Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DIABETIC MELLITUS PREDICTION WITH BRFSS DATA SETS

MARWA HUSSEIN MOHAMED [1,*] , MOHAMED HELMY KHAFAGY [2] , NESMA MOHAMED MAHMOUD KAMEL [2] , AND WAEL SAID [3]

[1] Faculty of Information system and computer science; Information System Department; October 6 University; Cairo, Egypt.
[2] Faculty of Computers & Artificial Intelligence; Computer science Department; Fayoum University; Cairo, Egypt.
[3] Faculty of Computers and Informatics; Computer science Department; Zagazig University; Cairo, Egypt.

E-mail: [1] Eng_maroo1@yahoo.com ; [1] Marwa.hussien.csis@o6u.edu.eg ;[2] Mhk00@Fayoum.edu.eg , [2] nesma710@gmail.com ; [3] wael.mohamed@zu.edu.eg

ABSTRACT

One of the chronic diseases that affect many people worldwide is diabetic mellitus. If the disease is predicted at an early stage, the risk and severity can both be significantly decreased. In this research, we need to predict the type 2 diabetic patients at an early stage to reduce the cost of treatment for countries because this is a long time disease we use many machine learning algorithms to find the accuracy for these diseases applied to BRFSS datasets for two years 2014 and 2015 with a different selection of features to predict the disease as decision tree, logistic regression, ADA Boost Classifier, extreme gradient boosting, Linear Discriminant Analysis, Light Gradient Boosting Machine, and catboost classifiers. While applying our experiments with the 2014 BRFSS data sets Neural network has the highest accuracy with 82%and with the 2015 BRFSS datasets the best accuracy model was 86% for CatBoost Classifier and Extreme Gradient Boosting where the lowest model was Linear Discriminant Analysis. Also, in our research we compare our results with others using the same datasets with different features selection and get high accuracy.

Keywords: *Chronic Diseases; Diabetic Mellitus; Machine Learning; Artificial Intelligence; Classification Models.*

1. INTRODUCTION

In the field of medicine, determining a patient's health status is a highly difficult task [1] One of the most important challenges in both developed and emerging countries. Medical history information includes several tests that are necessary to diagnose a specific disease, and the diagnosis is based on the doctor's experience; a doctor with less experience may diagnose a problem wrongly.

All parties involved in the healthcare industry can tremendously benefit from da-ta mining techniques. There is a large amount of data related to healthcare, but until it is transformed into knowledge and information, it is of little supervisory value [2]. Knowledge and information may assist in controlling expenses, increase profitability, and maintaining a high standard of patient care.

Diabetes Mellitus (DM) is a chronic illness brought on by an inadequate supply of insulin or an anomaly in the way that insulin is used to regulate the metabolism of carbohydrates, proteins, and fats. Insulin transfers energy [3]. It is an important hormone because it carries glucose from the blood into the cells of the body. Excessive thirst or urination, fatigue, weight loss, or blurred vision are some of the symptoms. The likelihood of developing polyuria (frequent urination), polydipsia (growing thirst), and polyphagia (hunger) increases as a result. DM, a common non-communicable dis-ease, is steadily rising to the top of the list of causes of death [4].

Diabetes is a chronic disease with a significant global growth rate; according to the International Diabetes Federation, 537 million people worldwide have diabetes, including 73 million in the Middle East and North Africa (MENA), and this figure is expected to grow to 135.7 million by 2045. Type 2 diabetes affects approximately 16% of persons in Egypt.

Type 2 diabetes can be predicted and can be delayed or even prevented because it usually develops due to several factors such as family history, age, unhealthy lifestyle, and more. Numerous models have been developed for estimating the risk of getting diabetes with type 2 using survey data, but their performance, particularly their sensitivity, might be improved. [5].

The researchers [6] findings may have implications in the public health domain, as it provides a potential technique for initial screening for diabetes mellitus type 2 at a low cost. This research paper is used for early prediction of the disease and early implementation that can decrease the risk of developing such a disease.

On the other hand, preventing or decreasing the prevalence of diabetes mellitus will have a great contribution in decreasing the healthcare cost, saving a lot of resources, and This will be accomplished by utilizing machine learning approaches to keep individuals' lives healthy.

This study employs machine learning to forecast diabetes mellitus. This work con-tributes significantly in the following ways:

•A key contribution of this work is the publication of a unique diabetes mellitus dataset covering 61,118 diabetes cases, 12,699 were pre-diabetic and 390,827 were healthy. In this paper, the dataset was collected from 2014 and 2015 dif-ferent patients' data for the two years.

•The dataset has 279 features we only select 27 variables which is the im-portant to determine the diabetic patients.

•We apply a pre- processing step to remove the missing values and then make the categorical to be ready to use with python and apply the machine learning techniques.

•SMOTE techniques are used to reduce the issue of class imbalance. Also, the data must be balanced amongst the three types.

•This method aids in interpreting which features were employed and how they affected the accuracy results.

•This technique aids in interpreting which features were employed and how they affected the accuracy results in predicting the three types of diabetic patient.

•The unique aspect of this work is that it predicts diabetes patients with high accuracy results while the size of data is huge and get the results of the model on 2014 and 2015 BRFSS datasets by using different machine learning techniques.

The rest of the paper organized as follow section 2 will show the previous work and results with different datasets to predict type 2 diabetic patients, Section 3 will describe the characteristics and qualities of the BRFSS datasets used in our experiments, section 4 new proposed algorithms steps, section 5 the experimental results and the last section for the conclusion and future work.

2. RELATED WORKS

Diabetes Type 1 symptoms come on more quickly and are more severe. Following are a few of type 1 and type 2 diabetes symptoms and signs:

- Ketones are present in the urine.
- thirst increases
- Urinating frequently
- To the point of death from hunger
- Frequently losing weight
- Fatigue
- Vision issues
- Infections that repeat frequently, such as vaginal infections and gum or skin infections
- A BMI of over 25 is considered obese.

G. Geetha, etl [7] propose a new model named (T2DDP) to alert people and patients has type 2 diabetes early to reduce the risk of this disease by the use of machine learning algorithms that are supervised like Naïve Bayes, Random Forest, and Ada-boost for decision tree they combine hybrid models to increase the accuracy results. The results of their models will send notifications to the patient's phone at an early stage to follow up with a doctor and take an immediate decision about the treatment. They use the Pima Indian datasets in the experimental results and divide the data into different classes like 85/15, 80/20, 70/30, and 60/40 with k10-fold cross-validation. Every time they calculate the accuracy. Their suggested model removes the outliers from the data and solves the missing values by using the average values than deleting these records the new model used stacking ensemble machine learning algorithms to combine the results from all models to get high accuracy, the final results are Naïve Bayes has 75.11%, Bagging with Random Forest has 93.08%, Adaboost for Decision Tree has

92.16% and Proposed T2DDP model has 96.56% accuracy results.

R Saxena,etl [8]they use Pima Indian datasets to find the early prediction of the diabetes mellitus they try different machine learning algorithms on this dataset such as Naïve Bayes, Support vector Machine, Random Forest, Neural Network, and logistic regression they get the results of the confusion matrix for every algorithm The patients were classified using the Nave Bayes Classification:

• 164 (True positive): a diabetic patient is predicted
• 104 (False Negative): are expected to have no diabetes yet are diabetic.
• 78 (False positive): are projected to have diabetes but are not diabetic. (Erroneous positive)
• 422 (true negative): the patient is predicted to be non-diabetic.

In the experiments, they use 10-fold cross-validation to calculate the accuracy of each model. We will list the outcomes for each method accuracy, Correctly Classified Instances, and Incorrectly Classified Instances, with a total of 768 datasets:

•Logistic Regression has 77.2 %,593, 175.
•Support Vector Machine 77.08 %,592, 176.
•Naïve Bayes 76.30 %, 586, 182.
•Random Forest 75.5 %, 580, 188.
•Neural Networks 75.1 %, 577, 191.

