© FEB 2023 | IRE Journals | Volume 6 Issue 8 | ISSN: 2456-8880

Adoption of Machine Learning and Data Mining Tools in the Identification and Prediction of Diabetes Disease in Patients Using Classification Mining Algorithm

OGUOMA IKECHUKWU STANLEY[1], OBIALOR COLLINS CHIMEZIE[2], OBICHERE CHIGOZIE DANIEL[3], NJOKU TOCHUKWU STANLEY[4], MAGNUS CHINONSO OKERE[5]

[1] Department of Computer Science, University of Agriculture and Environmental Science, Umuagwo, Owerri, Nigeria

[2, 3, 4, 5] Department of Computer Science, Imo State University Owerri, Nigeria

Abstract- The aim of this paper is to adopt Machine Learning and Data Mining tools in predicting if patient diabetes is positive or negative using classification mining algorithm. The objective of the research includes to analyze a dataset using classification mining tool to predict or identify if a patient has diabetes or not, to use the analyzed data result to improve the health standard of diabetic patients suffering from the disease, to recommend the perfect data mining technique best for analyzing and predicting data. The research was motivated due to the high increase of diabetic patient witnessed in Nigerian Hospitals because of the high intake of Carbohydrate based on report issued by world health organization (WHO) on a yearly basis. Data mining methodology called classification algorithm was adopted while decision tree was used as the modeling tool. The data was analyzed with R and SAS Enterprise Miner. The experiments are done on Pima Indians Diabetes Database (PIDD) sourced from UCI machine learning repository. The result after the experiment shows that the use of classification algorithm and decision model is the best and more accurate method suitable for data prediction and hence has more percentage acceptance level of performance when it comes to health issues, therefore it could be adopted for future use by medical practitioners to make decision on the subject matter.

Indexed Terms- Artificial Intelligence, Machine Learning, Diabetics, Classification Model and Data Mining

I. INTRODUCTION

Nigeria is said to be the giant of Africa and as such has their own different food recipe according to tribe and culture in the country. As an African man, majority of the food consumption are more in Glucose and carbohydrate which is major contribution of sugar and which was reported as one of the factors that causes diabetes. A report from World Health Organization (WHO) on the rate of diabetes patients in Nigeria shows a huge increase and hence advice that proper measures be deployed in other to control the situation. Therefore in other to adopt technology towards providing solution to the situation, there is need to carry a study and to look into the biotechnology and work with big data.

Advances in biotechnology in the quest to produce high throughput and inexpensive data production made it possible for the ushering in of applied biology into the scientific area of big data. (Kavakiotis*et al.,* 2017). Aside the introduction of high performance sequencing methods adopted in big data today, such as super-resolution digital microscopy, mass spectrometry, magnetic resonance imagery (MRI) etc. These technologies produces huge amount of data but cannot carry out data analysis, interpretation, or extraction of knowledge (Kavakiotis*et al.,* 2017), hence the area of biological data mining and knowledge discovery is highly of paramount important.

Insulin as stated by Diabetes UK, Understanding Diabetes-Your key to better health (2003) is a hormone that provides glucose from the digestion of carbohydrate process in food, which enters into the body cells and can be used as a form of energy, that is to say, lacking of insulin in a body cell may result loss of energy. Hence absents of insulin or is not effective, glucose take control of the body and build its up in the blood. If a Patients with the issue of diabetes problem means, there is a huge issue with their insulin and hence it is very much of importance to build up necessary steps to either create insulin in the body or to help the patient manage its insulin level.

The birth of data mining tool into data discovery, prediction and extraction has helped researcher and scientist in numerous ways to uncover hidden facts both in physical and biological sciences. Data mining (DM) is one of the hidden extractions of predictive tool in progression from a very big databases, is now a new great technology by means of large potential to provide huge support to companies looking into the direction of large data discovery from knowledge warehouses.

