SLU JST
Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Comparison of Data Mining Techniques for Diabetes Mellitus Prediction

*[1]Abdulhameed Ado Osi, [2]Salisu Mamman Abdurrahman and [2]Alhassan Adamu

[1.] Department of Statistics, Kano University of Science and Technology, Nigeria
[2.] Department of Computer Science, Kano University of Science and Technology, Nigeria.

*Corresponding author

Abstract

Background: *Diabetes mellitus is an autoimmune chronic disease that develops when the body is unable to make enough insulin or use it adequately. The incidence of the disease is growing at an alarming rate. Consequently, the medical research community is becoming more and more interested in the prevention and prognosis of diabetes mellitus.* ***Objective:*** *This work determined to predict the onset of TYPE_2 diabetes mellitus using patient's data gathered from several Northern Nigerian healthcare facilities.* ***Method:*** *We applied six different classification algorithms, including Naive Bayes (NB), Random Forest (RF), Neural Network (NN), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The models' predictive abilities were then tested using four different metrics.* ***Results:*** *Results of the study show that all six algorithms have outstanding performance. The RF method provided higher accuracy than NN, SVM, LDA, NB and KNN. The accuracy, sensitivity, specificity and kappa of RF were 99.6%, 100%, 99.4% and 99.1%, respectively.* ***Conclusions:*** *We conclude that when compared to the other five models, RF has been shown to provide the highest level of accuracy.*

Keywords: *Diabetes Mellitus, Data mining, Supervised Machine Learning, Health Care facilities.*

1. Introduction

Diabetes is a chronic, non-communicable, and incurable disease that develops when the pancreas is unable to create enough insulin or when the body is unable to effectively use the insulin that is produced. As a result, there is an increase in blood glucose levels (hyperglycemia), which can harm the body and cause the failure of many organs and tissues. Insulin is a hormone that allows glucose from food we eat, particularly carbohydrates, to enter body cells where it may be used to make energy. (Association American Diabetes, 2018).

Diabetes causes other diseases such as stroke and heart disease, renal disease, pneumonia, bacteremia, tuberculosis and recently discovered coronavirus are all diseases that are exacerbated by diabetes. Most of the critical COVID19 patients were diabetic. Diabetes significantly raises the risk of heart attack or stroke by two to three times and is the leading cause of blindness, non-traumatic amputation, and kidney failure. Many Nigerians with stroke, blindness, kidney failure, heart attack are linked to diabetes.

There are two primary forms of diabetes: type 1 (T1DM), which can develop at any age but is most common in children and adolescents, and Type 2 (T2DM), which is more prevalent in adults and makes up about 90% of all

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

cases. Type 1 (T1DM) diabetes patients must take daily insulin injections to keep their blood glucose levels under control (Kumar & Umatejaswi, 2017). A healthy lifestyle, which includes more physical activity and a balanced diet, is the cornerstone of treatment for type 2 diabetes.

Diabetes affects more than 460 million adults worldwide, or one out of every 11 adults aged 20 to 79. Diabetes is becoming increasingly prevalent. According to the (Association American Diabetes, 2018), this figure will rise to 700 million by 2045. More than 79% of adults with diabetes lived in low- and middle-income countries. Diabetes is estimated to affect 19 million people in Sub-Saharan Africa, with the figure expected to rise to 47 million by 2045. Diabetes financial burden was estimated to be 10% of global health expenditure and this cost may be reduced drastically if someone can predict his status. (Muniruzzaman, et al., 2017).

Nigeria is currently the most affected country in Africa, with an estimated 4 million Nigerians suffering from type 1 or type 2 diabetes. Many Nigerians are living with undiagnosed diabetes; studies show that more than half of the country's diabetics are unaware that they have the disease (Oputa & Chinenye, 2015). Diabetes mellitus prevalence in Nigeria varied by geopolitical zone, with 3.9% in the northwest, 5.8% in the northeast, 3.8% in the north-central, 5.5% in the southwest, 4.6% in the southeast, and 9.3% in the south-south zone..

