ISSN: 2714-3643
October, 2023

PREDICTIVE MODELS FOR DIABETES MELLITUS

Folake Akinbohun, Ph.D
Department of Computer Science; Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria
E-mail: folakeakinbohun@yahoo.com

Abstract
Diabetes mellitus is a disease that occurs when the body fails to utilize the excess sugar in the blood. Failure of insulin (a hormone produced in the pancreas) to internalize the cellular metabolism of glucose causes elevated blood sugar which leads to myriad of negative features in the body. The prevalence of diabetes is increasing among people. This study focuses on deployment of models to predict diabetes mellitus. The data set consists of 768 instances or records with 8 features and a class. The dataset was divided into training and testing phases in ratio 70:30. PART model was employed as a Rule-Based Algorithm and Random Forest was deployed as a Tree-Based model. It was found out that PART had 72.61% and Random Forest had accuracy of 78.70%. Random Forest showed a better accuracy result than PART. Hence, Random Forest is a good predictive algorithm for diabetes mellitus prediction.
Keywords: Diabetes, Random forest, Model, Insulin

1.0 Introduction

Diabetes mellitus is a growing public health disease that affects people worldwide both in developing and developed countries. Diabetes was the ninth-leading cause of death worldwide in 2020, accounting for almost 2 million fatalities yearly owing to diabetes and diabetes-related kidney disease (Deshpande, 2008).

According to 2021 International Diabetes Federation estimates, an estimated 537 million persons worldwide have diabetes, and 24 million people in the Africa region, 55 million will have diabetes by 2045 (International Diabetes Federation, 2021). Diabetes prevalence is fast increasing; prior 2019 estimates placed the population at 463 million individuals living with diabetes, with distributions being equal across both sexes with incidence peaking around the age of 55 (Khan *et. al*, 2020). It poses a challenge at social and economic levels. The key causes of type 2-diabetes include food and physical activity, which can lead to higher BMI, poor nutrition, hypertension, alcohol use, and smoking, as well as genetics (Deshpande, 2008).

Diabetes is a chronic disease that is caused when the body is not able to produce required amounts of insulin (the hormone that regulates blood sugar) or when the available insulin is ineffective. Normally, blood glucose levels are tightly controlled by insulin; a hormone produced by the pancreas. Insulin lowers the blood glucose level (www.webmd.com)

When the blood glucose increases (after eating food), insulin is released from the pancreas to normalize the glucose level by promoting the uptake of glucose into body cells. Common features or symptoms include blurring of vision, thirst, polyuria and weight loss. The effects of diabetes mellitus include long-term damage, dysfunction, and failure of various organs. If not well controlled, diabetes may cause some complications such as blindness, kidney failure and limb amputation due to chronic ulcer as a result of neuropathy.

Healthcare organizations are technologically deep into the big data analytics and clinical decision support environments to support health management care. Healthcare organizations are making use of data mining, predictive analytics and machine learning in order to extract hidden patterns from huge amount of patients' data so that clinical decision can be made. These data are structured by task; and it is called classification.

The classification algorithms are categorized as supervised learning that learns to classify new instances based on the knowledge learnt from a previously classified training set of instances. These data are capable of alleviating the challenges faced by medical personnel, through the development of smart medical equipment. The use of machine learning and other artificial intelligence for the analysis of medical data in order to assist prediction of diabetes is in place today.

Applying machine learning in diabetes mellitus is a means of utilizing available diabetes-related data for extracting knowledge. The social impact of the disease makes diabetes mellitus one of the main priorities in medical science research that needs to use machine learning/models for prediction. With increased

access to a large amount of patient data, healthcare providers are now focused on optimizing the efficiency and quality of their organizations by the use of predictive analytics. For a human being to do this, there is need for an automated system where algorithms are used to extract the information (Cheng, 2006).

