ISSN 2488-8648
http://www.ijbst.fuotuoke.edu.ng/

An Improved Model for Feature Selection on Type 2 Diabetes Risk Prediction in Nigeria

[1]Moko, A. and [2]Onuodu, F.E.
[1]Department of Computer Science and informatics, Federal University Otuoke, Bayelsa State, Nigeria,
[2]Department of Computer Science, University of Port Harcourt, Rivers State, Nigeria

Article Information

Article # 02016
Received: 9th May 2021
Revision: 16th July 2021
Acceptance:24th July 2021
Published: 26th July 2021

Key Words
Risk prediction, Models, Feature selection, Machine learning, Type 2 diabetes mellitus, Nigeria

Abstract

Diabetes mellitus (DM) is one of the world's fatal diseases, mostly in developed countries. It has become more predominant in developing countries such as Nigeria in recent years, posing more risks to individuals in the latter than the former. This study proposed a risk prediction model for Type 2 diabetes in Nigeria with feature selection adopting the qualitative and quantitative methodology. The Federal Medical Centre in Otuoke offered the dataset for this research used to develop the risk predictive machine learning models using python programming language in Collaboratory google based on logistic regression that uses dynamic regression technique to build predictors as a boolean combination of binary predictor variables, support vector machine a form of supervised learning technique widely used in classification and regression for medical diagnostic applications, gradient boosting builds a stadium additive model by executing a gradient descent in functional space, which is one of the most reliable and widely ensemble supervised learning methods, Decision Tree one of the most widely used data classification methods and Random Forest measures the quality of features by the impurity index, which means the average reduction from the division with the variable over all trees. The results attained revealed that in terms of accuracy and performance, the Gradient Boosting Algorithm predictive technique appears to be one of the optimally designed models, with 98%, followed by Decision Tree at 96% and Random Forest at 94% as their predictive accuracy score. The models can be incorporated into a digital system to help doctors better predict diabetes in patients and provide appropriate control measures. Sex, Age, Body Mass Index (BMI), Blood Pressure, Pulse rate, and Respiratory rate are the vital predictors in these models

***Corresponding Author**: Moko, A..; mokoaa@fuotuoke.edu.ng

Introduction

Diabetes mellitus is a disease problem associated with the human body caused by high blood glucose also called hyperglycemia, caused by poor eating habits, being overweight, and failing to exercise or engaging in physical activity etc, that can lead to other harmful risks. (Chou *et al*,.2019).
Diabetes is categorized into three types: type 1, type 2, and gestational diabetes, which can cause significant health problems if handled poorly. Type 1 diabetes, also recognized as insulin-dependent diabetes, is portrayed by the body producing a reduced amount of insulin. People with type 1 diabetes require insulin administration daily to regulate the glucose levels in their blood. If a person with type 1 diabetes does not have access to insulin, might end up dying. (Hassan *et al*,.2020) Type 1 diabetes cannot be prohibited because what causes it is still undiscovered. The symptoms of type 1 diabetes include loss of weight, frequent urination, exhaustion, changes in vision, and thirst. The symptoms of type 2 diabetes are like those

of type 1 diabetes because Type 2 diabetes was once prevalent in adults, but now some cases happen in children who have been diagnosed with the disease. Gestational Diabetes or type 3 diabetes occurs in women during pregnancy due to hormone changes. (Mahmud *et al*,. 2018)

Diabetes is becoming increasingly prevalent all around the world. The International Diabetes Federation (IDF) Diabetes Atlas predicts over 111 million people over the age of 65 with diabetes in 2019 showing an increasing rate as compared to previous years. As a result, diabetes is one of Nigeria's growing public health concerns with a 3.1% increase in prevalence It is estimated that diabetes will affect one out of five adults in this age group and is projected to affect 195 million people aged over 65 by 2030, and by 2045 the figure will be 276 million *(Nigeria diabetes report 2010 - 2045)*. The results show that the population of diabetes in aging societies will increase significantly for the next 25 years, and the resultant public and economic challenges. Increases in the population of the diabetes of aging societies and the resulting public health and economic challenges in the coming 25 years. Data shows that three out of four diabetics are in the working-age (i.e. 352 million people) between 20 and 64 years old. It is expected that this figure will rise by 2030 to 417 million and by 2045 to 486 million. This will have an escalating human impact and will make economic growth and development significantly more difficult over the decades to come. *(IDF Diabetes Atlas, 2019)*

