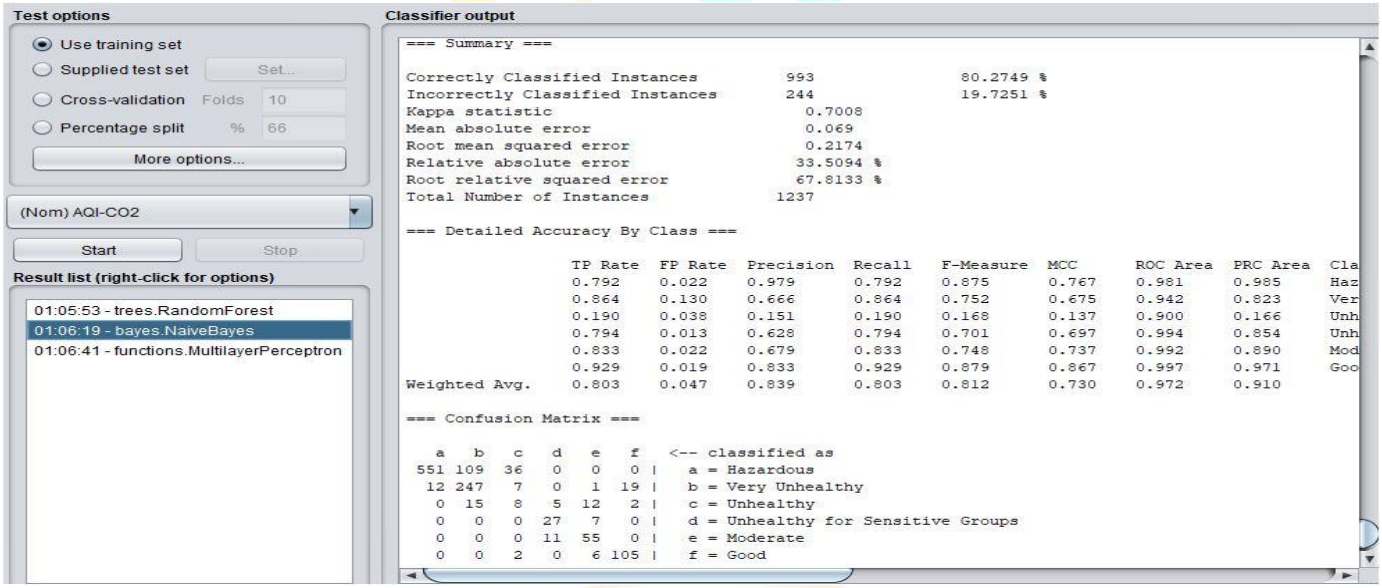

Figure 3.7: NB confusion matrix for CO_2

3.3.2 Prediction of PM10

In the prediction of PM_{10}, From Table 3.8 RF model showed a higher classification accuracy of 0.939 which is more accurate than the other four algorithms used, this agrees with the findings of Masih (2019) who applied ensemble learning techniques to model the atmospheric concentration of SO_2. Therefore, the RF model is preferred for prediction of PM_{10} based on AQI as target (Table 3.8). Predicted RF model versus experimental AQI is shown in Figure 3.8. From Figure 3.8, there is a strong and positive correlation between the RF model and experimental data. However, the seasonal variation on the scatter plot in Figure 3.9 shows that PM_{10} concentration is high during the dry season and low during rainy season.

Table3.8: Evaluation Results for PM_{10} Machine Learning Classification

Model	AUC	CA	F1	Precision
Neural Network	0.500	0.032	0.002	0.001
Random Forest	0.992	0.939	0.939	0.945
SVM	0.855	0.803	0.715	0.644
Logistic Regression	0.926	0.803	0.715	0.644
Naïve Bayes	0.956	0.767	0.804	0.863

CA=classification accuracy; AUC=area under the curve; F1=balanced weighted average of precision and recall

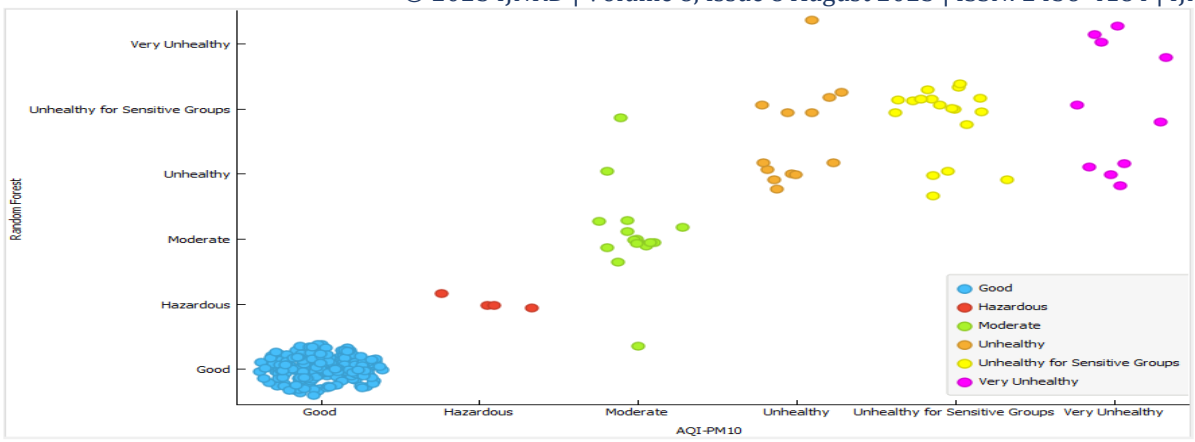

Figure3.8: RF Predicted versus Experimental AQI for PM_{10}

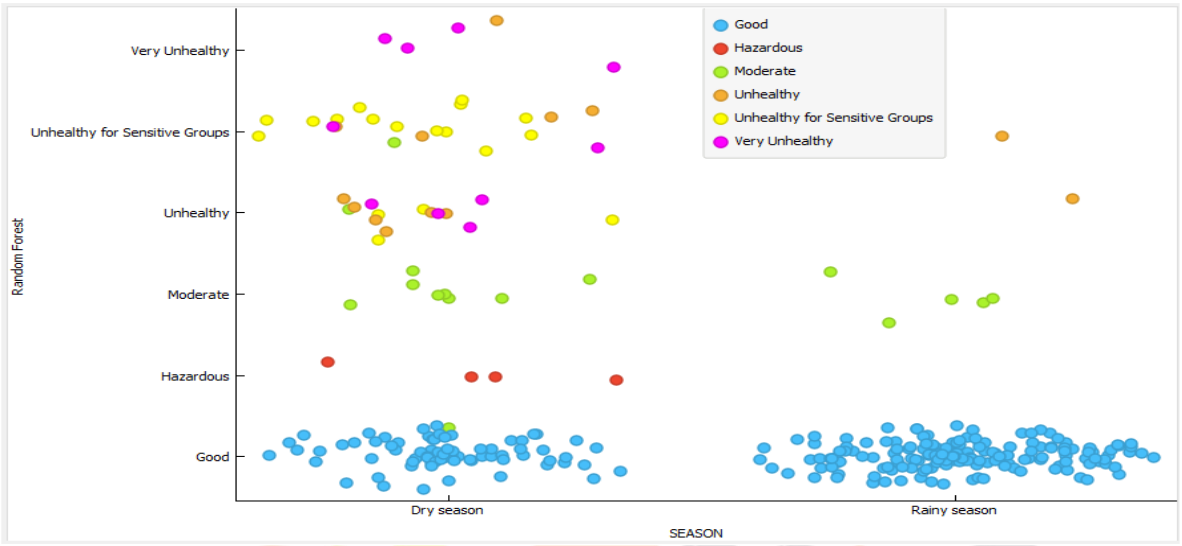

Figure3.9: Seasonal variation of PM_{10} distribution across different stations using RF prediction