Based on the accuracy of models' logistic regression classification has the best classifier to predict diabetic patients.

B A C Permana, etl [9] Diabetes manifests itself in a variety of ways. They must find the most important features that detect the disease using a data mining decision tree (C4.5) experimental results applied on secondary data obtained from the early-stage diabetes dataset, which can be accessed via https://www.kaggle.com/singhakash/early-stage-diabetes-risk-prediction-datasets with 520 records. They compute the entropy and gain values of each parameter to determine the decision tree's root. The most essential symptom is polydipsia, which has a 90.38% accuracy rate. Diabetes can be recognized early in people who have polydipsia symptoms.

Chollette C. Olisah, etl [6] created a twice-growth deep neural network (2GDNN) model to predict diabetic patients using neural networks and comparing it to other machine learning methods such as random forest (RF) and support vector machine (SVM). In their experiment, the novel model uses Spearman correlation as the first step and polynomial regression for feature selection and missing value imputation as the second phase. PIMA Indian and the Laboratory of Medical City Hospital (LMCH) datasets were used. The Pima dataset has two classes of diabetic and non-diabetic patients, whereas the LMCH dataset has three groups of diabetics, prediabetic, and non-diabetic patients. This will evaluate the new model's performance based on all features and select the most significant attributes to run the data. Based on the experimental results, the proposed 2GDNN achieves F1-score, train-accuracy, and test-accuracy scores of 97.34%, 97.24%, 97.26%, 99.01%, 97.25 and 97.28%, 97.33%, 97.27%, 99.57%, 97.33.

A.Sumathi, etl [10] The first hybrid prediction model for type 2 diabetes pattern (HPMT2D) and the second type 2 diabetes mellitus prediction model (T2DMPM) were constructed using the R tool with Bio Weka. They also employ the diabetes pattern detection approach with a tree ensemble clustering classifier (DDTEC) to predict type 1 and type 2 diabetes. The findings of the study were applied to the Australasian Diabetes in Pregnancy Society (ADIPS). The K-Means and LR classifier are the two machine learning algorithms employed in this model. Accuracy, recall, specificity, precision, and F-measure are used to evaluate the model. This dataset is divided into three categories: diabetic, prediabetic, and non-diabetic. The accuracy findings for HPMT2D, T2DMPM, and DDTEC are as follows: type 1, 89.10%, 84.60%, and 91.00%, respectively. Type 2 92.38%, 87.10%, and 93.40%. Gestational 87.50%, 85.20%, and 90.25%.

Victor Chang, etl [11]this paper used IOT technology to implement the Internet of Medical Things (IoMT) environment to diagnose type 2 diabetes. IOMT will help to collect data easily. The model uses Pima Indian datasets with two classes this data has many features to detect diseases by using the Weka program. The tree classifier models used are the J48 decision tree, Random Forest, and naïve Bayes. They try to run the model on all features data and on three features and the last experiment on five features to measure the accuracy. The random forest has

the highest accuracy 79.57% while selecting all features of diseases, naïve Bayes has the highest accuracy 79.13% while selecting the three factors of the datasets, and naïve Bayes 77.83% with selecting the five factors.

3. Dataset

We use the Behavioral Risk Factor Surveillance System (BRFSS) 2014 [12]to predict type 2 diabetes mellitus. This dataset has 464,644 diabetes cases and only has three classes diabetic and pre-diabetic patients and non-diabetic healthy people with 279 features. The link to download the data is data (https://www.cdc.gov/brfss/annual_data/annual_2014.html). this data has 61,118 diabetes cases, 12,699 were pre-diabetic and 390,827 were healthy.

Also, we use the BRFSS 2015 [12] Our study is based on the 2015 BRFSS which contains responses from 441,455 individuals in the USA and has 330 variables (https://www.cdc.gov/brfss/annual_data/annual_2015.html). These variables are either direct answers from participants or calculated variables based on participant answers. The researchers added many variables and features that may increase disease prediction in the 2014 and 2015 datasets, such as tobacco use, HIV/AIDS knowledge and prevention, exercise, immunization, health status, healthy days health-related quality of life, health care access, hypertension awareness, arthritis burden, chronic health conditions, alcohol consumption, fruits and vegetables, and seatbelt use.

4. NEW PROPOSED ALGORITHM

The Methodology to design a Multivariate Dataset in making a dataset with pictures to help people who can't see well in detecting objects and faces involves a few important steps. In the following, various steps were explained in making a set of data while highlighting the importance of training the model correctly for creating a Multivariate Dataset as shown in Fig.1.

Firstly we apply our architecture on BRFSS 2014 datasets [12]with limited pre-processing steps and machine learning algorithms Our new algorithm needs to predict the type 2 diabetic patients BRFSS datasets has collect data in 2014with 464,644 diabetes cases this data has three classes we make a preprocessing step to handling the missing data such as “do not know answers” and “refused to answer” and null values to use python to run different machine learning techniques and the difference between the diabetic and non-diabetic patients make the data unbalanced this may affect the accuracy results while using supervised machine learning. We employ smote to balance the data and support vector machine (SVM), logistic regression, Gaussian Naive Bayes, decision tree, random forest, and neural network to tackle the unbalancing in the dataset. We divided the data into two parts, two-thirds for the training set and one-third for the testing set, and chose 27 variables such as general health status, mental health status, health care insurance, checkup frequency, exercise, quality of sleep, presence of health problem that requires equipment, blind-ness, concentrating difficulty, angina or coronary heart disease presence, depression, presence of renal disease, receiving a flu shot, smoking status, physical activity, sex, ethnicity, body mass index, marital status, education level, employment status, annual income, and other factors.

The evaluation of the model performance was based on accuracy, specificity, sensitivity, and AUC table 1 lists all results output.

TABLE I: Predictive Model Performance for Type 2 Diabetes Using Data from the Behavioural Risk Factor Surveillance System, 2014.

algorithm	***Accuracy***	***Sensitivity***	***Specificity***	***AUC***
Neural network	*82.41%*	*37.81%*	*90.16 %*	*79.49 %*
Logistic regression	*80.68%*	*46.34 %*	*86.66 %*	*79.32 %*
Linear SVM	*80.82%*	*42.60%*	*87.46 %*	*78.07%*
Rbf SVM	*81.78%*	*40.14%*	*89.02 %*	*77.88 %*
Random forest	*79.27%*	*50.29%*	*84.31 %*	*76.08 %*
Naïve Bayes	*77.56%*	*48.76%*	*82.56 %*	*75.98 %*
Polynomial SVM	*79.62%*	*45.155*	*85.61 %*	*75.87 %*
Decision tree	*74.26 %*	*51.61 %*	*78.20 %*	*71.82 %*

The evaluation of eight predictive models yielded high area under the curve AUC values ranging from 0.72 to 0.79. The highest accuracy and specificity were for the neural network model 82.4% and 90.2%, and the highest sensitivity was for the decision tree model 51.6% for type 2 diabetes.
secondly, we apply our architecture on BRFSS 2015 datasets with the next steps.