The prevalence of Type 2 Diabetes Mellitus (T2DM) in Nigeria, Africa's largest country, has stood high and is still rising, with the country extensively discussed as having Africa's utmost diabetes risk. Diabetes prevalence in Nigeria as a percentage of the population aged 20 to 79 with type 1 or type 2 diabetes recorded in 1,00s is illustrated in Fig. 1. (Adeloye *et al*,. 2017)

According to IDF Diabetes Atlas 9th Edition, Nigeria has a prevalence Age-adjusted comparative prevalence of diabetes of 4.7% in 2010, 3.1% in 2019, and an estimate of 3.2% in 2030 and 3.2% in 2045 cases amid people aged 20–79 years and the ratio of people with undiagnosed diabetes 48% in 2019. Predicting the prevalence, epidemiology, and burden of type 2 diabetes mellitus (T2DM) in Nigeria, the few predictions identified could be based on enhanced modeling and projection of extremely limited data, and not invariably the actual hassle of the disease on the country.

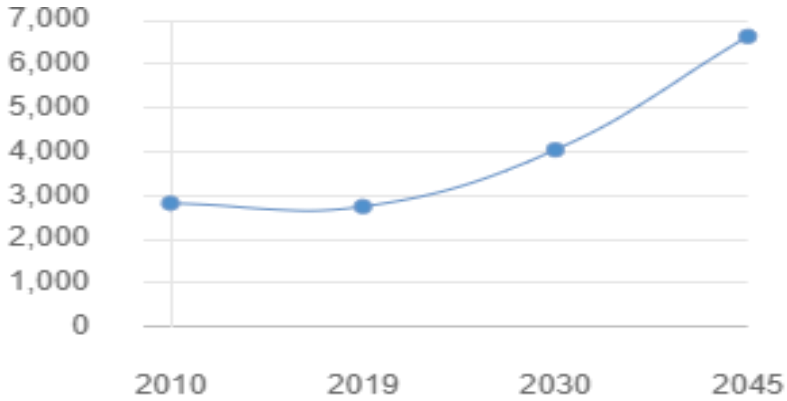


Fig 1. Nigeria Diabetes Prevalence Estimate in 1,000s (Source: IDF)

Type 2 Diabetes risk prediction has undergone various study and different models have been and are being developed to promptly predict the risk in Type 2 diabetes. However, due to the threat majorly it causes there are still some limitations that need improvement in predictive performance using the implementation of different approaches of the predictive method. This work aims to develop a predictive model using feature selection for the risk prediction of type 2 diabetes mellitus in Nigeria. The specific objectives are to design a model for type 2 diabetes risk prediction, implement with Python Programming Language on Google Colaboratory and evaluate our proposed system with the existing system performance.

In recent years T2DM different model prediction has been on the rise, many machine learning algorithms have been studied and used by researchers to optimally use to provide an early risk prediction of T2DM. this section studied a few relevant works related to the proposed issue.

Chen *et al.* (2017) proposed a hybrid prediction model for Type 2 Diabetes with the use of K-means and J48

decision tree for data reduction and classification. The study obtained experimental results from the UCI machine learning repository, and the results obtained a more optimal value of 90.04% when compared to other methods studied in the literature. Turi et al. (2017) introduced a general data representation method on diabetes risk factors and spline regression models to ascertain the strength of models with nonlinear and interaction terms. In adults aged 20 and older and multivariable adaptive regression spline (MARS) data waves from 2005–2006 to 2011–2012 were used. Results from the study show that a crucial risk aspect for type Diabetes increase in age for those above 69, alongside individuals with family records and lessened risk for younger than 45 individuals, both models used in the study had an accuracy score of 87% in classifying diabetes.

Based on the experience of several researchers for a valid comparison with other models, Wu et al (2017) proposed a new two-level regression algorithm model, improved K-means algorithm, and logistic regression. The model was centered on data mining methods to predict Type 2 Mellitus diabetes (T2DM). A series of preprocessing procedures were used to enhance the prediction model's accuracy and to make it adjustable to more than one dataset. The Pima Indian Diabetes Dataset and the Knowledge Analysis Waikato Environment Toolkit. The results demonstrated that the proposed model has achieved a 3,04% higher predictive accuracy than the others.

Zou *et al.* (2018) using a decision tree, random forest, and neural network for the prediction of diabetes mellitus, randomly selected 68,994 healthy individuals and diabetic patients as a framing set. Due to the unbalanced nature of the data, 5 times data were randomly extracted, and the outcome from the average of the 5 experiments using principal component analysis (PCA) and lowest redundancy maximum relevance (mRMR) to decrease the dimensionality displayed, the maximum accuracy (ACC=0.8084) with all features.

