University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

A Journal of the Faculty of Computing, University of Ibadan, Ibadan, Nigeria

Volume 16 No. 1, January 2026

journals.ui.edu.ng/uijslictr
http://uijslictr.org.ng/
uijslictr@gmail.com

University of Ibadan
Journal of Science and Logics in ICT Research

A Prediction Model for Cardiovascular Health Risk from Air Quality Index of Pollution laden Environment

Adeleke O. and Ayoola O. A.

Department of Computer Science,
University of Ibadan,
Ibadan, Nigeria.
Corresponding email: adfeleke4@gmail.com

Abstract
The cardiovascular health concerns are triggered by various causative agents, of which environmental inhaled pollutants such as CO, NO, NO_2, PM_{10}, $PM_{2.5}$, O_3 and SO_2 from is an important agents. The introduction of air pollutants is caused by human activities that introduce contaminants into the air. In Nigerian, people living in pollution laden environments are unknowingly exposed to these risks such as Asthma, cough, lung cancer etc. However, there is paucity of information on the health risk impacts on the people living in pollution laden environments. This is due to lack of predictive model to reveal the associated risk to enhance early detection and prevention. One of the methods to evaluate and predict the pollutant is the use of Air Quality Index (AQI) dataset. The quality of AQI data of an environment is a pointer to the degree of pollution and the health risk of the inhabitants. Existing predictive techniques such as Probability and Statistics model used to predict AQI were very complex with some level of uncertainty which necessitate an alternative approach for better accuracy. A Machine Learning (ML) approach combined with an associative decision rule was used to predict the air quality and to identify areas predominates with toxic air quality. Two datasets; open and locally sourced were used, data pre-processed and engineered implementation were done using python coding. The prediction models; Support Vector Classifier (SVC) and Random Forest Classifier (RFC) were employed. The performances of the models were evaluated using classification reports and confusion matrix metrics. The RFC gave an accuracy level of 99% and SVC an accuracy level of 83%. This results show that AQI predictions obtained through RFC is better in accuracy when compared with SVC.

Keywords: SVM, RFC, Development Goals, Health Risk, Contaminants

1.0 Introduction

The cardiovascular diseases (CVDs) are one of the leading causes of morbidity and mortality globally. There is growing evidence that their prevalence is associated with environmental influences with air pollution being the primary one. Airborne pollutants, including carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), particulate matter (PM 0 and PM 2.5), ozone (O3), and sulfur dioxide (SO 2) have a strong detrimental impact on the health of the heart and lungs. Human activities are the primary sources of these chemicals in the atmosphere: industrial production, vehicle emission, combustion of fossil fuels and urbanization. The outcome is a severe public-health risk especially in the developing countries where surveillance is normally restricted.

Cities are growing in Nigeria at a rapid rate, environmental regulations are not strong, and the population is dependent on energy sources based on fossil fuel. There are thus a great number of polluted neighborhoods. The harmful gases and particles often present in the atmosphere are inhaled by the residents over a long period without their awareness and increase their risks of asthma, chronic cough, lung cancer, high blood pressure and heart-related issues. However, in Nigeria, the number of local studies that measure the impact of air quality on the risk of cardiovascular is very limited. This deficiency is largely due to the fact that there are only weak predictive models that are able to transform

Adeleke O. and. Ayoola O. A (2026). A Prediction Model for Cardiovascular Health Risk from Air Quality Index of Pollution laden Environment. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 16 No. 1, pp. 89 - 99

pollution information into transparent health-risk indicators and trigger timely response.

Air Quality Index (AQI) is a regular method of categorizing the degree of danger of air pollution. It is useful in estimating the risk of exposure by comparing multiple levels of pollutants on a single scale. Nevertheless, classical AQI models, which are applied in probability or statistics models, were heavy to compute, used harsh assumptions and were uncertain which restricts their application in the real world. Due to such constraints, scientists are resorts to data-driven approaches, which can be more effective in revealing the many non-linear relationships between air quality and health. Recent machine-learning technologies are able to learn such relationships using large and diverse data sets. Our suggestion is to have a machine-based learning framework that utilizes associative decision rules to estimate the AQI, and identify the regions that have a considerable level of toxic pollution, to demonstrate the possible health risk of cardiovascular diseases.