4.1 Data preprocessing

The original BRFSS datasets have 330 features. According to prior study, high levels of serum uric acid, sleep quality/quantity, smoking, depression, cardiovascular dis-ease, dyslipidemia, hypertension, ageing, ethnicity, family history of diabetes, physical inactivity, and obesity are all risk factors. In our research we select only 21 features to predict the diseases high blood pressure, high cholesterol, smoking, obesity (BMI), age, sex, having a stroke, having coronary heart disease, physical activity, Consuming Fruit, Consuming Vegetables, Heavy drinking, health care coverage, ability to see a doctor, general health, mental health, physical health, difficulty walking, educational and income level.
We must clean the data firstly from missing values as “don’t know” or “refused” The datasets to be readable in Python need to be numerical data as 0 for no diabetes and 1 for pre-diabetes and diabetes. The variable name was changed to “Diabetes_binary” to be more readable.
The high blood pressure variable has been changed, as category 1 changed to 0 which represents No high blood and 2 changed to 1 to represent high blood pressure. In the variable of high cholesterol, category 2 which indicates no cholesterol changed to 0. Also, For the cholesterol check variable, category 3 which indicates “never check cholesterol” and category 2 which indicates “no cholesterol checks in past 5 years” have been changed to 0.
In the smoking variable, having a stroke, or coronary heart disease category 2 which indicates “no” changed to 0. The same was done in physical activity, fruit, and vegetable consumption.
In the high alcohol consumption variable, category 1 which indicates “no heavy drinking” has been changed to 0 and category 2 which indicates “heavy drinking” has been changed to 1. In health plan and medical cost, category 2 has been changed to 0 for no. In mental health status and physical health status, the scale of 0-30 days was kept, and number 88 which indicates no mental or physical disease has been changed to 0 days.
In the difficult walking variable category 2 has been changed to 0 “no”. In the sex variable, category 2 has been changed to 0 “female”.
All variables were renamed to be more readable, 'DIABETE3', '_RFHYPE5', 'TOLDHI2', '_CHOLCHK', '_BMI5', 'SMOKE100', 'CVDSTRK3', '_MICHD', '_TOTINDA', '_FRTLT1', '_VEGLT1', '_RFDRHV5', 'HLTHPLN1', 'MEDCOST', 'GENHLTH', 'MENTHLTH', 'PHYSHLTH', 'DIFFWALK', 'SEX', '_AGEG5YR', 'EDUCA' and 'INCOME2' have been changed to be respectively.
'Diabetes_012', 'HighBP', 'HighChol',’ CholCheck', 'BMI', 'Smoker', 'Stroke', ‘HeartDiseaseorAttack', 'PhysActivity', 'Fruits', "Veggies", 'HvyAlcoholConsump' 'AnyHealthcare', 'NoDocbcCost', ‘GenHlth', 'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age', 'Education’ and 'Income'.

4.2 Data visualization

We need to see the difference between diabetic and non-diabetic patients in the data sets as seen in Figure 1 the class of non-diabetic is not balanced with the records in the diabetic class. Also, need the range for data values in every selected feature by using a box plot to view the data values and make the values within the same range.

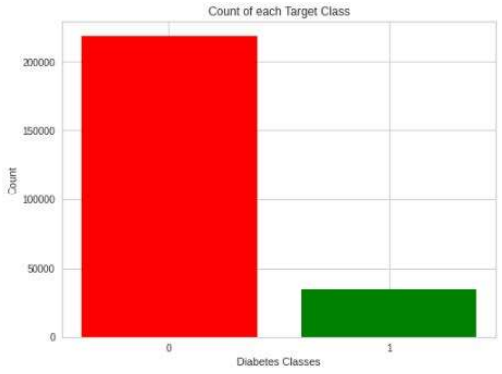

FIGURE 1: Diabetes classes

We also compared the different independent variables such as high cholesterol as in Figure 2, high blood pressure as in Figure 3, BMI, and others with the dependent variable “diabetes_binary”, we found some relations between them. There is an increase in the diabetes category with the increase in age

Figure 4.

To measure feature dependency, we used the chi-square test for categorical variables high blood pressure, high cholesterol, smoking, heart disease stroke, etc., and due to the large sample size, p values were all significant so we could not rely on this test to detect the correlation between variables. We also draw the correlation matrix and there was no significant correlation between any of the independent variables.

We need to check for the features if there are outlier values, we find them in the three variables (BMI, physical health, and mental health status) we drew the box plot and there were outliers for the three variables as seen in Figure 5(a, b, c). We make the data in the normal range and remove the outlier values.

