The African Journal of Information Systems

Volume 15 | Issue 1 Article 1

January 2023

A Predictive Model for Diabetes Mellitus Using Machine Learning Techniques (A Study in Nigeria)

Abraham Eseoghene Evwiekpaefe
Nigerian Defence Academy, Kaduna, Nigeria, aeevwiekpaefe@nda.edu.ng

Nafisat Abdulkadir
Nigerian Defence Academy, Kaduna, Nigeria, nafisahabdulkadir@gmail.com

A Predictive Model for Diabetes Mellitus Using Machine Learning Techniques (A Study in Nigeria)

Cover Page Footnote

Our appreciation goes to the management and staff of 44 Nigerian army reference Hospital, Kaduna and Yusuf Dan-Tsoho Memorial Hospial, Tudun Wada, Kaduna, Nigeria for their assistance during the data collection.

The African Journal of Information Systems

A Predictive Model for Diabetes Mellitus Using Machine Learning Techniques (A Study in Nigeria)

Research Paper

Volume 15, Issue 1, January 2023, ISSN 1936-0282

Abraham Eseoghene Evwiekpaefe
Nigerian Defence Academy
aeevwiekpaefe@nda.edu.ng

Nafisat Abdulkadir
Nigerian Defence Academy
nafisahabdulkadir@gmail.com

(Received January 2021, accepted May 2022)

ABSTRACT

Diabetes Mellitus (DM) is a metabolic disorder that occurs when the blood sugar level in the body is considered to be high, thereby resulting in inadequate insulin in the body leading to a myriad complications. The World Health Organization in 2021 indicated that in 2019, diabetes was the direct cause of 1.5 million deaths. Though some research has been carried out in the area of DM prediction in high-income countries, not much has been done in middle/low-income countries like Nigeria, using factors that are peculiar to their environment. This paper, therefore, aims to develop a machine learning model that predicts DM in individuals at an early stage. The study identified nine DM attributes and used three supervised learning algorithms of K Nearest Neighbors (KNN) decision trees, and artificial neural networks (ANN) to predict DM from a locally collected dataset in Nigeria. The results indicate that ANN produced the highest accuracy, at 97.40%.

Keywords

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

INTRODUCTION

The prevalence of diabetes is on the rise worldwide. The International Diabetes Federation notes that there are about 382 million people living with diabetes in the world and that by 2035, this will be doubled, as 592 million (Pradhan et al., 2020). This rise might be narrowed down to the age bracket of the populace, rate of urbanization, and continuous approval of unhealthy lifestyles. Some of these factors have inadvertently resulted in key health challenges around the world (Liu et al., 2013). In 2021 The World Health Organization noted that about 8.5% of adults aged 17 years and older are diabetic. In 2013, 1.5 million deaths were linked to diabetes, while high blood glucose resulted in 2.3 million deaths (Pradhan et al., 2020). In the last 10 years, diabetes patients have doubled across the world (Harz et al., 2020). Over 200 million people are diabetic, with an annual predominance of seven percent globally (Zahran, 2017). Various diseases are preventable when promptly diagnosed (Temurtas et al., 2009).

In the case of diabetes mellitus, or insulin deficiency diabetes, cells are irresponsive to insulin, which is the most common form of diabetes found in about 90% of diabetes disease patients (Mohamed et al., 2002). Compared to Type 1 diabetes, in DM the body produces insulin, but the level of insulin the pancreas produces is inadequate or in some cases, the body does not properly utilize the insulin, which may result in the build-up of glucose within the body, leading to various deficiencies. Unfortunately, there is no existent cure for diabetes mellitus, however, with a healthy lifestyle and diet, and regular exercise to keep fit, the chances of living healthily with the disease can be enhanced; however, the absence of any one of these may lead to medication or insulin treatment (Pei et al., 2014)

Artificial intelligence techniques are widely applicable in areas that are useful to human-related fields such as medical diagnoses, in which a physician has to analyze a multitude of factors before diagnosing an ailment, which makes a physician's job difficult and time-consuming. Machine learning and data mining techniques have been considered very helpful in the design of automatic diagnosis systems for various health conditions (Adeloye et al., 2017). Several studies have been conducted in the field of prediction for several diseases recently, where some clinicians now make use of machine learning models to predict different diseases (Kaur & Kumari, 2020; Modern, 2019; Pradhan et al., 2020). It is therefore imperative to design a diabetes classifier that is convenient, accurate, and cost-efficient.

In recent times, many methods and newer algorithms have been discovered that can be used to mine biomedical datasets for hidden information (Modern, 2019). This research develops a model with a high degree of accuracy for the prediction of DM in patients at an early stage, before it becomes escalated to a point of morbidity or mortality, using some supervised learning algorithms. The study was conducted on patients from some selected hospitals within the Kaduna metropolis in Nigeria. This research will contribute to the Nigerian health sector and the African health sector in general, by providing people with accurate prior knowledge about their health status as related to diabetes mellitus. This is likely to reduce the rate of complications, morbidity, and mortality being caused by this disease.

LITERATURE

Related Works

Zahran (2017) proposed an ANN predictive model for the appropriate quantity of insulin a diabetic patient would require, using a dataset of 180 patients. These were trained and tested, using data obtained from several patients. This contained factors such as length, weight, gender, and blood sugar. The proposed ANN model produced a good result in predicting the appropriate amount of insulin dosage with an average accuracy of 96.5% and an average prediction error of four percent. This research also showed that ANN can be used in the successful prediction of diabetes as a disease.

Uloko et al. (2018) researched the prevalence of risk factors for DM in Nigeria. In conducting this research, 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 a subgroup-specific DM prevalence was used to account for inter-study and intra-study heterogeneity. The results show that the frequency of DM in Nigeria has been on the increase in all affected regions of the country, with the south-south region having the highest degree in the geopolitical zones. Urbanization, physical inactivity, aging and unhealthy diet were identified as key risk factors for DM amongst Nigerians.

Chawan (2018) conducted research aimed at developing a system that can predict diabetes at an early stage in patients with high accuracy by combining the results of different machine learning techniques. The research predicted diabetes using two different supervised machine learning methods of support vector machine (SVM) and logistic regression. It considered seven patient features. They concluded that SVM showed a better performance, with an accuracy of 79% compared to logistic regression which had a performance accuracy of 78%. This research showed that a supervised learning algorithms can be useful in the prediction of diabetes.