In Figure 3.10 it was observed that STATION C has less PM_{10} pollution than A and B following the RF model prediction. It was observed that as solar radiation increases, the dispersion of PM_{10} also increases (Figure 3.11). Therefore, solar radiation is an important parameter in the pollutant dispersion process. Similarly, in Figure 2.12 there is high spread of PM_{10} as the air temperature increases.

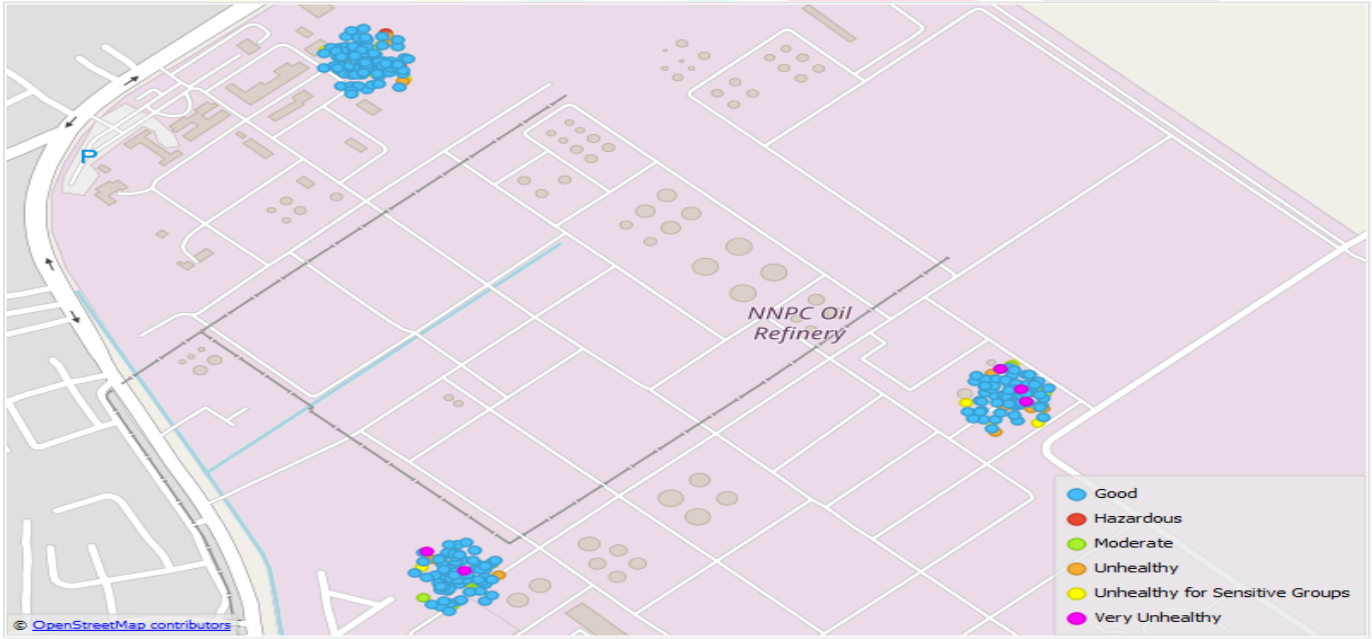

Figure 3.10: RF Prediction of PM_{10} distribution across different stations

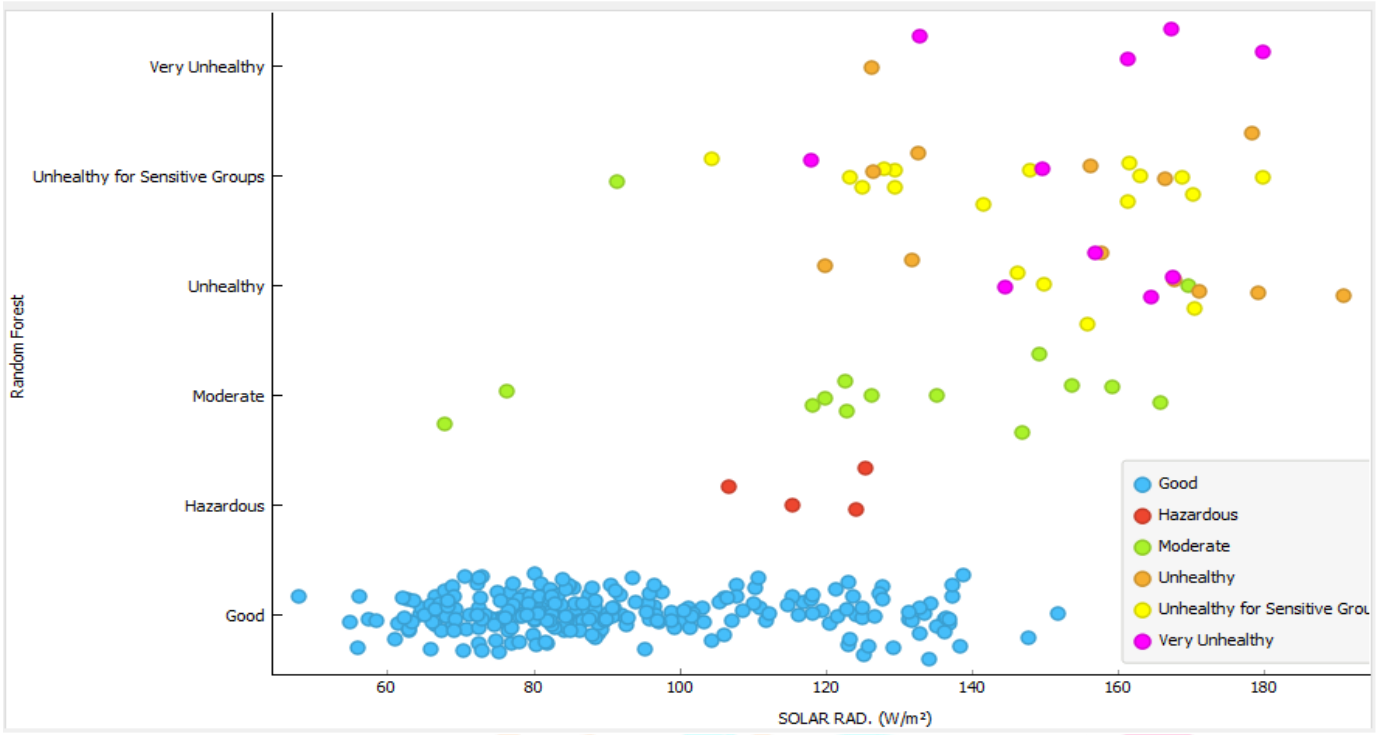

Figure3.11: RF prediction of the effect of solar radiation on PM_{10}

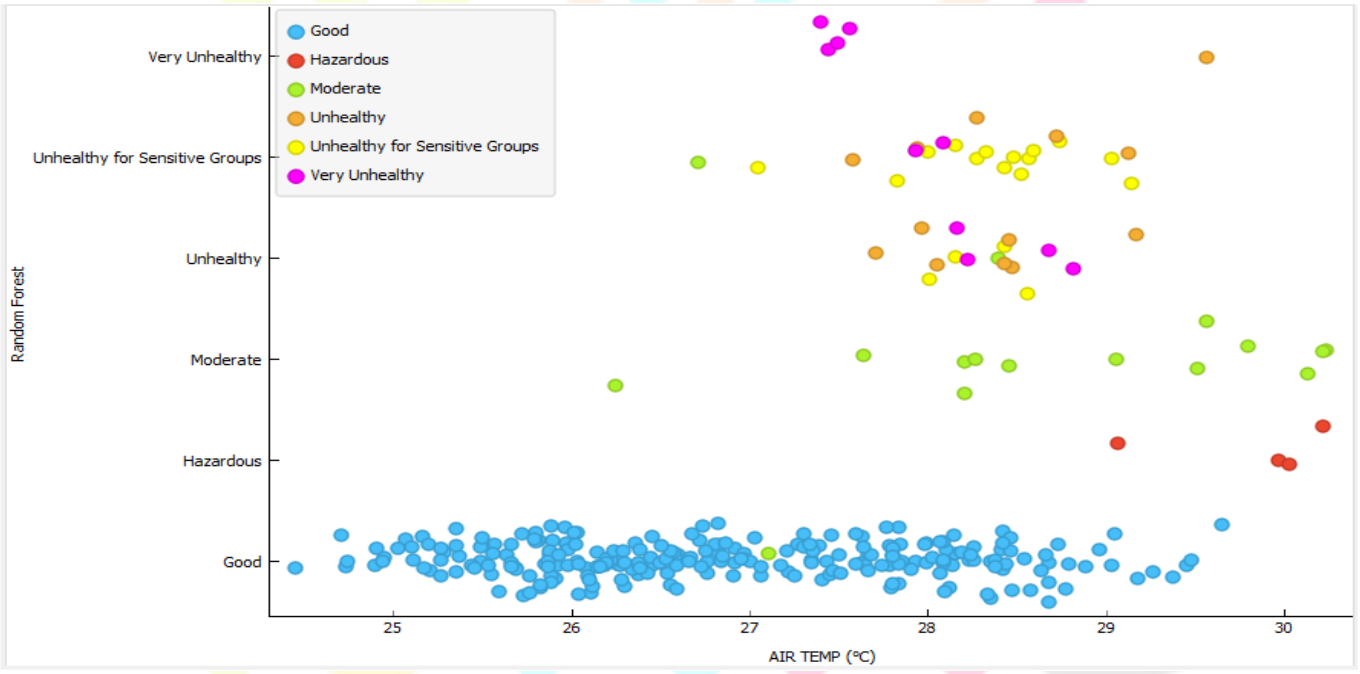

Figure 3.12: RF prediction of the effect of air temperature on PM_{10}

Based on the confusion matrix, the RF model using Orange Canvas software classified 248 out of 248 instances as Good, 4 out of 4 as Hazardous, 13 out of 16 as Moderate, 7 out of 13 as unhealthy, 14 out of 18 as unhealthy for sensitive groups, and 4 out of 10 for very unhealthy correctly (Figure 3.13). According to the proportion of the classifications on the test data, the RF was ahead of all other techniques in terms of accuracy. Similarly, Figure 3.14 shows the NB confusion matrix for PM_{10} using orange canvas, classified 210 out of 248 instances as Good, 4 out of 4 as hazardous, 13 out of 16 as moderate, 6 out of 13 as unhealthy, 8 out of 18 as unhealthy for sensitive groups, and 2 out 10 as very unhealthy. The correctly classified percentages are found in the confusion matrix using WEKA software, this is similar to the findings of Saravi et al. (2019) who used WEKA to predict flood events. Saravi highlighted the advantages of WEKA as it summarizes result and ease of interpretation.