The objective of the study is to compare two models for diabetes mellitus prediction in healthcare industry

2.0 Literature Review

Some related researches have been reviewed to indicate how much the disease is prevalent or how algorithms are used to predict the disease (Diabetes Mellitus). The burden of diabetes mellitus cannot be overemphasized as it is prevalent in most countries both at developing and developed countries as described below:

Qin (2022) compared the performance of five different machine-learning models for diabetes prediction using lifestyle data from the National Health and Nutrition Examination Survey data base. The data collected consisted of 17,833 people based on demographic traits and lifestyle-related variables from 1999-2000. Five machine-learning models (CATBoost, XGBoost, Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM)) were created to predict diabetes. Accuracy, sensitivity, specificity, precision, F1 score, and receiver operating characteristic (ROC) were used to assess model performance.

World Health Organisation (2016) gave account on the prevalence of diabetes in some African Countries which include: Nigeria had 4.3% diabetes prevalence in a population of 182 million, South Africa had 9.8 prevalence of diabetes in a population of 54.49 million. In a population of 46.05m, Kenya had 4% diabetes prevalence. Ethiopia had 3.80% prevalence of diabetes in a population of 99.39 million. In a population of 15.60 million, Zimbabwe had 4.60% diabetes prevalence. Zambia had 4.2% diabetes prevalence in a population of 16.2 million. In a population of 77.30 million, DR Congo possessed 4.3% diabetes prevalence. Mauritius had 14.20% diabetes prevalence in a population of 1.27 million and in a population of 27.40 millions, Ghana had 4.8% diabetes prevalence (www.africacheck.org). Type 1, Type 2 and gestational diabetes mellitus were prevalent (Anthonia and Chukwuma, 2014; Liu et al, 2020).

Tukur *et al* (2016) reviewed population-based studies on diabetes mellitus in Nigeria. Literature search of the PubMed data base was conducted of published research between 1990 and 2013.

Andrew *et. al.* (2018) determined the prevalence and risk factors for Diabetes Mellitus in Nigeria. They used a total of 23 studies equivalent to 14,650 persons. A random effects model was used to estimate the pooled prevalence of diabetes mellitus. The authors estimated the overall pooled prevalence of Diabetes Mellitus. They considered urban dwelling, unhealthy dietary habits, cigarette smoking, older age, physical inactivity and obesity as the risk factors of Diabetes Mellitus. They concluded in their study that there had been an increase in the prevalence of diabetes mellitus in Nigeria. Since there is an increase in diabetes mellitus, there is a need to predict the disease in medical domain.

Yukai (2018) proposed a study on the use of machine learning methods that combined feature selection and imbalanced process (SMOTE algorithm) to classify and predict diabetes follow-up control satisfaction data. The authors used input variables of Support Vector Machine (SVM), Decision Tree, and Integrated Learning Model (Adaboost and Bagging) for modeling and prediction.

Steffi *et. al*. (2018) worked on prediction of diabetes mellitus where the aim was to compare data mining algorithms for diabetes mellitus prediction. They employed Naïve Bayes, Logistic Regression, Artificial Neural Networks (ANNs), C5.0 Decision Tree and Support Vector Machine (SVM) models for predicting diabetes using common risk factors. The result showed that the decision tree model (C5.0) produced the best classification accuracy, followed by the logistic regression model, Naïve Bayes, ANN and the SVM gave the lowest accuracy.

Data mining models comparison for diabetes prediction was proposed by Amina *et. al*. (2018). Decision Tree, Naïve Bayes and K-Nearest Neighbors were deployed and compared. It showed that Decision Tree had the highest accuracy followed by Naïve Bayes and the least was K- Nearest Neigbours.

Deepti and Dilip (2018) presented a study on prediction of diabetes using classification algorithms. Three machine learning classification algorithms namely Decision Tree, Support Vector Machine and Naive Bayes were performed on Pima Indians Diabetes Database (PIDD). The results obtained showed that Naive Bayes outperformed other algorithms.

The study by Umar (2016) presented the pattern of presentation of Type 1 diabetic patients in Kano, Nigeria. The objective was to examine the prevalence of Type 1 diabetes in children at the Aminu Kano Teaching Hospital, Kano, Nigeria. The methods involved a retrospective study of case files of children with diagnosis of Type 1 diabetes mellitus for over 4 years. The most prevalent presenting symptoms were dehydration, fever, abdominal pain, polyuria and polydipsia. The highest prevalent presenting symptom was dehydration. The work was limited by the inability to use learning algorithms.