Sneha & Gangil (2019) conducted research aimed at selecting the significant attributes to design a prediction algorithm using machine learning and found the best classifier to give results closest to clinical results using WEKA, which is a predictive analysis tool. They reached a conclusion showing the successful use of supervised learning algorithms for the disease prediction after considering 11 significant attributes, where both the random forest algorithm, and decision tree (DT) algorithm produced the highest specificity, with 98.20% and 98.00%, respectively, while the naïve Bayesian outcomes showed the best results in terms of performance accuracy, with 82.30%.

Modern (2019), in a review article, presented the various machine learning types, different algorithms, and how these have proven to be successful in different areas of research over the years. They also discuss various applications of machine learning in various fields, ranging from self-driving cars to weather forecasting, to the diagnosis of cancer. The research concluded that machine learning has a significant ability to assist doctors and clinicians in the field of medicine and life sciences to quicken methods of diagnosis with a higher level of precision.

Pradhan et al. (2020) predict whether an individual had diabetes or not, identify the type of diabetes, and predict the survival rate of the patient when predicted positive, using the University of California Irvine Machine Learning (UCI) repository Pima Indian Diabetes Dataset (PIDD), which contains nine 9 different attributes. The research aimed at minimizing the error function in training a neural network. As a sequel to the ANN model, the neural network average error function was recorded at 0.01, with a prediction accuracy of 87.3%.

Kaur & Kumari (2020) developed five different models for the detection of diabetes using linear kernel support vector machine (SVM-linear), radial basis kernel support vector machine, KNN, ANN, and multifactor dimensionality reduction algorithms. Feature selection of dataset was done with the help of Boruta wrapper algorithm, considering some evaluation criteria. The experimental results indicated that all the models achieved good results, with the SVM-linear model providing a high degree of accuracy of 0.89 and precision of 0.88, respectively. This study also suggested the Boruta wrapper algorithm for feature selection, as it was able to achieve better accuracy. From the above-reviewed literature, it is important to note the prevalence of diabetes in Nigeria, although various research has been carried out on DM prediction in other countries using datasets that are peculiar to their environment. However, not much has been done to apply machine learning techniques in diabetes prediction, using dataset peculiar to the Nigerian environment. A hybrid method was used, which diverges from the extant literature. This study also narrowed this gap, by providing a machine learning model to assist in the accurate and timely prediction of diabetes mellitus.

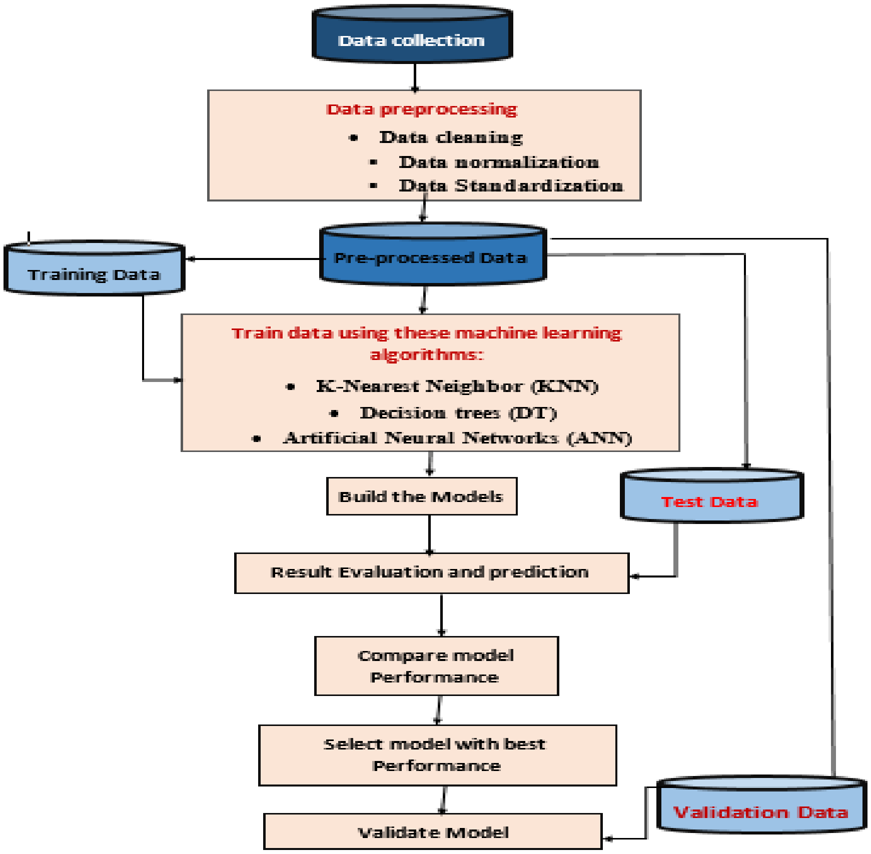
METHODOLOGY

The Adopted Method

This research is aimed at obtaining the best model that can predict diabetes in people at an early stage. The study therefore, adopted the knowledge discovery and data mining process model of data selection, pre-processing, transformation, data mining and interpretation/evaluation (Sharma & Saxena, 2018). The steps included: data selection (i.e. collection of data) carried out via oral interviews from the patients; pre-processing using label encoder; and transformation of the data using standard scaler. The data mining phase involved splitting of the data into 70% for training the models and 30% for testing the models, using three supervised learning algorithms namely, KNN, decision trees and ANN; and lastly interpretation and evaluation of data by comparison of the algorithms' respective performance considering some evaluation criteria like the F1 score, precision, recall, confusion matrix and prediction accuracy. The paper further selected the model with the best prediction accuracy, and prediction of future occurrence of DM from the dataset was done using the model with the best performance accuracy. This adopted method was visualized in the proposed model in Figure 1.