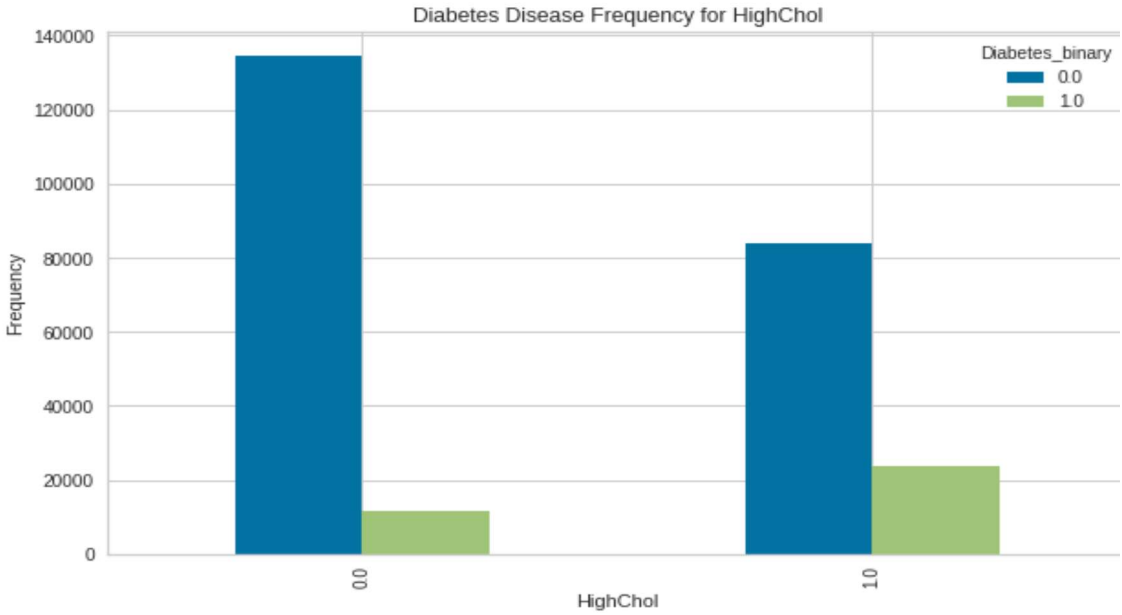

Figure 2: Diabetes frequency for Highchol

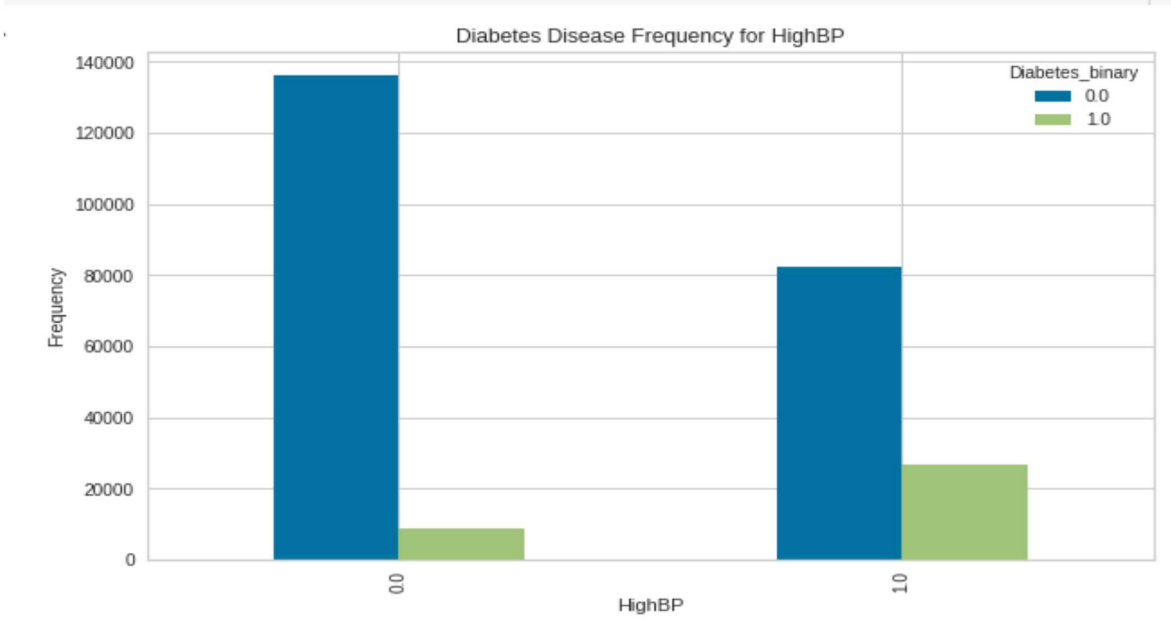

Figure 3: DIABETES frequency for HighBP

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

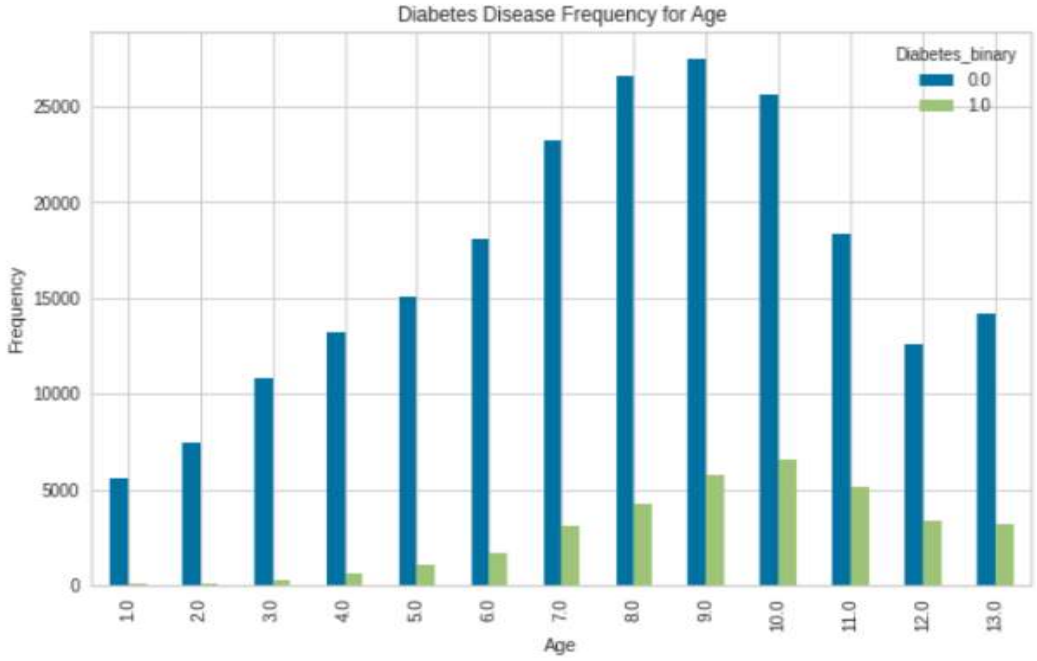

Figure 4 : Diabetes Frequency with Age

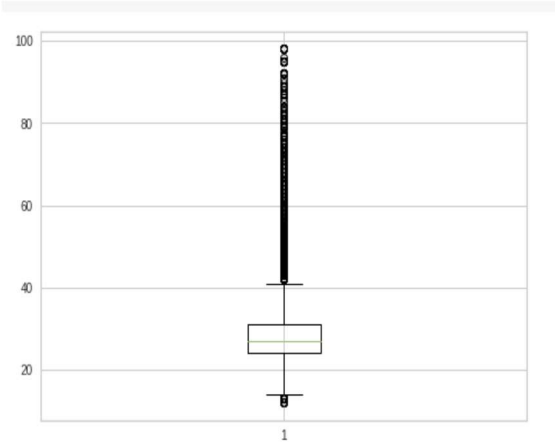

Figure 5.a: BMI for outliers

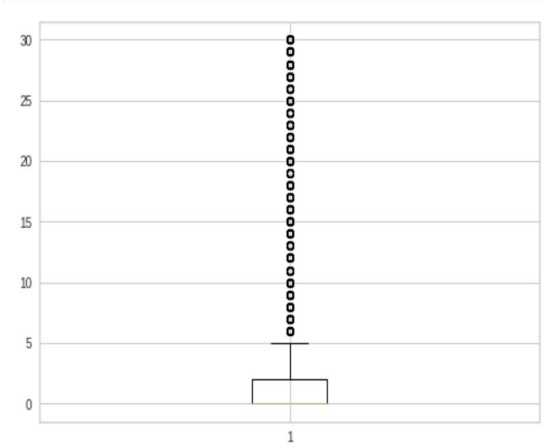

Figure 5.b Mental health

4.3 Balancing the dataset

Datasets have 218334 non-diabetic and 35346 records for pre-diabetic and diabetic so the data need to balance we use Synthetic Minority Over-sampling Technique (SMOTE) while training the model and using a decision tree, logistic regression, extreme gradient boosting, catboost classifier to classify the data and to measure the performance precision, recall, accuracy, and AUC.

5. EXPERIMENTAL RESULTS

We used Python and an online collab notebook to build and compare the predictive models. We applied several classifiers to build our model, we first trained our data using the decision tree algorithm, and we defined the metrics in the variable "scoring" which include accuracy, balanced accuracy- calculated by sum sensitivity and specificity and divide the result by 2- precision macro, recall a macro- average of the 2 class 0 and 1- and roc AUC.

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

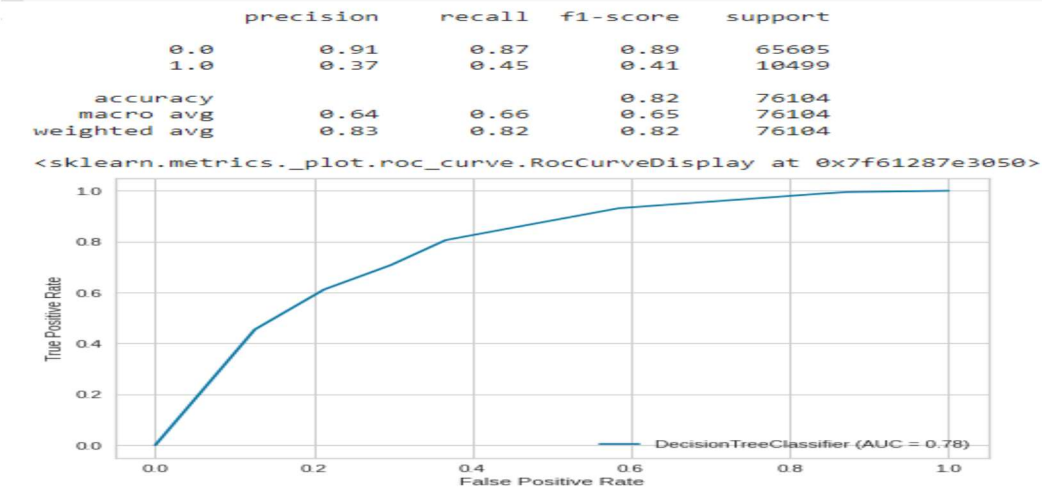

Figure 6 Performance of Decision Tree Predictive Models for Diabetes.

The classification report Figure 6 showed high accuracy of 82% and a high AUC value of 78% but low sensitivity "recall" for class 1 "prediabetes and diabetes" 45% with high specificity of 87%. On the other hand, class 0 "no diabetes" had high sensitivity 87% and low specificity 45%. The precision macro and recall macro in testing were 64% and 66% respectively, these were so low in comparison with the training results which were 90% and 86% respectively. The testing accuracy did not change so much concerning the training accuracy, which was 86%.