2.0 Literature Review

The World Health Organization (WHO, 2022) reveals that air contamination is one of the most important issues of the 21 st century. These pollutants involve man-made particles in the atmosphere that actually impact negatively on climate, humanity, nature and vegetation. Climatic pollution has now been a major cause of untimely deaths among the general population leading to a lot of deaths annually [1]. It was reported by the WHO that in 2016, approximately 58 per cent of deaths related to open-air contaminations were associated with the ischemic coronary disease and stroke, 18 per cent with persistent obstructive pulmonary disease and severe lower respiratory diseases, respectively, and 6 per cent with cellular breakdown in the lungs [2].

To determine the overall effect of the contaminated air to cardiovascular risk, Air Quality Index (AQI) was used to find out the nature of the air as far as pollution is concerned. The methods that were in existence, including the likelihood measurement models, were able to ascertain the nature of air but could not give the necessary information to predict poisonous air within an environment. It is, therefore, necessary to present a model that may supply the necessary information to identify toxins in the air in reference to health safety [3].

The quality of air in Nigeria is something to be concerned about because it has been ranked the 152 nd (out of 180 countries) on the Environmental Performance Index of Air Quality [4]. This model is aimed at delivering information and data to assist the Nigerian government to monitor the AQI within a polluted environment towards the achievement of the Sustainable Development Goals (SDGs) according to the WHO AQI bucket list depicted in Table 1 [5]. It conducted a research on the ability of an autoregressive integrated moving averaging model (ARIMA) to estimate estimates of months to months on the air contamination file. In the study, it has been revealed that ARIMA was capable of producing forecasts that were within the 95% confidence interval. It has been reported that the application of machine learning models offers more precise data to use in the analysis required to conduct AQI forecasts in different logical regions and areas [6]. This is an insight on how to enhance current statistical models, which are time-series forecasting in nature [8].

The predictions of air quality with Taiwan data between 2012 and 2016 were based on a Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) model. The model had a prediction accuracy of the next four hours of PM2.5 concentration in 66 stations throughout Taiwan.

Air Quality in India was also forecasted by developing a model to predict the AQI using historic data on the past years and projecting this upcoming year as a gradient-descent-boosted multivariate regression problem. The model achieved 96 % accuracy [9]. Two different models including Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel, which was implemented with and without the Principal Component Analysis (PCA) were used to predict the quality of the air in California: SVR with RBF kernel filtered with PCA (PCA-SVR-RBF). They compared the two models using Pearson correlation, Mean Squared error (MSE), Root Mean Squared error (RMSE), and Normalized Root mean squared error (nRMSE). The PCA-SVR-RBF model had an accuracy score of 88% on the training set and 92.7 on the validation set; SVR-RBF had a marginally higher

accuracy score on the two sets with the accuracy score of 90.02 on the training set and 94.1 on the validation set. The PCA-SVR-RBF and SVR-RBF showed comparable results of predicting the AQI [10].

Another experiment to forecast air pollution on a due date was conducted on monthly AQI. In the prediction of air pollution, the dataset was processed and analysed with the use of big-data analytics methods. It was calculated that the value of the difference between the presence or absence of air pollution against a calculated threshold value [11]. The weakness was that no machine-learning algorithm was tested with many machine-learning algorithms of prediction algorithm including: Linear Regression, SVM, Decision Trees and Random Forest. The accuracy of the machine-learning algorithms was measured by comparing the metrics after the experimentation and implementation of the algorithms through MSE and RMSE. Random Forest provided a better and more promising output than other machine-learning methods, but neural networks algorithms were not consulted in their study [3]. Four algorithms namely SVM, Random Forest, Adaptive Boosting, ANN, stacking ensemble and linear regression were used in predicting air quality to control and mitigate it adverse effects. Three evaluation measures were used to assess the performances of the models and these include RMSE, Mean Absolute Error and R-squared. It was determined that the AdaBoost model, as well as the stacking ensemble, was more competitive than SVM, Random Forest and ANN in predicting AQI after the implementation and evaluation [12].