Figure 1

Framework of the Proposed Model


Data collection
Data preprocessing
• Data cleaning
▪ Data normalization
▪ Data Standardization
Training Data
Pre-processed Data
Train data using these machine learning algorithms:
• K-Nearest Neighbor (KNN)
• Decision trees (DT)
• Artificial Neural Networks (ANN)
Build the Models
Test Data
Result Evaluation and prediction
Compare model Performance
Select model with best Performance
Validation Data
Validate Model

Data Description

The dataset used in carrying out this research was obtained from the 44 Nigerian army reference Kaduna and Yusuf Dan-Tsoho Memorial Hospital, in Tudun Wada, Kaduna, Nigeria. The data was obtained from each of the individuals (patients) via oral interview, with the help of research doctors from Barau Dikko Hospital Kaduna. The data amounted to a total of 255 samples, which consist of two parts, namely: non-diabetic people and diabetic people; with 105 diabetic samples and 150 non-diabetic samples. The dataset includes nine physical examination indexes: age, sex, number of pregnancies, glucose level, blood pressure level, body mass index, height, weight and how regularly they exercise. These attributes are tabulated in Table 1. The data were entered into an excel sheet and saved in CSV format to make it accessible to the training algorithms.

Table 1

Tabular View Showing the Attribute, their Variable Type with a Specified Range

Attribute Number	Attribute	Variable Type	Attribute Description
A1	Age	Integer	24 years – 80 years

A2	Sex	Binary	Male = 1
			Female = 0

A3	Number of pregnancies	Integer	0 – 15
A4	Glucose level (mmol/l) or (mg/dl)	Real	5.0 mmol and above = High GL Below 5.0 mmol = Low GL
A5	Height (m)	Real	1.4m – 1.83m
A6	Weight (kg)	Real	37kg – 114kg
A7	BMI (kg/m^2)	Real	< 18.5 kg/m^2 = Underweight Between 18.5 - 24.9 = Normal Between 25.0 - 29.9 = Overweight > 30 = Obese
A8	Blood pressure (mm/hg)	Real	< 120/80 = Normal Between 120 - 129/ 80 = Elevated Between 30-139/80-80= Hypertension Stage 140/90 = High Blood Pressure
A9	Regular Exercise	Binary	Twice or less exercise in a week is (Irregular) = 0 Thrice or more exercise in a week is (regular) = 1

A10	Diabetes	Binary	Diabetic = 0
			Non-diabetic = 1

Note. mmol/l = millimole per liter; mg/dl = milligram per deciliter; mmol = millimole; GL = glucose level; m = meters; kg= kilogram; kg/m^2 = kilogram per meter squared; BMI = body mass index; mm/hg = millimeters of mercury.

Significance of Attributes

Age (A1): Although diabetes tends to have a higher prevalence in older people. There are some scenarios in which even young people can acquire this disease, necessitating their inclusion in the study cohort.

Sex (A2): According to extant literature, this disease is said to be more prevalent in women, though it is also found in both sexes.

Number of Pregnancies (A3): As discussed in A2, women are more prone to diabetes based on the reviewed literature, one of the reasons being gestational diabetes, which occurs when the glucose level increases during pregnancy (Wokoma et al., 2001). Although it decreases again after delivery, such women are prone to DM in the future.

Glucose level (A4): This is the major predictor for diabetes because it indicates the blood sugar level of an individual measured in milligram mole per liter or milligram per deciliter. There are different ways it could be measured, including: oral glucose test, random plasma glucose test, or fasting plasma glucose. In the case of this research, fasting plasma glucose was considered.

Height (A5): This attribute helps to identify the trend of diabetes as it relates to height and thereafter is utilized in the body mass index (BMI) calculation.

Weight (A6): This attribute helps to identify the trend of diabetes as related to various weights of the people, thereafter it is used in the computation of the BMI.

Body Mass Index (BMI) (A7): This is an attribute that is used to measure obesity. According to extant literature, this constitutes one of contributing factors to diabetes, and is measured in kilograms per meter squared (kg/m^2).

Blood Pressure Level (A8): This attribute is measured in millimeter per mercury, where the number at the top indicates the maximum pressure the heart applies while beating (systolic pressure), and the number below indicates the level of pressure in the arteries in-between heartbeats (diastolic pressure). The difference between the numbers captures pulse pressure.

Regular Exercise (A9): This attribute reveals how active a person is, where according to literature, a sedentary lifestyle may lead to diabetes. This was established via an oral interview, where exercise three times and above per week is considered regular exercise (1) exercise twice or less in a week is considered irregular exercise (0).

Class of diabetes (A10): This attribute, which is also the target variable, reveals the diabetes status of an individual based on the existing dataset. It is indicated by the variable of 1 or 0 for either diabetic or not diabetic, respectively.

Data Evaluation Criteria

Various deciding factors can be implemented to evaluate how well each of the algorithms performed, and to decide which algorithm outperforms the other, including factors such as the confusion matrix, recall, accuracy, and a few others which will be discussed accordingly.

Confusion Matrix

A table that is usually used to evaluate a classification model’s performance on a dataset for which the true output values are known is referred to as a confusion matrix. The four parameters provided by the confusion matrix table can be used to measure all performance, except for accuracy, which is obtained after prediction (Ting, 2017). These parameters consist of true positive (TP) and true negative (TN), which are correctly predicted observations, indicated in green; and false positives and false negatives, which are the observations that are wrongly predicted, indicated in red (as shown in Table 2).

Table 2

Confusion Matrix Table

		Predicted Class	
		Class= Yes	**Class = No**
Actual Class	**Class= Yes**	**True Positive (TP)**	**False Negative (FN)**
	Class = No	**False Positive (FP)**	**True Negative (TN)**

Note. Adapted from Ting, 2017.

True Positives (TP) - These refer to the values that have been correctly predicted. Given that the value of the actual class is positive, while the value of the predicted class is also positive (Saxena et al., 2009). For example, if the actual class indicates that a person is diabetic, and the predicted class value indicates that it is true, then it is a true positive.

True Negatives (TN)–This happens when the predicted value, as well as the actual value, are both negative. For instance, if the real value of the class shows that the individual is not diabetic, and the predicted class indicates it is true, then this is a true negative.

False Positives (FP) – This occurs when the predicted class is positive, while the actual class is negative. Given that the value of the actual class show that the individual is not diabetic, and the predicted class indicates otherwise, this would be tagged as a false positive.