It is now clear that the introduction of DM tool in data processing has help to predict the future trends and behavious, enabling various business owners create practical knowledge-driven selections in such that it could help to answer business queries that factually remained too time irresistible to resolve (Pravarti and Santosh; 2017). Nevertheless, the area of biological science has witnessed huge transformation because of the introduction of new technological tools, though there are still areas that have some issues when it comes to data discovery and making decisions. Hence the aim of this paper is to adopt Machine Learning and Data Mining tools in predicting if patient diabetes is positive or negative using classification mining technique. while the objectives of the research includes to analyze the diabetes dataset using classification mining tool in other to predict or identify if a patient has diabetes or not, to use the analyzed data result to improve the health standard of diabetic patients suffering from the disease, to recommend the perfect data mining technique best for analyzing and predicting data. This paper will make use of approaches to uncover or create models from data after a critical data analysis on the dataset has been concluded. It is far much important to note that abundance of data has helped so much to strengthened significantly data-oriented research in biological science as such researching in the field such as in the areas of application of prognosis and diagnosis relating to human life's, or human life quality in other to reduce diseases such as diabetes. Furthermore, for this study framework, effort were made to uncover so much literatures in the application of machine learning and data mining tools for discovery, management, identification and diagnosis of diabetic disease in patients. This paper is organized in as follows: Introduction: presents general introduction of machine learning (ML) and the importance of it in biological sciences, definition of diabetes and what happens when insulin is high or low in human body, it also looked at data mining (DM) and how it is an important tool for effective prediction and making discovery for future use, it also highlighted the objectives of the study and gave significant facts why adoption of machine learning and data mining tools are good for decision making. Literature Review: looks at generally the literature review on related works, machine learning modeling tools and technique, Methodology: methodology adoption for the study, proposed system diagram, system algorithm while Results: present experiment of R and SAS, experiment output, the model rules using R, the decision tree model rules using SAS, conclusion and recommendation of the study and then appendices.

II. LITERATURE REVIEW

According to Sajida*et al.*, (2016) reviewed the role of Adaboost and Bagging collaborative machine learning approaches (Nai-Arun and Sittidech., 2014) the researcher adopted J48 decision tree as the classifying modeling tool for diabetes mellitus. The adopted approach classifies if a patient has diabetics or not based on the diabetic risk factors. After the research, result shows that, Adaboost machine learning collaborative technique outperforms very much better comparatively bagging as well as a j48 decision tree.

Orabi, Kamal and Rabah (2016) worked on a system for diabetes prediction where their researches aim at predicting the rate of diabetic level in a patient at a particular age. Their system was achieved by the adoption of machine learning, by application of decision tree modeling tool. After the experiment, result obtained was quite satisfactory as the new system was able to perfectly predict the rate of diabetes incidents at a particular age with a higher accuracy using Decision tree (Priyam*et al.,* 2013) and (Esposito *et al.,* 1997). In same way, Pradhan*et al.,* (2014) developed a system using the Genetic Programming (GP). This system was used to carry out training and testing on a particular database for prediction of diabetes with the use of diabetic dataset sourced from UCI repository. (Sharief and Sheta 2014) and (Pradhan*et al.,* 2012) were able to provide optimal accuracy after a thorough compares of other related study on techniques were done. Rashid et al (2016) worked on a prediction model with two sub-modules to predict diabetes-chronic disease. The researchers adopted ANN (Artificial Neural Network) for the first module and the second was FBS (Fasting Blood Sugar). The system employed the decision tree technique which was used to detect symptoms of diabetes on patient's health. Nongyao*et al.,* (2015) designed an algorithm used to classify risk of diabetic mellitus patients. There work employed four different machine learning classification methods, namely: Decision tree, artificial neural network, logistic regression and naïve bayes. The robustness of the designed work was improved by employing a model known as Bagging and Boosting technique. After the experiment by the researcher, result shows that random forest algorithm gives very much optimum results among all other algorithms applied in the study.

III. METHODOLOGY

- CLASSIFICATION ALGORITHMS

Classification algorithm is stated by (Sidana 2017) as a supervised learning method were a computers are programmed to learn from set inputs as data and then uses the learning approach to classify new observations. The researcher further explained that the observed data set could be bi-class (example: identifying if a patient have diabetic test result of a patient is positive or negative or say a mail in an inbox is spam male or non-spam). Sometime, the class might be multi-class (Sidana 2017). Classification modeling can also be applied in areas like, speech recognition, handwriting identification, bio-metric identification or verification process, document classification (Sidana; 2017). The scholar outlined some of the few classification algorithms in machine learning as follows:

1. Linear Classifiers: Logistic Regression, Naive Bayes Classifier
2. Nearest Neighbor
3. Support Vector Machines
4. Decision Trees
5. Boosted Trees
6. Random Forest
7. Neural Networks

- ADOPTED ALGORITHM FOR THE STUDY

- DECISION TREES

Decision tree (Han and Kamber; 2001) could be seen as a type of tree structure typically in a form of flowchart design. These tree structures are used to carry out classification and prediction modeling of objects in a class in a form of nodes and internodes. Both root and the internal nodes are taken as the test cases in the modeling process which in terms used as a separator with different features(Han and Kamber; 2001). According to (Sidana; 2017), these decision trees uses a classification or regression models to form a tree structure. The structure breaks down a particular data set into various smaller and smaller subsets as the associated decision tree development id in progress. The researcher further noted that the decision tree is build up with the nodes and leaf nodes, where the decision nodes has two or more different branches while leaf nodes shows the classification or decision results(Sidana; 2017) stated. Figure 1: Illustrate the structure of a decision tree

