Spring 1-18-2021

A PREDICTIVE MODEL FOR DIABETES USING MACHINE LEARNING TECHNIQUES (A CASE STUDYOF SOME SELECTED HOSPITALS IN KADUNA METROPOLIS)

A E. EVWIEKPAEFE
NIGERIAN DEFENCE ACADEMY

NAFISAT ABDULKADIR
NIGERIAN DEFENCE ACADEMY

NOTE: WHEN USING THIS TEMPLATE FOR THE REVIEW VERSION OF YOUR PAPER, YOU MUST REMOVE ALL THE AUTHOR INFORMATION IN THE LINES BELOW AS WELL AS ANY AUTHOR IDENTIFYING INFORMATION WITHIN THE TEXT. THE REVIEW PROCESS IS DOUBLE BLIND (reviewers do not know who authors are and authors do not know who reviewers are). INCLUDE ALL AUTHOR INFORMATION WHEN USING THIS TEMPLATE FOR THE CAMERA READY VERSION (you submit the camera ready version once the review process is completed and after you received confirmation that your paper is accepted for publication).

The African Journal of Information Systems

A PREDICTIVE MODEL FOR DIABETES USING MACHINE LEARNING TECHNIQUES (A CASE STUDY OF SOME SELECTED HOSPITALS IN KADUNA METROPOLIS)

Research Paper

Volume X, Issue X, Month YYYY, ISSN 1936-0282 **(will be set by editors)**

Dr. AE EVWIEKPAEFE
Nigeria Defence Academy, Kaduna
aeevwiekpaefe@nda.edu.ng

ABDULKADIR Nafisat
Nigeria Defence Academy, Kaduna
nafisahabdulkadir@gmail.com

(*Received month year, accepted month year*)

ABSTRACT

Diabetes Mellitus (DM) which refers to a metabolic disorder that occurs when the level of blood sugar in the body is considered high, which could be a resulting effect of inadequate availability of insulin in the body. It is a chronic disease which may lead to myriads of complications in the body system. Statistics by the World Health Organization (WHO) in 2013, indicated that DM was the cause of death of over 1.5 million people around the world and in 2016, 8.5% of adults within age seventeen (17) and above were reported to be diabetic and diabetic patients have continued to increase in recent years. It is therefore very glaring that these alarming figures calls for very urgent and effective attention. There has been a recent proliferate increase in studies relating to machine learning in the healthcare sector, hence the motivation for this research work. The research was based on the prevalence of diabetes amongst the masses of Kaduna metropolis using some selected hospitals as a case study after which a predictive model was designed for diabetes, using some selected supervised learning algorithms like Decision tree algorithm, K- Nearest Neighbour algorithm and Artificial Neural Networks on a dataset gotten from 44 Army Reference Hospital and Yusuf Danstoho Memorial Hospital Kaduna which constitutes of nine (9) attributes that was considered. The results indicated that ANN produced the highest accuracy with 97.40% followed by decision tree algorithm with 96.10% accuracy then K-NN algorithm with 88.31%

accuracy. This result was further validated using fifty (50) dataset out of which forty-eight results were rightly predicted.

Keywords

Diabetes mellitus, metabolic disorder, healthcare sector, artificial neural network.

INTRODUCTION

The prevalence of diabetes is on the increase across the world, the International Diabetes Federation have noted that; there are about three hundred and eighty-two (382) million people living with diabetes in the world and by 2035, this will be doubled as five hundred and ninety-two (592) million (Pradhan et al., 2020). The increase might be due to suburbanization, gradual adoption of unhealthy lifestyles and aging of the population without preparedness for prevention and control, throwing up so many challenges to the diabetes care which has now become a major health problem in most countries around the world (Liu et al., 2013). The world health organization (WHO) in 2016, noted that about 8.5% of adults aged seventeen (17) years and above are diabetic patients. In the year twenty thirteen (2013) diabetes 1.5 million deaths were linked to diabetes, while high blood glucose resulted to 2.3 million deaths (Pradhan et al., 2020). In the last ten years, diabetes patients have doubled across the world (Harz et al., 2020). Over two hundred (200) million people are diabetic with an annual predominance of seven percent in the world (Zahran, 2017). For a long time, people have suffered from various diseases that could have been prevented in some cases, but due to lack of prompt diagnosis of symptoms in patients, this may lead to very dilapidating consequences (Temurtas et al., 2009).

Several studies have been done in the field of prediction for several diseases recently, to the level that some of today's clinicians now make use of machine learning models to predict different diseases. It is therefore, imperative to design a diabetes classifier that is convenient, accurate and cost efficient. Artificial Intelligence techniques provide a wide range of ideas that are useful to human related fields of application like, medical diagnosis which is a process where a physician has to analyze lot of factors before diagnosing diabetes which makes the physician's job difficult and time consuming. Machine learning and data mining techniques have been considered very helpful in the design of automatic diagnosis system for various health conditions (Adeloye et al., 2017). In recent times, many methods and algorithms have been discovered which can be used to mine biomedical datasets for hidden information including supervised learning techniques like Neural networks (NNs), K- Nearest Neighbour (KNN), Support Vector Machines (SVM), Fuzzy Logic Systems, Decision Trees (DT), Naive Bayes, and logistic regression; unsupervised learning techniques like clustering analysis, pattern recognition and image analysis; and reinforcement algorithms which is applied in the field of game theory, control theory decision theory a (Modern, 2019). This research work intends to develop a prediction model with high degree of accuracy for diabetes in people at an early stage, before it becomes escalated to a point of morbidity or mortality using some supervised learning algorithms and a case study of selected hospitals within Kaduna metropolis. This research will also contribute to the health sector by providing people with accurate prior knowledge about their health status as related to diabetes hence, reducing the rate of complications, morbidity and mortality being caused by this disease.

LITERATURE

Related Works

(Uloko et al., 2018), conducted a research on the prevalence of risk factors for diabetes mellitus in Nigeria. In conducting this research work, a total of 23 studies (n = 14,650 persons) were considered. In estimating the pooled prevalence of DM, a random effects model was implemented and subgroup specific DM prevalence was used to account for inter-study and intra-study heterogeneity. From the results achieved they concluded that, the prevalence of DM in Nigeria has been on the increase in all regions of the country affected, with south-south with the having the highest prevalence seen in the geopolitical zones. Urbanization, physical inactivity, aging, and unhealthy diet are key risk factors for DM among Nigerians. They recommended the urgent need for a national diabetes care and prevention policy scheme.