3.0 METHODOLOGY

The study was carried out using the proposed model stages as shown in the conceptual frame work in Figure 1.0.

3.1 Data collection and Combination

The model starting stage was the collection of local Air pollution data were recorded at various locations in Ibadan, Moniya, Oluyole, the University of Ibadan, and Agbowo using sensors. Pre-processing of the data and the development of predictive model engineered features was performed. In this study, two sources were used to get the data. The former had been a local dataset acquired with the use of sensors of different locations in Ibadan in a period of four years (2016-2020), consisting of 18,367 records. The second source was a web-based dataset of Kaggle.com and it had 93,450 records. The two sets were combined to make a cumulative set of 111,817 records.

3.2 Data Pre-Processing and Feature Engineering

The preprocessing of data and feature engineering took place on the combined dataset that was a combination of the local and online sources. The data that was initially combined had many undesired features, outliers, and gaps, which were cleaned. Based on Kaggle dataset, which contains the following fields: date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, benzene, toluene, and xylene, only SO2, O3, NO2, NO, PM10 and CO were useful in the current study hence the other variables were thrown away. The outliers, which were mainly due to malfunctions in the sensors or transmission errors, were detected and eliminated in the locally obtained dataset.

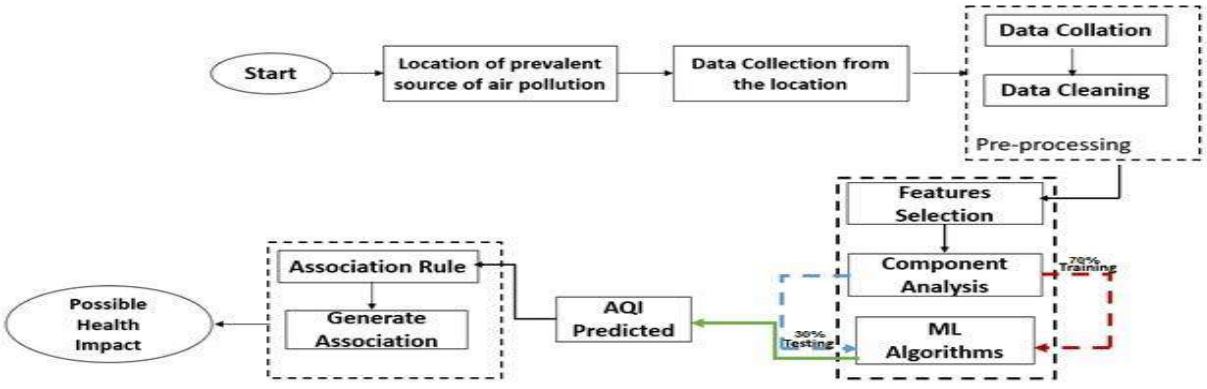


Figure 1.0 The conceptual frame work for the model

When using the standard pollutant benchmarks to analyze the area and the value of the boundary, some of the missing characteristics in the combined dataset were not estimable, and were excluded. Following feature engineering, it was found that the number of clean records was 85,880, which included the columns of average pollutant concentrations in different locations, namely, ozone (O3), sulfur dioxide (SO2), nitrogen dioxide (NO2), nitrogen oxide (NO), particulate matter (PM10), and carbon monoxide (CO).

3.3 Component Analysis and Pollutant index calculation

The second step was to compute the pollutants index based on the level of concentration as obtained by the sensors. The AQI index of each pollutant is calculated by hand, by transforming the concentration (value 3) of each pollutant into an index (I 3) according to the following formula in Equation 1 and Table 1:

$$I_i = LB_j + (value_i - lbi) \times (LB_j / lb_i) \quad (1)$$

Where **i** = PM_{10}, NO_2, NO, O_3, CO, SO_2; **j** = the level in the AQI system of the pollutant concentration such as good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy and hazardous; **LB** = Lower bound of j; **lb** = lower bound of i; $\mathbf{Value_i}$ = actual value of the pollutant. For example, based on the AQI bucket list of Nigeria in Table 1.0.