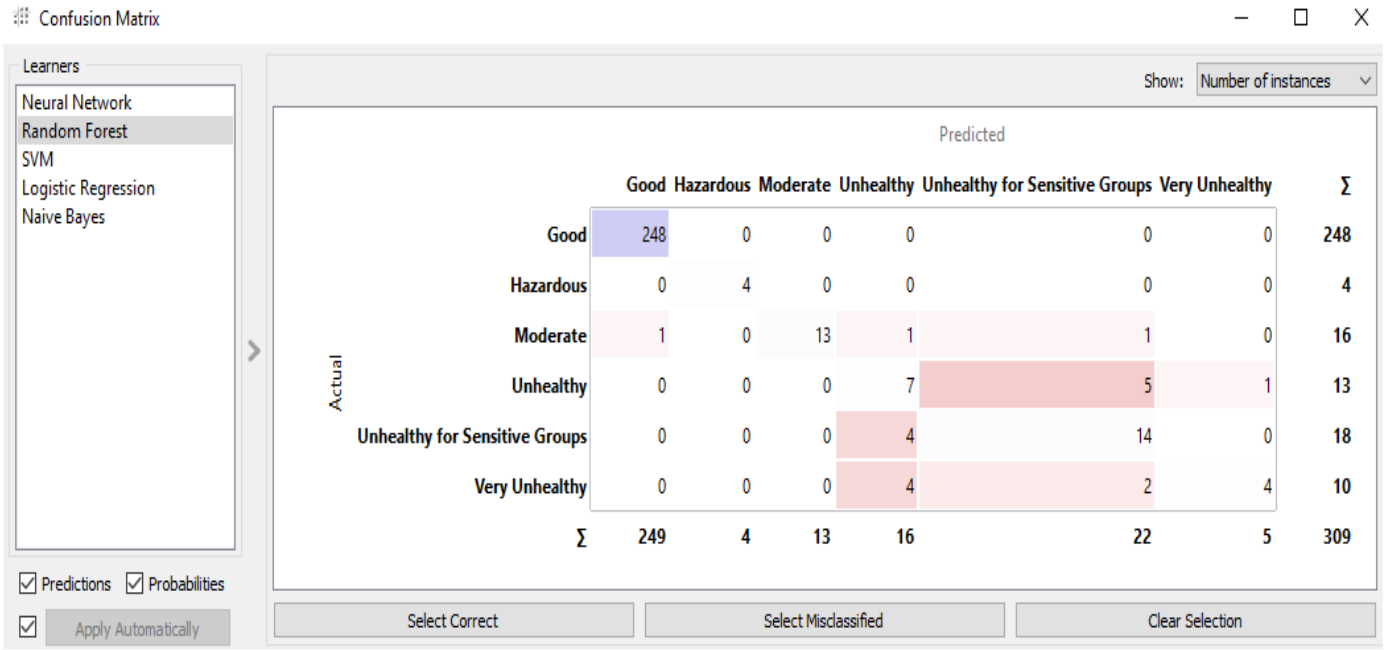

Figure 3.13: RF confusion matrix for PM_{10} using Orange Canvas

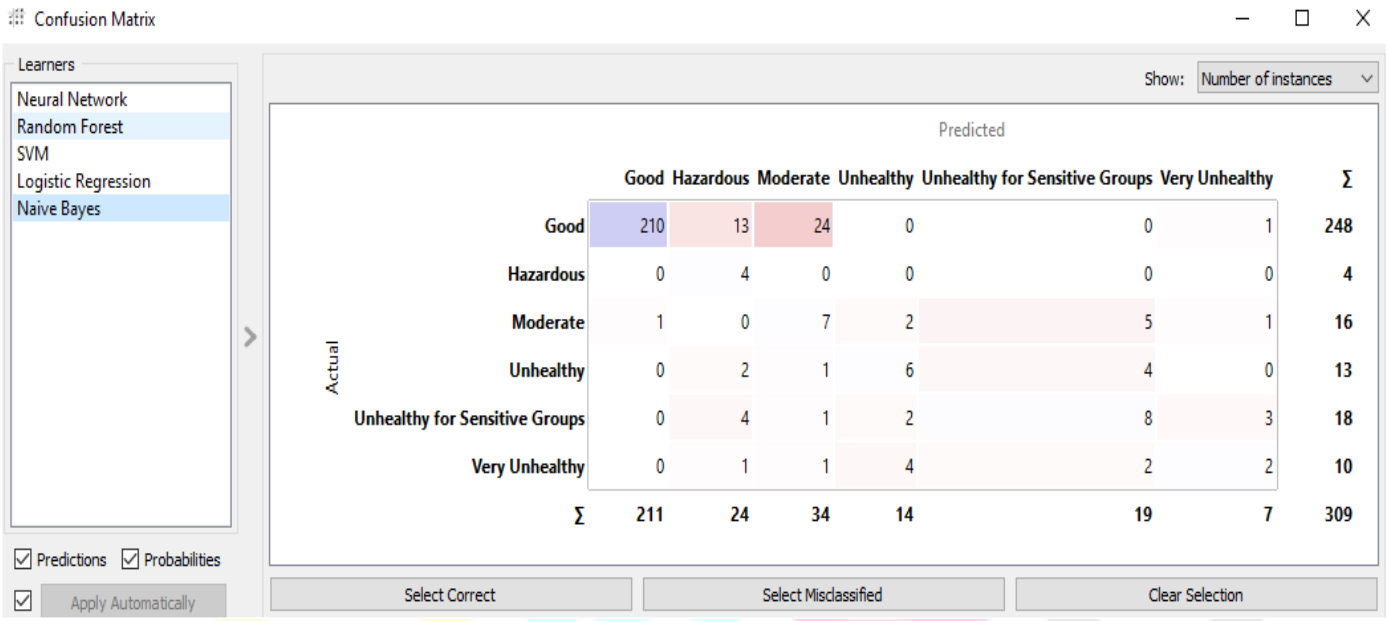

Figure 3.14: NB confusion matrix for PM_{10} using Orange Canvas

Following the confusion matrix, the RF model using WEKA software classified 86 out of 86 as unhealthy for sensitive groups, 54 out of 54 as unhealthy, 40 out of 40 as very unhealthy, 83 out of 83 as moderate, 946 out of 946 as good, and 28 out of 28 as hazardous correctly. The correctly classified instances in total 100% (Figure 3.15). Similarly, in Figure 3.16 showing the NB confusion matrix the WEKA software classified 63 out of 63 as unhealthy for sensitive groups, 37 out of 54 as unhealthy, 25 out of 40 as very unhealthy, 57 out of 83 as moderate, 837 out of 1016 as good and 26 out of 28 as hazardous. The correctly classified instance is 84.5%.