(Chawan, 2018) conducted a research aimed at developing a system which can predict diabetes at an early stage in patients with a high accuracy by combining the results of different machine learning techniques. The research predicts diabetes using two (2) different supervised machine learning methods including SVM and Logistic Regression. It considered seven (7) features of the patients. They reached a conclusion that SVM showed a better performance with accuracy of seventy-nine percent (79%) compared to logistic regression which had a performance accuracy of seventy-eight percent (78%).

(Sneha & Gangil, 2019) conducted a research that was aimed at selecting the attributes that aid in early detection of diabetes mellitus using WEKA which is a predictive analysis tool. They were able to reach a conclusion which shows that decision tree algorithm and Random Forest Algorithm has the highest predictive analysis by 98.20% and 98.00% respectively. While Naïve Bayesian outcomes states the best in performance accuracy with 82.30%.

(Modern, 2019) stated that, there are an enormous amount of data available in the world today, but very few are there for the analysis of it because of which nowadays many new fields are emerging starting from Data Science to Bioinformatics and Cheminformatics. It can be assured that this world of AI is going to benefit a lot to humanity, converting the toughest jobs to the simplest ones. Machine learning has led to minimizing the errors Involved with the co-relation of different kinds of attributes. Most importantly, it has transformed the approach of hit and trial method into a way with full of logic and simulations. Today, using various simulations several required properties and the after effects of many materials can be predicted, which has led us to the maximization of a lot of resources. In this review article, they presented the machine learning types, different algorithms and along with their uses in several in different ways.

(Kaur & Kumari, 2019) developed five different models for the detection of diabetes using, linear kernel support vector machine (SVM-linear), radial basis kernel support vector machine (SVM-RBF), K Nearest Neighbour (k-NN), Artificial Neural Networks (ANN) and Multifactor Dimensionality Reduction (MDR) algorithms. Feature selection of dataset was done with the help of Boruta wrapper algorithm, considering some evaluation criteria namely; accuracy, recall, precision, F1 score, and Area Under the Curve (AUC). The experimental results indicated that all the models achieved good results with SVM-linear model providing a very good accuracy of 0.89 and precision of 0.88. From the results of this study, it can be concluded that on the basis of all the parameters linear kernel support vector machine (SVM-linear) and k-NN are the two (2) most accurate predictive models for diabetes. This work also suggested that Boruta wrapper algorithm can be used for feature selection as they were able to achieve a better accuracy with its use.

Research Gap

From the above reviewed literatures, it is important to note that, although various research work have been carried out in the area of diabetes prediction in other countries using various risk factors that are peculiar to their environment but not much have been done in applying any of the machine learning techniques in diabetes prediction, using risk factors that are peculiar to the Nigerian environment.

It is also evident from the reviewed literatures that supervised learning algorithms overtime, produced very good prediction accuracy in research works where they were applied though not much work have been done in comparing the prediction accuracy of these three (3) supervised learning algorithms i.e K-NN, decision trees and ANN.

METHODOLOGY

Analysis of the New System

This research work is aimed at getting the best model that is able to predict diabetes in people at an early stage, subsequent to collection and pre-processing of data, the dataset will be trained using some supervised learning algorithms like K-NN, decision trees and ANN after which their respective prediction accuracies would be compared and the best model would be picked and used for Implementation on the dataset to predict future occurrence. How the result is to be achieved would be visualized in Figure 1.

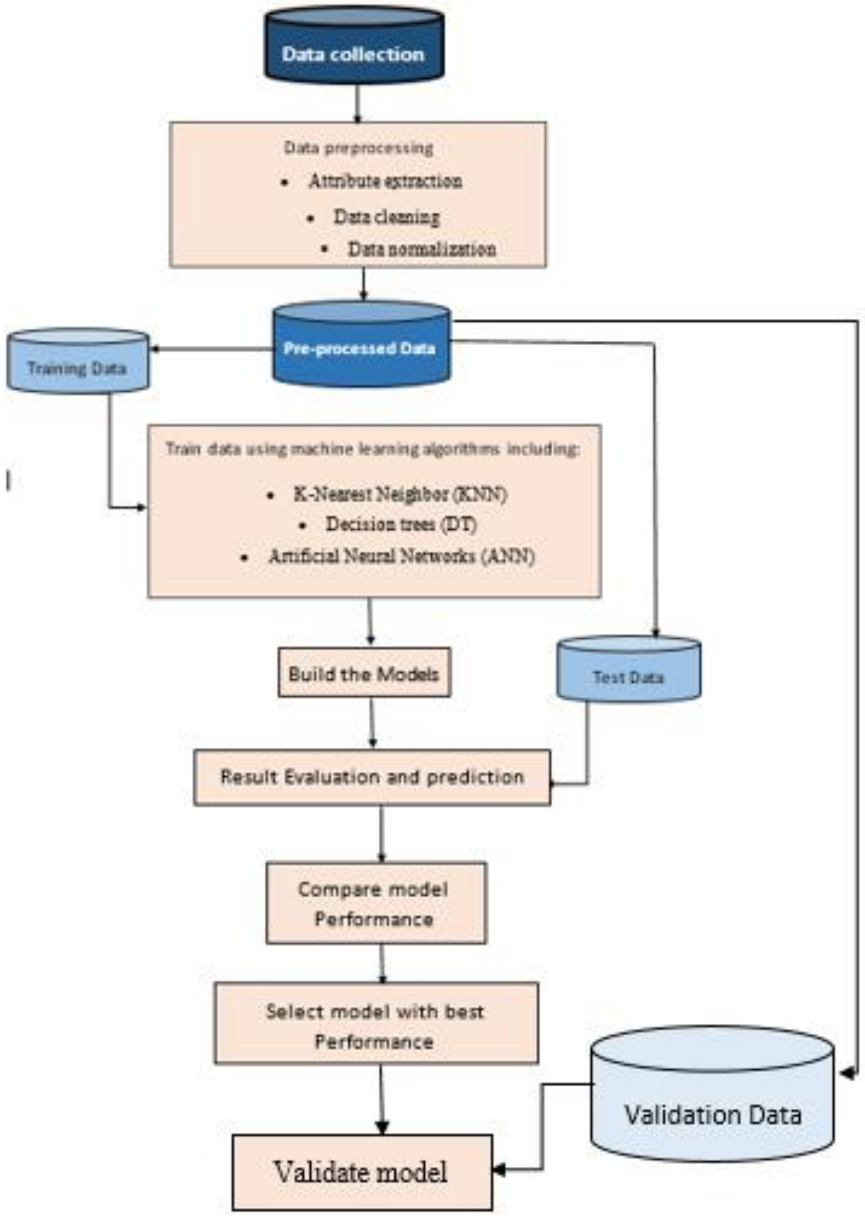


Figure 1 Framework of the New Model

Data Description

The dataset used in carrying out this research was obtained from the 44 Nigerian army reference hospital Kaduna and Yusuf Dan-Stoho Memorial Hospial Tudun Wada Kaduna where we met each of the individuals to take their data with the help of some doctors form Barau Dikko Hospital Kaduna. These data amounted to a total of two hundred and fifty-five (255) samples, which consist of two parts: the non-Diabetic people and the diabetic people, with one hundred and five (105) diabetic samples and one hundred and fifty (150) non-diabetic samples. The dataset includes nine (9) physical examination indexes: age, sex, number of pregnancies, glucose level, blood pressure level, body mass index, height, weight and how regularly do they exercise, these attributes are tabulated in the Table 1.