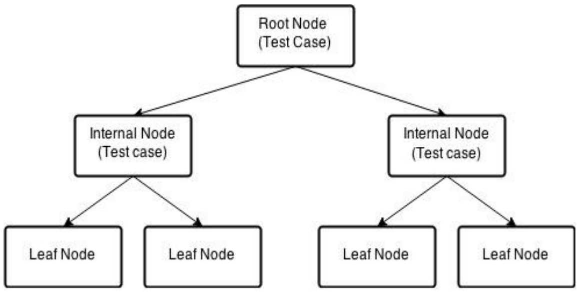


Figure 1: Illustrate the structure of a decision tree (Source: Pilani, Dubai, and Sumbaly; 2015)

Adoption of decision tree for this study did not just come but it was adopted because of its powerful technique for classification and prediction ability on a particular data set. Hence the identification and prediction of diabetes disease in a patient will have a very significant outcome after the analysis of the data set has been concluded and presented for future use. The proposed system diagram is shown in figure 2 below:

THE PROPOSED SYSTEM DIAGRAM

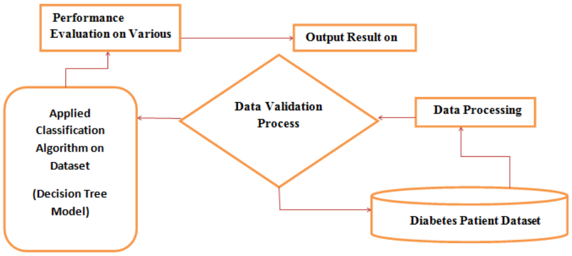


Figure 2: Diagram of the proposed model (Fieldwork 2022)

The above diagram illustrates how the proposed system loads the diabetes dataset in database from where the data processing and validation phase is done, if there is an error in after the data validation process, the system redirects the system back into the database to restart the process again but if the data validation process is successful, the application of the classification algorithm is applied on the dataset in this case (decision tree model), after decision tree process has been concluded, the system checks for Performance Evaluation on Various Measures and then display result of analysis on screen.

In summary, the researcher adopted the R and SAS Enterprise Miner for the analysis while dataset was gotten from Pima Indians Diabetes Database (PIDD) sourced from UCI machine learning repository https://data.world/data-society/pima-indians-diabetes-database.

Table 1: SYSTEM ALGORITHM

INPUT	Pima Indians Diabetes Database (PIDD) sourced from UCI machine learning repository https://data.world/data-society/pima-indians-diabetes-database.
OUTPUT	Decision Tree Predictive Model with leaf node either tested-positive or tested-negative on patient

IV. RESULTS

- EXPERIMENTS ON THE DATASET USING R

The first process was launching of the RStudio IDE after a successive launching, the following steps were done to design the model.

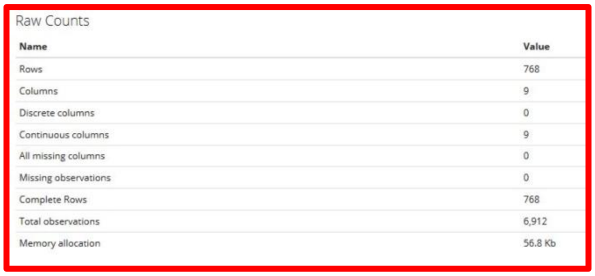
Step1: Loading packages to be used (that is libraries)
Step 2: Loading My Dataset to R Dataframe
Step 3: Exploring the data, at this stage, skimr::skim(diabetes) was used
Step 4: Converting outcome from numeric to factor and renaming them for easy understanding
Step 5: Plot Observations
Step 6: Checking For Missing Values in Each Variable
Step 7: Due to excessive missing values in skinthickness and insulin, the two variables where removed
Step 8: Replacing the missing values with the mean of each variable
Step 9: Converting to numeric, integer to avoid missing decimal values in the data set
Step 10: Normalizing the dataset

- MODEL BUILDING

Step 11: The data set was Split into two with the percentage of (75% = training and 25%=testing) respectively
Step 12: applying decision tree algorithm
Step 13: Using GINI MODEL
Step 14: Making Prediction from The Model built With Gini Model
Step 15: Generating Frequency Table to Create Tabular Results of Categorical Variable
Tabular results enable the researcher to know if the result produce from the prediction by the computer is correct when compared with the original one on the dataset.
Step 16: Confusion Matrix
Hence the confusion matrix is used for more accuracy on the rate at which the model predicts user data.
Step 17: Validating the Model On the test Dataset

Step 15, step 16 was carried out again for a more accurate prediction of the model.