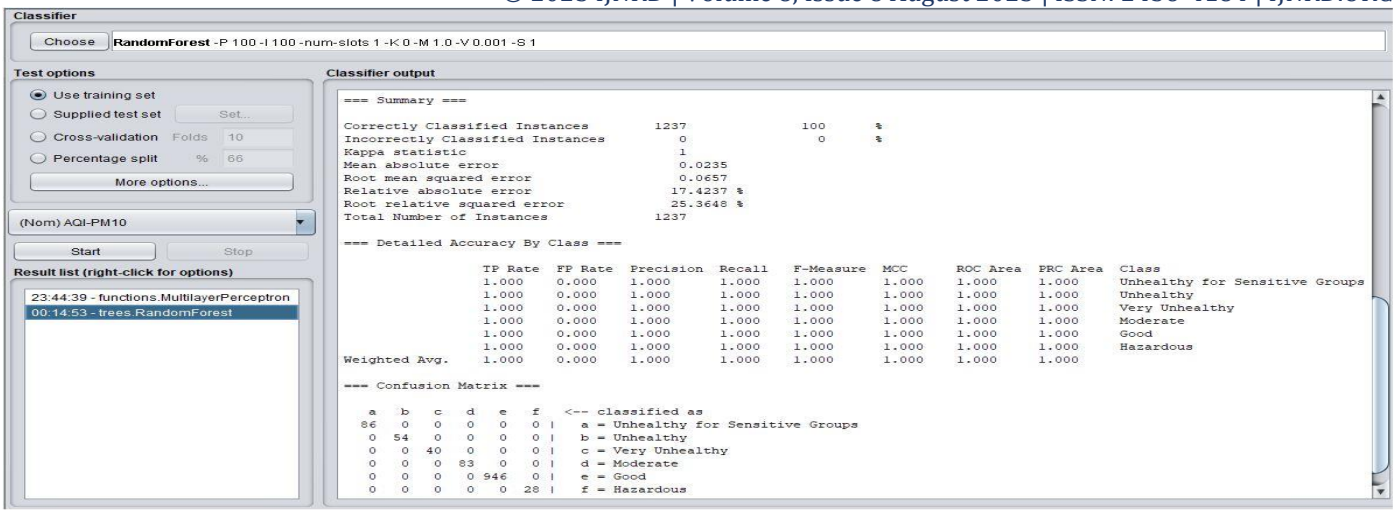

Figure 3.15: RF confusion matrix for PM_{10} using WEKA

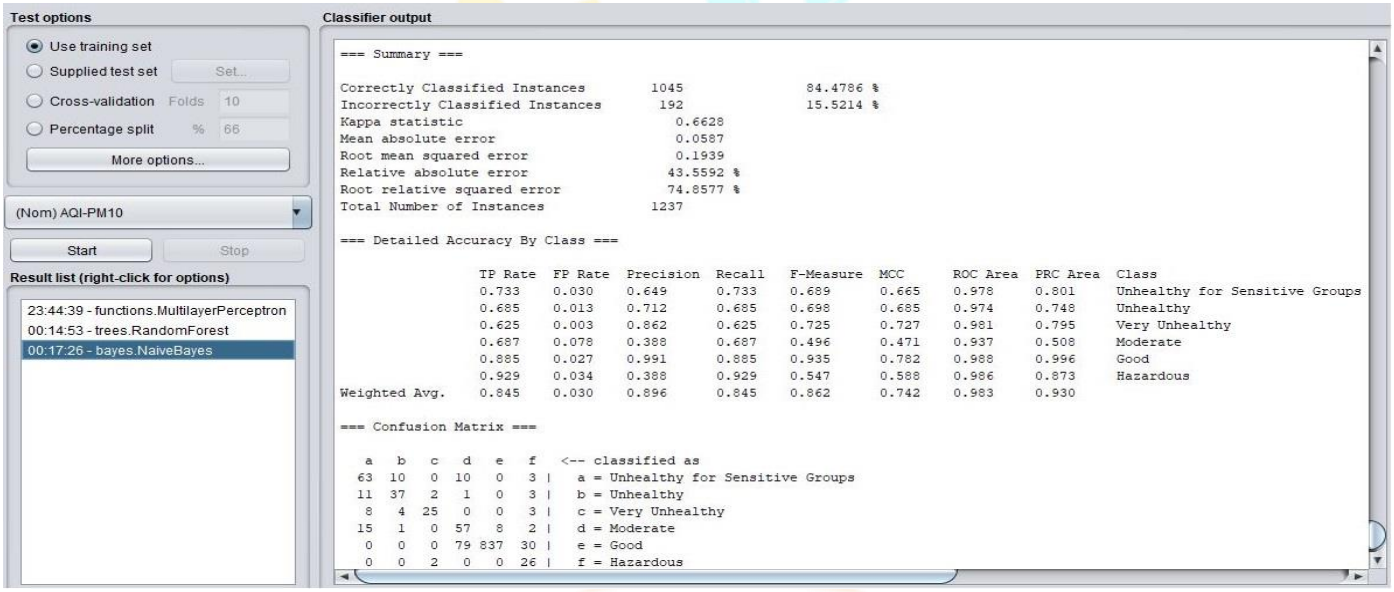

Figure 3.16: NB confusion matrix for PM_{10} using WEKA

3.3.3 Prediction of VOC

The evaluation result for VOC in Table 4.14 shows that RF model classification is more accurate than the other four models with classification accuracy of 0.955. In addition, the RF model prediction versus experimental values in Figure 3.17 shows a good correlation. Concentration of VOC was observed to be Good for all the monitoring stations except STATION B which has combination of Good, Moderate and Hazardous categories of AQI as shown in Figure 3.18.

Table 3.9: Evaluation Results for VOC

Model	AUC	CA	F1	Precision
Neural Network	0.500	0.000	0.000	0.000
Random Forest	0.944	0.955	0.942	0.955
SVM	0.551	0.942	0.913	0.942
Logistic Regression	0.514	0.942	0.913	0.942
Naïve Bayes	0.899	0.489	0.644	0.489