3.0 Methodology

Literatures related to Diabetes mellitus and the researches by some authors were reviewed in order to experiment the study. For the purpose of the study, the diabetes mellitus data set was collected from machine learning repository. Two algorithms were applied. They are Projective Adaptive Resonance Theory (PART) and Random Forest.

3.1. System Architecture

The architecture of the comparison of models on diabetes mellitus prediction is presented in Figure 1. The major components of the architecture are preprocessing, models, model evaluation.

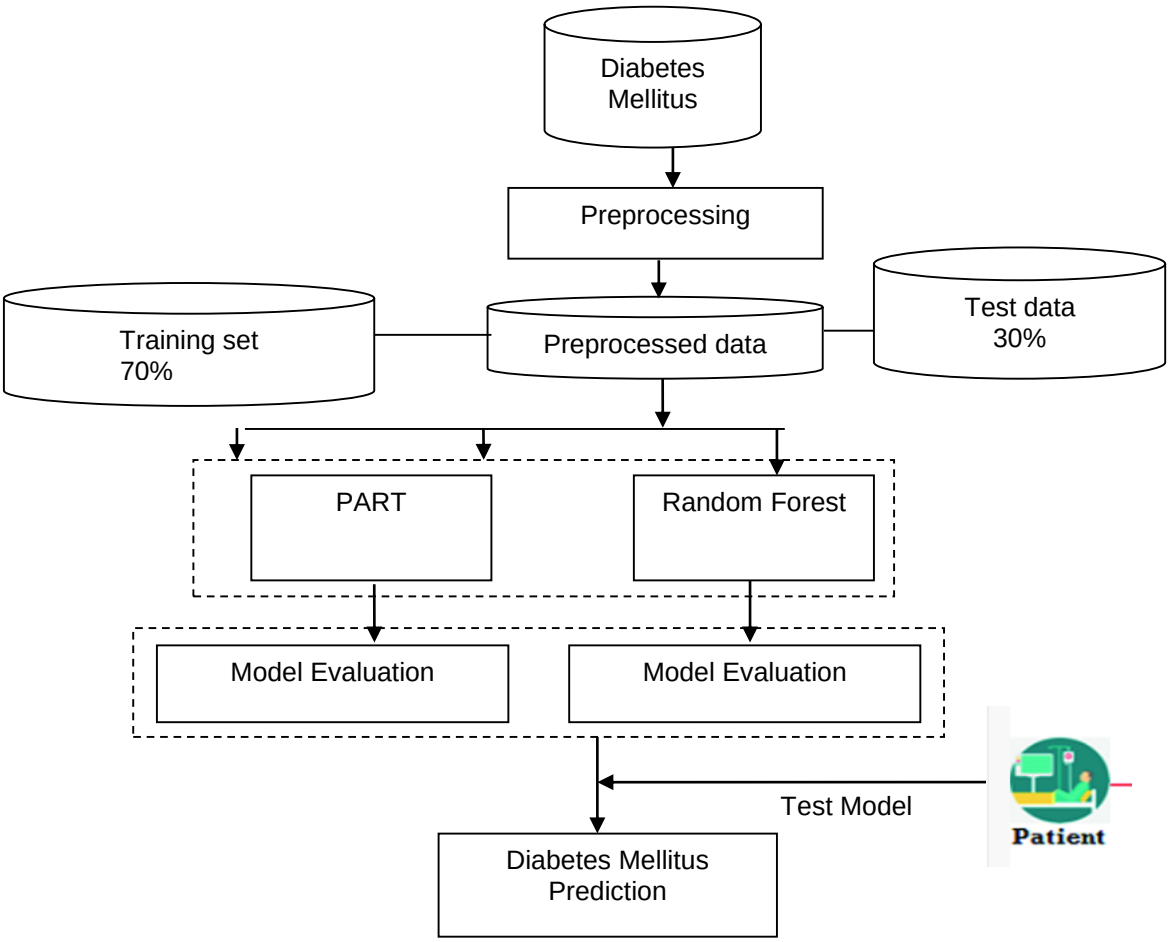


Figure 1: Architecture of the proposed system for prediction of diabetes mellitus

Pre-processing: Pre-processing consists of data collection, data cleaning and transformation. The Pima Indians Diabetes Data set was collected from uci repositories and preprocessed. The description of the dataset is presented in Table 1. The table contains 768 records or instances and 8 features (attributes) and 1 class or target.