Once the index of individual pollutants in a point has been calculated, the AQI of that particular data point will be determined. The air quality index of a specific information point is the aggregate of maximum indexed pollutants in that particular area. That pollutant concentration value $\mathbf{SO_2}$ = 55, will fall in the interval with $\mathbf{lb_{SO2}}$ = 40 and $\mathbf{ub_{SO2}}$ = 80, corresponding with moderate pollutant level with $\mathbf{LB_{moderate}}$ = 50 and $\mathbf{UB_{moderate}}$ = 100 with the maximum index is taken as the air quality index of that particular location using the formula equation 2 and Figure 2.0;

$$\mathbf{AQI} = \max(I_{PM10}, I_{CO}, I_{SO2}, I_{NO2}, I_{O3}, I_{NO}) \quad (2)$$

Where I_{PM10}= Pollutant Index of PM_{10}
I_{CO}= Pollutant Index of CO
I_{SO2}= Pollutant Index of SO_2
I_{NO2}= Pollutant Index of NO_2
I_{O3}= Pollutant Index of O_3
I_{NO}= Pollutant Index of NO

Table 1.0 AQI Index and Health Implication (AQI-BUCKET)

AQI	Air Pollution Level	Health Implications	Cautionary Statement (for PM2.5)
0 - 50	Good	Air quality is considered satisfactory, and air pollution poses little or no risk	None
51 -100	Moderate	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
101-150	Unhealthy for Sensitive Groups	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
151-200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects	Active children and adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else, especially children, should limit prolonged outdoor exertion
201-300	Very Unhealthy	Health warnings of emergency conditions. The entire population is more likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.
300+	Hazardous	Health alert: everyone may experience more serious health effects	Everyone should avoid all outdoor exertion

```
In [11]: def conversion(val):
             #convert to float
             val = float(val)
             if(val <= 40):
                 return val*(5/4)
             elif(val>40 and val <=80):
                 return 50+(val-40)*(5/4)
             elif(val>80 and val <=130):
                 return 100+(val-80)*(10/8)
             elif(val>130 and val <=180):
                 return 150+(val-130)*(15/13)
             elif(val>180 and val <=280):
                 return 200+(val-180)*(20/18)
             else:
                 return 300+(val-280)*(30/28)
```

```
In [14]: def cal_aqi(row):
             aqi=0
             NO = row['NO']
             NO2 = row['NO2']
             CO = row['CO']
             O3 = row['O3']
             SO2 = row['SO2']
             PM10 = row['PM10']
             if(NO2>NO and NO2>CO and NO2>O3 and NO2>SO2 and NO2>PM10):
                 aqi=NO2
             elif(CO>NO and CO>NO2 and CO>O3 and CO>SO2 and CO>PM10):
                 aqi=CO
             elif(NO>CO and NO>NO2 and NO>O3 and NO>SO2 and NO>PM10):
                 aqi=NO
             elif(SO2>CO and SO2>NO2 and SO2>O3 and SO2>NO and SO2>PM10):
                 aqi=SO2
             elif(O3>CO and O3>NO2 and O3>SO2 and O3>NO and O3>PM10):
                 aqi=O3
             elif(PM10>CO and PM10>NO2 and PM10>SO2 and PM10>NO and PM10>O3):
                 aqi=PM10
             return aqi
```

Figure 2.0: AQI Calculation

3.4 Machine Learning Model Prediction of Air Quality Index

The data were divided into a 70 test set and a 30 training set. We tested two predictive models, one of a Linear SVM classifier and a Random Forest classifier having ten estimators. The models were deployed in the open-source distribution, Anaconda, in the Jupyter Notebook environment, and on Google Colab, the online platform of Python run at Google. Using a collection of association rules, we compared the approximate AQI and concentration of pollutants with the potential health-implications on the local residents.

3.5 Model Performance Evaluation

The performance of the models for accuracy was evaluated using two evaluation metrics for classification analysis, these are; classification report and confusion matrix. There are four ways to check if the predictions are right or wrong:

1. **TN / True Negative**: the case was negative and predicted negative
2. **TP / True Positive**: the case was positive and predicted positive
3. **FN / False Negative**: the case was positive but predicted negative
4. **FP / False Positive**: the case was negative but predicted positive

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making. The confusion matrix model is shown in Figure 3.0.