Due to the rise in the possibility of diabetes and hypertension, Ijaz *et al*. (2018) suggested a hybrid prediction model (HPM) that provides a prediction for T2D, and hypertension centered on key concerns. The projected model comprises of density-based spatial clustering of application with noise (DBSCAN)- based outliner discovery to eliminate the outliner data, class synthetic minority oversampling technique was used to stabilize the spread and the classification of the disease was done with Random Forest (RF). When compared to other models the HPM performed better in envisaging diabetes and hypertension. The research highlighted that HPM can be embedded into an IOT-based health surveillance system.

Lai *et al.* (2019) to properly recognize patients at risk of diabetes mellitus using demographic data and laboratory findings of patients in Canada, survey produced a potent predictive model using Logistic Regression and Gradient Boosting Machine (GBM) techniques to delicately collate the current records of 13309 patients aged 18 to 90 of age. Gauging the proposed model using the area under the receiver operating feature curve (AROC) results showed that AROC is 84,7% with an average sensitivity of 71,6% for the projected model of GBM, while AROC is 84,0% with a sensitivity of 73.4% for the projected model of logistic regression. The proposed models performed better when assessed to other machine learning technologies. An automated tool to predict the development of Type 2 Mellitus diabetes by using the techniques proposed by Abbas *et al*. (2019) for T2DM, produced data from the OGTT to develop a predictive model based on the support vector machine, to accomplish this goal the models were taught and certified using OGTT and 1.492 healthy people's demographics. At three time points of 30.60- and 120-min plasma glucose and insulin concentrations were taken before glucose intake. Findings show higher possible forecasting performance for prospective T2DM developments for plasma glucose levels and data obtained, with an average accuracy of 96.80% and a sensitivity of 80.09% in a validation set.

Xu and Wang (2019) suggested an ensemble learning methods risk prediction model for Type II diabetes, for an optimized feature selection and extreme gradient boost (Xgboost) classification, the proposed model used a weighted selective feature algorithm based on random forest (RF-WFS). The performance of the method proposed was validated with different metrics.

Xie *et al*. (2019) evaluated cross-sectional data from the 2014 behavioral system of a monitoring risk factor for 138,146 people, including 20,467 people with type 2 diabetes. To predict diabetes type 2, several machine-learning models were. For investigation of the association of possible risk factors with type 2 diabetes, the study employed single and multivariable weighted logistic regression models. Results from the study indicate that the predictive models for type 2 diabetes ranged from 0.7182 to 0.7949 with a large area under the curve (AUC). Although the most accurate (82,4 percent), specific (90,2 percent) and AUC (0,7949) model was present in the neural network model, the decision tree model was the most sensitive (51,6 percent) for type 2 diabetes.

Nguyen *et al.* (2019) applied a broad and deep learning model, which integrates a generalized linear model with a wide range of features and a deep feed-forward network to help enhance the prediction of type 2 diabetes (T2DM). Methods and Materials were done by training different models in a logistic loss function using stochastic gradient descent. Synthetic Minority Oversampling Technique (SMOTE) was used to analyze the performance of each cross-validation fold for the imbalance class in which synthetic examples are created for the minority class. The proposed method was implemented results showed that the ensemble model not with SMOTE achieved 84.28% overall accuracy under the recipient functional curve (AUC) of 84.13 percent, 31.17 percent sensitivity, and 96.85 percent specificity. Utilizing the 150 and 300 percent SMOTE, there was no improvement in AUC (83.33 and 82.12 percent) but a significant reduction in specificities (49.40 and 71.57 percent, respectively).
Nair and Bhagat (2019) show how feature selection works and classifies using a select best and select percentile, the paper showed how accuracy is improved using classification algorithms through feature selection used in machine learning. The outcome of the study showed that the enabled algorithms, Naïve Bayes, Support Vector Machine, Logistic Regress, and K- Nearest Neighbour to be better classified, with Logistic Regression achieving a greater precision with 96.9 percent compared to all other algorithms.
Muhammed *et al.* (2020) with a diagnostic dataset obtained from Murtala Mohammed Specialist Hospital, Kano proposed predictive machine learning models based on support vector machine, logistic regression, K-nearest neighbor, Gradient Boosting, and Naive Bayes algorithm. The study results showed that the random forest predictive model obtained the highest accuracy score with 88.76%, while the gradient boosting, and random forest predictive models got the highest predictive score of 86.28% respectively.
Kopitar *et al.*(2020) study equate machine learning-based prediction models for the prediction of diagnosed T2DM (i.e. Glmnet, RF, XGBoost, LightGBM) with frequently used regression models. The prediction efficacy was evaluated with 100 bootstrapping iterations in various data subsets simulating new input data in six months. A simple regression model with the smallest average RMSE of 0.838, preceded by RF (0.842), Light GBM (0.846), Glmnet (0.859), and XGBoost with six months of data available (0.881). The study results showed no clinically significant progress when further advanced prediction models were applied and an increased permanence of selected models over time helps to simplify model understanding and model calibration must also be taken into consideration in the growth of clinical prediction models.
Hassan *et al*. (2020) opined that diabetes a disease affecting people metabolically leads to severe problems such as premature death, kidney failure, stroke etc. the study proposed using classifications techniques decision tree, K-nearest neighbour, and support vector machine (SVM) to sort patients with diabetes. The techniques were utilized to determine the performance using the factors of accuracy sensitivity, precision, and specificity. The findings showed that SVM produced the highest accuracy performance of 90.23% against the other used classifiers.
Deberneh and Kim (2021) proposed a model of T2D for the prediction of prevalence that may predict T2D as regular, prediabetes, or diabetes in the preceding year. The prediction models were generated using LR, RF, XGBoost, SVM, and ensemble classifiers (CIM, ST, SV). The selection of functions has been used to select major features that can effectively differentiate between the three classes. FPG, HbA1c, triglycerides, IBM, gamma-GTP, sex, age, uric acid, smoking, beverage, physical exercise, and family history were the features selected. Experimental findings showed that the prediction model performance in forecasting T2D in the Korean population was good.