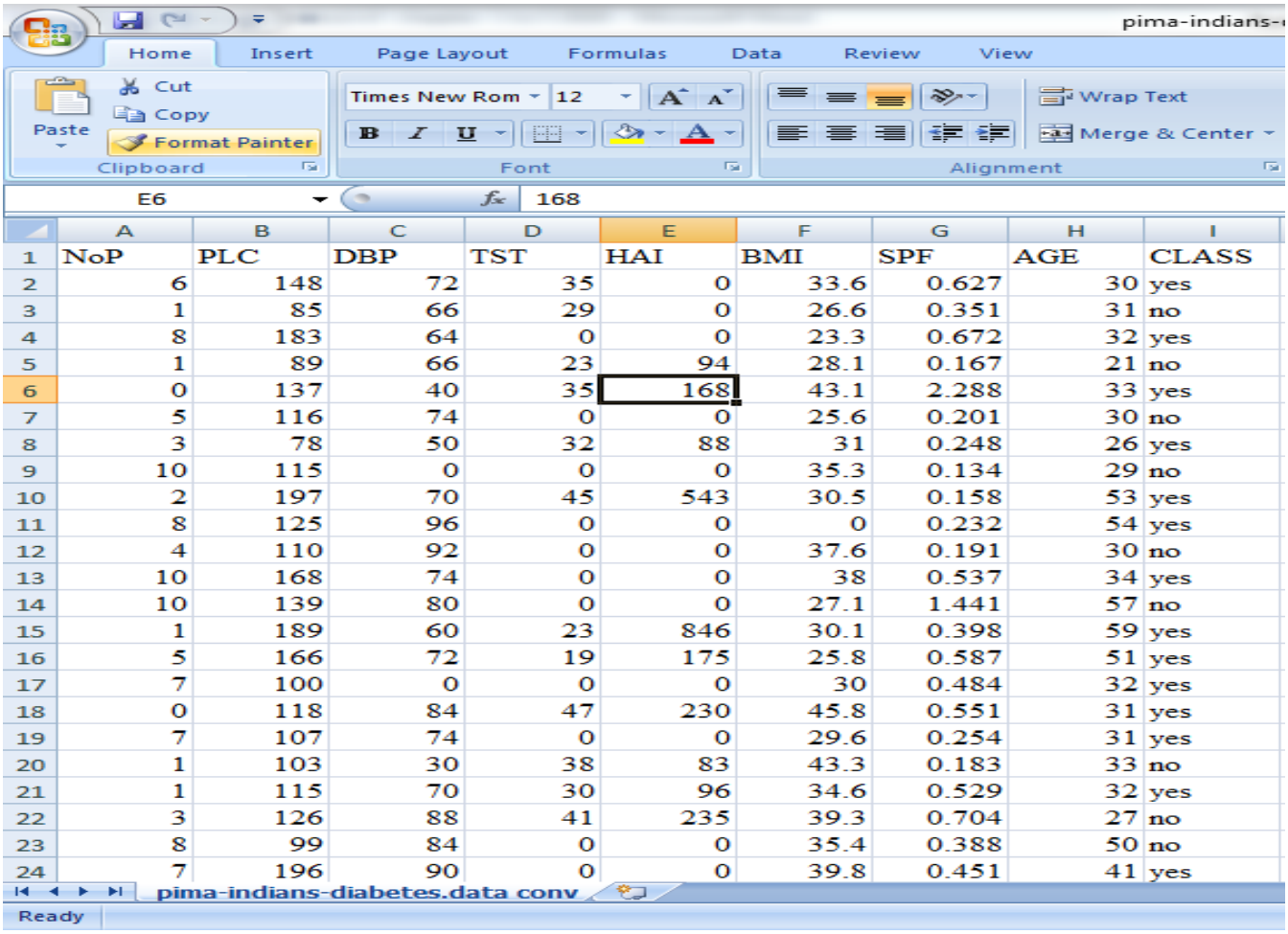
Table 1: Attributes and description of the diabetes mellituss