False Negatives (FN) – This occurs when the actual class is positive, but the predicted class indicates that is negative. In an instance where the actual class value indicates that the individual is diabetic, and the predicted class indicates that the individual is non-diabetic, this is a false negative.

Accuracy

Accuracy refers to the proportion of correctly predicted output to the total outputs, where for the most part, it is assumed that a model can be tagged as the best model when it produces a high accuracy. There is no doubt that accuracy is a great measure of performance, but only in symmetric datasets, where there is a close margin between the values of the false positives and negatives (Ting, 2017). This necessitates the consideration of the other four parameters, which are provided by the confusion matrix table in deciding the performance of a model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

The precision metric refers to the ratio of the correctly predicted positive observations to that of the total predicted positive observations (Visa et al., 2011). This metric answers how many were actually diabetic from amongst all the diabetic samples. A high degree of precision defines a low false-positive rate.

$$Precision = \frac{TP}{TP + FP}$$

Recall/Sensitivity

The recall metric refers to the ratio of the correctly predicted positive observations to other observations in the Yes class. Recall shows how many were labelled diabetic from amongst all the passengers that were truly diabetic.

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

F1 score

The F1 score can be referred to as the average weight of precision over recall. It considers false positives and false negatives, which makes it difficult to understand when compared to accuracy, but the F1 Score is considered more useful than accuracy, particularly in cases where the class has uneven distribution (Visa et al., 2011). Accuracy is useful where both false positives and negatives have the same values, where it is advisable to use both Precision and Recall.

$$F1\ Score = \frac{2 * (Recall * Precision)}{(Recall + Precision)}$$

Data Pre-Processing

After the dataset has been collected, the collected data must be pre-processed to train the network efficiently. This procedure involves: (1) solving the problem of missing data; (2) data normalization; and (3) data standardization.

The problem of missing values is usually solved by taking the average of neighboring values, but in the case of this research, there was no missing data recorded. It is pertinent to carry out the data normalization procedure before passing the input data to the learning algorithm, due to the fact that to some degree, the mixing of variables could influence the learning algorithm, leading to a rejection of the variables and poor prediction accuracy (Tymvios et al., 2008). The library provided by Python, Scikit-Learn, which comes with a variety of data pre-processing modules, including a Label Encoder, was used to normalize the values of Sex, BP, and Output, so as to make them more understandable to the learning algorithm. Another Scikit-Learn pre-processing module, Standard Scaler, was adopted on the attributes Age, NOP, GL, BMI, Weight and BP so as to enable all the input values to fall within the range of -1 to 1, thereby making the input values appear easier to the training algorithm.

RESULT AND DISCUSSION

The results obtained from developing the models for the diabetes prediction system were achieved using the KNN, ANN and the decision tree algorithms. The models were trained using ten input variables (viz. sex, number of pregnancies, glucose level, blood pressure level, body mass index, height, weight, and regular exercise). The diabetic status of the individuals was used as the target output and 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 participants. Before training the models, data normalization and standardization were carried out on some of the attributes, using the pre-processing label encoder, and standard scaler module from the Python library Scikit-Learn in the Jupyter notebook, available on the Anaconda navigator. This was done to normalize and standardize some of the values to make them understandable to the learning algorithm.

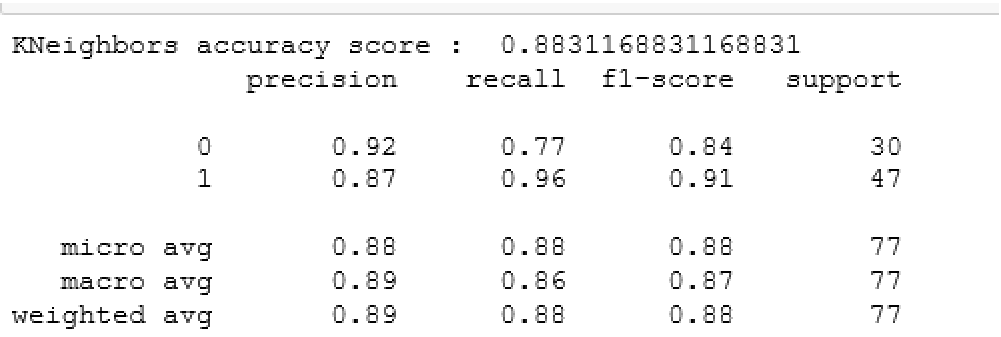
Results of KNN

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

In the experiment (as depicted in Figure 2), 70% of the data was used for training, while the remaining 30% was used for testing. This resulted in an accuracy of 88.31%, with a precision of 0.87 for non-diabetic (ND) and 0.92 for diabetic (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.

Figure 2

Results for the KNN Model



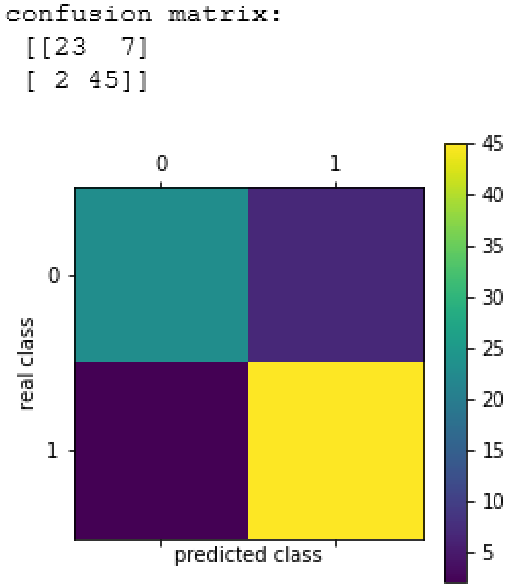
```
KNeighbors accuracy score :  0.8831168831168831
              precision    recall  f1-score   support

           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
```

Note. KNN = K Nearest Neighbors.

Additionally, the KNN produced a fair confusion matrix (see Figure 3), which showed that of the 30% of the diabetic dataset used for testing, 23 were true positive i.e., they were correctly predicted to be diabetic; and seven were false negative; meaning that they were wrongly predicted to be non-diabetic while being diabetic. It also indicated two as false positive, meaning that they were wrongly predicted to be non-diabetic; while 45 of the dataset were true negative, viz. correctly predicted to be non-diabetic.