Methodology

This section involves specifying the methods for collecting and analyzing data required to define or solve the problem for which the research is being conducted. During the development process, the Hybrid Methodology: Qualitative and Quantitative methods is integrated in this research. This method adopts a systematic approach that combines and analyzes with more contextualized insights into statistical information. Using mixed methods, allows data from two or more sources are also triangulated or substantiated.

Analysis of Existing System

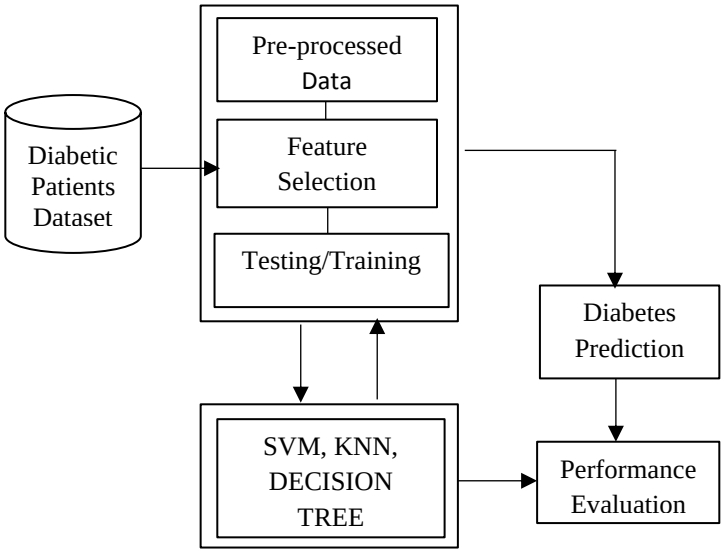


Fig 2. The architecture of the Existing System (Hassan *et al*., 2020)

Algorithm for the Existing System
Step 1: Input Datasets; Step 2: Preprocess Data; Step 4: Feature selection; Step 5: Test and train dataset.;
Step 6: Apply prediction models on the trained dataset;
Step 7:Do performance evaluation.

Analysis of Proposed System
Data Collection and Analysis
One hundred patients were investigated, and Six features were collected for each patient with approval given from Federal Medical Centre Outreach, Otuoke with approval. The dataset of training was achieved by sampling 70% of all patients and a test set with the rest of the 30%. This shows that from the data collected, 70 patients were used for the training set and 30 patients in the testing. The training dataset was used to train the model and to evaluate how well the model performs. Each record is used as a predictor variable. Sex (1 for male and 2 for female), Age (years), Body Mass Index (BMI), Blood Pressure (systolic) (mm Hg), Blood Pressure (diastolic) (mm Hg), Pulse rate (bpm), and Respiratory rate (bpm) are the features collected. The features were gathered from Federal Medical Center Otuoke, Nigeria thereby forming a general basis for T2DM prediction tools.