Table 1 Tabular view showing the attribute, their variable type with specified range

Attribute Number	***Attribute***	***Variable Type***	***Range***
A1	Age	integer	18 above
A2	Sex	Binary	0 or 1
A3	Number of pregnancies	integer	0 and above
A4	Glucose level (mmol/l) or (mg/dl)	real	0 and above
A5	Height	Real	0 and above
A6	Weight	Real	0 and above
A7	BMI (kg/m^2)	Real	0 and above
A8	Blood pressure (mm/hg)	Real	0 and above
A9	Regular Exercise	binary	0 or 1
A10	Class of diabetes	Binary	0 or 1

Data Pre-Processing

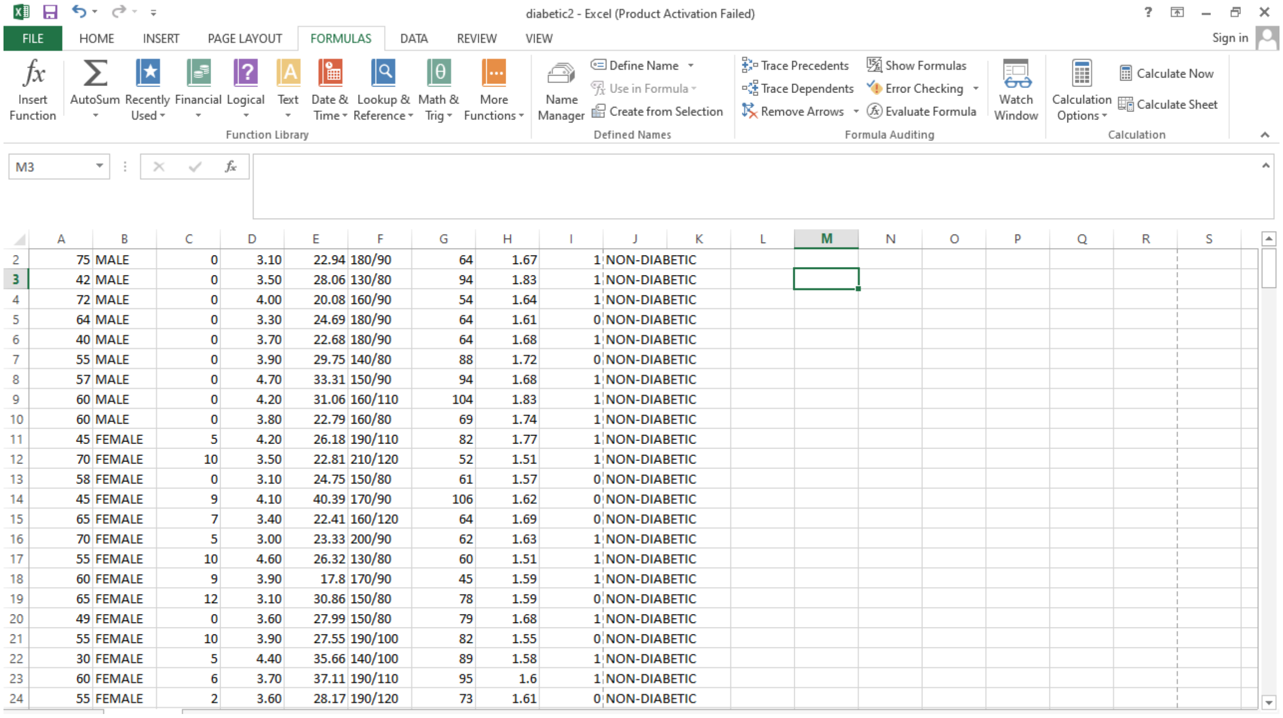
After the complete dataset has been collected, it is of high importance the data being collected is pre-processed to train the network efficiently. The procedure involves: (1) solving the problem of missing data; (2) data normalization; and (3) data standardization

To solve the problem of missing data, the missing values are solved by the average of neighbouring values, but in the case of this research they was no missing data recorded. It is imminent to carry out the data normalization procedure before parsing the input data to the learning algorithm, as mixing of variables with a variety of magnitude could confuse the learning algorithm. Therefore leading to a rejection of the variables and poor prediction accuracy (Tymvios et al., 2008). Python provides some data pre-processing libraries that comes with a variety of features like skLearn, which has embedded in it the LabelEncoder and the StandardScaler that will be adopted.

RESULTS AND DESCRIPTION

Th results obtained from developing the models for the diabetes prediction system using a case study of Kaduna metropolis were achieved using the KNN, ANN and the decision trees algorithm. The models were trained using ten input variables (sex, number of pregnancies, glucose level, blood pressure level, body mass index, height, weight and regular exercise) then the diabetic status of the individual was used as the target output that was compared with the predicted output.

All of these attributes were included in the dataset with no missing values as the data was collected directly from the people, see Figure 2 for a screen shot of the dataset after inputting into excel spreadsheet in the CSV format.