No	Attribute	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Blood Pressure	Diastolic blood pressure (mmHg)
4	Triceps	skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body Mass Index (BMI)
7	Diabetes Pedigree Function	Diabetes Pedigree Function
8	Age	Age (in years)
	Class	0 (No) or 1(Yes)

3.2 Diabetes Mellitus Data Set

Diabetes Mellitus data set comprises of the features and class. The class is patient will have diabetes (1) or patient not having diabetes (0). The features for prediction are the following: No of pregnancy, glucose, blood pressure, skin thickness, triceps, insulin, body mass index, age and diabetes pedigree function as presented in Table 2.

The data set consists of 768 records and 8 features. The data set was divided into: training data and testing data. 538 records are used for training data while 230 records are used for test data. The training data or cases are assumed to be represented as a pair [x_1, x_2, x_3, …x_n → y] where x_1, x_2, x_3 …x_n are vectors of attribute values describing some cases while y is the appropriate class or target.



	A	B	C	D	E	F	G	H	I
1	NoP	PLC	DBP	TST	HAI	BMI	SPF	AGE	CLASS
2	6	148	72	35	0	33.6	0.627	30	yes
3	1	85	66	29	0	26.6	0.351	31	no
4	8	183	64	0	0	23.3	0.672	32	yes
5	1	89	66	23	94	28.1	0.167	21	no
6	0	137	40	35	168	43.1	2.288	33	yes
7	5	116	74	0	0	25.6	0.201	30	no
8	3	78	50	32	88	31	0.248	26	yes
9	10	115	0	0	0	35.3	0.134	29	no
10	2	197	70	45	543	30.5	0.158	53	yes
11	8	125	96	0	0	0	0.232	54	yes
12	4	110	92	0	0	37.6	0.191	30	no
13	10	168	74	0	0	38	0.537	34	yes
14	10	139	80	0	0	27.1	1.441	57	no
15	1	189	60	23	846	30.1	0.398	59	yes
16	5	166	72	19	175	25.8	0.587	51	yes
17	7	100	0	0	0	30	0.484	32	yes
18	0	118	84	47	230	45.8	0.551	31	yes
19	7	107	74	0	0	29.6	0.254	31	yes
20	1	103	30	38	83	43.3	0.183	33	no
21	1	115	70	30	96	34.6	0.529	32	yes
22	3	126	88	41	235	39.3	0.704	27	no
23	8	99	84	0	0	35.4	0.388	50	no
24	7	196	90	0	0	39.8	0.451	41	yes

3.3. Models Construction

The data set was divided into training and test data. The preprocessed data were divided into training and test data. 70% of the data set was used for training and 30% of the data set was used for testing. This was trained using two models such as PART, and Random Forest. The two models were considered to discover and extract the patterns in the diabetes data set. The two models are described as follows:

i. PART (Projective Adaptive Resonance Theory) Model

PART is a type of rule-base algorithm that uses the principle of separate-and-conquer rule learner. The algorithm combines C4.5 and RIPPER rule learning. The algorithm works on the sets of rules being produced which are called decision lists. A new data is compared to each rule in the list in turn, and the item is assigned the class of the first matching rule.

Rule-based classifier is a type of classifier which makes the class decision by using various "if...else" rules. The condition with "if" is where the attributes or features are used to test if conditions are true. The features/attributes in this study are: No of pregnancy, glucose, blood pressure, skin thickness, triceps, insulin, body mass index, age and diabetes pedigree function. The features are represented by $[x_1, x_2, x_3, \ldots x_n]$ and the predicted class of each rule is represented by y in this form $[x_1, x_2, x_3, \ldots x_n \rightarrow y]$ and this classifies if a person or patient will have diabetes or not.

ii. Random Forest

Random Forest is a supervised learning algorithm that builds multiple decision trees and merges them together to get a more accurate prediction. It builds a forest which is an ensemble of decision trees, usually trained with the "bagging" method. The idea of the bagging method is that a combination of learning models increases the overall result.

The Random Forest algorithm determines the outcome based on the decision trees' predictions. It forecasts by averaging or averaging the output of various trees. Increasing the number of trees improves the outcome precision.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.