Table 1. Sample of the Dataset

Sex	Age	Systolic	Diastolic	BMI	PR	RR
2	28	150	90	26.64	72	20
2	46	120	60	23.23	74	20
1	24	90	60	23.44	80	18
2	33	100	60	27.98	72	22
1	68	140	100	33.06	104	20

1	36	100	60	23.61	124	32
1	29	110	70	38.1	72	20
2	19	90	60	24.86	78	20

Feature Selection

To enhance the performance of a model and reduce the number of input variables for the computational cost of modeling, the procedure by which input variables can be reduced when the predictive model is developed is known as feature selection. The methods of feature selection are thought of in terms of supervised and unsupervised methods (Brownlee,2019).

Feature Selection Attributes; it allows the learning algorithm for machines to train more quickly, it reduces a model's complexity, and facilitates interpretation, if the right subset is chosen, it improves the accuracy of the model and, it lessens excess capacity.

Prediction Techniques

In the previous section, the data set and feature selection were presented. In addition, data pre-processing was carried out, which involves the removal of inaccurate, contrary, and inconsistent data. This section gives a brief explanation of the different predictive models employed.

Logistic Regression is a Machine Learning algorithm that uses the dynamic regression technique to build predictors as a Boolean combination of binary predictor variables. The algorithm is applied to a training set to find a single Boolean expression that predicts a binary outcome. Many Boolean expressions can be investigated and simultaneously embedded into a linear regression model in the case of a regression task. (Muhammed et al,. 2020).

Support Vector Machine (SVM) This is a form of supervised learning technique widely used in classification and regression for medical diagnostic applications. SVM minimizes the empirical classification error while maximizing the percentage, which is strongly preferred with reduced computational power due to its significant accuracy. (Maniruzzaman et al,. 2018)

A Decision Tree is one of the most widely used classification methods used for data classifying, depending on the importance of each attribute, the attribute with greater importance which is from the root will be given precedence and the procedure will be repeated until the leaf node attains via the internal node (Hassan et al,. 2020)

Random Forest is when a group of decision trees uses a bagging approach. It generally results in a prediction that is precise and reliable. RF measures the quality of features by the impurity index, which means the average reduction from the division with the variable over all trees.

Gradient Boosting (GBM) It builds a stadium additive model by executing a gradient descent in functional space, which is one of the most reliable and widely ensemble supervised learning methods. GBM calculates the functionality of the splitter, weighed by the mean square improvement to the model, and averaged over all trees accordingly, according to the number of times a variable is chosen. (Chou et al,.2019)

Algorithm of the Proposed System

Step 1: Input Data; Step 2:Preprocess Data: Step 3; Train Data 70%: Step 4; Compute feature Selection with the importance; Step 5:Select feature for splitting with a high score; Step 6:Test Data 30%:Step 7; Predict data using prediction models: Step 8; Do classification: Step 9: Do Performance Evaluation: Step 10;End

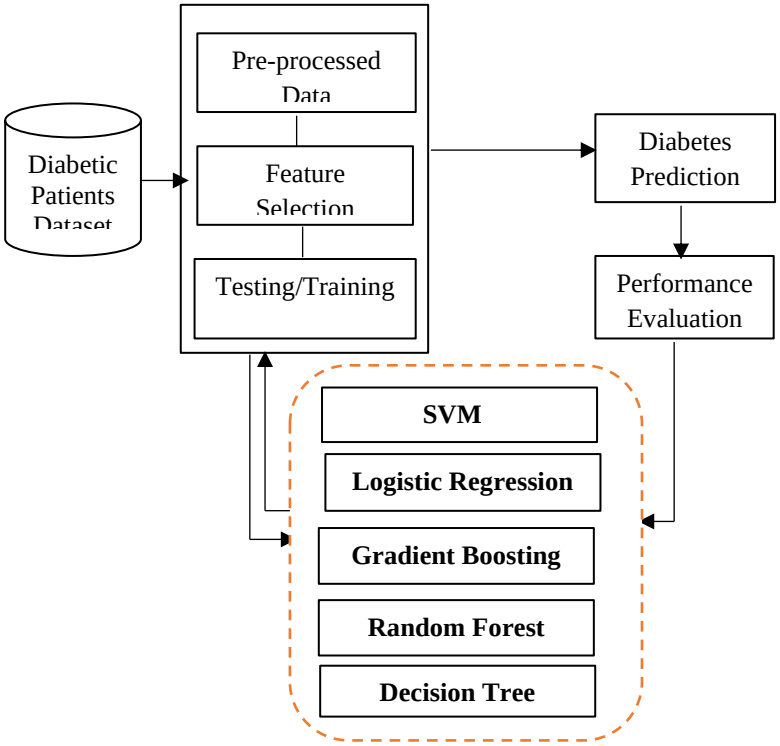
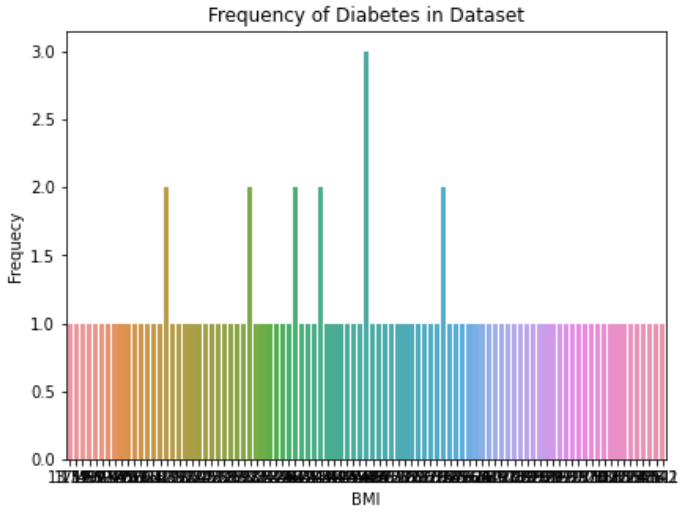