5.1 C atBoost Classifier

We used the model evaluation function, to display many types of plots to evaluate the performance of this model as follows:

The first plot is for the classification reports as seen in Figure 7-a, the sensitivity for class 1 "prediabetes and diabetes" is 0.190 which is very low. The model specificity for class 1 based on the confusion matrix figure 7-b, c is equal to TN/TN+FP = 63913/63913+1609 = 0.975

We retrieved ROC curves for the model which were equal to 83% using the predicted model function we obtained 86% accuracy.

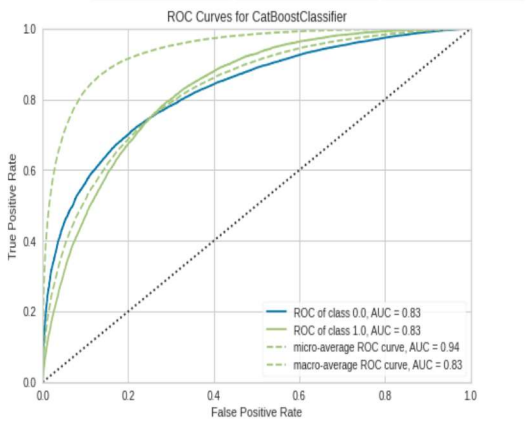

Figure 7-a catboost classifier classification report

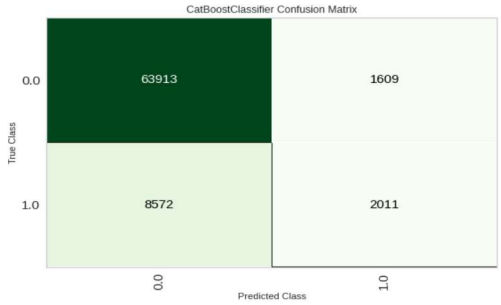

Figure 7-b ROC curves for catboost classifier

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

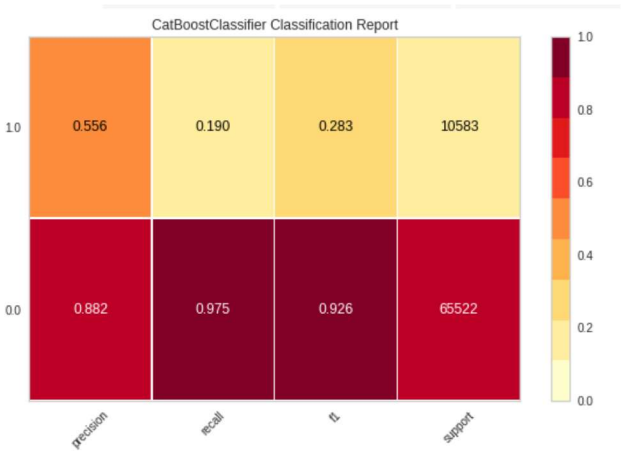

Figure 7-c Catboost Classifier confusion matrix

5.2 Extreme Gradient Boosting

The plot of the classification report seen in Figure 8-a showed a very low sensitivity of 18% for class 1 "prediabetes and diabetes". We calculated the specificity for class 1 from the confusing matrix in Figure 8-a,b,c and it was equal to 97.5%. The ROC curve for the model was 82% and using predict model function, we obtained 86% accuracy.

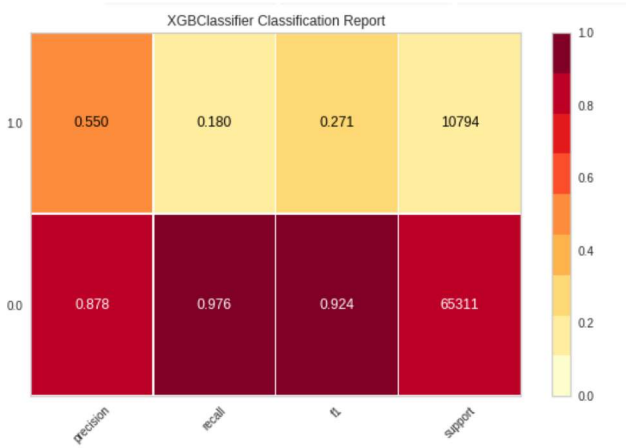

Figure 8-a xgboost- classification report

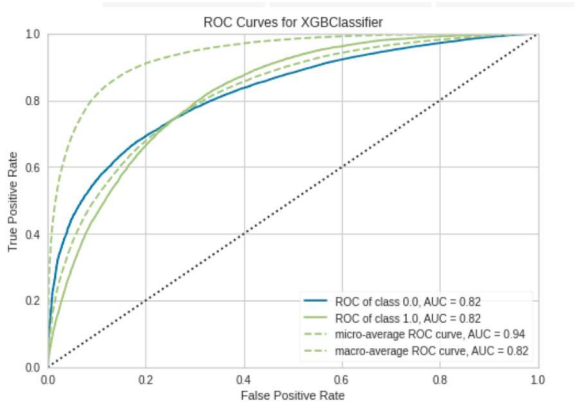

Figure 8-b xgboost ROC curves

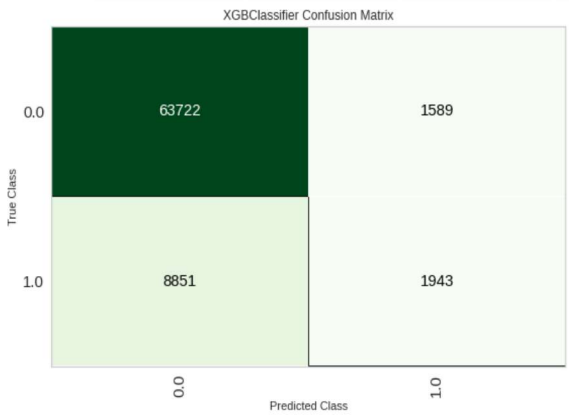

Figure 8-c xgboost - confusion matrix

5.3 Light Gradient Boosting Machine

The plot of the classification report seen in Figure 9-a showed a very low sensitivity of 23% for class 1 "prediabetes and diabetes". We calculated the specificity for class 1 from the confusing matrix Figure 9-b,c and it was equal to 96.7%. The ROC curve for the model was 82% and using predict model function, we obtained 86% accuracy.