To categorize a new object based on attributes in Random Forest, each tree provides a classification and the tree "votes" for that class. In the case of regression, the forest selects the classification with the most votes (across all trees in the forest), and it takes the average of the outputs from distinct trees.

Random Forest operates as follows. The following is how each tree is planted and grown:

1. Assume that the training set has N cases. Then, a random sample of these N cases is chosen with replacement. This sample will be used as the tree's training set.
2. If there are M input variables, a number m<M is supplied such that at each node, m variables are chosen at random from the M. The best split from these m is used to split the node. While we are growing the forest, the value of m remains constant.
3. Each tree is developed to its full potential without being pruned.
4. Forecast new data by aggregating the tree forecasts (i.e. majority votes for classification, average for regression).

4.0 Result and Discussion

4.1 Results of the Performance Evaluation Metrics

The two models: PART and Random Forest were evaluated using some model evaluation metrics such as Accuracy, Precision, Recall and F Measure. The performance metrics are presented:

Results of PART and Random Forest

When PART algorithm was employed to the test data which were equivalent to 230 instances, the confusion matrix shows that 167 instances were correctly classified while 34 instances were incorrectly classified as shown in Table 3.

Random Forest (RF) algorithm was employed on the test data which were equivalent to 230 instances, the confusion matrix showed that 181 instances were correctly classified while 49 instances were incorrectly classified as shown in Table 4. The comparison on the basis of performance metrics between the two models: PART and Random Forest are presented in Table 5. The graphical representation of the models is displayed in Figure 2

Table 3: Confusion matrix using PART

	Yes	No
Yes	27	45
No	18	140

Table 4: Confusion matrix using Random Forest

	Yes	No
Yes	43	29
No	20	138

Table 5: Comparison between PART and Random Forest

Metrics	**PART**	**Random Forest**
Accuracy	72.61 %	78.70 %
Precision	0.708	0.781
Recall	0.726	0.787
F-Measure	0.705	0.783

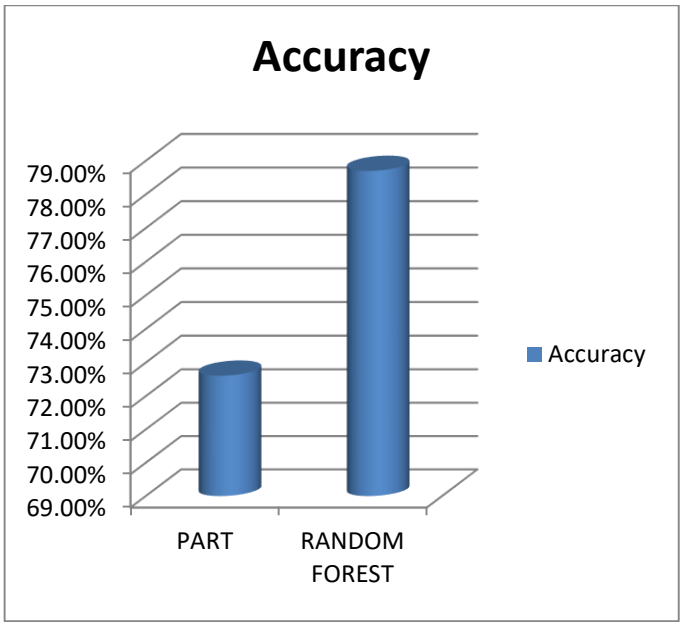


Figure 2: Comparison of the Accuracies between PART and Random Forest

4.2 Discussion on the Performance Metrics of Diabetes Mellitus Evaluation

The results of the predictive models of Diabetes Mellitus data set using learning algorithms: PART and Random Forest are stated as follows: The result showed that the accuracy of PART was 72.61% and Random Forest (RF) with accuracy of 78.70% respectively. This means that Random Forest had the higher accuracy than PART in predicting or classification of Diabetes Mellitus.

Using the precision and recall as performance metrics of the two models (PART and RF), the precision of PART was 0.708 and RF was 0.781 respectively, which means that RF Precision's value was higher than PART, hence RF is better in prediction of Diabetes Mellitus.

The Recall value for PART was 0.726 and RF was 0.787 respectively. The Recall metric value of RF was higher than PART. This results that RF is a better model for prediction of Diabetes Mellitus.