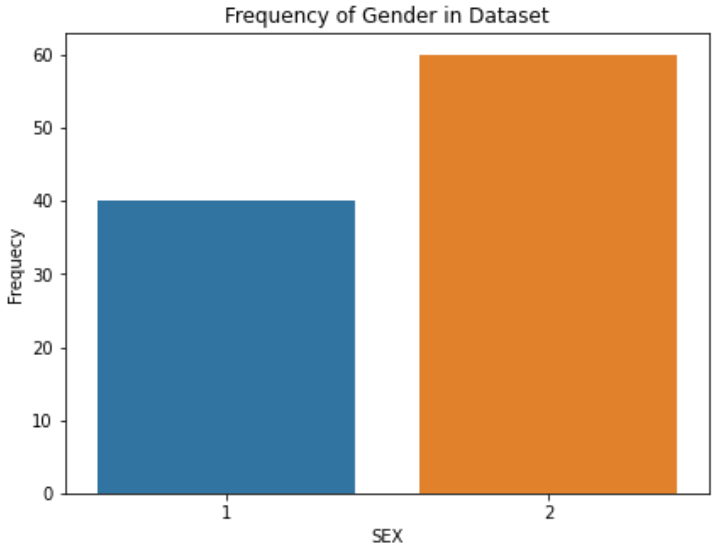
Fig 3. The architecture of the Proposed System

Results and Discussions

Following the preparation of the data set and the use of the feature selection technique, the sub-models were loaded with input data, and finally, a few other experiments were conducted on the sub-models. Producing the below-displayed graphs.



(a)



(b)

Figure 4. (a) Frequency of Diabetes in Dataset (b) Frequency of Gender in Dataset

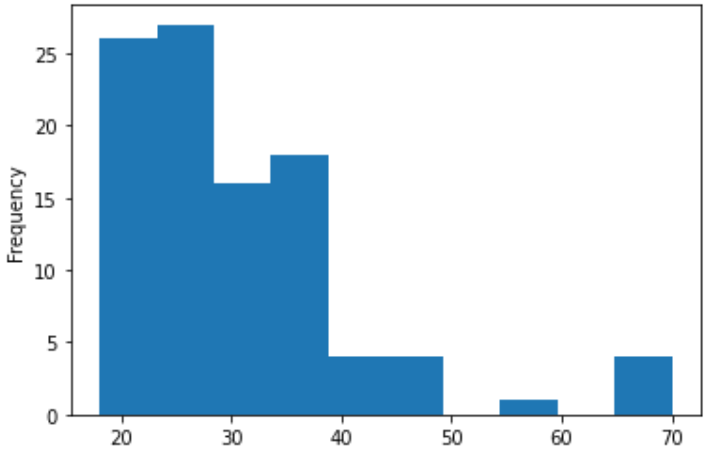


Figure 5. Age histogram Plot

The main problems which we have resolved are to improve prediction performance and adapt the model to various datasets. Figure 4a. shows the frequency of diabetes in the dataset from the Body Mass Index of patients with the lowest frequency at 1.0Hz and the highest at 3.0Hz, while Figure 4b. depicts the frequency of gender in the dataset with 60% of females (2) and 40% for males (1) and Figure 5 shows the Age histogram plot of patients which ranges from ages 18 to 70.