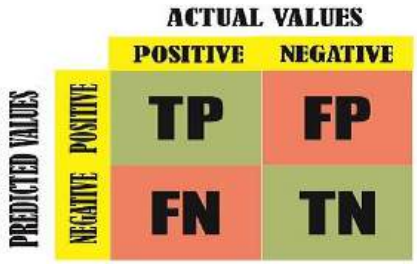


Figure 3.0: Confusion Matrix

A classification report of a performance evaluation metric in machine learning, precision, recall, F1 Score, and support of a trained classification **model was done.** It provides a better understanding of the overall performance of the trained model. The rate of accuracy is the proportion of the total number of correct predictions. It is calculated using the following equation 3

Accuracy Rate = (TP + FN) / (TP + TN + FP + FN) (3)

3.6 Association Rule Generation for Health Impact of the Air Quality

A number of association rules were used to correlate the predicted AQI to causative pollutants and related health risk. The analysis was conducted to identify the possible health effect of AQI on the inhabitants of the identified place. The regulations demonstrate the impact of the combinations of AQI rates and certain pollutants on human health. Table 4.0 illustrates some of these rules.

Association Rules

Rule 1: {AQI = Good ⇨ Health = Good}
Rule 2: {(AQI = Moderate) Λ (Pollutant = PM_{10}) ⇨ Health = Mild cough, throat and lung irritation}
Rule 3: {(AQI = Unhealthy) Λ (Pollutant = PM_{10}) ⇨ Health = Worsen Asthma condition}
Rule 4: {(AQI = Very Unhealthy) Λ Pollutant = PM_{10}) ⇨ Health = Lung Cancer, Heart problems}
Rule 5: {(AQI = Hazardous) Λ (Pollutant = $PM_{2.5}$) ⇨ Health = High mortality rate}

Figure 4.0: Health Impact Rules

4.0 Results

This section is divided into four sections general description of the dataset is given, engineered data results are discussed, AQI prediction models are formulated with a discussion of the parameters settings and approaches used to impute missing data, and AQI model performance is assessed.

4.1 Data visualization

Python Jupyter notebook was used to summarise and describe the pre-processed data. Table 2.0 shows the count, the mean, the minimum and maximum values of the dataset.

Table 2.0: Health Impact Rules

```
In [7]: df.describe()
```

Out[7]:

	NO2	NO	CO	SO2	O3	PM10
count	85899.000000	85899.000000	85899.000000	85899.000000	85899.000000	85899.000000
mean	43.645138	28.098431	2.121917	13.828817	42.474529	130.811500
std	35.151973	40.082234	5.529152	15.700175	44.737614	136.409281
min	0.010000	0.010000	0.000000	0.000000	0.000000	0.000000
25%	18.980000	6.000000	0.760000	4.890000	18.180000	0.000000
50%	34.040000	12.830000	1.200000	10.100000	34.830000	103.150000
75%	58.460000	30.660000	1.880000	17.560000	55.980000	192.930000
max	480.050000	503.570000	202.860000	217.390000	1031.790000	1071.430000

4.2 AQI Prediction Model

The features of the dataset were engineered as shown in Table 3.0 and Table 4.0 to derive their respective pollutant index and air quality index respectively. The results of the engineered features are shown in Table 3.0

The calculated AQI shown in Table5.0 was used to create bucketlist using the AQI bucketlist for Nigeria presented Table 1.0 as the target value for the prediction model. The engineered dataset which contains the pollutants in a region and the bucketlist of that region is shown in Tables6.0 and 6.0b. This dataset were used in the training and validation of our prediction models.