Figure 3

Confusion Matrix for KNN



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

Note. KNN = K Nearest Neighbors.

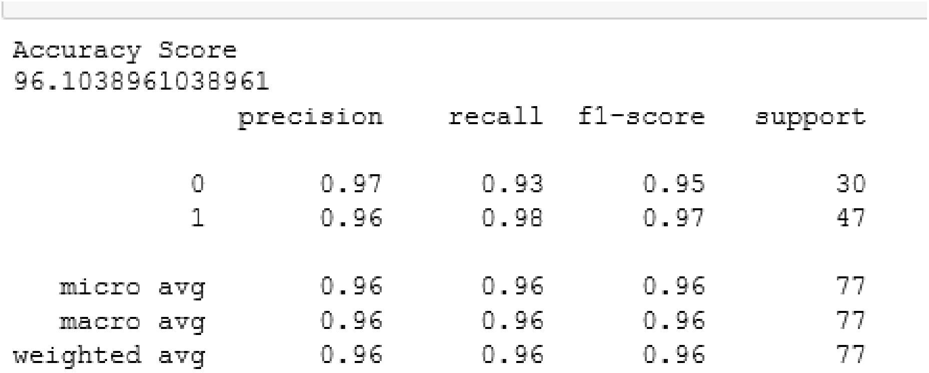
Results of Decision Tree Algorithm

The second model developed for predicting the diabetes status of the participating individuals used 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 state of 100.

This amounted to an accuracy of 96.10%, with a 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, as shown in Figure 4.

Figure 4

Decision Tree Results



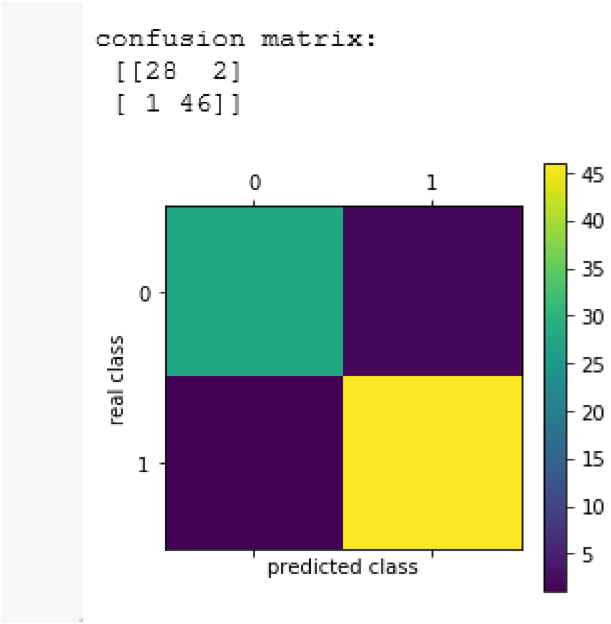
```
Accuracy Score
96.1038961038961
              precision    recall  f1-score   support

           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
```

Note. The F1-score is the average weight of precision over recall.

The decision tree algorithm produced a better confusion matrix as illustrated in Figure 5, which showed that of the 30% of the diabetic dataset used for testing, 28 were true positive, or predicted to be diabetic; and two were false negative, or wrongly predicted to be non-diabetic. It also indicated that of the 30% non-diabetic dataset, one was a false positive, or wrongly predicted to be non-diabetic; while 46 were true negative, or correctly predicted to be non-diabetic.

Figure 5

Decision Tree Algorithm Confusion Matrix



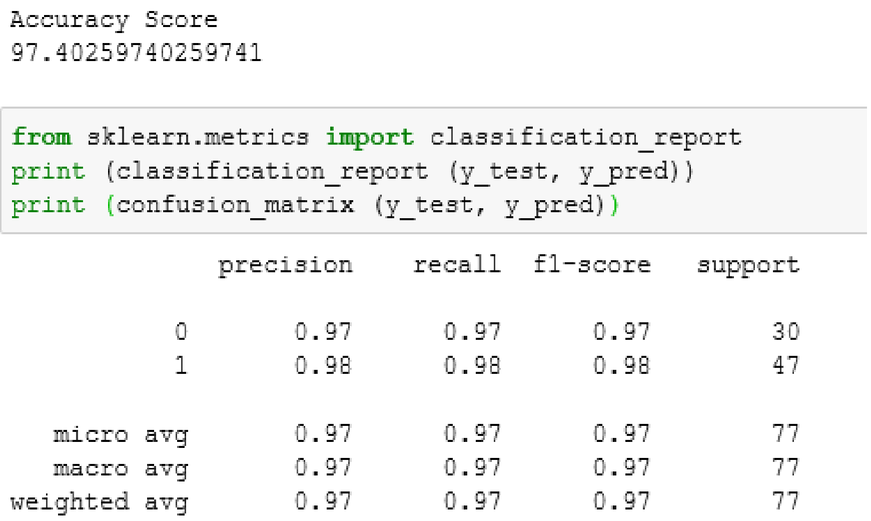
```
confusion matrix:
 [[28  2]
 [ 1 46]]
```

Results for ANN

The third model was developed for predicting the diabetes status of the participating individuals used the ANN. This experiment was conducted using the MLP Classifier from the SK learn 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. The experiment considered three hidden layers, with 12 neurons in each of the layers, using a maximum iteration of 600. The results (as shown in Figure 6) gave an accuracy of 97.40%, a 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.

Figure 6

Results for ANN Classifier



```
Accuracy Score
97.40259740259741
```

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

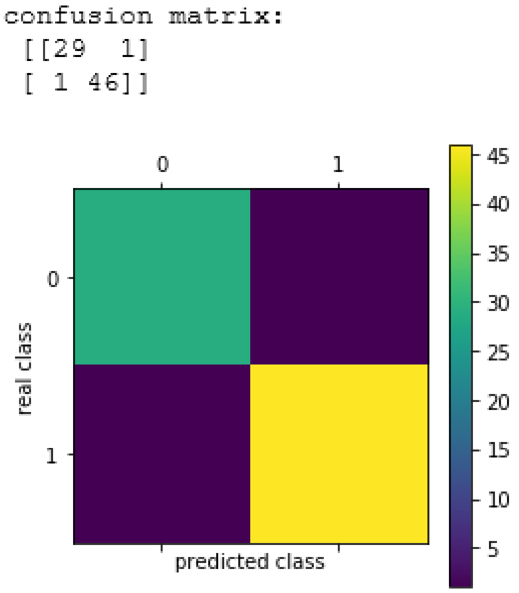
```
              precision    recall  f1-score   support

           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
```

Note. ANN = Artificial Neural Network.