This report shows that the algorithms we proposed contributed significantly to the prediction model, with greater predictability than the experimental results of other researchers. Decision Tree with an accuracy score of 96%, 44% for Logistic Regression, Support Vector Machine 38% Random Forest 94%, and Gradient Boosting with the highest accuracy score of test data at 98%. When compared to other existing predictive models our model gave a more optimal predictive accuracy score.

Performance Analysis

Table 2. below shows the performance analysis of our proposed method for risk prediction of type 2 diabetes to the existing method proposed by Hassan et al. (2020). Often different studies use cross-validation methods to fix prediction model classification performance. This study adopted the confusion matrix a tool for the performance analysis classification results of the model. The four confusion matrix parameters used to determine the classifier's predictions of results are true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). Our model obtained a higher predictive result using the Accuracy Score which is known as the percentage of cases correctly predicted to the total number, Precision, Recall and F-1 Score as seen in equations below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F-1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Table 2. Performance Evaluation Table

S/N	Parameters	Existing System	Proposed System	Parameters
1.	Accuracy Score	90.23%	98%	Accuracy Score
2.	Number of Models	3	5	Number of Models
3.	Precision (P)	77%	97%	Precision(P)
4.	Recall (R)	87%	98%	Recall (R)
5.	F-1 score	83%	97%	F-1 score

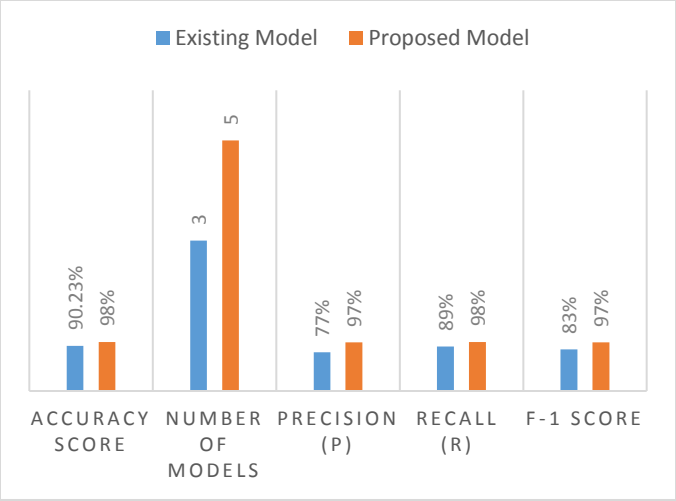


Figure 7. Performance Evaluation Chart

Conclusion

Diabetes mellitus is a disease problem related to the human body caused by high blood glucose also called hyperglycemia, caused by poor eating habits, being overweight, and failing to exercise or engaging in physical activity etc, that can lead to other harmful risks. In this study, a risk predictive learning model for type 2 diabetes mellitus has been based largely on logistic regression, vector support system, random forest, decision tree, and algorithms of gradient boosting. However, 98% of the proposed model showed the gradient boosting prediction-based learning model to be the suitable classifier, with other models like Decision tree with an accuracy score of 96% and Random Forest with a score of 94% in diagnosing and forecasting diabetes mellitus type 2 in suspected patients, this model will be helpful and useful to researchers and health personnel. This research faced so many limitations, to mention few, in respect to dataset acquisition it was realized that there is a poor and incomplete method of data collection in Nigeria, of which there is a need for proper data collection and storage method to help in making accurate and life-changing predictions.

Contributions to Knowledge

An improved model for feature selection on type 2 diabetes risk prediction in Nigeria has been proposed in this study and the risk prediction method recommended gives a higher accuracy score than other presented risk prediction models.

Suggestions for Future Work

Future studies should be geared towards achieving a higher Type 2 diabetes mellitus risk prediction model, using more features and models.

References

Abbas, H. T., Alic, L., Erraguntla, M., Ji, J. X., Abdul-Ghani, M., Abbasi, Q. H., and Qaraqe, M. K. (2019). Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. *PLOS ONE*, *14*(12).

Adeloye, D., Ige, J. O., Aderemi, A. V., Adeleye, N., Amoo, E. O., Auta, A., and Oni, G. (2017). Estimating the prevalence, hospitalization and mortality from type 2 diabetes mellitus in Nigeria: A systematic review and meta-analysis. *BMJ Open*, *7*(5),

Brownlee, J. (2019, November 26). How to Choose a Feature Selection Method for Machine Learning. *Machine Learning Mastery*. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Chen, W., Chen, S., Zhang, H., and Wu, T. (2017). A hybrid prediction model for type 2 diabetes using K-means and decision tree. *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 386–390.