EXPERIMENT OUTPUT



Raw Counts

Name	Value
Rows	768
Columns	9
Discrete columns	0
Continuous columns	9
All missing columns	0
Missing observations	0
Complete Rows	768
Total observations	6,912
Memory allocation	56.8 Kb

Figure 3: Basic Statistics of Diabetes data set (Fieldwork 2022)

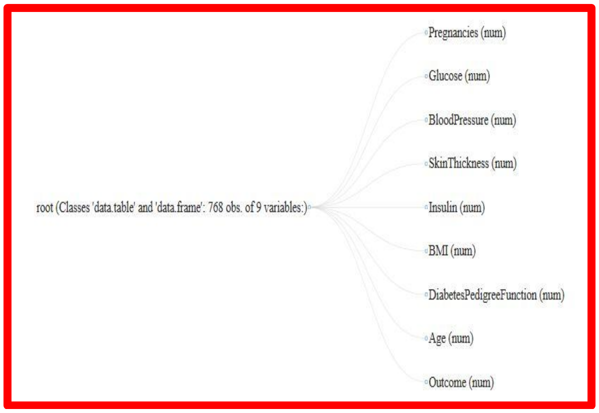


Figure 4: Diabetes data structure (Fieldwork 2022)

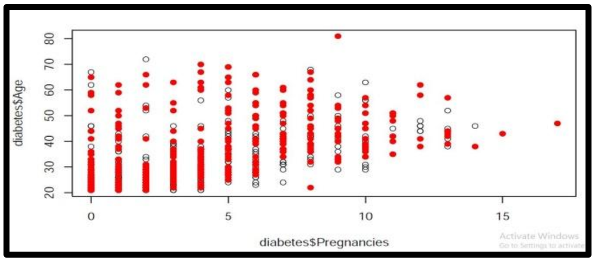
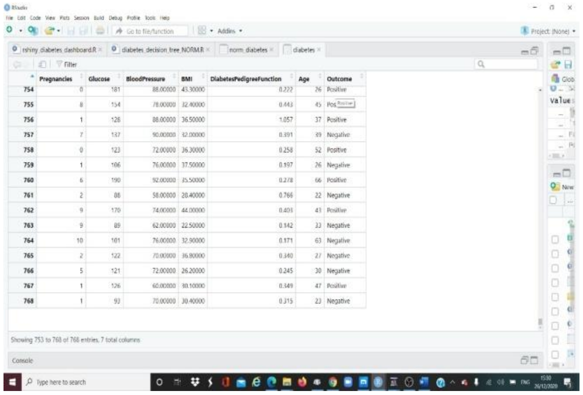


Figure 5: Frequency Plot of Age against Pregnancy (Fieldwork 2022)


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
rshiny diabetes dashboard.R
diabetes decision tree NORM.R
norm diabetes
diabetes
Filter
Console
Type here to search

	Pregnancies	Glucose	BloodPressure	BMI	DiabetesPedigreeFunction	Age	Outcome
754	0	181	88.00000	43.30000	0.222	26	Positive
755	8	154	78.00000	32.40000	0.443	45	Positive
756	1	128	88.00000	36.50000	1.057	37	Positive
757	7	137	90.00000	32.00000	0.391	39	Negative
758	0	123	72.00000	36.30000	0.258	52	Positive
759	1	106	76.00000	37.50000	0.197	26	Negative
760	6	190	92.00000	35.50000	0.278	66	Positive
761	2	88	58.00000	28.40000	0.766	22	Negative
762	9	170	74.00000	44.00000	0.403	43	Positive
763	9	89	62.00000	22.50000	0.142	33	Negative
764	10	101	76.00000	32.90000	0.171	63	Negative
765	2	122	70.00000	36.80000	0.340	27	Negative
766	5	121	72.00000	26.20000	0.245	30	Negative
767	1	126	60.00000	30.10000	0.349	47	Positive
768	1	93	70.00000	30.40000	0.315	23	Negative

Showing 753 to 768 of 768 entries, 7 total columns

Figure 6: exploration of the dataset (Fieldwork 2022)

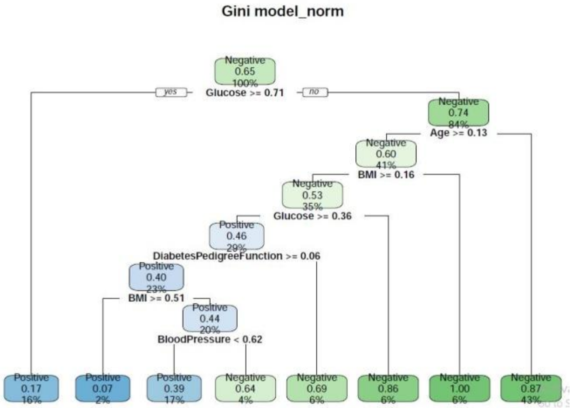


Figure 7: Gini model result on the data set (Fieldwork 2022)

Figure 7 above shows the decision tree model on how to predict if patient diabetes is positive or negative.