Data mining is the process of extracting useful, valid, unexpected, and understandable knowledge from data. Other disciplines, such as statistics, machine learning, and pattern recognition, obviously share these broad goals. (Torgo, 2017). Data mining tries to extract useful information from large data sets by exploring and analyzing large amounts of data in order to find meaningful patterns (Ledolter, 2013). Many studies recently around the universe focused on data mining and machine learning in health care sector. (Azeem, Kamal, Hamid, & Ali, 2018) Applied six machine learning algorithms on diabetes patient's medical record of PIMA India, prediction accuracy of the applied methods was compared and evaluated in which K Nearest Neighbor and Support Vector Machine gives highest accuracy for predicting diabetes. Also, glucose concentration is discovered to have highest importance among other features. In (Nai-arun & Moungmai, 2015) four machine learning models were used to classify diabetes mellitus, they are Artificial Neural Network, Decision Tree, Logistics Regression and Naïve Bayes. From the result Random Forest has best performance, they incorporated Bagging and boosting techniques for improving the robustness of the models. (Alic, Gurbeta, & Bandjevic, 2017)An overview of machine learning techniques in diabetes and cardiovascular disease classification using Artificial Neural Network and Nave Bayes is presented; the results show that NN has the highest accuracy value of 99.5% and 97.9% for diabetes and CVD classification, respectively (Muniruzzaman, et al., 2017) Adapted Gaussian Process based classification methods for classification of Diabetes Mellitus in comparison with commonly used techniques, such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Naïve Bayes (NB). Their performance was evaluated using accuracy, sensitivity, specificity and receiver operation characteristic (ROC). The results show the performance of GP-based model is highest compared with other techniques. (Zou, et al., 2018) Employed Principal Component Analysis (PCA) to reduce the dimensionality of the dataset before applying decision tree, random forest and neural networks to predict diabetes mellitus. It was found that the accuracy of using all features has better results than using PCA. (Kavakiotis, et al., 2017) Focus on reviewing machine learning and data mining approaches applied in the field of diabetes research with respect to prediction and diagnosis, Support Vector Machine (SVM) was discovered to be the most successful and widely used algorithm. (Samer & Samy, 2018) And (Pradhan & Kumar, 2011) used Artificial Neural Network (ANN) to predict whether a person is diabetic or not.

SLU JST
Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

The goal of this study is to predict the development of type 2 diabetes for a given patient on basis of information collected about the patient, without the need of blood test or going to the hospital. In this work, six algorithms were employed to develop predictive models on diabetes mellitus data in Northern Nigeria. We evaluate and compare the performance of the algorithms using performance evaluation metrics. We also test the effects of the various predictor sets on the accuracy of the predictive models. All data manipulation and modeling were implemented in R (Team, R Core, 2019).

2. Materials and methodologies

Dataset and attributes

This study uses dataset extracted from diabetes patients. The records was obtained using special designed record forms from five States in northwest geopolitical zone of Nigeria, comprises Kano State, Jigawa State, Katsina State, Zamfara State and Sokoto State. The distribution of responders according to States are as follows: Kano 341 respondants (27.1%), Jigawa 217 respondants (17.3%), Katsina 240 respondants (19.1%), Zamfara 215 respondants (17.1%) and Sokote 244 respondants (19.4%). The dataset contains Demographic, clinical and diagnostic information. Data entry forms was designed and filled with the help from the hospital staff through verbal interview with patients under the research surveillance. The dataset comprises 1257 observations and 19 instances. The instances were decribed in the table 1 below. The attributes were selected for the study based on literature and expert consultation.

Table 1: Variable descriptions

S/N	Variable name	Description	Type
1.	Age	Patient's age	Numeric
2.	Diastolic	Patient's diastolic blood pressure	Numeric
3.	Body Mass Index	Patient's body mass index	Numeric
4.	Glucose	Patient's glucose level	Numeric
5.	Hospital visit	Number of times patient visited hospital for last six months	Numeric
6.	Sex	Patient's sex	Categorical (M=male and F=female)
7.	Occupation	Patient's occupation status	Categorical (1=Business, 2=Civil servant, 3=Farming, 4=Housewife. 5=Retired, 6=student, 7=Others.)
8.	Diet	Diet took by the patient	Categorical (NBD—not a balanced diet; BLD—balanced diet)
9.	Exercise	Method adapted for exercise	Categorical (1=Football, 2=Stretching, 3=Walking, 4=None)
10.	Marital Status	Patient's marital status	Categorical (1=single, 2=married, 3=divorced, 4=separated and 5=widow).

Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

11.	Residential address	Patient's residential Address	Categorical (1=village, 2=town and 3=city).
12.	Educational attainment	Patient's level of education	Categorical (1=primary school, 2= high school, 3=college/university, 4=postgraduate, 5=Islamic school and 6= Not formal education)
13.	Excessive	Excessive thirst	Binary (yes or no).
14.	urin	Frequent urination	Binary (yes or no).
15.	Weight	Having Weight loss or gain	Binary (yes or no).
16.	Flu	Having Flulike symptoms	Binary (yes or no).
17.	Blurred	Having Blurred vision	Binary (yes or no).
18.	irritability	Having irritability	Binary (yes or no).
19.	healings	Slow healings on cut or bruise	Binary (yes or no).
20.	Type	Type of diabetes diagnosed patient with	Binary (type_1 or type_2).

Training and Test Dataset

Part of evaluating data mining models is dividing a dataset into training and testing sets. After cleaning and processing, the data was divided into two parts: 80% for training and 20% for testing. The models were trained using the training data, and then we applied our "trained" model to the test data, which was previously unseen.

K-fold cross validation

The most common method for validating a model's performance is K-fold cross-validation. The 10-fold cross validation method was used in this article. The initial sample was divided into ten sub-samples. The validation data was kept for each subsample, while the remaining 9 samples were used to train the model.

Naive Bayes

The Naive Bayes (NB) Algorithm is a supervised classification technique for predicting the likelihood of class membership. NB is a relatively simple to understand and build data mining algorithm that outperforms other data mining algorithms.
NB is based on the Bayes theorem and has classification capabilities comparable to decision trees and neural netw orks. When applied to large data sets, Naive Bayes demonstrated high accuracy and speed. The Naive Bayes algorithm gets its name from the fact that it makes some "naive" assumptions about the data. The naive Bayesian classifier assumes attribute conditional independence with respect to the class and no hidden or latent attributes. In most real-world applications, these assumptions are rarely true. If the features are independent, it will find a solution much faster than discriminative models like logistic regression. This means that less training data is required to build a highly accurate prediction model.

K-Nearest Neighbor

K-nearest neighbor (KNN), also known as lazy learning, is a supervised learning algorithm that uses a proximity search method to find the closest match between a new and old case based on a predetermined amount. KNN is an

SLU JST
Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

instance-based learning technique. It calculates the distances between samples using distance measurements such as Euclidean, Hamming, and Manhattan. For example, the train data is searched for an instance that most closely resembles the new instance in question when predicting any new instance. For a classification problem, KNN does this by looking at the closest points to the nearest neighbors to determine the proper class. The k factor comes into play by determining how many neighbors the algorithm should examine.

Random forest.

Random Forest is a decision tree ensemble that combines the results of unpruned decision trees. Each tree is developed using a bootstrap resample with replacement. Each decision tree node is split using a random selection of variables. Prediction is made by aggregating all trees' predictions by "majority vote". The number of predictive variables to randomly choose at each node for splitting (mtry) and the number of trees (ntree) to grow in the forest are two important parameters in Random Forest.

Linear Discriminant Analysis

LDA is one of the commonly use technique for classifying pattern between two classes. However, it can also be extended for multiple classes. LDA assumes that the distributions of features within each tribe follow a multivariate Normal distribution with a common covariance matrix. LDA divides the classes into decision regions and attempts to maximize the ratio of between-class variance to within-class variance..

Support Vector Machine

Support vector machines (SVM) are one of the most successful machine learning methods, capable of solving both regression and classification problems. SVM can handle both linear and nonlinear data. Nonlinear mapping is used by SVM to transform the original training tuple into a higher dimension. It seeks the best linear separation hyperplane, which is the set of decision boundaries that separates the tuple based on their class labels.

Neural networks

Neural networks comprise of two-stage classification models. In the first stage, Q transforms input features from the matrix X into a matrix of derived features Z with dimension n. Then, in these models, each tribe is modeled as a function of the matrix Z, where Z is often referred to as the hidden layer.

Performance evaluation

It is required that a classifier to give high prediction accuracy. In order to evaluate the prediction rate, several standard performance evaluation metrics exits in literature. To evaluate the performance of the six different classifiers, we used the following parameters: accuracy, specificity, sensitivity, and kappa index. The percentage of correctly classified instances out of all instances is referred to as accuracy. Kappa or Cohen's Kappa is similar to classification accuracy, but it is normalized at the random chance baseline on your dataset; the true positive rate, also known as sensitivity, is the true positive rate. It is the number of instances from the positive (in our case, TYPE 2 diabetes) class that were correctly predicted. The true negative rate or specificity. It is the number of instances from the negative class (in our case, TYPE 1 diabetes) that were correctly predicted. These metrics are derived from the confusion matrix, which is a straightforward cross-tabulation of the data's observed and predicted classes. Table 2 shows an example of an outcome with two classes. Off-diagonal cells represent the number of errors for each possible case, while diagonal cells represent cases where the classes are correctly predicted.

Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Table 2: Confusion Matrix

		True class	
		Positive	**Negative**
Predicted classes	**Positive**	**True Positive Count (TP)**	**False Positive Count (FP)**
	Negative	**False Negative Count (FN)**	**True Negative Count (TN)**

$$Sensitivity = \frac{TP}{TP+FN} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (3)$$

$$Kappa = \frac{O-E}{1-E} \quad (4)$$

For Kappa, where O and E represent observed and expected accuracy, respectively, which can be obtained from the confusion matrix's marginal totals. This statistic has a value between 0 and 1; a value of 0 indicates that there is no agreement between the observed and predicted classes, while a value of 1 indicates that the model prediction and the observed classes are perfectly concordant.

Results

We enrolled 1257 diabetics in the study. There were 574 males (46.4%) and 683 females (54.3%), for a female/male ratio of 1.19:1. The patients' average age (standard deviation (SD)) was 50.77(14.62) years. Of the 1257 patients, 849 (67.5%) had T2DM and 408 (32.5%) had T1DM. Tables 3 and Table 4 show the characteristics of both categorical and continuous variables. The mean ages of those with T2DM and those with T1DM were 54.80 (SD, 13.27) and 42.39 (SD, 13.72), respectively. Table 3.1 and Fig. 3.2 show that patients with T2DM were significantly older than those with T1DM (p=0.000). T1MD and T2MD associations with categorical variables are shown in Table 3.

Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

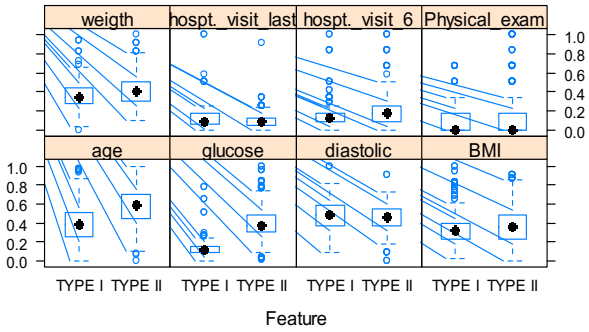

Figure 1: Box and Whisker Plots of continuous variables by diabetes type.

The box whisker plot in Fig. 1 aids in understanding the overlap and separation of attribute-class groups, as well as checking for the presence of outliers. The plot clearly shows that the distributions of glucose, age, and weight for patients with T2DM and T1DM differ.

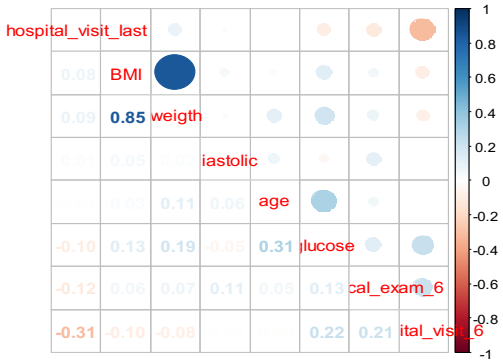

Fig 2 Correlation plot for continuous variables

A dot-representation was used, with blue representing positive correlation and red representing negative correlation. The greater the dot, the greater the correlation. The matrix is symmetrical, and the diagonal attributes are perfectly positively correlated (because it shows the correlation of each attribute with itself). Some of the attributes are highly correlated, as we can see. For example, the correlation between BMI and Weight was 0.85, indicating a strong relationship between them. To put it another way, one can represent the other, so weight was removed from the covariates prior to modeling.

Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

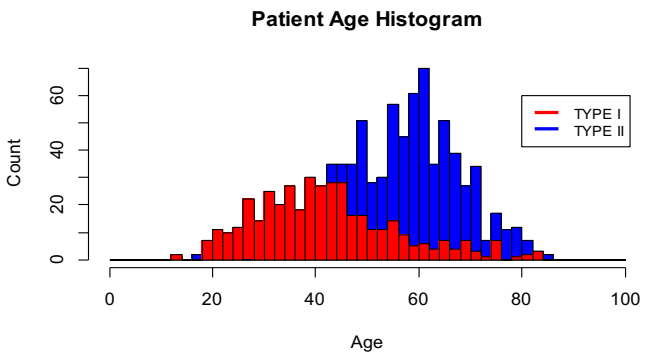

Figure 3.: distribution of age based on diabetes status

Fig. 3 depicts the distribution of age among diabetes patients; it is clear that as age increases, there are more patients with T2DM than T1DM. This means that those who were older were more likely to develop T2DM than those who were younger. That is, age has a significant influence on diabetes status.