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

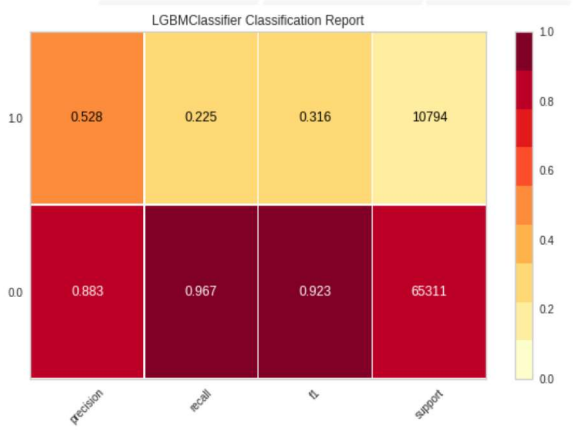

Figure 9-a lightgbm –classification

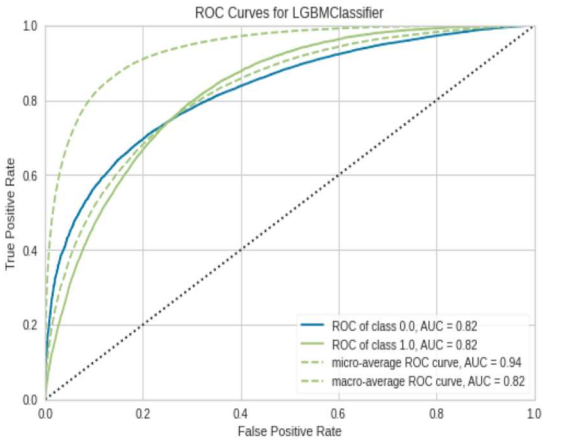

Figure 9-b lightgbm- ROC curves

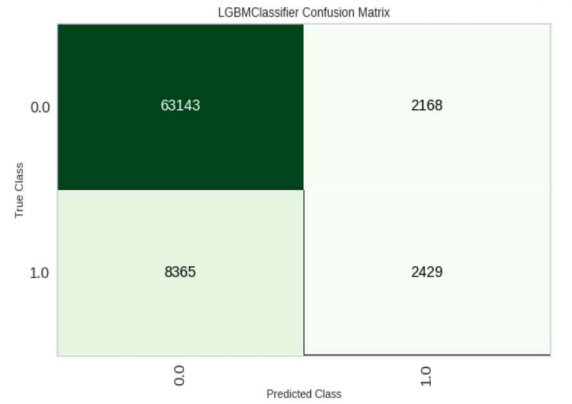

Figure 9-c lightgbm - confusion matrix

5.4 Ada Boost classifier

The plot of the classification report seen in Figure 10-a showed moderate sensitivity of 64% for class 1 "prediabetes and diabetes". We calculated the specificity for class 1 from the confusing matrix figure 10-b,c and it was equal to 81%. The ROC curve for the model was 81% and using predict model function, we obtained 79% accuracy.

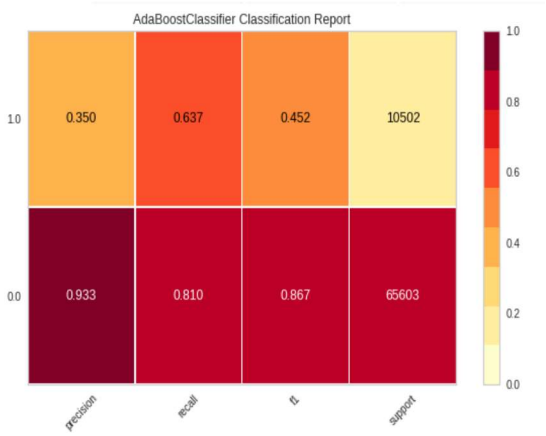

Figure 10-a Classification using the Ada boost classifier.

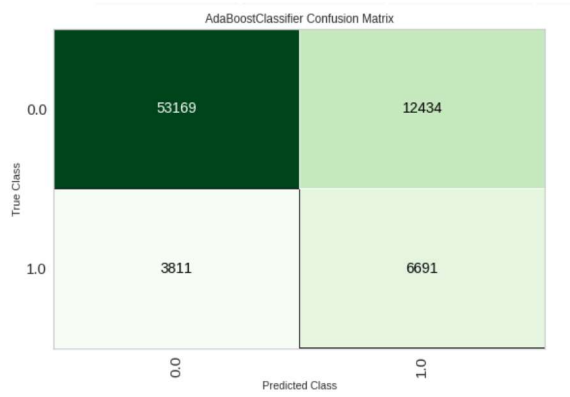

Figure. 10-b Classifier Ada Boost-Confusion matrix

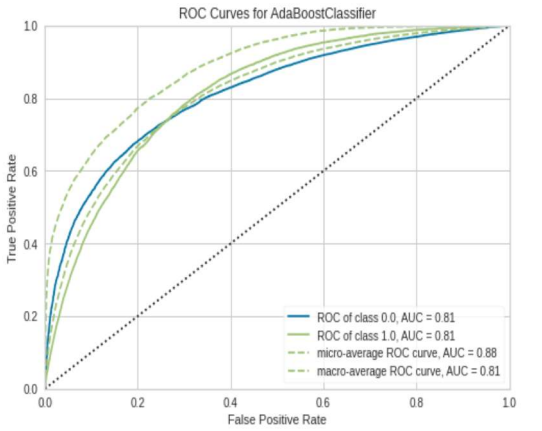

Figure 10-c Classifier Ada boost-Confusion matrix

5.5 Logistic Regression

The plot of the classification report seen in Figure 11-a showed a high sensitivity of 78% for class 1 "prediabetes and diabetes" and 72% for class 0 "no diabetes". We calculated the

specificity for class 1 from the confusing matrix figure 11-b,c and it was equal to 72%, and for class 0 and it was 77%. The ROC curve for the model was 82% and using predicting model function, we obtained 73% accuracy.

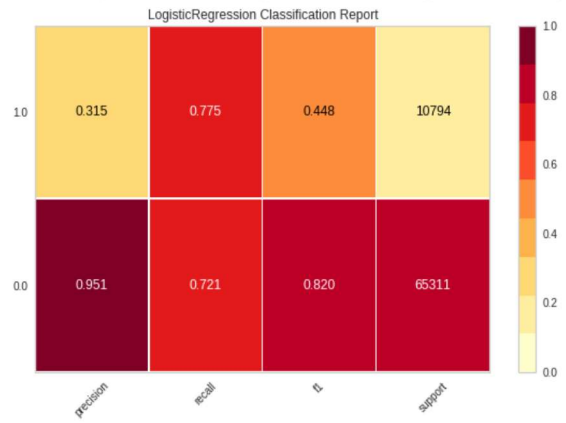

Figure 11-a Report on Logistic Regression and Classification

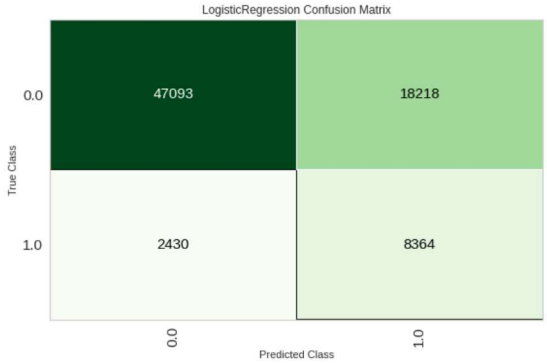

Figure 11-b Confusion matrix for Logistic Regression

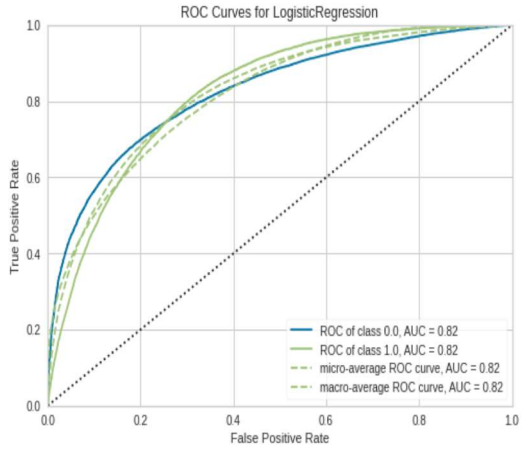

Figure 11-c ROC Curves for Logistic Regression

We retrieved the features importance plot shown in Figure 12 which showed the importance of the general health status variable, BMI, age, and mental health status in predicting diabetes mellitus.