- THE MODEL RULES USING R

If (YES) patient Glucose is >= 0.71 it means the patient has diabetes (Positive with 16%) but if (NO) we have (Negative with 84%) else If (NO) patient Age is >= 0.13 it means the patient diabetes is (Negative with 43%) else if (NO) patients BMI >=0.16 it means the patient diabetes is (Negative with 6%) else if (NO) patient Glucose >=0.36 it means the patient diabetes is (Negative with 6%) but if (YES) it means patient has diabetes (positive with 29%) else if (NO) patient DiabetesPedigreeFunction>=0.06 it means the patient diabetes is (Negative with 6%) but if (YES) patient diabetes is (Positive with 23%) else if (NO) patient BMI >=0.51 it means patient diabetes is (Positive with 20%) else if (YES) patient diabetes is (Negative with 2%) but if (YES) patient BloodPresure< 0.62 it means patient diabetes is

(Positive with 17%) else if (NO) patient diabetes is (negative with 4%).

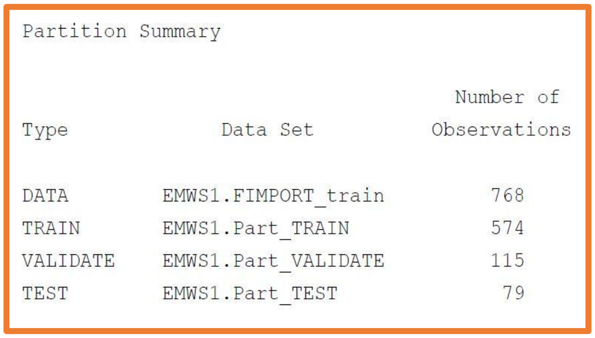
- SUMMARY ON R

This is summary on how the experiment was done using the RStudio. The Basic Statistics of the dataset (diabetes) was show in figure 3 above listing all the roll counts and their various names and value while figure 4 shows the data structure of the dataset and their variables. The dataset was loaded into the environment after which exploring of the dataset was done shown in figure 6. A plot of observations was done which shows various frequencies of all the variables shown in figure 5. Then there is need to check for missing values in each variable after which it was observed that there are excessive missing values in variable (skinthickness and insulin) which were removed and replaced with mean of each variable. So as to ensure that there is no missing decimal values in the dataset, conversion needs to be done where numeric values where converted to integer values. Because the data set needs to be transformed to 0 and 1 so as to enable easy scaling, hence normalization of the dataset was done by using the function(x). The Gini model was used in creating the model but first, the dataset was Split into two with the percentage of (75% = training and 25%=testing) respectively then a decision tree algorithm was applied on the training dataset of (75%), note: The GINI Model was adopted because of it presented a more clear model and hence make it easier for understanding of result when used to predict if patients has diabetes or not shown in figure 7 above after the model has been achieved, then Generating of the Frequency Table which will create tabular results of categorical variable of positive or negative throughput after testing on diabetes dataset and model. Then after the prediction result by the computer is on the dataset and model is made, there is need for more accuracy of the prediction on both dataset and model by applying a confusion matrix to guarantee the accuracy on the rate at which the model and predicts the dataset. To achieve a perfect result, a calculation of the confusion matrix is done by adding the dataset prediction positive value with negative value from the model then divide with the total number of the dataset. 142 +338/583 = 0.82345, this result shows that the applied testing on the trained (75%) dataset using the model is 82.3 accurate while the classification error is 0.176. Then testing the model on the test dataset (25%), then the model is rerun again on the test dataset, and it shows the prediction, then a confusion matrix is calculated using same formula 34 +102/185 = 0.7351 which is 74% accurate on test dataset.

In summary, R was used to create a model where the diabetes dataset was split into two. The model was applied on train dataset and test dataset and accuracy was determined on both before prediction of positive or negative patients on diabetic disease are achieved.