Chou, J., Flory, J., and Wang, F. (2019). Feature Selection in Predictive Modeling: A Systematic Study on Drug Response Heterogeneity for Type II Diabetic Patients. *AMIA Summits on Translational Science Proceedings*, *2019*, 295–304.

Dahiru, T., Aliyu, A., and Shehu, A. (2016). A review of population-based studies on diabetes mellitus in Nigeria. *Sub-Saharan African Journal of Medicine*, *3*(2), 59.

Deberneh, H. M., and Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *International Journal of Environmental Research and Public Health*, *18*(6), 3317.

Elavarasan, D., Vincent P M, D. R., Srinivasan, K., and Chang, C.-Y. (2020). A Hybrid CFS Filter and RF-RFE Wrapper-Based Feature Extraction for Enhanced Agricultural Crop Yield Prediction Modeling. *Agriculture*, *10*(9),

Feature Selection Methods | Machine Learning. (2016, December 1). *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Hassan, A. S., I. Malaserene, and A. Anny Leema. (2020). Diabetes Mellitus Prediction using Classification Techniques. *International Journal of Innovative Technology and Exploring Engineering*, *9*(5), 2080–2084.

IDF Diabetes Atlas 9th edition (2019). (n.d.). Retrieved April 28, 2021, from https://www.diabetesatlas.org/en/

Ijaz, M. F., Alfian, G., Syafrudin, M., and Rhee, J. (2018). Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest. *Applied Sciences*, *8*(8), 1325.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., and Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, *10*(1), 11981.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., and Gao, X. (2019). Predictive models for diabetes

mellitus using machine learning techniques. BMC Endocrine Disorders, 19(1), 101.

Larabi-Marie-Sainte, Aburahmah, Almohaini, and Saba. (2019). Current Techniques for Diabetes Prediction: Review and Case Study. Applied Sciences, 9(21), 4604.

Mahmud, S. M. H., Hossin, M. A., Ahmed, Md. R., Noori, S. R. H., and Sarkar, M. N. I. (2018). Machine Learning-Based Unified Framework for Diabetes Prediction. *Proceedings of the 2018 International Conference on Big Data Engineering and Technology - BDET 2018*, 46–50.

Maniruzzaman, Md., Rahman, Md. J., Al-MehediHasan, Md., Suri, H. S., Abedin, Md. M., El-Baz, A., and Suri, J. S. (2018). Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. *Journal of Medical Systems*, *42*(5), 92.

Muhammad, L. J., Algehyne, E. A., and Usman, S. S. (2020). Predictive Supervised Machine Learning Models for Diabetes Mellitus. *SN Computer Science*, *1*(5), 240. https://doi.org/10.1007/s42979-020-00250-8

Nair, R., and Bhagat, A. (2019). *Feature Selection Method to Improve the Accuracy of Classification Algorithm*. *8*(6), 4.

Nguyen, B. P., Pham, H. N., Tran, H., Nghiem, N., Nguyen, Q. H., Do, T. T. T., Tran, C. T., and Simpson, C. R. (2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer Methods and Programs in Biomedicine*, *182*, 105055.

Rezaee, M., Putrenko, I., Takeh, A., Ganna, A., and Ingelsson, E. (2020). Development and validation of risk prediction models for multiple cardiovascular diseases and Type 2 diabetes. *PLOS ONE*, *15*(7), e0235758. https://doi.org/10.1371/journal.pone.0235758

Snášel, V., Abraham, A., Krömer, P., Pant, M., and Muda, A. K. (Eds.). (2016). Innovations in Bio-Inspired Computing and Applications: Proceedings of the 6th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2015) held in Kochi, India during December 16-18, *2015* (Vols. 323–335). Springer International Publishing.

Turi, K. N., Buchner, D. M., and Grigsby-Toussaint, D. S. (2017). Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors. *Preventing Chronic Disease*, *14*, 160244. https://doi.org/10.5888/pcd14.160244

World Development Indicators (WDI)—Knoema.com. (n.d.). Knoema. Retrieved April 28, 2021, from https://knoema.com//WBWDI2019Jan/world-development-indicators-wdi

Wu, H., Yang, S., Huang, Z., He, J., and Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100–107.

Xie, Z., Nikolayeva, O., Luo, J., and Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, *16*, 190109.

Xu, Z., and Wang, Z. (2019). A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier. *2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI)*, 278–283.

Zhang, L., Wang, Y., Niu, M., Wang, C., and Wang, Z. (2020). Machine learning for characterizing the risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. *Scientific Reports*, *10*(1), 4406.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontiers in Genetics*, *9*, 515.