Table 3.0: Actual values of individual pollutants gotten from local data

```
In [13]: df.head()
```

Out[13]:

	NO2	NO	CO	SO2	O3	PM10
0	0.0125	0.0500	0.6250	11.1125	70.5750	58.825000
1	0.0125	146.1250	1.3125	27.0000	2.0750	95.075000
2	0.0125	0.0250	1.1375	9.1625	39.9625	116.025000
3	0.0250	31.4875	2.3250	22.8625	10.7750	489.171429
4	0.0375	29.4375	2.9250	38.8125	13.0625	454.585714

Table4.0: Calculated Index of individual pollutants

```
In [22]: df
```

Out[22]:

	NO2	NO	CO	SO2	O3	PM10	AQI
0	0.012500	0.050000	0.6250	11.1125	70.575000	58.825000	70.575000
1	0.012500	146.125000	1.3125	27.0000	2.075000	95.075000	146.125000
2	0.012500	0.025000	1.1375	9.1625	39.962500	116.025000	116.025000
3	0.025000	31.487500	2.3250	22.8625	10.775000	489.171429	489.171429
4	0.037500	29.437500	2.9250	38.8125	13.062500	454.585714	454.585714
5	0.050000	0.875000	2.1250	2.9875	10.225000	0.000000	10.225000
6	0.050000	154.925000	0.3750	31.1875	8.537500	48.750000	154.925000
7	0.062500	14.450000	0.0250	1.7250	1.750000	1.437500	14.450000

Table 5.0: The Calculated AQI values

```
In [8]: df.head()
```

Out[8]:

	NO2	NO	CO	SO2	O3	PM10
0	0.01	0.04	0.50	8.89	56.46	47.06
1	0.01	116.90	1.05	21.60	1.66	76.06
2	0.01	0.02	0.91	7.33	31.97	92.82
3	0.02	25.19	1.86	18.29	8.62	456.56
4	0.03	23.55	2.34	31.05	10.45	424.28

Table 6.0a**: Engineered dataset for modelling.**

```
In [5]: df.head()
```

Out[5]:

	NO2	NO	CO	SO2	O3	PM10	OUTCOME
0	0.01	0.05	0.63	11.11	70.58	58.83	Moderate
1	0.01	146.13	1.31	27.00	2.08	95.08	U.S.G
2	0.01	0.03	1.14	9.16	39.96	116.03	U.S.G
3	0.03	31.49	2.33	22.86	10.78	489.17	Hazardous
4	0.04	29.44	2.93	38.81	13.06	454.59	Hazardous

Table 6.0b.**: Engineered dataset for modelling.**

```
In [5]: df.head()
```

Out[5]:

	NO2	NO	CO	SO2	O3	PM10	OUTCOME
0	0.01	0.05	0.63	11.11	70.58	58.83	Moderate
1	0.01	146.13	1.31	27.00	2.08	95.08	U.S.G
2	0.01	0.03	1.14	9.16	39.96	116.03	U.S.G
3	0.03	31.49	2.33	22.86	10.78	489.17	Hazardous
4	0.04	29.44	2.93	38.81	13.06	454.59	Hazardous

The prediction model was developed using two different algorithms. The algorithms deployed were Random Forest Classifier and Support Vector Classifier. The model was developed using 70 – 30 training/testing split. Of the 85,880 datasets, 60,116 was used for training while 25,764 was used for the testing.

4.3 Performance Evaluation

In performing the evaluation of machine learning models, we check how the real target values relate to the predictions that we have made and this allows us to find the error rate particularly in the case of outliers. A confusion matrix is a summary of the performance of a model. It shows the amount of correct predictions (TP + TN) and the one that are incorrect (FP and FN). Table 7.0 illustrates the matrices of two models, SVC, and RFC.

From the confusion matrix Table 7.0, the number of values in the diagonal column represents the correctly predicted values (TP + TN). For the prediction of **Good AQI,** it was observed that out of the possible right prediction value of **5566** (5561+0+2+2+1+0), the SVC was able to correctly predict **4830** and a confused prediction of **736** while RFC was able to correctly predict **5561** and a confused prediction of **5**. For the prediction of **U.S.G AQI**, it was observed that of the possible right prediction value of **3966**, the SVC was able to correctly predict **2889** and a confused prediction of **1077** while RFC was able to correctly predict **3951** and a confused prediction of **15**. In general, it was observed that for the Support Vector Classifier, the number of False Positives (FP) and False Negative (FN) is higher compared to that of the Random Forest Classifier. The matrix shows that RFC has a higher rate of prediction for True Positive (TP) and True Negative (TN) with few FP and FN compared to that of the Support Vector Classifier. Furthermore, the rate of accuracy of the two models was evaluated using equation 3.