- EXPERIMENTS ON THE DATASET USING SAS ENTERPRISE MINER

The dataset was first cleaned in R and saved in the working directory as SaS_Diabetes. This is done because SAS cannot clean the dataset before it could be used to carry out another experiment in SAS.
Step 1: Launch the SAS enterprise miner
Step 2: Create a project
Step 4: Create a Diagram
Step 5: Import the data set (SaS_Diabetes) from the sample on the enterprise miner software menu
Step 6: Editing the variable (set outcome = target)
Step 7: Go to sample on the enterprise software menu to drag partition for the partitioning of the dataset. Then connect the imported data set to the partitioning and set the ratio of the partitioning (Training= 75.0, Validation= 15.0, Test=10.0). After setting the partition variable, the dataset result was checked and it displayed all the dataset current statistic reports shown in figure 8.



```
Partition Summary

                                          Number of
Type            Data Set                 Observations

DATA        EMWS1.FIMPORT_train              768
TRAIN       EMWS1.Part_TRAIN                 574
VALIDATE    EMWS1.Part_VALIDATE              115
TEST        EMWS1.Part_TEST                   79
```

Figure 8: Partition Summary of the current dataset (Fieldwork 2022)

Step 8: Performing the statistics explore by dragging it from the menu of the enterprise miner. Then connect it to the data partition node then run the statexplore which will show an output and a bar chart graph containing all the worth and variables in the dataset shown in figure 9.

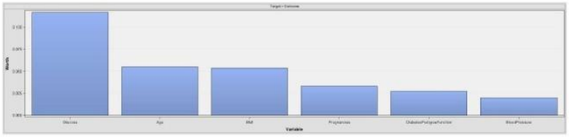


Figure 9: showing the variable worth of the diabetes dataset in SAS (Fieldwork 2022)

Step 9: Applying the decision tree on the data set by dragging from the sample on the enterprise miner software. Then connect the decision tree to the statexplore node, then change the ordinal target criteria to GINI model then run which will shows a decision tree as the model in figure 10, while the leaf statistic and score ranking overlay of the output model was shown in figure 11 and figure 12 respectively.