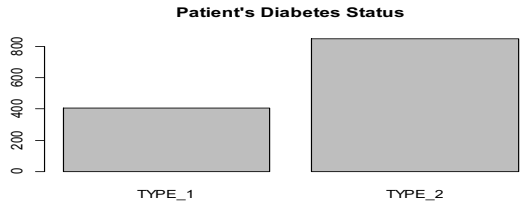

Figure. 4 Patient's diabetes status plot

The bars in Fig. 4 represent the counts of patients with T1DM and T2DM diabetes; we see that the majority of patients enrolled in the study were diagnosed with T2DM diabetes.

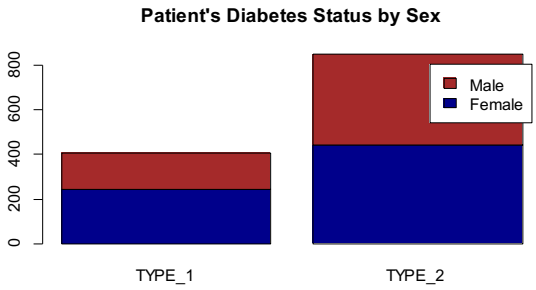

Figure 5 Distribution of diabetes status by sex

Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

According to Fig. 5, there are more females than males in the target population for both T1DM and T2DM diabetes patients.

Table 3 Characteristics of study participants (categorical variables) by diabetes status

SNo.	Characteristics	Levels	All (n=1257) n(%)	TYPE I (n=849) n(%)	TYPE II (n=408) n(%)	p-value
1.	Sex	Female	683(54.3)	441(51.9)	242(59.3)	p = 0.017
		Male	574(45.7)	408(48.1)	166(40.7)	
2.	Occupation	Business	188(15.0)	107(12.6)	81(19.9)	p=0.000
		Civil Servant	214(17.0)	147(17.3)	67(16.4)	
		Farming	192(15.3)	166(19.6)	26(6.40)	
		Housewife	439(34.9)	272(32.0)	167(40.9)	
		Retired	52(4.1)	39(4.60)	13(3.20)	
		Student	55(4.40)	35(4.10)	20(4.90)	
		Others	117(9.3)	83(9.80)	34(8.30)	
3.	Diet	Balanced	556(44.2)	392(46.2)	164(40.2)	p=0.052
		Non-balanced	701(55.8)	457(53.8)	244(59.8)	
4.	Exercise	Football	3(0.2)	6(0.4)	0(0.0)	p=0.000
		Stretching	777(61.8)	517(60.9)	260(63.7)	
		Walking	443(35.2)	314(37.0)	129(31.6)	
		None	34(2.8)	15(1.7)	19(4.7)	
5.	Marital Status	Single	67(5.3)	37(4.4)	30(7.4)	p=0.008
		Married	1026(81.6)	704(80.9)	322(78.9)	
		Divorcee	15(1.2)	14(1.6)	1(0.2)	
		Separated	59(4.6)	41(4.8)	17(4.2)	
		Widow	91(7.2)	53(602)	98(9.3)	
6.	Residential address	City	459(36.5)	291(34.3)	168(41.2)	p=0.038
		Town	442(35.2)	303(35.7)	139(34.1)	
		Village	356(28.3)	255(30.0)	101(24.8)	
7.	Educational attainment	Primary	116(9.2)	85(10.0)	31(7.6)	p=0.075
		Secondary	228(18.1)	142(16.7)	86(21.1)	
		Collage/University	227(18.1)	168(19.8)	59(14.5)	
		Postgraduate	13(1.0)	11(1.3)	2(0.5)	
		Islamic Education	654(52.0)	426(50.2)	228(55.9)	
		No formal Educ.	19(1.6)	17(2.0)	2(0.5)	
8.	Excessive Thirst	YES	664(52.8)	557(65.6)	107(26.6)	p=0.000
		NO	593(47.2)	292(34.4)	301(73.8)	
9.	Frequent Urine	YES	904(71.9)	714(84.1)	190(46.6)	p=0.000
		NO	353(28.1)	135(15.9)	218(53.4)	
10.	Weight lost	YES	733(58.3)	550(64.8)	183(44.9)	p=0.000
		NO	549(41.7)	299(35.2)	225(55.4	
11.	Flu	YES	720(57.3)	497(5835)	223(54.7)	p=0.214
		NO	537(42.7)	352(41.5)	135(45.3)	
12.	Blurred Vision	YES	528(42.4)	371(43.7)	157(38.5)	p=0.090
		NO	729(58.0)	478(56.3)	251(61.5)	
13.	Irritability	YES	327(26.0)	207(24.4)	120(29.4)	p=0.067