F1 score is a metric that evaluates the harmonic mean of precision and recall and measures a test's accuracy; hence it is a good measure which determines the best model. From the results, the F1 score for PART was 0.705 and Random Forest had F1 score of 0.783. This indicates that Random Forest had a better F1 score than PART. This shows that Random Forest predicts Diabetes Mellitus better than PART model.

From the models considered, the better model for prediction of Diabetes Mellitus in this study is demonstrated below in a compact representation:

$$\text{Model}_{(\text{Better model})} = RF > PART$$

If F1 Score > Accuracy in a model, then the model is good. From the results, the better model among the two models is the model with higher F1 score which is RF.

5.0 Conclusion

From the study, the following are the conclusions that were drawn:

1. The comparison of two predictive models Random Forest and PART were considered for Prediction of diabetes mellitus.
2. Random Forest was considered as a better model for predicting diabetes mellitus in medical domain.
3. It is also noted from this study that other machine learning algorithms or models can be employed for prediction of Diabetes Mellitus in medical domain.

References

Andrew, E. U., Baba, M. M., Mansur, A. R., Ibrahim, D. G., Fabian, H. P., Ayekame, T. U., Musa, M. B. & Kabiru, B. S. (2018). Prevalence and Risk Factors for Diabetes Mellitus in Nigeria: A Systematic Review and Meta-Analysis. *Journal of Diabetes Therapy*. 9(3): 1307–1316.

Anthonia, O. O. & Chukwuma, E. (2014). Diabetes mellitus in Nigeria: The past, present and future. *World Journal of diabetes.*

Amina, A., Yasir, A., Muhammad, A & Khurram, Z. (2018). Data Mining Models Comparison for Diabetes Prediction. *International Journal of Advanced Computer Science and Applications, Vol. 9, No. 8, 320-323.*

Cheng, T.H., Wei, C.P. & Tseng, V.S. (2006), Feature Selection for Medical Data Mining: Comparisons of Expert Judgment

and Automatic Approaches, *Proceedings of the 19th IEEE Symposium on Computer Based Medical Systems (CBMS'06)*.

Deepti, S. & Dilip, S. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*. Volume 132, 1578-1585.

Deshpande, A. D., Harris-Hayes, M. & Schootman, M. (2008). Epidemiology of diabetes and diabetes-related complications. *Physical Therapy*. 88(11): 1254–1264.

International Diabetes Federation. *IDF Diabetes Atlas, 10th edn.* Brussels, Belgium: 2021. Available at: https://www.diabetesatlas.org.

Khan, M. A; Hashim, M. J; King, J. K; Govender, R. D.; Mustafa, H. & Al Kaabi, J. (2020). Epidemiology of type 2 diabetes- global burden of disease and forecasted trends. *Journal of Epidemiology and Global Health.* 10(1): 107–111. *doi*:*10.2991/jegh.k.191028.001.*

Lin, X., Xu, Y., & Pan, X. (2020). Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. *Sci Rep* 10. https://doi.org/10.1038/s41598-020-71908-9

Qin, Y, Wu, J, Xiao, W, Wang, K, Huang, A, Liu, B, Yu, J, Li, C, Yu, F, & Ren Z. (2022). Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. *International Journal Environmental Research and Public Health.* 19(22):15027. doi: 10.3390/ijerph192215027.

Steffi J., R. Balasubramanian, R. & Aravind, K. K. (2018). Predicting diabetes mellitus using data mining techniques: *Comparative analysis of data mining classification algorithms. International Journal of Engineering Development and Research* (www.ijedr.org). volume 6. 460-467.

Tukur, D., Aliyu, A & Shehu, A. U. (2016). Review of population-based studies on diabetes mellitus in Nigeria. *Sub-Sahara Journal of medicine*. 3(2). 59-64.

Umar, I. U. (2016). Pattern of presentation of Type 1 diabetic patients in Kano, Nigeria. Nigerian *Journal Basic Clinical Sciences*;13, 85-88.

Yukai, L., Huling, L. & Hua, Y. (2018). Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017. *Computational and Mathematical Methods in Medicine.* www.webmd.com.