The ANN produced the best confusion matrix (see Figure 7), which showed that of the 30% of the diabetic dataset used for testing, 29 of them were true positive i.e., they were correctly predicted to be diabetic, and only one of them where false negative, meaning they were wrongly predicted to be non-diabetic while diabetic. It also indicated that of the 30% of the non-diabetic dataset used for testing, only one was a false positive, meaning that it was wrongly predicted to be non-diabetic, while 46 of the dataset were true negative, meaning they were correctly predicted to be non-diabetic.

Figure 7

ANN Confusion Matrix



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

Note. ANN = Artificial Neural Network

Comparison of the Models

From the results obtained for each of the models, considering the different performance evaluation techniques, it is evident that they produced different performance levels, as shown in Table 2.

Table 3

Results Comparison

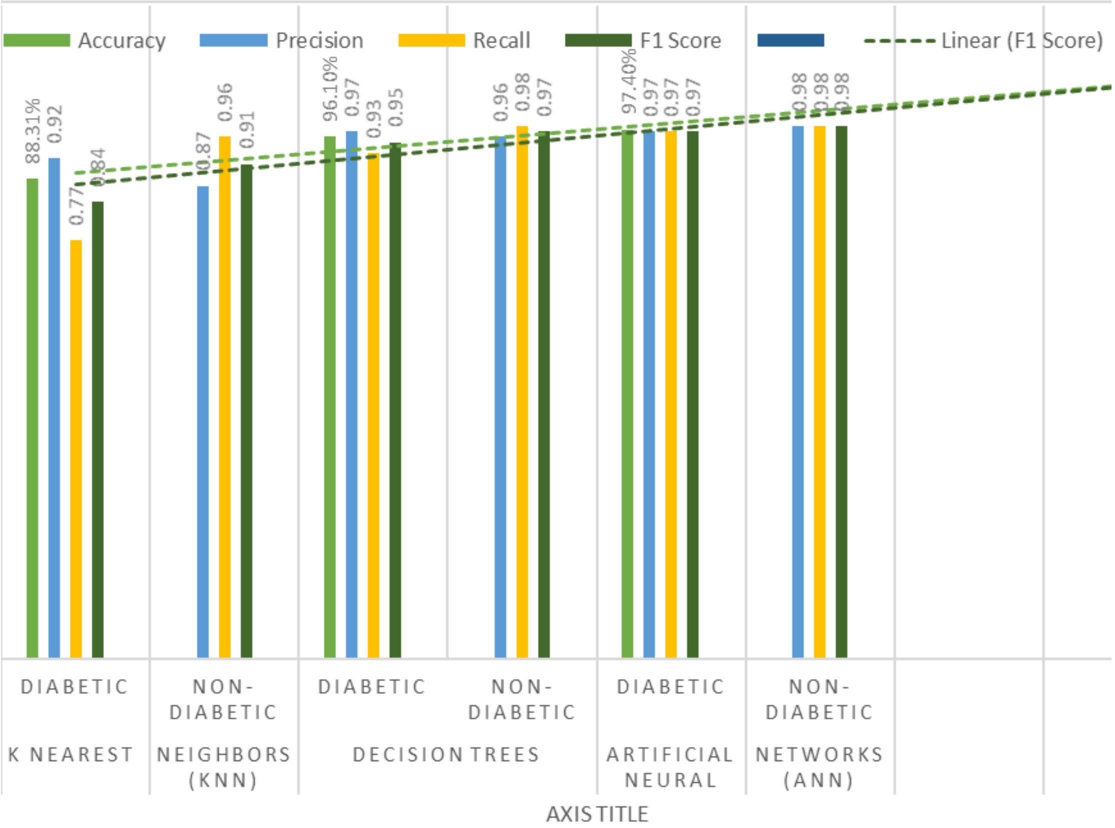
Model	Class of diabetes	Accuracy (%)	Precision	Recall	F1 Score	Support
K Nearest Neighbors (KNN)	Diabetic	88.31	0.92	0.77	0.84	30
	Non-Diabetic		0.87	0.96	0.91	47
Decision Trees	Diabetic	96.10	0.97	0.93	0.95	30
	Non-Diabetic		0.96	0.98	0.97	47
Artificial Neural Networks (ANN)	Diabetic	97.40	0.97	0.97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

Note. The F1-score is the average weight of precision over recall.

The results in Table 3 shows the performance of the three models, with ANN having the highest accuracy, precision, recall and F1 Score, followed by the decision tree model, and lastly the KNN model.

Figure 8

Performance of the Models



Note.The F1-score is the average weight of precision over recall.

From the results obtained in Table 2 and the chart displayed in Figure 9, which shows the various performance levels, it is evident that ANN produced the best performance, with an accuracy of 97.40 percent. Therefore, ANN was used to predict diabetes using some of the data.

