Research Article GU J Sci 35(3): 834-852 (2022) DOI: 10.35378/gujs.931760

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

Machine Learning Model to Diagnose Diabetes Type 2 Based on Health Behavior

Haithm ALSHARI *, Alper ODABAS

Department of Mathematics and Computer Science, Eskisehir Osmangazi University, 26040, Eskişehir, Turkey

Highlights

- This study objective is to develop a reliable ML model to diagnose incidence of type 2 diabetes.
- The model uses the relationship between a healthy lifestyle and type 2 diabetes.
- We used relatively new machine learning algorithms to build the model, like XGBoost and ANN.
- The dataset that was used to build this model was collected from 6 waves of NHANES datasets.
- The best model was built with the XGBoost algorithm with the histogram method.

Article Info

Received: 03 May 2021
Accepted: 15 Sep 2021

Keywords

Artificial intelligence
Diabetes
Health behavior
Gradient boosting
ANN

Abstract

Diabetes, in 2016, was the 7th death-causing disease in the world. It was the direct cause of 1.6 million deaths. In 2019, the number of adults (20-79 years) that were living with diabetes was approximately 463 million and is expected to rise to 700 million in 2045. The early diagnosis of diabetes will help treat it and prevent its complications. The need for an easy and fast way to diagnose diabetes is crucial. In this study, we are proposing a method to diagnose diabetes with the help of machine learning algorithms and tools. The proposed method utilizes the power of machine learning to create a model that can predict diabetes based on the health behavior of the patient. The model uses the relationship between a healthy lifestyle and diabetes. Our goal is to build a reliable machine learning model to predict diabetes, which will help significantly in easing and speeding up the diagnosing procedure of diabetes. We used modern machine learning algorithms like XGBoost, LightGBM, CatBoost, and artificial neural networks, and the dataset was obtained from the National Health and Nutrition Examination Survey (NHANES). In our study, the XGBoost algorithm performed the best with a Cross-Validation (10-fold) score of 0.864, and an overall accuracy of 87.7% for the validation dataset and 84.96% for the test dataset.

1. INTRODUCTION

Diabetes mellitus, commonly known as diabetes, is one of the metabolic diseases which causes blood glucose levels to rise above the normal limits. The movement of sugar from the blood into our cells to be stored or used for energy is controlled by the insulin. Diabetes makes our bodies unable either to produce enough insulin or to use the insulin efficiently [1]. There are many types of diabetes: type 1 diabetes, type 2 diabetes, prediabetes, and gestational diabetes. The most prevalent type is type 2 diabetes, around 90% of all diabetes cases are type 2 diabetes. Moreover, this type has a strong relationship with the health behaviors of people, such as physical activity and nutrition. Thus, we will focus our study on it. Diabetes can be diagnosed with several tests. Each one of these tests is done using a blood test and requires to be done more than once on the same day or multiple days to diagnose diabetes. Moreover, these tests must be carried out in a healthy environment, such as a lab or hospital [2].

Risk factors of type 2 diabetes**,** the most common ones are family history, age, obesity, distribution of fat in the body, lack of activity, race, and gestational diabetes.

Health Behaviors and Diabetes**,** health behaviors have been defined by scientists in many ways. For instance, according to Conner and Norman, health behaviors are “any activity undertaken for the purpose of preventing or detecting disease or for improving health and wellbeing” [3]. And according to Gochman,

*Corresponding author, e-mail: haitham.alshari@gmail.com

in the Handbook of Health Behavior Research, health behaviors are "behavior patterns, actions and habits that relate to health maintenance, to health restoration and to health improvement" [4]. Within these definitions, behaviors include health service usage (like physician visits, screening, vaccination), following medical regimens (like dietary and diabetic regimens), and self-directed health behaviors (like exercise, diet, smoking, and alcohol consumption). There are many studies show that health behaviors affect the prevalence of diabetes, with healthier behaviors the prevalence of diabetes decreases. One study shows that with the increase in dietary fibers intake, the risk of diabetes is reduced [5]. The same study also associated a healthy lifestyle with a low risk of diabetes. Another study proved that lifestyle interventions have similar effects in preventing or delaying type 2 diabetes in people with impaired glucose tolerance to pharmacological interventions [6-8]. These studies reveal the relationship between diabetes and health behaviors with evidence to support this relationship.