	A	B	C	D	E	F	G	H	I	J
2	75	MALE	0	3.10	22.94	180/90	64	1.67	1	NON-DIABETIC
3	42	MALE	0	3.50	28.06	130/80	94	1.83	1	NON-DIABETIC
4	72	MALE	0	4.00	20.08	160/90	54	1.64	1	NON-DIABETIC
5	64	MALE	0	3.30	24.69	180/90	64	1.61	0	NON-DIABETIC
6	40	MALE	0	3.70	22.68	180/90	64	1.68	1	NON-DIABETIC
7	55	MALE	0	3.90	29.75	140/80	88	1.72	0	NON-DIABETIC
8	57	MALE	0	4.70	33.31	150/90	94	1.68	1	NON-DIABETIC
9	60	MALE	0	4.20	31.06	160/110	104	1.83	1	NON-DIABETIC
10	60	MALE	0	3.80	22.79	160/80	69	1.74	1	NON-DIABETIC
11	45	FEMALE	5	4.20	26.18	190/110	82	1.77	1	NON-DIABETIC
12	70	FEMALE	10	3.50	22.81	210/120	52	1.51	1	NON-DIABETIC
13	58	FEMALE	0	3.10	24.75	150/80	61	1.57	0	NON-DIABETIC
14	45	FEMALE	9	4.10	40.39	170/90	106	1.62	0	NON-DIABETIC
15	65	FEMALE	7	3.40	22.41	160/120	64	1.69	0	NON-DIABETIC
16	70	FEMALE	5	3.00	23.33	200/90	62	1.63	1	NON-DIABETIC
17	55	FEMALE	10	4.60	26.32	130/80	60	1.51	1	NON-DIABETIC
18	60	FEMALE	9	3.90	17.8	170/90	45	1.59	1	NON-DIABETIC
19	65	FEMALE	12	3.10	30.86	150/80	78	1.59	0	NON-DIABETIC
20	49	FEMALE	0	3.60	27.99	150/80	79	1.68	1	NON-DIABETIC
21	55	FEMALE	10	3.90	27.55	190/100	82	1.55	0	NON-DIABETIC
22	30	FEMALE	5	4.40	35.66	140/100	89	1.58	1	NON-DIABETIC
23	60	FEMALE	6	3.70	37.11	190/110	95	1.6	1	NON-DIABETIC
24	55	FEMALE	2	3.60	28.17	190/120	73	1.61	0	NON-DIABETIC

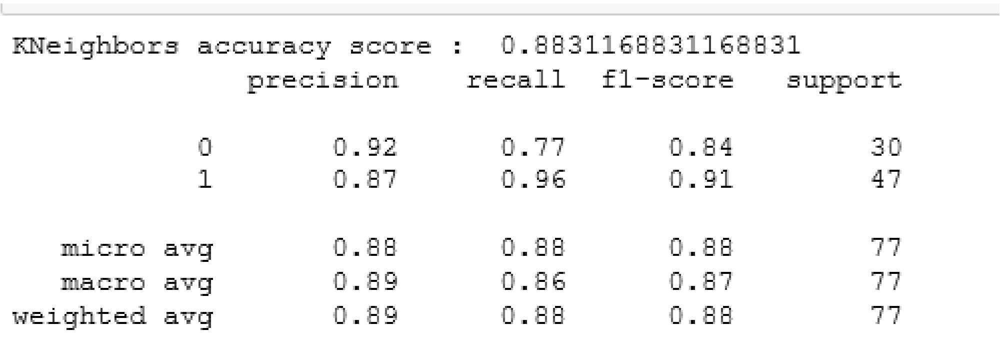
Figure 2 Excel Spreadsheet for Datasets

While training this model, data normalization included using the pre-processing label encoder function from the python library using the jupyter notebook available on the anaconda navigator to normalize some of the values to make it understandable to the learning algorithm

Results of KNN

The first model was designed was for predicting the diabetes status of an individual using the KNN algorithm. This experiment was conducted using the K Nearest Neighbours Classifier function from the python programming library while the number of neighbours considered were, K = 8 at point 2 using the Euclidean distance technique for the computation.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to accuracy of 88.31%, with precision of 0.87 for ND and 0.92 for D, Recall of 0.96 for ND and 0.77 for D, F1 Score of 0.91 for ND and 0.84 for D and Support of 47 for ND and 30 for D.



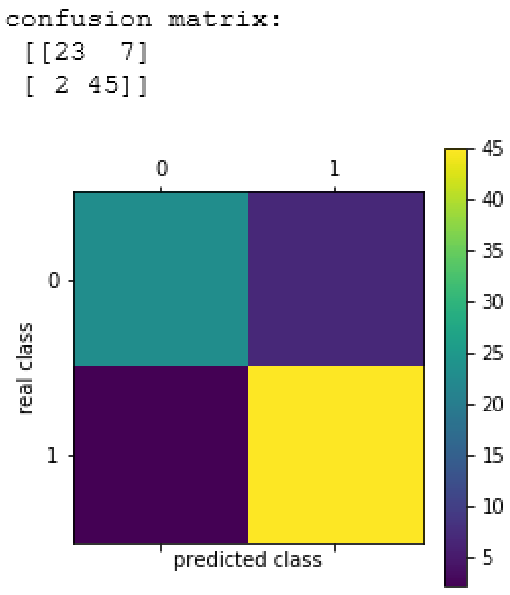
```
KNeighbors accuracy score :  0.8831168831168831
              precision    recall  f1-score   support

           0       0.92      0.77      0.84        30
           1       0.87      0.96      0.91        47

   micro avg       0.88      0.88      0.88        77
   macro avg       0.89      0.86      0.87        77
weighted avg       0.89      0.88      0.88        77
```

Figure 3 Results for the KNN Model

Also, the KNN produced a fair confusion matrix which showed that from the 30% of diabetic dataset used for testing twenty-three (23) of them where true positive i.e they were predicted to be correctly diabetic and seven (7) of them where False Negative, meaning they wrongly predicted to be non-diabetic while they are actually diabetic. It also indicates that 30% of non-diabetic dataset used for testing two (2) of them where False Positive, meaning they were predicted to be wrongly non-diabetic while forty-five (45) of the dataset where True Negative, meaning they were correctly predicted to be non-diabetic.



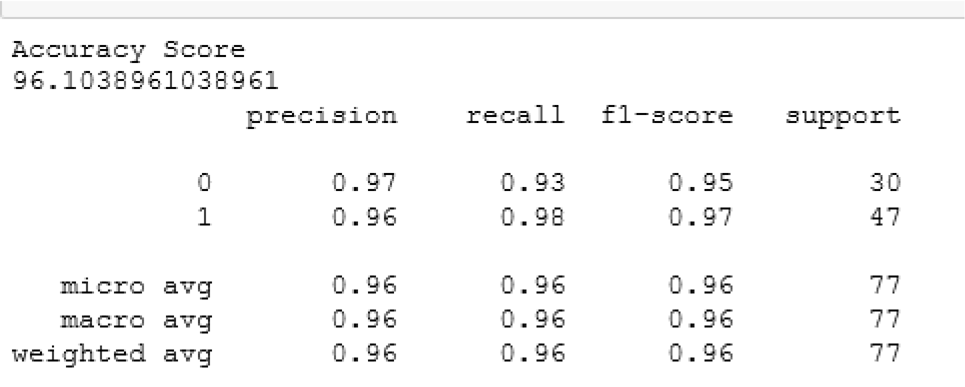
```
confusion matrix:
 [[23  7]
 [ 2 45]]
```

Figure 4 Confusion Matrix for KNN

Results for Decision Tree Algorithm

The second model was also designed was for predicting the diabetes status of an individual using the Decision Tree algorithm. This experiment was conducted using the Decision tree Classifier Function from the python programming library. During the experiment, 70% of the data was used for training while the remaining 30% was used for testing considering a random sate of 100.