It was observed that Random Forest Classifier accomplished a superior accurate score of 99% compared to the Support Vector Classifier Model which accomplished an accuracy of 83% on the dataset as shown in Table 8.0

Table 7.0: Confusion matrix for the AQI prediction Obtained with both models for the datasets

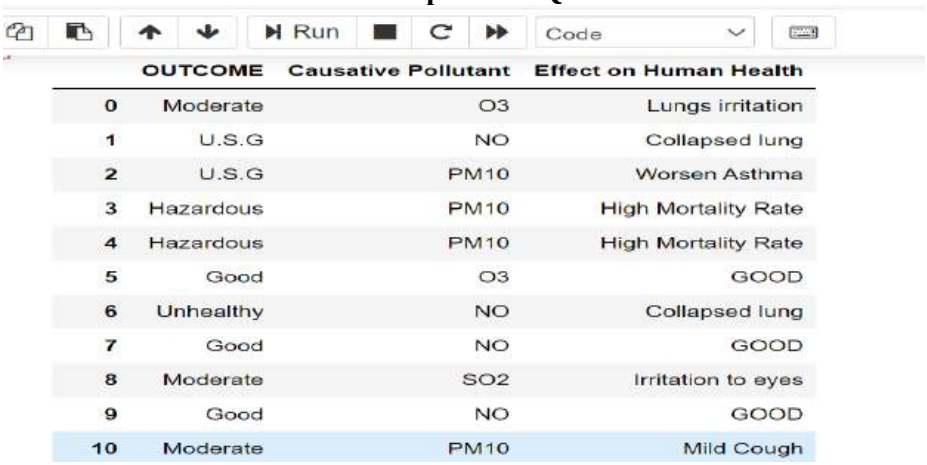
		Support Vector Classifier						Random Forest Classifier					
		G	H	M	U.S. G	Un.	V. Un	G	H	M	U.S. G	Un.	V. Un
PREDICTED VALUES	Good	4830	0	734	2	0	0	5561	0	2	2	1	0
	Hazardous	0	2965	0	6	4	131	0	3094	0	0	3	9
	Moderate	721	0	4870	582	1	0	2	0	6171	1	0	0
	U.S. G	4	0	685	2889	388	0	1	0	8	3951	6	0
	Unhealthy	0	0	55	318	3074	230	0	0	2	16	3657	2
	V. Unhealthy	0	156	1	93	211	2820	0	8	0	6	22	3245

***G=Good, H=Hazardous M=Moderate, U.S. G= Unhealthy for sensitive groups, Un=Unhealthy, V.Un=Very Unhealthy**

Table 8.0 Level of Accuracy for the AQI prediction

METRICS	SVC	RFC
Accuracy	0.83 (83%)	0.99(99%)

Table 9.0: The health impact of AQI on the inhabitants



	OUTCOME	Causative Pollutant	Effect on Human Health
0	Moderate	O3	Lungs irritation
1	U.S.G	NO	Collapsed lung
2	U.S.G	PM10	Worsen Asthma
3	Hazardous	PM10	High Mortality Rate
4	Hazardous	PM10	High Mortality Rate
5	Good	O3	GOOD
6	Unhealthy	NO	Collapsed lung
7	Good	NO	GOOD
8	Moderate	SO2	Irritation to eyes
9	Good	NO	GOOD
10	Moderate	PM10	Mild Cough

4.4 Health Impact of the Air Quality

The predicted AQI by the models combined with the causative pollutant was, in turn, being used to determine the possible health impact of pollution on the inhabitants of a particular region. This was carried out using some sets of rules as shown in figure 4.0.