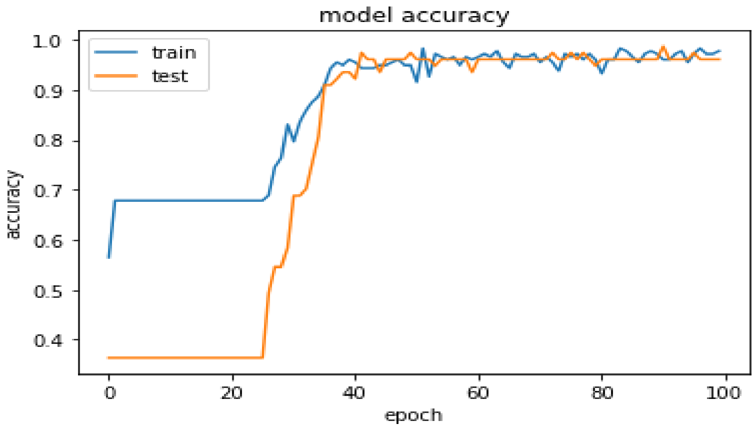
More Metrics Considered for the ANN Model

Model Accuracy

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

Figure 9

Illustrating ANN Model Accuracy Against Epoch During Training and Testing



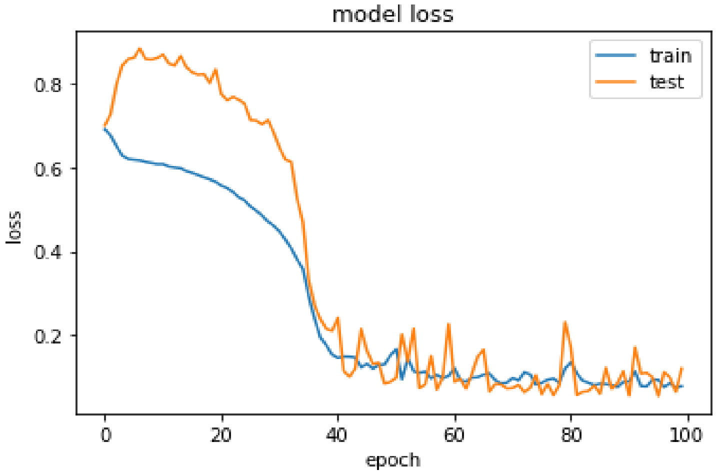
Note. ANN = Artificial Neural Network

Loss

The loss also being one of the metrics for the evaluation of the model (Figure 10) shows a plot of loss against epoch where 100 epochs were also considered in this case. The Figure illustrates how the loss of the model reduced with the increased epochs during both training and testing.

Figure 10

Illustrating ANN Model Loss Against Epoch During Training and Testing



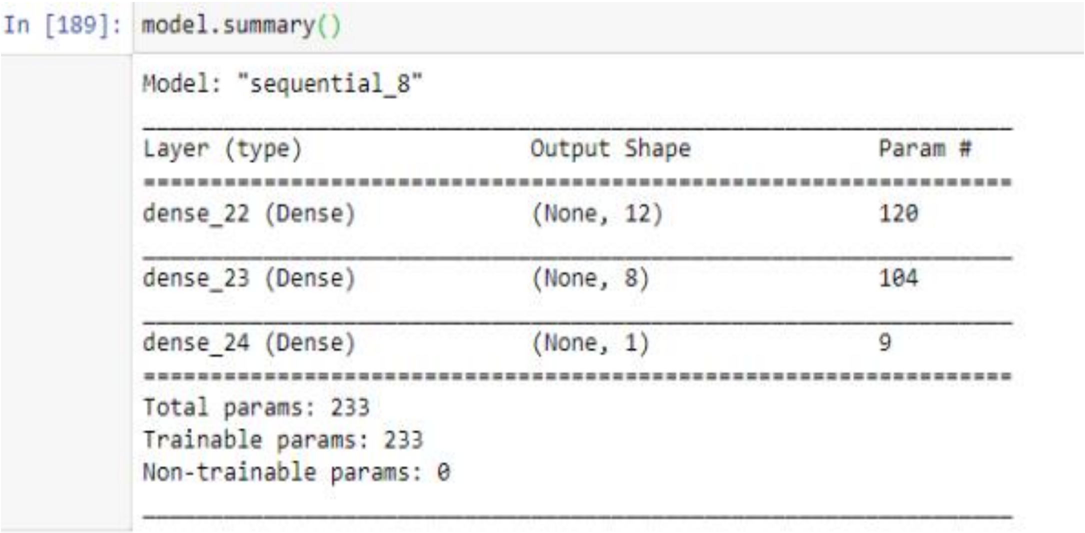
Note. ANN = Artificial Neural Network

Predicting New Outputs from the Model

After training the models and considering the evaluation criteria where the ANN outperformed the other supervised learning algorithms used, the model was saved and loaded for further predictions using a test dataset to predict its respective output. This was done in order 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 outputs are shown in Appendix A.

Figure 11

Summary of the ANN Model



```
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
          _________________________________________________________________
```

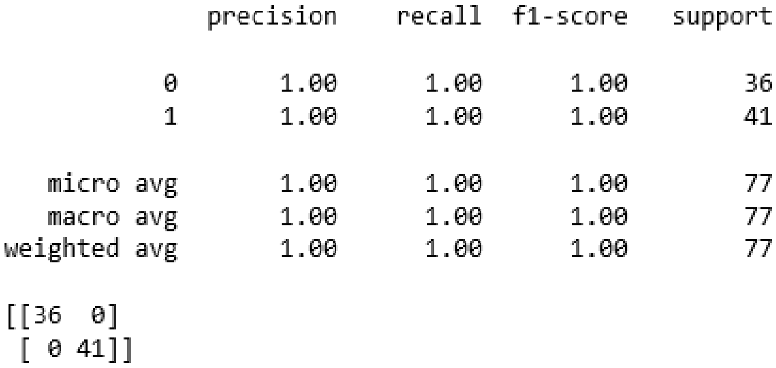
Note. ANN = Artificial Neural Network

K-Fold Cross-Validation and Hybridization of Models

In order to improve the performance of the models, a 10 K-Fold split was carried out with a random state of 10, where a hybrid model was further developed by implementing the Voting Classifier ensemble method to improve the performance of the prediction model, using the two best-performing models, i.e., the ANN and decision tree models. The ensemble produced a slightly more accurate model, with an accuracy of approximately 97.75%, and a precision, recall, F1-score and support of 100%, as shown in Figure 12 after 10 folds cross-validations.