Which amounted to accuracy of 96.10%, with precision of 0.96 for ND and 0.97 for D, Recall of 0.98 for ND and 0.93 for D, F1 Score of 0.97 for ND and 0.95 for D and Support of 47 for ND and 30 for D.



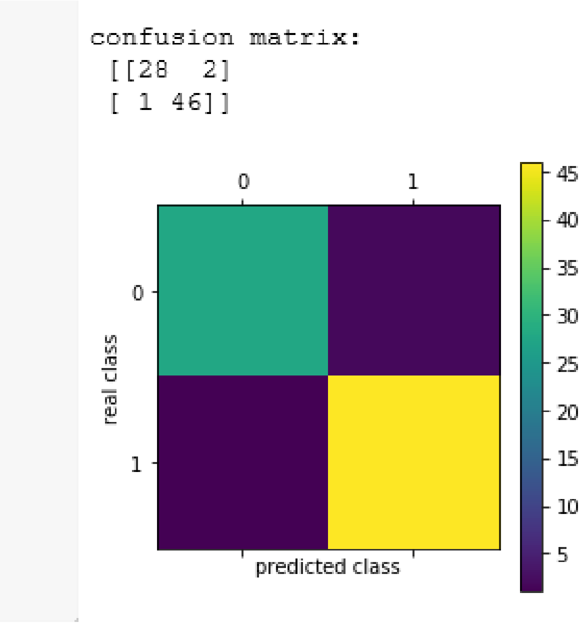
```
Accuracy Score
96.1038961038961
              precision    recall  f1-score   support

           0       0.97      0.93      0.95        30
           1       0.96      0.98      0.97        47

   micro avg       0.96      0.96      0.96        77
   macro avg       0.96      0.96      0.96        77
weighted avg       0.96      0.96      0.96        77
```

Figure 5 Decision Tree Results

The Decision Tree Algorithm produced a better confusion matrix which showed that from the 30% of diabetic dataset used for testing 28 of them where true positive i.e they were predicted to be correctly diabetic and 2 of them where False Negative, meaning they wrongly predicted to be non-diabetic while they are actually diabetic. It also indicates that 30% of non-diabetic dataset used for testing 1 of them where False Positive, meaning they were predicted to be wrongly non-diabetic while forty-six (46) of the dataset where True Negative, meaning they were correctly predicted to be non-diabetic.



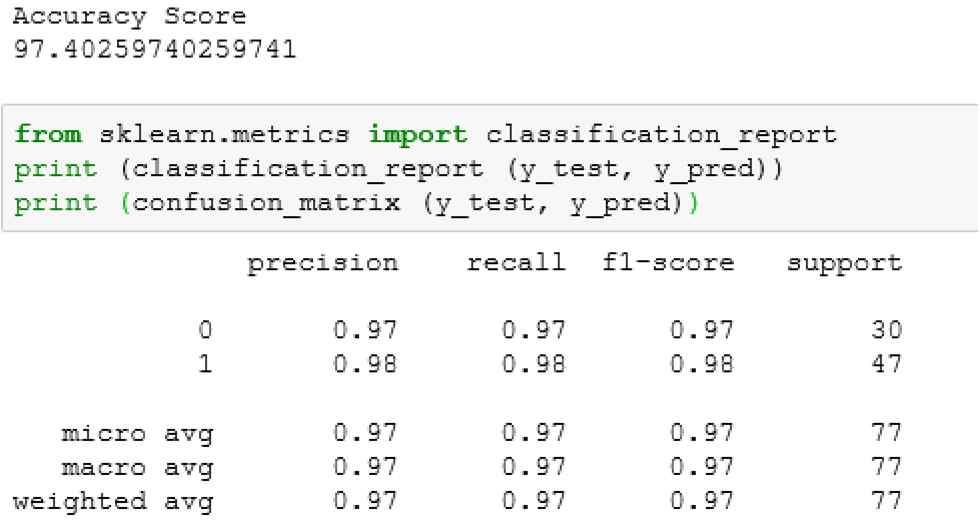
```
confusion matrix:
 [[28  2]
 [ 1 46]]
```

Figure 6 Decision Tree Algorithm Confusion Matrix

Results for ANN

The third model was designed for predicting the diabetes status of an individual using the Artificial Neural Networks (ANN). This experiment was conducted using the MLP Classifier from the SKlearn.Neural network function in the python programming library. During the experiment, 70% of the

dataset were used for training while the remaining 30% was used for testing, while considering three (3) hidden layers with twelve (12) neurons in each of the layers using a maximum iteration of 600.
Which produced an accuracy of 97.40%, with precision of 0.98 for ND and 0.97 for D, Recall of 0.98 for ND and 0.97 for D, F1 Score of 0.98 for ND and 0.97 for D and Support of 47 for ND and 30 for D.



```
Accuracy Score
97.40259740259741
```

```
from sklearn.metrics import classification_report
print (classification_report (y_test, y_pred))
print (confusion_matrix (y_test, y_pred))
```

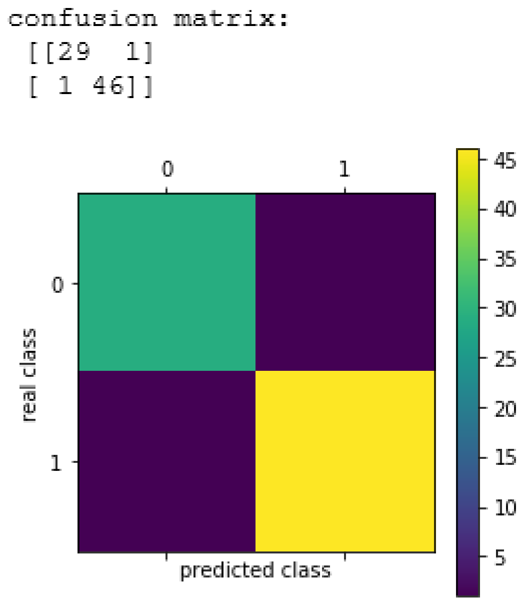
```
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        30
           1       0.98      0.98      0.98        47

   micro avg       0.97      0.97      0.97        77
   macro avg       0.97      0.97      0.97        77
weighted avg       0.97      0.97      0.97        77
```

Figure 7 Results for ANN Classifier

The ANN produced the best confusion matrix which showed that from the 30% of diabetic dataset used for testing twenty-nine (29) of them where true positive i.e they were predicted to be correctly diabetic and only one (1) of them where False Negative, meaning they wrongly predicted to be non-diabetic while they are actually diabetic. It also indicates that of the 30% of non-diabetic dataset used for testing only one (1) of them where False Positive, meaning they were predicted to be wrongly non-diabetic while forty-six (46) of the dataset where True Negative, meaning they were correctly predicted to be non-diabetic.



```
confusion matrix:
 [[29  1]
 [ 1 46]]
```

Figure 8 ANN Confusion Matrix

Comparison of the Models

From the results obtained for each of the model. Considering the different performance evaluation techniques, it is evident that they produced different performance levels as shown in the table below.