SLU JST
Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

		NO	930(74.4)	642(75.6)	288(70.6)	
14.	Slow healing	YES	992(78.9)	623(73.4)	369(90.4)	p=0.000
		NO	265(21.1)	226(26.6)	39(9.6)	

We used the chi-square test to determine the statistical significance of categorical variables used to identify diabetes status, specifically their association with diabetes progression. Table 3 displays the results 0.05 level of significance was considered. The findings revealed a highly significant relationship between some demographic variables and diabetes status, including gender, marital status, and occupation. The results for patient condition variables show a link between exercise, excessive thirst, frequent urination, weight loss, slow healing, and diabetes level.

Table 4 Characteristics of study participants(continuous variables) by diabetes status

S/No.	Characteristics	All (n=1257)	TYPE I (n=849)	TYPE II (n=408)	p-value
1.	Age (Years) Mean(SD)	50.77(14.62)	42.39(13.72)	54.80(13.27)	**0.000**
2.	Diastolic Mean(SD)	82.93((13.60)	81.70(11.70)	85.50(16.60)	**0.000**
3.	Body Mass index Mean(SD)	22.33(4.47)	21.76(4.53)	22.60(4.29)	**0.001**
4	Glucose level Mean(SD)	7.5(3.0)	4.5(1.3)	8.9(2.5)	**0.000**
5.	Hospital visit Mean(SD)	34.25(33.12)	40.88(48.72)	31.07(26.00)	**0.000**

The p-values of the t-tests performed for each of the continuous predictors are shown in Table 4. The p-values in bold are less than the level of significance used in the test (= 0.05), indicating that the variable was highly significant in determining the patient's diabetes type.

Table 5: Performance of the classification models

S/No.	*Classifier*	*Accuracy*	*Sensitivity*	*Specificity*	*Kappa*
1.	KNN	0.824	0.728	0.870	0.598
2.	NN	**0.992**	**1.000**	**0.988**	**0.982**
3.	LDA	0.952	0.914	0.970	0.890
4.	SVM	0.940	0.951	0.935	0.866
5.	NB	0.940	0.877	0.970	0.861
6.	RF	**0.996**	**1.000**	**0.994**	**0.991**

We need to compare the models and choose the most accurate. As a result, six algorithms were compared: LDA, RF, NN, SVM, NB, and KNN. The classifiers' prediction results were summarized. Table 5 shows the accuracy, sensitivity, specificity, and kappa values for each of the six classifiers. The summary statistics were generated using a 10-fold cross-validation procedure. Equations (1)-(3) were used to calculate the values in the table 5. The RF achieved 0.996 classification accuracy, 1.000 sensitivity, 0.994 specificity, and 0.991 kappa. The classification accuracy of NN was 0.992, with a sensitivity of 1.000, a specificity of 0.988, and a kappa of 0.982. On all metrics,

Sule Lamido University Journal of Science & Technology Vol. 8 No. 1 [June, 2024], pp. 125-138
https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

NN and RF outperformed the remaining models. It is critical to have a high sensitivity in order to properly identify patients with a specific potentially T2DM condition. Aside from accuracy, these two models have a higher sensitivity.

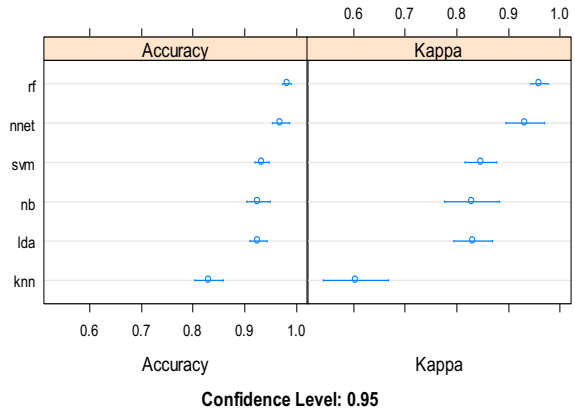

Figure 6 Model performance in terms of Accuracy and Kappa

We can compare the spread and mean accuracy of each model by plotting the model evaluation results. Fig. 6 depicts the mean estimated accuracy, kappa index, and 95% confidence interval. Based on the accuracy values, we can clearly see that the RF model produces the highest classification accuracy when compared to the other methods. The RF also has the highest sensitivity, specificity, and kappa values. NN was the second best model despite having 100% sensitivity. Based on this, it is possible to conclude that the RF method outperforms the other five methods.