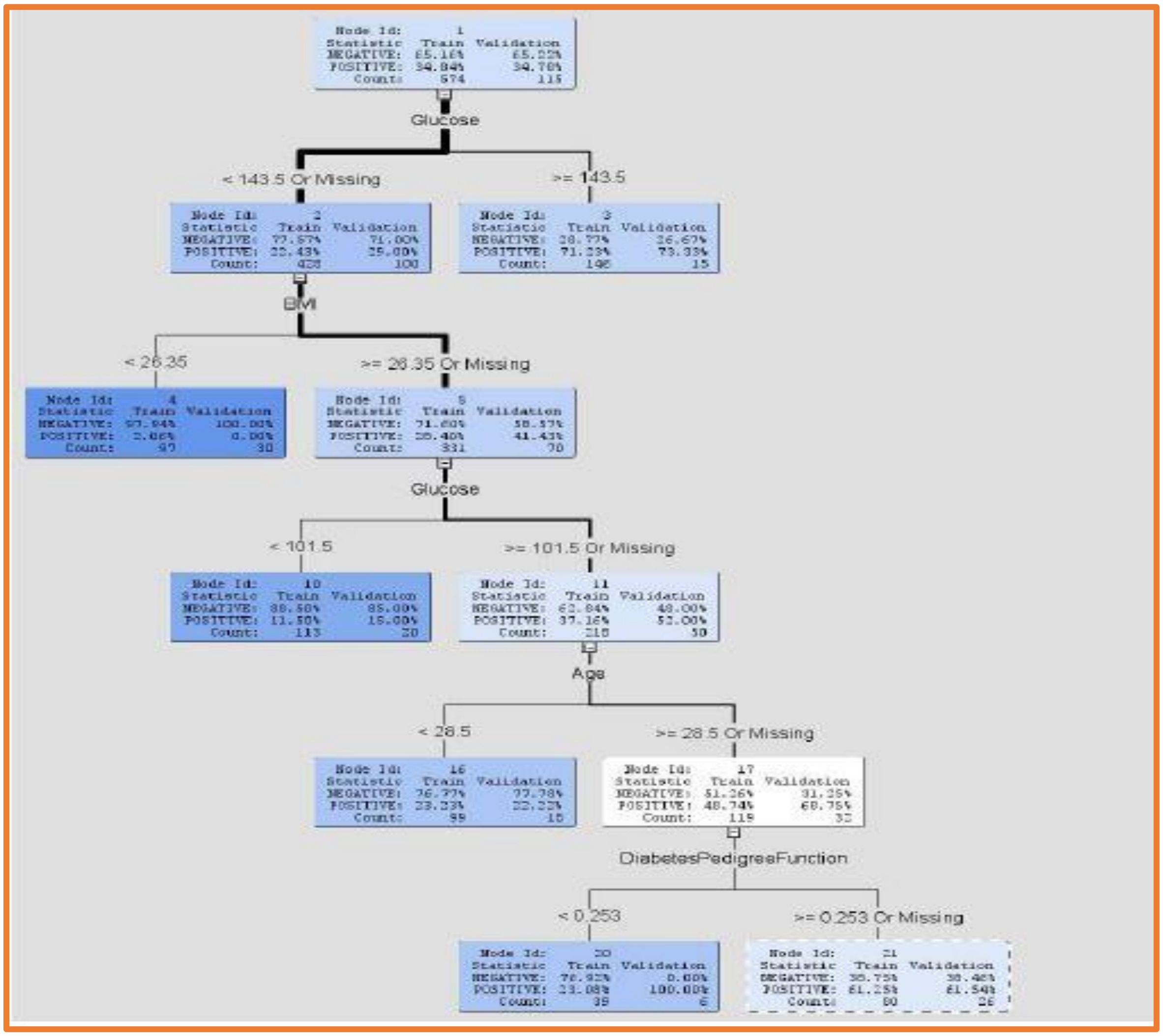


Figure 10: Decision Tree of the model (Fieldwork 2022)

TABLE 2: THE DECISION TREE MODEL RULES USING SAS

if Glucose >= 143.5 then Tree Node Identifier = 3 Number of Observations = 146 Predicted: Outcome=Positive	if Glucose < 143.5 AND Glucose >= 101.5 or MISSING AND BMI >= 26.35 or MISSING AND Age < 28.5 then Tree Node Identifier = 16

= 0.71 Predicted: Outcome=Negative = 0.29	Number of Observations = 99 Predicted: Outcome=Positive = 0.23 Predicted: Outcome=Negative = 0.77
if Glucose < 143.5 or MISSING AND BMI < 26.35 then Tree Node Identifier = 4 Number of Observations = 97 Predicted: Outcome=Positive = 0.02 Predicted: Outcome=Negative = 0.98	if Glucose < 143.5 AND Glucose >= 101.5 or MISSING AND DiabetesPedigreeFunction>= 0.253 or MISSING AND BMI >= 26.35 or MISSING AND Age >= 28.5 or MISSING then 66 Tree Node Identifier = 21 67 Number of Observations = 80 68 Predicted: Outcome=Positive = 0.61 69 Predicted: Outcome=Negative = 0.39
if Glucose < 101.5 AND BMI >= 26.35 or MISSING then Tree Node Identifier = 10 Number of Observations = 113 Predicted: Outcome=Positive = 0.12 Predicted: Outcome=Negative = 0.88	if Glucose < 143.5 AND Glucose >= 101.5 or MISSING AND DiabetesPedigreeFunction< 0.253 AND BMI >= 26.35 or MISSING AND Age >= 28.5 or MISSING then Tree Node Identifier = 20 Number of Observations = 39 Predicted: Outcome=Positive = 0.23 Predicted: Outcome=Negative = 0.77

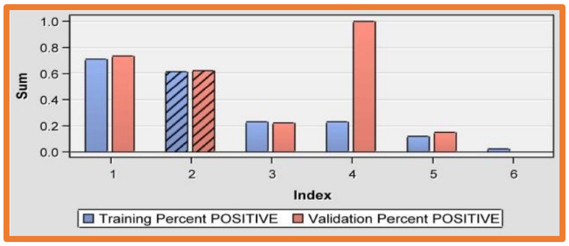


Figure 11: Leaf statistic

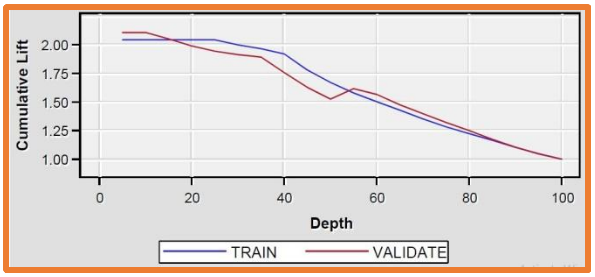


Figure 12: Score ranking overlay

CONCLUSION

As earlier stated, that the aim of this work is to adopt Machine Learning and Data Mining tools in the identification and prediction of Diabetes Patients Using Classification Mining Algorithm. This research was able to show clearly how diabetes disease could be managed using prediction models. These models were able to predict the status of a patient's diabetes state efficiently and accurately.

RECOMMENDATION

The researcher therefore recommends the following:

1. Full adoption of machine learning tools should be used in solving real life challenging problems more especially in health related problems more especially in Nigeria.
2. Different models showed be employed and then compared against each other for accurate data prediction.
3. Other organizations should be encouraged to apply machine learning tools for easy decision making.
4. Other researches can be done in the area of predicting if a patient with diabetes can die within a specific date because of high increase in glucose in the system or not using other modeling tools like entropy.

REFERENCES

[1] Pravarti Jain AndSantosh Kr Vishwakarma (2017)A Case Study on Car Evaluation and Prediction: Comparative Analysis using Data Mining Models, International Journal of Computer Applications (0975 – 8887) Volume 172 – No.9

[2] Diabetes UK, Understanding Diabetes-Your key to better health (2003), Prediction of Diabetes using Classification algorithm accessed from http://www.diabetes.or.uk/infocenter/pubs/understand.doc, Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82,115–121.doi:10.1016/j.procs.2016.04.016.