Table 2 Results Comparison

Model	*Class of diabetes*	*Accuracy*	*Precision*	*Recall*	*F1 Score*	*Support*
K Nearest Neighbors (KNN)	Diabetic	88.31%	0.92	0.77	0.84	30
	Non-Diabetic		0.87	0.96	0.91	47
Decision Trees	Diabetic	96.10	0.97	0.93	0.95	30
	Non-Diabetic		0.96	0.98	0.97	47
Artificial Neural Networks (ANN)	Diabetic	97.40	0.97	0.97	0.97	30
	Non-Diabetic		0.98	0.98	0.98	40

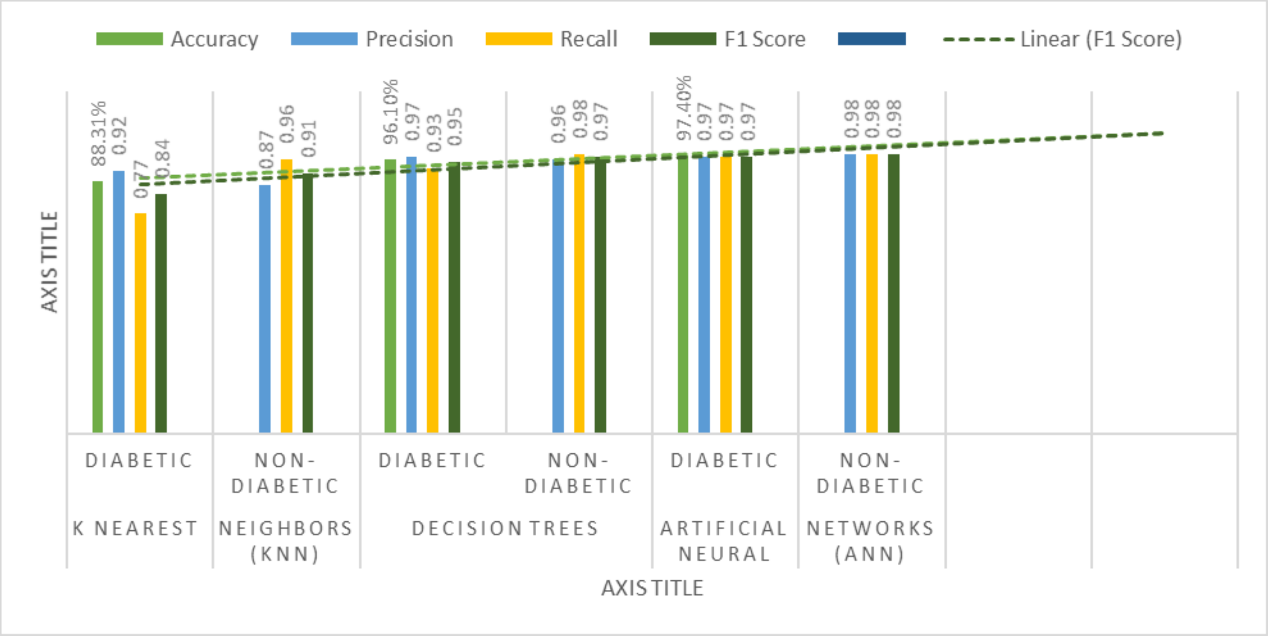


Figure 9 Chart Showing the Performance of the Models

From the results gotten in Table 2 and the chat displayed in Figure 9 which shows the various performance levels it shows that ANN produces the best performance with an accuracy of 97.40%. Therefore, ANN would be used to predict diabetes using some of the data.

More Metrics Considered For the ANN Model

After training the models and considering the evaluation criteria where the ANN outperformed other supervised learning algorithms. The model was saved and loaded for further predictions using a test dataset to predict respective output.

Model Accuracy

It is one of the metrics for the evaluation of the model Figure 9 showing a plot of accuracy against epoch where 100 epochs were considered in this case. The figure illustrates how the accuracy of the model improved both during training and testing, as the number of epochs increased.

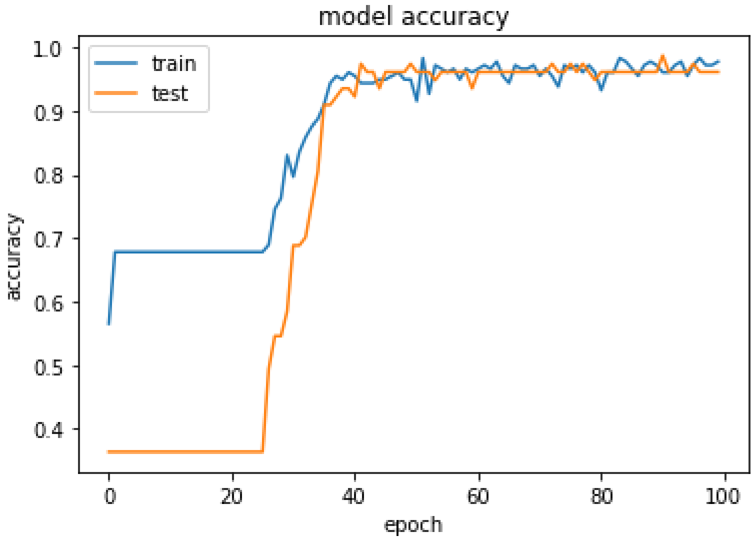


Figure 10 Illustrating ANN Model Accuracy against epoch During Training and Testing

Loss also being one of the metrics for the evaluation of the model figure (4.31) shows a plot of loss against epoch where 100 epochs were considered in this case. The figure illustrates how the loss of the model reduced with the increased epochs both during training and testing.