3. Discussion

In this study, we created models to predict the type and risk of diabetes. Data from five northern Nigerian states were used. Our study's predictive models can distinguish between T2DM and T1DM, and the health centers in our study were chosen based on population representation and geographic location. The following models are compared in this paper: Nave Bayes (NB), Random Forest (RF), Neural Network (NN), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). We used four metrics to assess their predicted performance. Based on our data set results, all six predictive models in our study have very good predictive ability. The results, however, revealed that the RF classifier is the best model, followed by the NN.

Many academics have recently expressed an interest in developing and comparing diabetes prediction models using data mining and machine learning methods (Alghamdi et al., 2017; Perveen et al., 2016; Santhanam & Padmavathi, 2015; Sneha & Gangil, 2019; Wu et al., 2018). And RF has been reported as the best performing model by (Alghamdi et al., 2017; Sneha & Gangil, 2019). None of these studies were achieved more accurate model than ours. Many studies have found that ensemble methods outperform single models, but our best model, RF, outperforms the results of ensemble methods used in other studies (Das et al., 2020).

Tables 3 and Table 4 investigated the impact of categorical and continuous covariates on diabetes progression. Age, blood pressure, BMI, hospital visit, glucose level, sex, occupation, physical activity, frequent urine, excessive

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

thirst, and weight loss were discovered to be important factors in determining diabetes status. Several studies, including (Dehaki & Akbarzadeh, 2018; Perveen et al., 2016), discovered that a patient's age, gender, ethnicity, glucose level, family history, blood pressure, and BMI were all significantly related to his diabetes status.

Females 682 (54.3%) were the sex with the most T2DM patients in this study, which is similar to (Dehaki & Akbarzadeh, 2018) results where 52.8% of the subjects were female. The overall average age of the patients was 50.77(SD, 14.62), and the average age of T2DM patients was 54.80(SD, 13.27), which was older than 42.39(SD, 13.72) for T1DM patients. T2DM was diagnosed in 67.5% of the participants in our study. T2DM was discovered in 90% of the patients in (Adeleke & Ayenigbara, 2019; Nayak et al., 2017). Like (Perveen et al., 2016), our findings revealed a highly significant age difference between T2DM and T1DM patients.

This study has a number of limitations. First, the models were evaluated solely on the basis of their accuracy, specificity, sensitivity, and kappa statistic. Second, participants were drawn from various secondary/tertiary healthcare facilities. As a result, respondents were more likely to be civilized and educated, which may have an impact on the generalizability of the findings. Finally, the study did not account for the effect of some important covariates such as disease history and genetics, alcohol consumption, and so on.

4. Conclusions

Diabetes is the most common noncommunicable disease. Diabetes prevention and management are critical because it can lead to or exacerbate other health issues. Diabetes, both type 1 and type 2, can cause heart problems. We developed data mining prediction models in this paper that accurately identify people who are at a higher risk of developing T2DM. To create the model, we first evaluated predictors' predictive power. These include patient demographics such as age, gender, educational attainment, marital status, and other clinical information. Blood pressure and BMI. Using statistical tests, we demonstrated that age, glucose level, BMI, and blood pressure have the greatest predictive power for determining T2DM patients. Furthermore, evaluation of predictive model results in terms of accuracy, sensitivity, specificity, and kappa shows that the RF method performs significantly better. When compared to the other five models, RF has been shown to provide the highest level of accuracy. The accuracy of the RF classifier is 99.6%, while the accuracy of the NN classifier, which is the second best, is 99.2%. This study discovered that data mining techniques can be very useful for predicting the type and risk levels associated with diabetes. This research can help doctors and health organizations improve disease diagnosis, which can lead to earlier disease cure in patients.

Acknowledgement

Tertiary Education Trust Fund (TETFund) provided funding for this study through an institutional-based research grant (IBR).