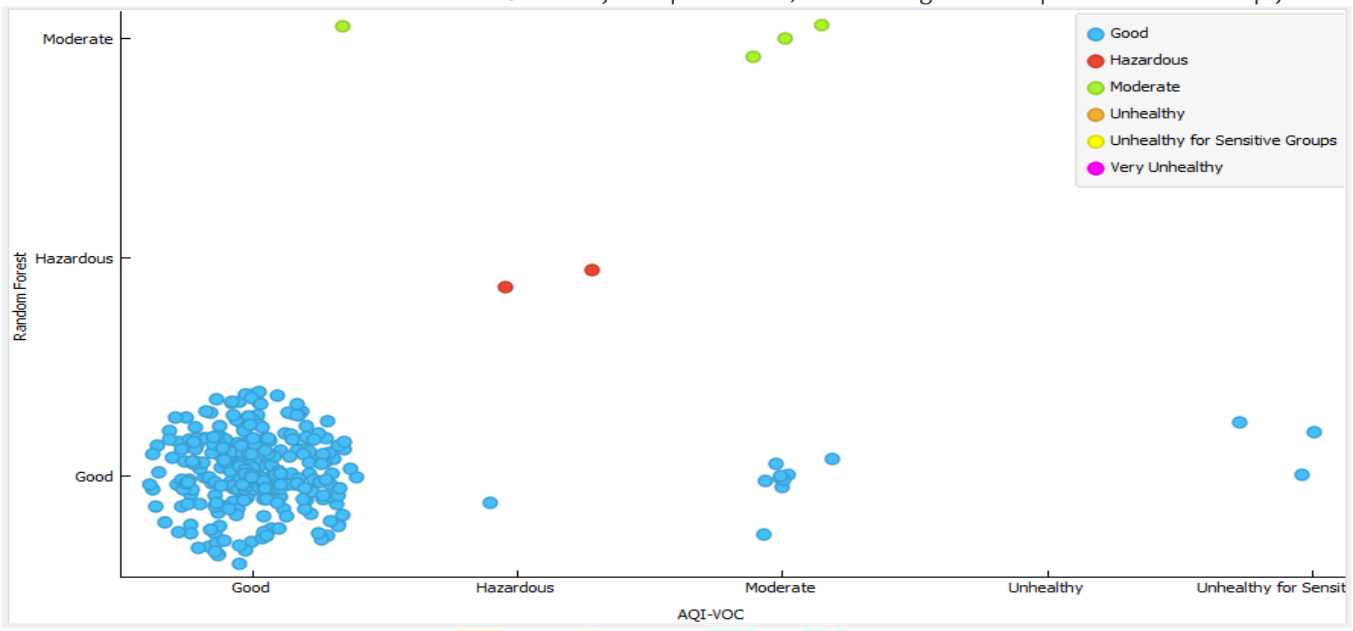

Figure 3.17: RF Predicted vs Experimental AQI for VOC

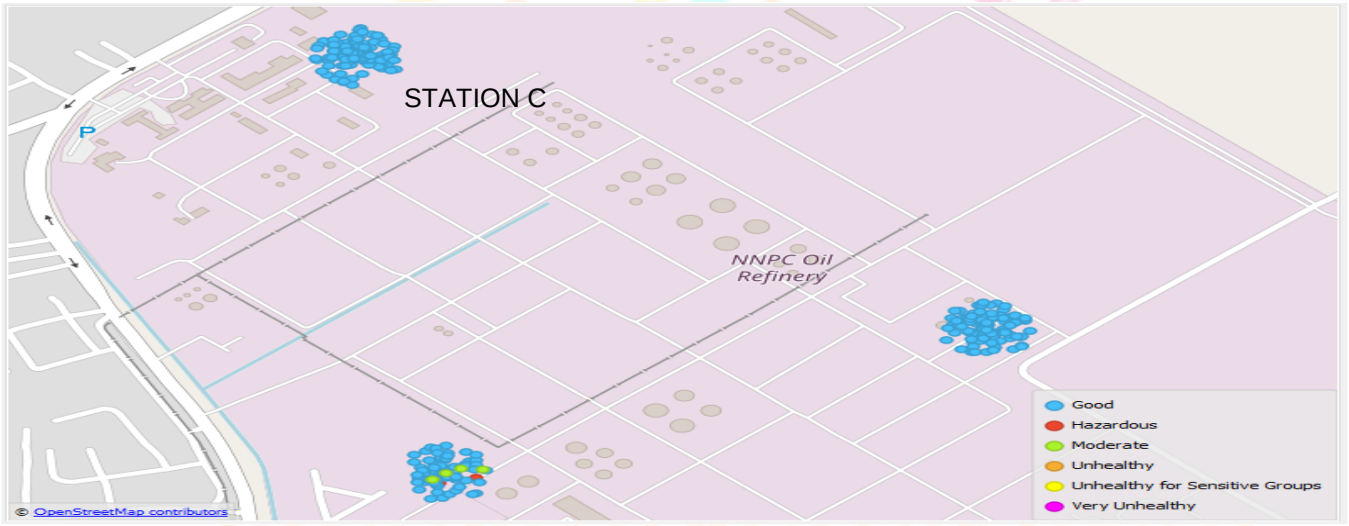

Figure 3.18: RF Prediction of VOC distribution across different stations

Based on the confusion matrix for VOC prediction, the RF model using Orange Canvas software classified 290 out of 291 instances as Good, 2 out of 3 as Hazardous, 3 out of 12 as Moderate, 0 out of 0 as unhealthy, 0 out of 3 as unhealthy for sensitive groups, and 0 out of 0 for very unhealthy correctly (Figure 3.19). According to the proportion of the classifications on the test data, the RF was ahead of all other techniques. Similarly, the NB confusion matrix classified 147 out of 291 as good, 2 out of 3 as hazardous, 2 out of 12 as moderate, 0 out of 3 as unhealthy for sensitive groups and 0 out of 0 as very unhealthy (Figure 3.20).