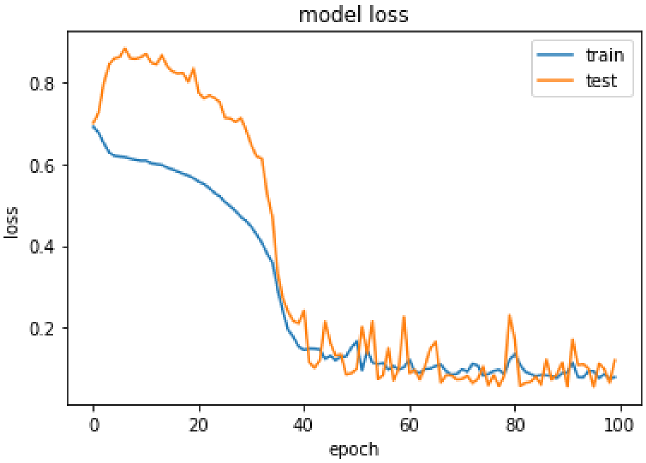
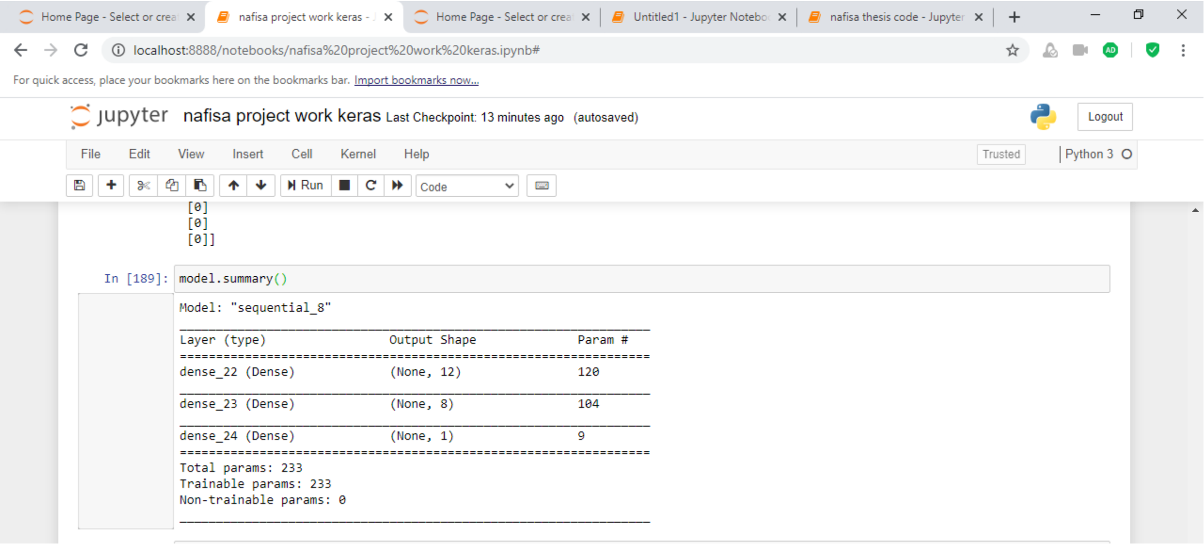


Figure 11 Illustrating ANN Model Loss against epoch during Training and Testing

Predicting New Outputs from the Model

After considering different metrics for the ANN model, the model was saved and summarized in other to be implemented in making fresh predictions to validate its accuracy in the prediction of the possible diabetic status of a person based on data inputs fed into the neural network model. The schematic of the model is shown in Figure 11 and the predicted output are also shown in the table in Table 2.


Home Page - Select or crea
nafisa project work keras -
Home Page - Select or crea
Untitled1 - Jupyter Notebo
nafisa thesis code - Jupyter
localhost:8888/notebooks/nafisa%20project%20work%20keras.ipynb#
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
jupyter
nafisa project work keras Last Checkpoint: 13 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Help
Trusted
Python 3
Run
Code

```
        [0]
        [0]
        [0]]

In [189]: model.summary()

Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_22 (Dense)             (None, 12)                120
_________________________________________________________________
dense_23 (Dense)             (None, 8)                 104
_________________________________________________________________
dense_24 (Dense)             (None, 1)                 9
=================================================================
Total params: 233
Trainable params: 233
Non-trainable params: 0
_________________________________________________________________
```

Figure 12 Summary of the ANN Model

Table 3 Test Data Showing the Given Output and the Predicted Output for Model Validation

AGE	SEX	NOP	GL	BMI	BP	WEIGHT	HEIGHT	RE	GIVEN OUPUT	PREDICTED OUPUT
75	MALE	0	3.1	22.94	180/90	64	1.67	1	NON-DIABETIC	[1]
42	MALE	0	3.5	28.06	130/80	94	1.83	1	NON-DIABETIC	[1]
72	MALE	0	4	20.08	160/90	54	1.64	1	NON-DIABETIC	[1]
64	MALE	0	3.3	24.69	180/90	64	1.61	0	NON-DIABETIC	[1]
40	MALE	0	3.7	22.68	180/90	64	1.68	1	NON-DIABETIC	[1]
55	MALE	0	3.9	29.75	140/80	88	1.72	0	NON-DIABETIC	[1]
57	MALE	0	4.7	33.31	150/90	94	1.68	1	NON-DIABETIC	[1]
60	MALE	0	4.2	31.06	160/110	104	1.83	1	NON-DIABETIC	[1]
60	MALE	0	3.8	22.79	160/80	69	1.74	1	NON-DIABETIC	[1]
45	FEMALE	5	4.2	26.18	190/110	82	1.77	1	NON-DIABETIC	[1]
70	FEMALE	10	3.5	22.81	210/120	52	1.51	1	NON-DIABETIC	[1]
58	FEMALE	0	3.1	24.75	150/80	61	1.57	0	NON-DIABETIC	[1]
45	FEMALE	9	4.1	40.39	170/90	106	1.62	0	NON-DIABETIC	[1]
65	FEMALE	7	3.4	22.41	160/120	64	1.69	0	NON-DIABETIC	[1]
70	FEMALE	5	3	23.33	200/90	62	1.63	1	NON-DIABETIC	[1]
55	FEMALE	10	4.6	26.32	130/80	60	1.51	1	NON-DIABETIC	[1]
60	FEMALE	9	3.9	17.8	170/90	45	1.59	1	NON-DIABETIC	[1]
65	FEMALE	12	3.1	30.86	150/80	78	1.59	0	NON-DIABETIC	[1]
49	FEMALE	0	3.6	27.99	150/80	79	1.68	1	NON-DIABETIC	[1]
55	FEMALE	10	3.9	27.55	190/100	82	1.55	0	NON-DIABETIC	[1]
30	FEMALE	5	4.4	35.66	140/100	89	1.58	1	NON-DIABETIC	[0]
60	FEMALE	6	3.7	37.11	190/110	95	1.6	1	NON-DIABETIC	[1]