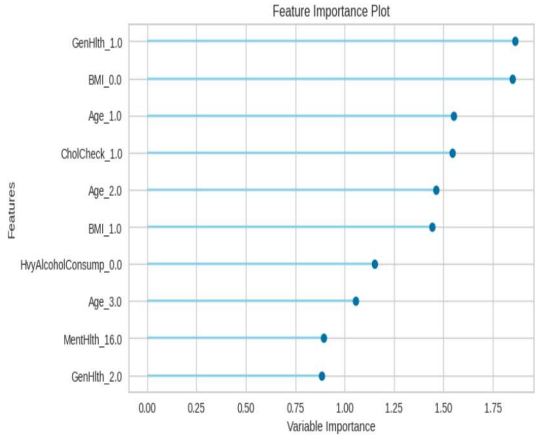

Figure 12 Logistic Regression –Features importance

5.6 Linear Discriminant Analysis

The plot of the classification report seen in Figure 13-a showed a high sensitivity of 78% for class 1 "prediabetes and diabetes" and 71% for class 0 "no diabetes". We calculated the specificity for class 1 from the confusing matrix figure 13-b,c and it was equal to 71%, and for class 0 and it was 78%. The ROC curve for the model was 82% and using predicting model function, we obtained 72% accuracy. The features importance chart showed that the BMI has a high predictive value, general health status, Age, and cholesterol check in Figure 14.

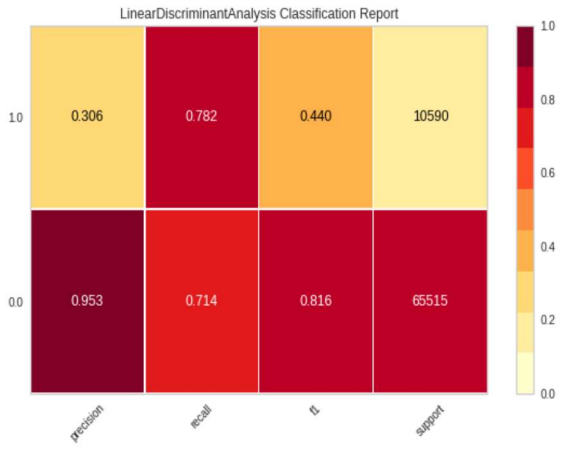

Figure. 13-a linear discriminant analysis – Classification report

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

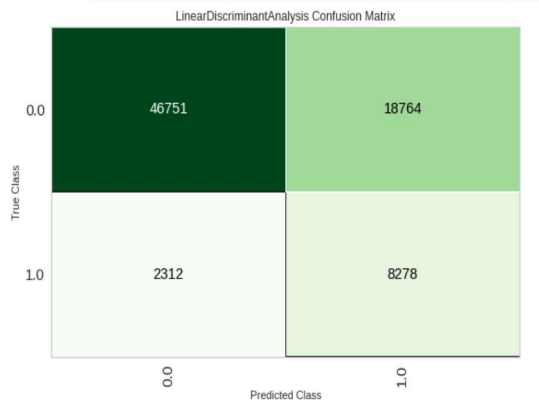

Figure. 13-B Linear Discriminant Analysis – Confusion Matrix

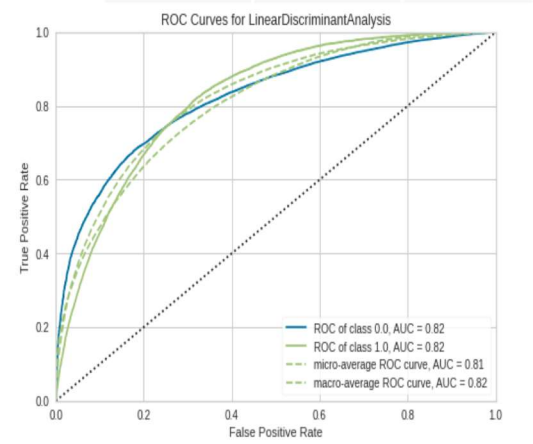

Figure. 13-C Linear Discriminant Analysis –ROC Curves

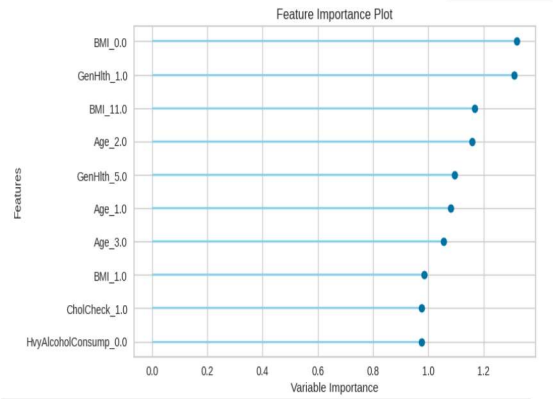

Figure. 14 Linear Discriminant Analysis –Features Importance

5.7 Combine the Ada boost classifier with the Logistic regression model.

Using the blend models tool, we mixed two models (Ada boost classifier and Logistic regression). This function allows us to train another model that takes the outputs from the first model and generate a new output [14,15].

We evaluate the blender model and draw the performance plots as follows: The plot of the classification report showed a high sensitivity of 77% for class 1 "prediabetes and diabetes" and 72% for class 0 "no diabetes".

We calculated the specificity for class 1 from the confusion matrix and it was equal to 0.72%.

The ROC curve for the model was 82% and using predicting model function, we obtained 73% accuracy. We increased the probability threshold to 55%, this returned an accuracy of 78%, a sensitivity of 67%, and an AUC of 82%.

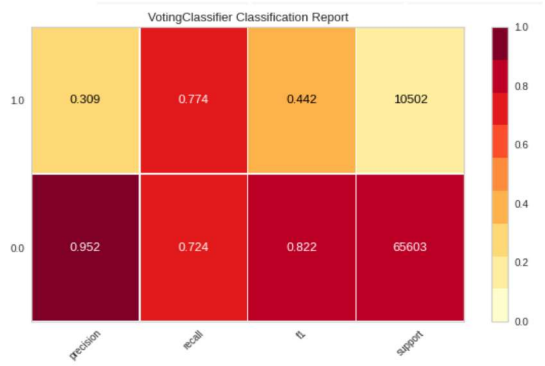

Figure. 15-a Voting classifier Classification Report

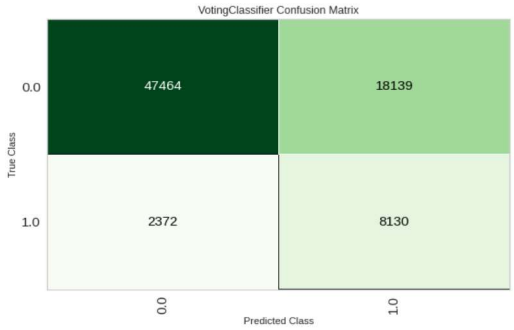

Figure. 15-B Voting Classifier Confusion Matrix

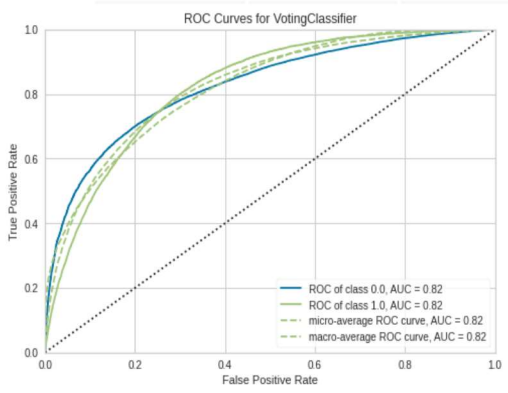

Figure 15-C ROC Curves For Voting Classifier

5.8 Comparing models Performances

After training and testing the six models [16], we found that the highest accuracy 86%, and The AUC values for the catboost classifier, extreme gradient boosting, and mild gradient boosting machine models were (83%, 82%, and 82%, respectively). These models also achieved the lowest sensitivity (19% - 23%). Catboost classifier and extreme gradient boosting achieved the highest specificity 97.5%.

On the other hand, Ada boosts classifier achieved high accuracy 79 %, moderate sensitivity 64% high specificity 81% and Auc 81% The logistic regression model achieved moderate accuracy 73%, and high sensitivity 78%, while linear discriminant analysis achieved lower accuracy 72% and same sensitivity 78%. The last two models had high AUC values of 82% and moderate specificity of 72% and 71% respectively. The results of the various models are summarized in Table (2).