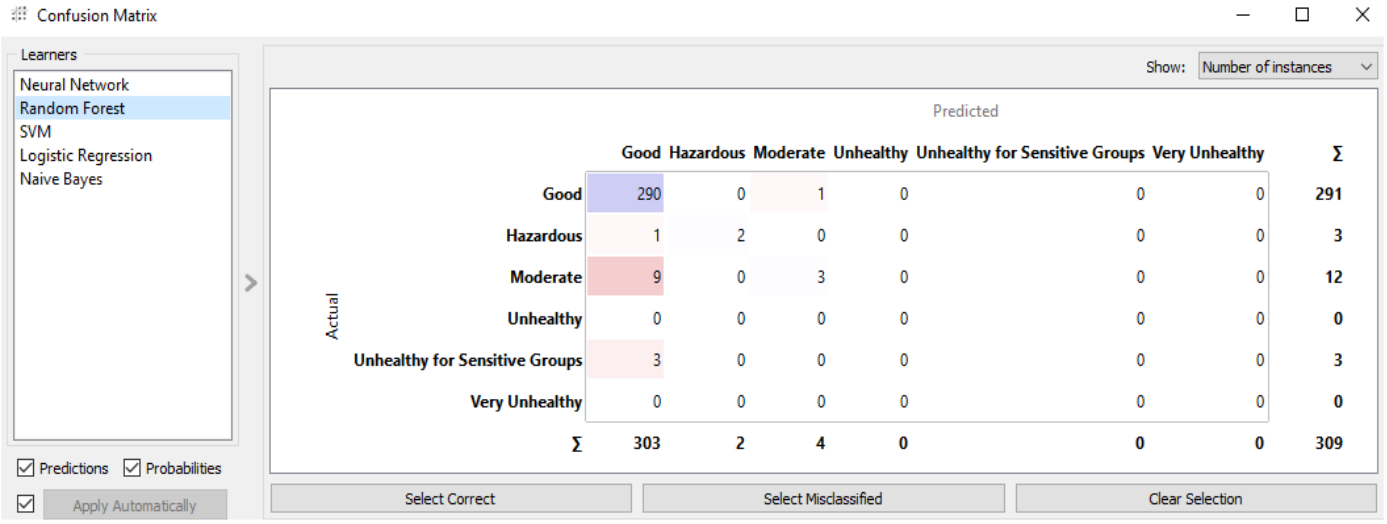

Figure 3.19: RF confusion matrix for VOC using Orange Canvas

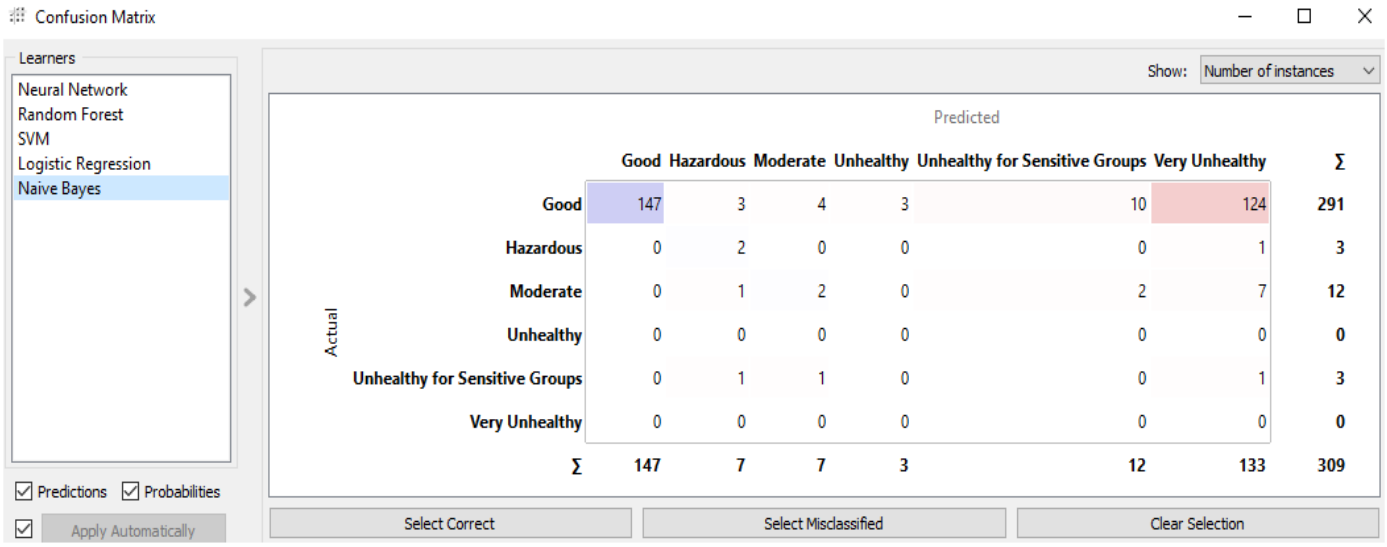

Figure 3.20: NB confusion matrix for VOC using Orange Canvas

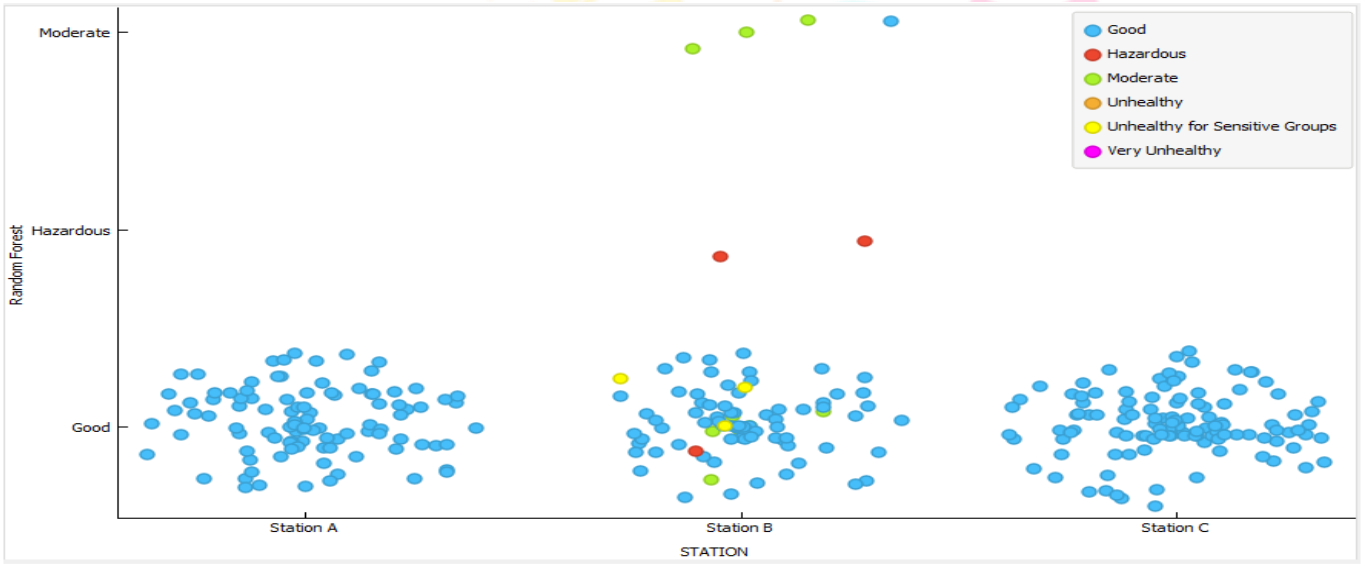

Figure 3.21: RF prediction of VOC distribution across different stations

From the confusion matrix, the RF model using WEKA software classified 1167 out of 1167 as good, 41 out of 41 as moderate, 9 out of 9 as unhealthy for sensitive groups, 6 out of 6 as unhealthy, 2 out of 2 as very unhealthy, and 12 out of 12 as hazardous. The correctly classified instances in total 100% (Figure 3.22). Similarly, in Figure 3.23, NB model classified 961 out 1167 instances as good, 39 out of 41 as moderate, 1 out of 9 as unhealthy for sensitive groups, 0 out of 6 as unhealthy, 0 out of 2 as very unhealthy and 12 out of 12 as hazardous. The correctly classified instance in total 81.9%.