Table 9.0, shows that in a location marked as an unhealthy AQI, it means the prevalent of Nitrogen oxide (NO) pollutant in the air, which means that the prolonged exposure of inhabitants of that location to such conditions might lead to collapsed lungs. Also, in a location with moderate AQI, it means SO2 gas is prevalent in the air, the prolonged exposure of inhabitants in such a location to the such gas could lead to irritation to the eyes. But the moderate AQI is caused by PM10 particle, it could lead to mild cough and if moderate AQI is caused by O3, it leads to lung irritation as presented in Table 9.0.

4.5 Discussions

This study developed a prediction models of AQI with the use of two data sources Kaggle.com and local AQI sensors to predict the air quality index in particular locations and categorize these locations into pre-existing buckets. The findings indicate that the locations under study are largely the Moderate and U.S.G. (Unhealthy for Sensitive Groups) categories though they occasionally slide into the Unhealthy category. This inconsistency has been explained by the fact that the areas were urban and semi-urban, with the emissions of various manufacturing industries continuously releasing a good amount of pollutants into the atmosphere. Wearing face masks by the residents whenever they are outdoors is therefore recommended to guard their health as well as generate long-term health. The performance of the model is rather close to the existing methods: it is doing fairly well in most scenarios and fails slightly less in the cases of predicting unhealthy conditions. The book utilizes two algorithms that are supervised learning, namely, Support Vector Machine (SVM) and Random Forest (RF). Findings show that RF makes better predictions of AQI than SVM besides consuming fewer resources and taking shorter times to execute.

5.0 Conclusion

The risk of cardiovascular health in the places of study was moderate and the risk of mild cough following long exposure is high. This work predicts the AQI of chosen locations, which forms the basis of monitoring and controlling the air pollutants levels in an attempt to protect the residents. Random Forest (RF) model performed better than the other models in the prediction of AQI with a high accuracy of 99.0. The research suggests creating health awareness among the health workers and the population on health hazards caused by pollution in their locality.

REFERENCES

[1] WHO (2022, April) How air pollution is destroying our health. Retrieved from https://www.who.int/news-room/spotlight/how-air-pollution-is-destroying-our-health.

[2] WHO (2021, November) Air Pollution. Retrieved from https://www.who.int/thailand/health-topics/air-pollution#.

[3] Madhuri VM, Samyama, Gunjal GH, Savitha Kamalapurkar (2020). "Air Pollution Prediction using Machine Learning Supervised Learning Approach" IJSTR volume 9, issue 04, 2020

[4] Aqicn. (2022, April). Air Quality Monitoring in Nigeria. Retrieved from https://aqicn.org/country/nigeria/.

[5] L. Y. Siew, L. Y. Chin, P. Mah, and J. Wee, (2008). "Arima and integrated Arima models for forecasting air pollution index," He Malaysian Journal of Analytical Science, vol. 12, no. 1, pp. 257–263.

[6] T. M. Mitchell, (2009). "Machine learning," in Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Pasadena, CA, USA, July 2009.

[7] M. Castelli, I. Goncalves, P. Ales, and L. Trujillo (2016). An evolutionary system for ozone concentration forecasting, pp. 1123–1132, Springer Science Business Media, New York, NY, USA,2016.

[8] Yi-Ting Tsai, (2018). Dept. of Computer Science and Information Engineering National Taipei University." Air pollution forecasting using RNN with LSTM", 2018 IEEE 16th Int. Conference.

[9] Gnana Soundari, Gnana Jeslin, Akshaya A.C. (2019). "Indian Air Quality Prediction and Analysis Using Machine Learning". IJAER Volume 14, Number 11, 2019.

[10] Mauro Castelli, Fabiana Martins, Clemente, Ales Popovic, sara Silva and Leonardo Vanneschi (2020). "A Machine Learning to Predict Air Quality in California" Hindawi Complexity Volume 2020, Article ID 8049504.

[11] S. Suganya, Dr. T. Meyyappan (2020)." Forecasting And Prediction Of Air Pollution LevelsTo Protect Human Beings From Health Hazards" International Journal Of Scientific & Technology Research Volume 9, Issue 01, January 2020

[12] Yun-Chia Liang, Yona Maimury, Angela Hsiang-Ling Chen, Josue Rodolfo Cuevas Juarez. (2020). "Machine Learning-Based Prediction of Air Quality" Applied Science 2020, 10, 9151