[3] Sharief, A. A., Sheta, A., (2014). Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3,54–59.doi:doi:10.14569/IJARAI.2014.031007.

[4] Pradhan, P. M. A., Bamnote, G. R., Tribhuvan. ,Jadhav, Chabukswar., Dhobale, V., (2012). A Genetic Programming Approach for Detection of Diabetes. International Journal Of Computational Engineering Research2, 91–94.

[5] Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract,(2016). An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing 424,323–335.doi:10.1007/978-3-319-28031-8.

[6] Nai Arun, N., Moungmai, R.,(2015). Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69,132 142.doi:10.1016/j.procs.2015.10.014.

[7] Nai Arun,N.,Sittidech,P.,(2014). Ensemble Learning Model for Diabetes Classification. Advanced Materials Research 931-932,1427–1431.doi:10.4028/www.scientific.net/AMR.931-932.1427.

[8] Orabi, K.M., Kamal, Y. M.,Rabah, T.M., (2016). Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer.pp.420–427.

[9] Priyam, A., Gupta, R., Rathee, A., Srivastava, S., (2013). Comparative Analysis of Decision Tree Classification Algorithms. International Journal of Current Engineering and TechnologyVol.3,334–337.doi:JUNE2013,arXiv:ISSN2277-4106.

[10] Bamnote, M.P., G.R., (2014). Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770.doi:10.1007/978-3-319-11933-5.

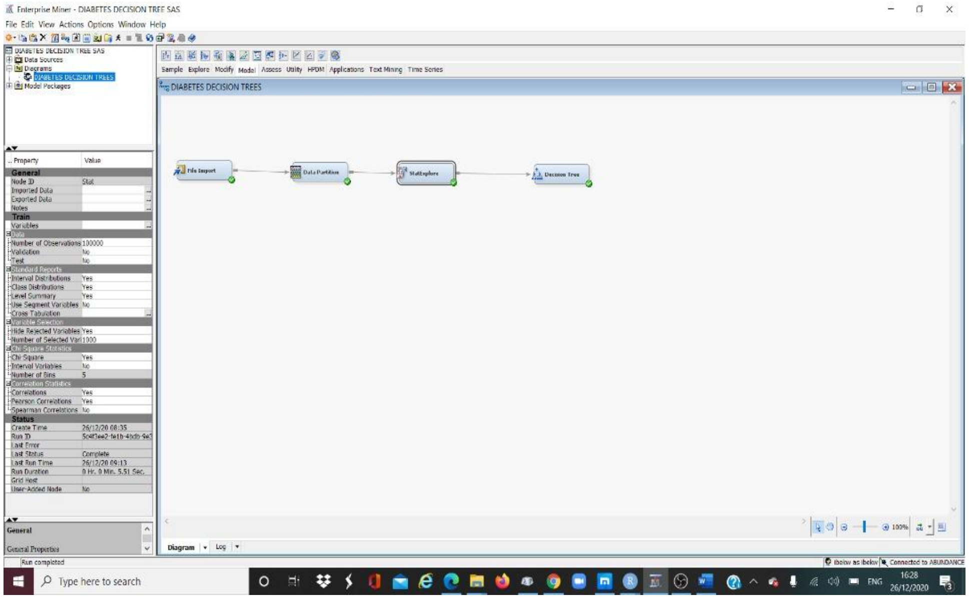
[11] Esposito, F., Malerba, D., Semeraro, G., Kay, J., (1997). A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491. doi:10.1109/34.589207.

[12] BITS Pilani, Dubai, BITS Pilani, Dubai and Ronak Sumbaly (2015) Diagnosis of Diabetes Using Classification Mining Techniques accessed from

[13] International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1)

[14] Jiawei Han and Micheline Kamber, (2001) "Data Mining Concepts and Techniques”, Morgan Kauffman Publishers.

APPENDICES A



Property	Value
General	
Node ID	Stat
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Data	
Number of Observations	100000
Validation	No
Test	No
Standard Reports	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross Tabulation	...
Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Vari	1000
Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	No
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No
Status	
Create Time	26/12/20 08:35
Run ID	5c4f3ee2-fe1b-4bdb-9e3
Last Error	
Last Status	Complete
Last Run Time	26/12/20 09:13
Run Duration	0 Hr. 0 Min. 5.51 Sec.
Grid Host	
User-Added Node	No

Run completed

Connected to ABUNDANCE

Figure 13: Screen shot of the Enterprise Miner on Diabetes Data set Decision tree in SAS