Figure 12

Results of Ten (10)-Fold Cross-Validation and Model Hybridization



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        36
           1       1.00      1.00      1.00        41

   micro avg       1.00      1.00      1.00        77
   macro avg       1.00      1.00      1.00        77
weighted avg       1.00      1.00      1.00        77

[[36  0]
 [ 0 41]]
```

Note.The F1-score is the average weight of precision over recall.

CONCLUSION

Research has shown that in recent times, DM is one of the leading causes of death in both developed and developing countries. This has further been predicted to double in number in a few decades' time. Machine learning possesses an excellent ability to change the risk of diabetes for the better, as a result of advanced machine learning techniques, and the availability of a large diabetes dataset, which would assist in the prompt and precise prediction of the disease before it becomes escalated. Early-stage detection of diabetes is a major key to treatment.

This work developed three supervised machine learning models considering nine attributes, which included: age, gender, number of pregnancies, blood pressure level, glucose level, weight, height and how regularly they exercise. After training these models, using the pre-processed dataset, ANN outperformed the other supervised learning algorithms with an accuracy of 97.40%, recall of 0.97, precision of 0.97, and F1 score of 0.97. It also demonstrated an excellent confusion matrix predicting incorrectly only two of the sample data. The model further used a 50 validation dataset, out of which 48 results were accurately predicted. Thus, this paper believes that the ANN model would assist healthcare centers in taking precise and prompt decisions with regards to DM disease status at a quite early stage.

Research Contribution to Theory

In theory, a lot of research has been carried out to in the area of disease prediction using various supervised learning techniques such as KNN, support vector machines, and convolutional neural networks, while considering datasets that are peculiar to different parts of the world. Not much has been done in the application of these supervised learning techniques on a dataset that is peculiar to the Nigerian environment, thus preventing patients from knowing the possibility of being diabetic in the future, as this will help in reducing the high diabetes mortality rate caused by late diagnosis.

This research contributed to theory by narrowing the existing research gap from reviewed historical and recent literatures on the implementation of machine learning techniques considering a Nigerian data set for diabetes disease prediction in Nigeria. This research developed three supervised machine learning models based on: KNN, decision trees algorithm and ANN techniques on 250 datasets that comprised both diabetic and non-diabetic individuals. The result showed that the ANN model had the best prediction accuracy, at 97.40%. The model was further validated using 50 datasets.

Research Contribution to Practice

Currently, the Nigerian health system is generally poor, as there are a dearth of qualified doctors and quality equipment, mismanagement of funds, unqualified doctors in practice, misdiagnosis of patients, amongst other factors. Problems like these make the system less dependable by the day and makes patients reluctant to go to the hospital. This research provides some form of solution for diabetes in particular in Nigeria, as it provides a highly accurate model that will help clinicians project the possibility of patients being diabetic in the nearest future, considering a dataset that has been collected over time. This would solve the problem of misdiagnosis and late diagnosis, and consequent health risks.

This research contributes to practice by producing a 97.40% accurate and validated ANN model, which can help to predict the future occurrence of diabetes in Nigerians, based on their current health records. This model will help mitigate the diabetes death rate among Nigerians due to late clinical diagnosis. This model will be helpful to academicians, policy makers, government officials, medical personnel, and researchers in Nigeria as well as Africa at large.

Recommendation for Future Work

It is recommended that:

- more attributes can be included in the data set to test for a wider range of possible outcomes, and
- that the model's user interface be made more user-friendly.

LIMITATIONS OF THE WORK

Limitations of the work are:

- the system could produce more accurate results with more data, and
- not all the DM causing factors were added as a result of inaccessibility to the machinery required for the collection of such information.

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. BMJ open, 7(5), e015424.

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). Artificial neural network for diabetes using JNN. *International Journal of Academic Engineering Research (IJAER), 4*(10), 14–22.

Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*.

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), 3-23. https://doi.org/10.1186/1472-6947-13-80

Mohamed, E. I., Linder, R., Perriello, G., Di Daniele, N., Poppl, S. J., & De Lorenzo, A. (2002). Predicting type 2 diabetes using an electronic nose-based artificial neural network analysis. *Diabetes, Nutrition & Metabolism, 15*(4), 215–221.

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.

Pei, E., Li, J., Lu, C., Xu, J., Tang, T., Ye, M., Zhang, X., & Li, M. (2014). Effects of lipids and lipoproteins on diabetic foot in people with type 2 diabetes mellitus: A meta-analysis. *Journal of Diabetes and Its Complications, 28*(4), 559–564. https://doi.org/10.1016/j.jdiacomp.2014.04.002

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

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2009). Evaluating algorithm performance metrics tailored for prognostics. In *2009 IEEE Aerospace Conference,* (pp. 1–13). IEEE. https://doi.org/10.1109/AERO.2009.4839666

Sharma, N. and Saxena, A. (2018). A survey of data mining and knowledge discovery process model and its applications in database. *International Journal of Research and Analytical Reviews*, *5*(3), 1210 – 1214.

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/10.1016/j.eswa.2008.10.032

Ting, K. M. (2017). Confusion matrix. In Sammut, C. & Webb, G. I. (Eds.) *Encyclopedia of Machine Learning and Data Mining*, (p. 260). Springer. https://doi.org/10.1007/978-1-4899-7687-1_50

Tymvios, F. S., Michaelides, S. C., & Skouteli, C. S. (2008). Estimation of surface solar radiation with artificial neural networks. In Badescu, V. (Eds), *Modeling Solar Radiation at the Earth's Surface: Recent Advances*, (pp. 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.

Visa, S., Ramsay, B., Ralescu, A., & Knaap, E. (2011). Confusion matrix-based feature selection. *CEUR Workshop Proceedings*, *7*(10), 221-245.

Wokoma, F. S., John, C. T., & Enyindah, C. E. (2001). Gestational diabetes mellitus in a Nigerian antenatal population. *Tropical Journal of Obstetrics and Gynaecology*, *18*(2), 13–20. https://doi.org/10.4314/tjog.v18i2.14430

World Health Organization. (2021). Diabetes. Retrieved April 20, 2021, from https://www.who.int/news-room/fact-sheets/detail/diabetes

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.

Appendix A

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

AGE	SEX	NOP	GL	BMI	BP	WEIGHT	HEIGHT	RE	GIVEN OUTPUT	PREDICTED OUTPUT
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]

AGE	SEX	NOP	GL	BMI	BP	WEIGHT	HEIGHT	RE	GIVEN OUTPUT	PREDICTED OUTPUT
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]

Note. NOP = number of pregnancy; GL = glucose level; BMI = body mass index; BP = blood pressure; RE = regular exercise.