TABLE 2 : the performance of all models

Model	***Accuracy***	***AUC***	***Sensitivity***	***Specificity***
CatBoost Classifier	*86 %*	*83 %*	*19%*	*97.5 %*
Extreme Gradient Boosting	*86 %*	*82 %*	*19 %*	*97.5 %*
Light Gradient Boosting Machine	*86 %*	*82 %*	*23 %*	*96.7 %*
Ada Boost Classifier	*79 %*	*81 %*	*64 %*	*81 %*
Logistic Regression	*73 %*	*82 %*	*78 %*	*72 %*
Linear Discriminant Analysis	*72 %*	*82 %*	*78 %*	*71 %*
Blender Model	*73 %*	*82 %*	*77 %*	*72 %*
Blender Model (0.55)	*78 %*	*82%*	*67 %*	

TABLE 3 : Comparing Performance of Predictive Models for Diabetes for our study and Xie et al study

Study	Model	Accuracy	Sensitivity	Specificity
Z. Xie ,etl [13]	Neural network	82.41 %	37.81 %	90.16 %
	Logistic regression	80.68 %	46.34 %	86.66%
	Decision tree	74.26 %	51.61 %	78.20 %
New model	Ada Boost Classifier	79 %	64 %	81 %
	Logistic Regression	73 %	78 %	72%
	Blender Model(0.55)	0.78	0.67	0.72

6. CONCLUSION

In our research we predict type 2 diabetic patients by using BRFSS datasets to run our model with many machine learning techniques and test the accuracy this data has many features that can be used to detect the disease we select only 21 features which very important and make cleaning to the dataset and balancing using SMOTE technique [17,18]to solve the difference between the diabetic and non-diabetic class based on these results we have high accuracy and specificity with catboost, extreme gradient boosting with 82% compared to others researches. Predicting this chronic disease early solve many problems to save people life in many countries.

In future work, we need to run our model with more selected features of the dataset and different datasets to get high accuracy while making preprocessing steps for the data.

7. AUTHOR CONTRIBUTIONS

Nesma Mohamed Mahmoud Kamel, Conceptualization; Data curation; Investigation; Methodology; Software; Validation; Marwa Hussein Mohamed: Visualization; Writing–original draft. Wael Said, Marwa Hussien Mohamed Data curation; Methodology; Visualization. Mohamed Helmy Khafagy, Wael Said: Project administration; Supervision; Writing – review & editing.

CONFLICTS OF INTEREST: The authors declare no conflict of interest.

FUNDING INFORMATION: The authors received no specific funding for this work.

DATA AVAILABILITY STATEMENT: The private dataset patients are available at the following link: https://chronicdata.cdc.gov/browse?category=Behavioral+Risk+Factors.

REFERENCES:

[1]. V. Rawat and Suryakant, 'A classification system for diabetic patients with machine learning techniques', International Journal of Mathematical, Engineering and Management Sciences, vol. 4, no. 3, pp. 729–744, Jun. 2019, doi: 10.33889/IJMEMS.2019.4.3-057.

[2]. H. Kaur and V. Kumari, 'Predictive modelling and analytics for diabetes using a machine learning approach', Applied Computing and Informatics, vol. 18, no. 1–2, pp. 90–100, Jan. 2022, doi: 10.1016/j.aci.2018.12.004.

[3]. A. Mohanty, S. Parida, S. C. Nayak, B. Pati, and C. R. Panigrahi, 'Study and Impact Analysis of Machine Learning Approaches for Smart Healthcare in Predicting Mellitus Diabetes on Clinical Data', in Intelligent Systems Reference Library, Springer Science and Business Media Deutschland GmbH, 2022, pp. 75–101. doi: 10.1007/978-981-16-5304-9_7.

[4]. G. Cardozo, G. B. Pintarelli, G. R. Andreis, A. C. W. Lopes, and J. L. B. Marques, 'Use of Machine Learning and Routine Laboratory Tests for Diabetes Mellitus Screening', Biomed Res Int, vol. 2022, 2022, doi: 10.1155/2022/8114049.

[5]. B. Farran, R. AlWotayan, H. Alkandari, D. Al-Abdulrazzaq, A. Channanath, and T. A. Thanaraj, 'Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait', Front Endocrinol (Lausanne), vol. 10, Sep. 2019, doi: 10.3389/fendo.2019.00624.

[6]. C. C. Olisah, L. Smith, and M. Smith, 'Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective', Comput Methods Programs Biomed, vol. 220, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.

[7]. G. Geetha and K. M. Prasad, 'An Hybrid Ensemble Machine Learning Approach to Predict Type 2 Diabetes Mellitus', Webology, vol. 18, no. SpecialIssue2, pp. 311–331, 2021, doi: 10.14704/WEB/V18SI02/WEB1807

[8]. R. Saxena, S. K. Sharma, and M. Gupta, 'Analysis of machine learning algorithms in diabetes mellitus prediction', in Journal of Physics: Conference Series, IOP Publishing Ltd, May 2021. doi: 10.1088/1742-6596/1921/1/012073.

[9]. B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto, and I. Gunawan, 'Classification of diabetes disease using decision tree algorithm (C4.5)', in Journal of Physics: Conference Series, IOP Publishing Ltd, Apr. 2021. doi: 10.1088/1742-6596/1869/1/012082.

[10]. A. Sumathi and S. Meganathan, 'Machine learning based pattern detection technique for diabetes mellitus prediction', Concurr Comput, vol. 34, no. 6, Mar. 2022, doi: 10.1002/cpe.6751.

[11]. V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, 'Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms', Neural Comput Appl, 2022, doi: 10.1007/s00521-022-07049-z.

[12]. N. M. Silva, 'The behavioral risk factor surveillance system', Int J Aging Hum Dev, vol. 79, no. 4, pp. 336–338, Oct. 2014, doi: 10.1177/0091415015574184.

[13]. Z. Xie, O. Nikolayeva, J. Luo, and D. Li, 'Building risk prediction models for type 2 diabetes using machine learning techniques', Prev Chronic Dis, vol. 16, no. 9, Sep. 2019, doi: 10.5888/pcd16.190109.

[14]. Mohamed, Marwa & Ibrahim, Mohamed & Khafagy, Mohamed. (2020). Two recommendation system algorithms used SVD and association rule on implicit and explicit data sets. International Journal of Scientific & Technology Research. 9. 17-24.

[15]. Ilyasova, Nataly, Nikita Demin, and Nikita Andriyanov. 2023. "Development of a Computer System for Automatically Generating a Laser Photocoagulation Plan to Improve the Retinal Coagulation Quality in the Treatment of Diabetic Retinopathy" *Symmetry* 15, no. 2: 287. https://doi.org/10.3390/sym15020287

[16]. Hemamalini, Selvamani, and Visvam Devadoss Ambeth Kumar. 2022. "Outlier Based Skimpy Regularization Fuzzy Clustering Algorithm for Diabetic Retinopathy Image

Segmentation" *Symmetry* 14, no. 12: 2512. https://doi.org/10.3390/sym14122512

[17]. Mohamed, M.H.; Ibrahim, L.F.; Elmenshawy, K.; Fadlallah, H.R. Adaptive Learning Systems based on ILOs of Courses. WSEAS Trans. Syst. Control. 2023, 18, 1–17. https://doi.org/10.37394/23203.2022.1.

[18]. Mohamed, M.H.; Khafagy, M.H.; Elbeh, H.; Abdalla, A.M. Sparsity and Cold Start Recommendation System Challenges Solved by Hybrid Feedback. Int. J. Eng. Res. Technol. 2019, 12, pp: 2734–2741