References

Adeleke , O. R., & Ayenigbara, G. O. (2019). Preventing Diabetes Mellitus in Nigeria: Effect of Physical Exercise, Appropriate Diet, and Lifestyle Modification. *Int J Diabetes Metab*, 1-5. doi: 10.1159/000502006

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Ahmad, A., Mustapha, A., Eliza, D. Z., Masah, N., & Yasmin, N. Y. (2011). Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus. In V. Snasel, J. Platos, & E. El-Qawasmeh, *Digital information Processing and Communications* (pp. 537-545). Berlin, Heidelberg: Springer. Retrieved from https://doi.org/10.1007/978-3-642-22389-1_47

Aishwarya, R., Gayathri, P., & Jaisankar, N. (2013). A Method for Classification Using Machine Learning Tachnique for Diabetes. *International Iournal of Engineering and Technology, 5*, 2903-2908.

Alic, B., Gurbeta, L., & Bandjevic, A. (2017). Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases. *6th Mediterranean Conference on Embedded Computing.* Bar, Montenegro.

Association American Diabetes. (2018). Classification and Diagnosis of Diabetes: Standard of medical. *CareDiabetes, 41*, S13–S27.

Azeem, M. S., Kamal, N., Hamid, W., & Ali, M. S. (2018). Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. *24th International conference on Automation & Computing* (pp. 6-11). Newcastle: Newcastle University.

Dahiru, T., Aliyu, A. A., & Shehu, A. U. (2016). A review of population-based diabetes mellitus in Nigeria. *Sub-Saharan African Journal of Medicine, 3*(2), 59-64. doi:10.4103/2384-5147.184351

International Diabetes Federation. (2019). *IDF Diabetes Atlas. 9th Edition.* Brussels, Belgium,: International Diabetes Federation. Retrieved from International Diabetes Federation: http://www.idf.org/diabetesatlas

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process, 5*(1), 1-14. doi:10.5121/ijdkp.2015.5101

Kavakiotis, I., Tsave, O., Salifoglou, A., Nicos, M., Loannis, V., & Chouvarda, L. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal, 15*, 104-116. Retrieved from http://dx.doi.org/10.1016/j.csbj.2016.12.006

Ledolter, J. (2013). *Data Mining and Business Analytics With R.* Hoboken, New Jersey: John Wiley & Sons, Inc., .

Marcano-Cedeno, A., Torres, J., & Andina, D. (n.d.). A Prediction Model to Diabetes Using Artificial Metaplasticity. In J. M. Ferrandez, J. R. Alvarez Sanchez, F. de la Paz, & F. J. Tolado, *New Challenges on Bioinspired Applications* (Vol. 6687). Berlin, Heidelberg: Springer. Retrieved from https:///doi.org/10.1007/978-3-642-21326-7_45

Muniruzzaman, M., Kumar, N., Menhazul Abedin, M., Shykhul Islam, M., Suri , H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative Approaches for Classification of Diabetes Mellitus Dtat: Machine Learning Paradigm. *Computer Methods and Programs in Biomedicine*, 1-29. doi:10.1016/j.cmpb.2017.09.004

Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifier for the Risk of Diabetes Prediction. *Procedia Computer Science, 69*, 132-142. doi:10.1016/j.procs,2015.10.014

Oputa , R. N., & Chinenye, S. (2015). Diabetes in Nigeria – a translational medicine approach. *African Journal of Diabetes Medicine, 23*(1), 7-10.

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Pradhan, M., & Kumar, R. S. (2011,). Predict the onset of diabetes disease using Artificial Neural Network (ANN). *International Journal of Computer Science anf Emerging Technologies, 2*(2), 303-311.

Samer, N. E., & Samy, S. A.-N. (2018). Diabetes Prediction Using Artificial Neural Network. *Inataernational Journal of Advanced Science and Technology, 121*, 65-64. Retrieved from http://dx.doi.org/10.14257/ijast.2018.121.05

Sisodia, D., & Singh, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science, 132*, 1578-1585.

Suresh Kumar, P., & Umatejaswi, V. (2017,). Diagnosing Diabetes using Data Mining Techniques. *International Journal of Scientific and Research Publications, 7*(6), 705-709.

Team, R Core. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Torgo, L. (2017). *Data Mining WIith R* . Broken Sound Parkway NW: CRC Press Taylor & Francis Group .

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 Diabetes Mellitus Prediction Model Based on Data Mining. *Informatics in Medicine Unlocked*, 1-11. doi:10.1016/j.imu.2017.12.006

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). predicting Diabetes Mellitus with Machine Learning Techniques. *Frontiers in Genetics, 9*(515), 1-10. doi:10.3383/fgene.2018.00515