55	FEMALE	2	3.6	28.17	190/120	73	1.61	0	NON-DIABETIC	[1]
72	MALE	0	7.4	30.45	140/80	88	1.7	0	DIABETIC	[0]
43	MALE	0	15.6	17.44	120/90	51	1.71	0	DIABETIC	[0]
65	FEMALE	15	6.5	22.22	170/110	52	1.53	0	DIABETIC	[0]
55	FEMALE	10	6.8	32.84	160/80	82	1.72	0	DIABETIC	[0]
50	FEMALE	10	7.6	21.72	150/70	57	1.62	1	DIABETIC	[0]
41	FEMALE	0	10.2	27.55	130/70	62	1.5	1	DIABETIC	[0]
42	FEMALE	9	6.5	25.81	130/70	62	1.55	1	DIABETIC	[0]
45	FEMALE	3	5	31.63	140/90	83	1.62	1	DIABETIC	[0]
70	FEMALE	4	6.07	25.34	110/60	54	1.46	1	DIABETIC	[0]
55	FEMALE	8	10.8	30.06	150/90	76	1.59	1	DIABETIC	[0]
45	FEMALE	9	6.6	22.3	180/100	60	1.64	0	DIABETIC	[0]
45	FEMALE	10	4.2	44.13	110/80	100	1.55	0	DIABETIC	[1]
49	FEMALE	11	9.5	31.25	180/100	79	1.59	0	DIABETIC	[0]
59	FEMALE	9	6.2	36.45	180/110	91	1.58	0	DIABETIC	[0]
35	FEMALE	7	11.1	26.06	120/70	61	1.53	0	DIABETIC	[0]
57	FEMALE	10	6.8	26.57	170/100	68	1.6	0	DIABETIC	[0]
55	MALE	0	7.3	19.88	150/90	63	1.78	0	DIABETIC	[0]
65	MALE	0	4.9	20.95	190/90	55	1.62	0	DIABETIC	[0]
55	MALE	0	5.9	19.95	180/130	57	1.69	0	DIABETIC	[0]
58	MALE	0	9.5	28.73	180/100	91	1.78	0	DIABETIC	[0]
35	MALE	0	17.7	19.37	90/60	54	1.67	0	DIABETIC	[0]
40	FEMALE	0	8.5	26.02	140/60	45	1.51	0	DIABETIC	[0]
55	FEMALE	9	6.5	26.02	160/80	57	1.48	0	DIABETIC	[0]
46	FEMALE	13	12.2	41.92	170/90	110	1.62	0	DIABETIC	[0]
62	FEMALE	14	13	24.98	140/80	60	1.55	0	DIABETIC	[0]
51	FEMALE	9	11.08	24.52	150/90	62	1.59	1	DIABETIC	[0]
62	FEMALE	14	13	24.98	140/80	60	1.55	0	DIABETIC	[0]

CONCLUSION

Research has shown that in recent times that DM is the leading cause of death in both developed and developing countries. This has further estimated the number to double in a few decades to come. Machine learning possesses a very good ability to change the risk of diabetes for better, as a result of advanced machine learning techniques and availability of huge diabetes dataset which would assist in prompt and precise prediction of the disease before it gets escalated. Early stage detection of diabetes, is a major key for treatment.

This work developed a supervised machine learning model (Artificial Neural Network) considering nine (9) attributes which includes; age, gender, number of pregnancies, blood pressure level, glucose level, weight, height and how regularly do they exercise after training the model, using pre-processed dataset an accuracy of 97.40%, recall of 0.97, precision of 0.97, F1 Score of 0.97 with a good confusion matrix using the python programming language which was implemented on Anaconda Jupyter notebook. The model further used fifty (50) validation dataset out of which forty-eight (48) results were accurately predicted. This Artificial Neural Network Model (ANN) would assist healthcare centres in taking precise and prompt decisions about the disease status at a quite early stage.

REFERENCES

Adeloye, D., Ige, J. O., Aderemi, A. V, Adeleye, N., Amoo, E. O., Auta, A., & Oni, G. (2017). *Estimating the prevalence , hospitalisation and mortality from type 2 diabetes mellitus in Nigeria : a systematic review and meta-analysis*. 1–16. https://doi.org/10.1136/bmjopen-2016-015424

Chawan, P. M. (2018). Logistic Regression and Svm Based Diabetes. *International Journal For Technological Research In Engineering*, *5*(6), 4347–4350.

Harz, H. H., Rafi, A. O., Hijazi, M. O., & Abu-Naser, S. S. (2020). Artfical Neural Network for Diabetes Using JNN. *International Journal of Academic Engineering Research (IJAER)*, *4*(10), 14–22.

Kaur, H., & Kumari, V. (2019). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, *xxxx*, 1–6. https://doi.org/10.1016/j.aci.2018.12.004

Liu, J., Tang, Z. H., Zeng, F., Li, Z., & Zhou, L. (2013). Artificial neural network models for prediction of cardiovascular autonomic dysfunction in general Chinese population. *BMC Medical Informatics and Decision Making*, *13*(1). https://doi.org/10.1186/1472-6947-13-80

Modern, S. (2019). A critical review on machine learning algorithms and their applications in pure sciences. *Research Journal of Recent Sciences*, *8*(1), 14–29.

Pradhan, N., Rani, G., Dhaka, V. S., & Poonia, R. C. (2020). Diabetes prediction using artificial neural network. *Deep Learning Techniques for Biomedical and Health Informatics*, *121*, 327–339. https://doi.org/10.1016/B978-0-12-819061-6.00014-8

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*(1), 1–19. https://doi.org/10.1186/s40537-019-0175-6

Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications*, *36*(4), 8610–8615. https://doi.org/https://doi.org/10.1016/j.eswa.2008.10.032

Tymvios, F. S., Michaelides, S. C., & Skouteli, C. S. (2008). Estimation of surface solar radiation with artificial neural networks. *Modeling Solar Radiation at the Earth's Surface: Recent Advances*, 221–256. https://doi.org/10.1007/978-3-540-77455-6_9

Uloko, A. E., Musa, B. M., Ramalan, M. A., Gezawa, I. D., Puepet, F. H., Uloko, A. T., Borodo, M. M., & Sada, K. B. (2018). Prevalence and Risk Factors for Diabetes Mellitus in Nigeria: A Systematic Review and Meta-Analysis. *Diabetes Therapy*, *9*(3), 1307–1316. https://doi.org/10.1007/s13300-018-0441-1

Zahran, B. (2017). A Neural Network Model for Predicting Insulin Dosage for Diabetic Patients. *International Journal of Computer Science and Information Security (IJCSIS)*, *14*(6), 770–777.