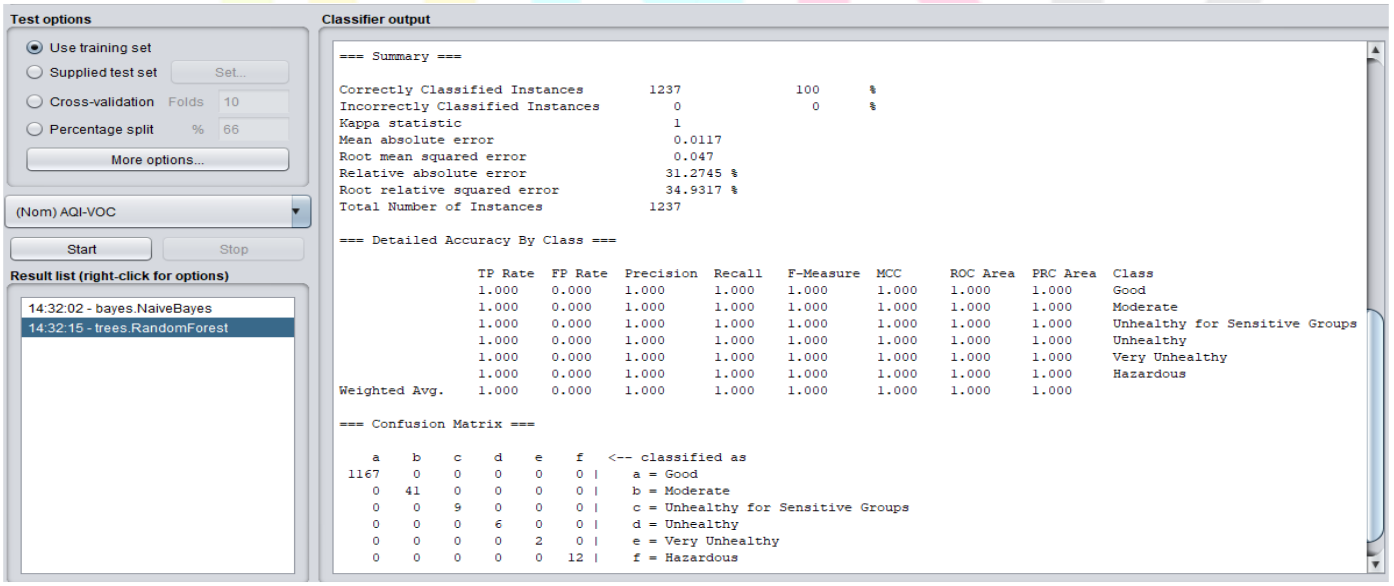

Figure 3.22: RF confusion matrix for VOC using WEKA

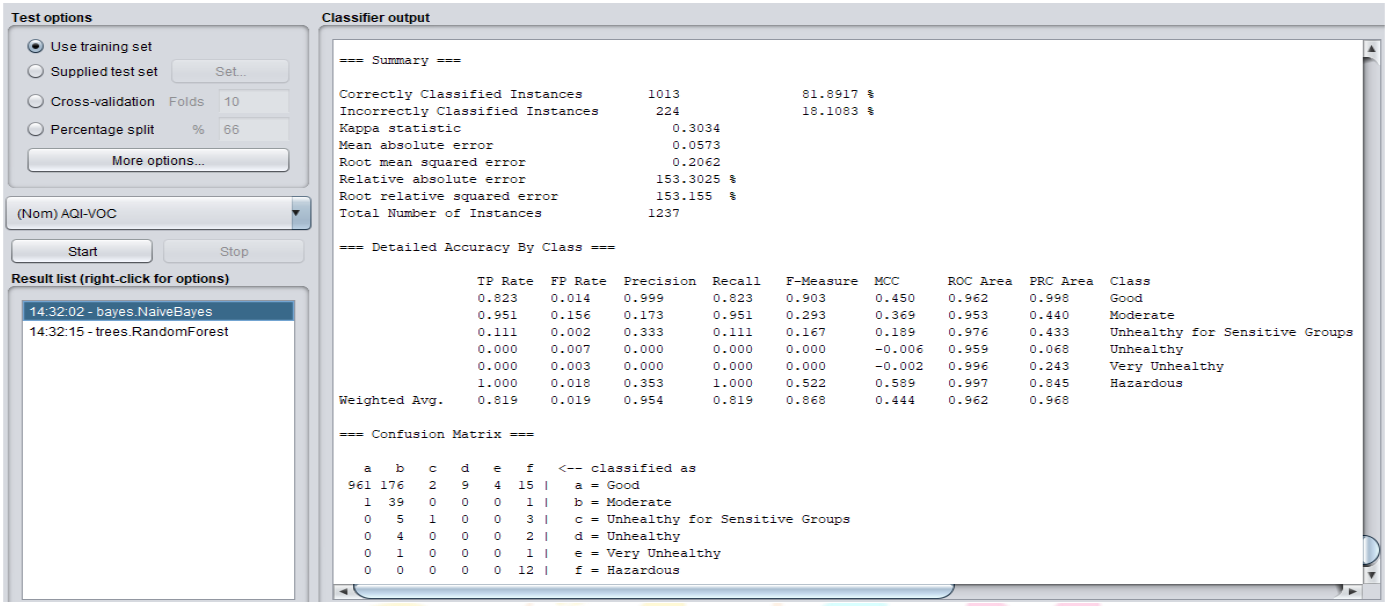

Figure 3.23: NB confusion matrix for VOC using WEKA

4. CONCLUSIONS

Based on this finding, increase in solar radiation increases the dispersion of CO_2 in the atmosphere. Based on CO_2 RF model prediction, it was observed that STATION A is hazardous, STATION C looks very unhealthy and STATION B is mixed with unhealthy, and good condition based on AQI level. Since, CO_2 is not a pollutant of health concern for the purpose of this study, however, a significant implication of this outcome is greenhouse effect due to high levels of CO_2.

There is a strong and positive correlation between the RF model and experimental data. However, the seasonal variation on the scatter plot shows that PM_{10} concentration is high during the dry season. Therefore, the rainy season has low level of PM_{10}. It was observed that as solar radiation increases, the dispersion of PM_{10} also increases. Therefore, solar radiation is an important parameter in the pollutant modelling process. Similarly, it was concluded that there is high spread of PM_{10} as the air temperature increases, also, It was observed that VOC disperses faster as temperature increases

Finally, this work has proven that, the application of ML concept using high quality and accurate data can bring more advances in Nigeria not only for air quality prediction, but any type of environmental monitoring (based on provided data type) to help preparedness, raise awareness and build resilience Environmental Management System, especially in areas more prone to industrial pollution.

REFERENCES

Bin, W. (2008). Analyse the spatial-temporal characteristics of air pollution in China by using air pollution index (API)," M.S. thesis, *Ocean Univ. of China*, Qingdao, China.

Breiman L. (2001) ``Random forests," *Mach. Learn., vol.* 45, no. 1, pp. 5_32.

Chen, Y., Wang, L., F. Li, B. Du, K. K. R. Choo, H. Hassan, and W. Qin (2016). ``Air quality data clustering using EPLS method," *Inf. Fusion*, vol. 36, pp. 225_232

Geer LA, W. J. (2012). Ambient air pollution and term birth weight in Texas from 1998 to 2004. *PubMed.*

Graves, A. (2012). Supervised Sequence Labelling with Recurrent Neural Net- works (Studies in Computational Intelligence), vol. 385. *Springer, 2012*, pp. 5_13.

Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., & Xie, G. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. *International Journal of Environmental Science and Development*, *9*(1), 8–16

Kujaroentavon, K., S. Kiattisin, A. Leelasantitham, and S. Thammaboosadee (2014), ``Air quality classification in Thailand based on decision tree," in *Proc. 7th Biomed. Eng. Int. Conf.*, Fukuoka, pp. 26_28.

Lipton, Z. C., Berkowitz, J., C. Elkan (2015). ``A critical review of recurrent neural networks for sequence learning," *arXiv:1506.00019*. [Online]. Available.

Liu, Jen-Hao, Chen, Yu-Fan, Lin, Tzu-Shiang, (2017). Developed urban air quality monitoring system based on wireless sensor networks. In: Sensing technology (icst), *2011 fifth international conference on. IEEE*. p. 549-554.

Masih, A. (2019). Application of ensemble learning techniques to model the atmospheric concentration of SO_2. *Global Journal of Environmental Science and Management*, *5*(3), 309–318.

Muhammad S. Y., Makhtar M., Rozaimee A., A. Abdul, and A. A. Jamal (2015). ``Classification model for water quality using machine learning techniques," *Int. J. Softw. Eng. Appl.*, vol. 9, no. 6, pp. 45_52.

Niharika, V. M. & Rao, P. S. (2014)., ``A survey on air quality forecasting techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 103_107.

Punia M., P. K. Joshi, and M. C. Porwal (2011). ``Decision tree classification of land use land cover for Delhi, India using IRS-P6 AWiFS data," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5577_5583Sammut and Webb, 2011.

Sanchez, N. P., Saffari, A., Barczyk, S., Coleman, B. K., Naufal, Z., Rabideau, C., & Pacsi, A. P. (2019). Results of three years of ambient air monitoring near a petroleum refinery in Richmond, California, USA. *Atmosphere*, *10*(7).

Saravi, S., Kalawsky, R., Joannou, D., Casado, M. R., Fu, G. and Meng, F. (2019). Use of Artificial Intelligence to Improve Resilience and Preparedness Against Adverse *Flood Events, Water 2019*, 11, 973.

Simpson, I.J.; Marrero, J.E.; Batterman, S.; Meinardi, S.; Barletta, B.; Blake, D.R. (2013). Air quality in the Industrial Heartland of Alberta, Canada and potential impacts on human health. *Atmos. Environ*., 81, 702–709.

Wang, Y., & Kong, T. (2019). Air Quality Predictive Modeling Based on an Improved Decision Tree in a Weather-Smart Grid. *IEEE Access*, *7*, 172892–172901