The conventional methods of diagnosing diabetes are done using a blood test [9]. Those methods are not intuitive when we have to test numerous people, or we don't have the required equipment. Therefore, the need for an easy and fast way to diagnose diabetes is crucial. On the other hand, machine learning has been a valuable tool in the medical field. It has made many contributions to diagnosing many diseases with an impressive level of accuracy, such as the prediction of skin cancer [10]. Therefore, to make the diagnosing process of diabetes easier and more intuitive, we implement machine learning techniques to predict diabetes based on health behaviors.

The main objective of this study is to develop a reliable machine learning model that can efficiently diagnose the incidence of type 2 diabetes and prediabetes that uses only features that can be acquired without lab testing. To this end, we will test various machine learning algorithms while focusing on the most common algorithms, namely XGBoost, LightGBM, CatBoost, and ANN.

Many related works tried to tackle this problem with many machine learning techniques from logistic regression to ANN. Many of them use the PIMA Indian Diabetes dataset which includes some features that require lab tests.

Zou, Q., et al. in [11] used a costume dataset that they collected from hospitals in China. This dataset includes some features that need to be performed in lab like low-density lipoprotein (LDL), and high-density lipoprotein (HDL).

Juneja, A., et al. in [12] used the PIMA Indian Diabetes dataset which as we mentioned before includes some features that require lab tests.

Muhammad, L. J., et al. in [13] used a dataset they collected from the Murtala Mohammed Specialist Hospital, Kano State, in Nigeria, and this dataset also included some features that are lab-related like HDL and triglyceride.

Tigga, N. P., et al. in [14] used dataset that they collected from participants and this dataset does not include any feature that needs a lab test, but the dataset size is small with only 952 observations.

2. LITERATURE REVIEW

2.1. Machine Learning

Machine learning (ML) is a field of computer science that provides systems with the ability to learn and improve from experience without explicitly programmed. Machine learning aims to develop computer programs that are able to access data and use it for learning and draw insights. The learning process is done by analyzing the data through building and adapting models, which allow programs to learn the patterns in data and make decisions based on them. The main goal of machine learning is to allow computers to learn automatically without human assistance or intervention [15,16]. Machine learning can be categorized into four categories:

A. Supervised machine learning: The most used and practical form of machine learning. In supervised learning, we know both the inputs and the desired outputs. In other words, the ML system is presented with data that is labeled, i.e., each data is tagged with the correct label. The goal is to create a mathematical function that maps each input data with its desired output. The two main tasks of supervised machine learning are classification and regression. In classification, the ML system uses statistical classification methods to output a categorization, for example, "rainy day" or "sunny day". In regression, the ML system uses statistical regression analysis to output a numerical value [15].
B. Semi-supervised machine learning is supervised learning with the exception of not all of the training data is labeled, only a partial amount of them is. A good example of semi-supervised learning is image recognition. Here we might provide the ML system with many labeled images that contain the objects we want to identify, then in the training process the system processes many more unlabeled images [15].
C. Unsupervised machine learning: In unsupervised learning, all the outputs are unknown, i.e., unlabeled. The ML system tries to create a structure based on the relationships between the inputs only. The main task of unsupervised learning is clustering. Clustering is the process of grouping the dataset inputs into groups with similar attributes. Consumer trends and patterns in stock data are examples of unsupervised learning. Unsupervised learning problems can be solved using various algorithms, such as K-Means and Neural networks [15].
D. Reinforcement machine **learning** is an area of machine learning where the ML system is provided with feedback in the form of rewards and punishments, rather than explicitly told, “True” or “False”. This comes into play when finding the correct answer is important but finding it in a timely manner is also important. This technique becomes handy when both finding the correct answer and finding the answer in a timely manner are important [15].

Since our task is classifying, which is a task of supervised machine learning, we will concentrate the next part on classification algorithms. One of the most widely used algorithms for classification in ML is *decisions trees*.

2.2. Gradient Boosting Decision Trees

Decision trees are a machine learning method that uses a tree-like graph or model of decisions to split the dataset into either classifying or predicting based on features. Another way to think of them as a flow chart-like structure in which each internal node represents a test on a feature and each branch divides the data into one of two groups. In the prediction process of the decision tree, the data is assigned to the suitable node, and the result of the nodes’ test is the prediction of that node. Figure 1 shows a simple decision tree.

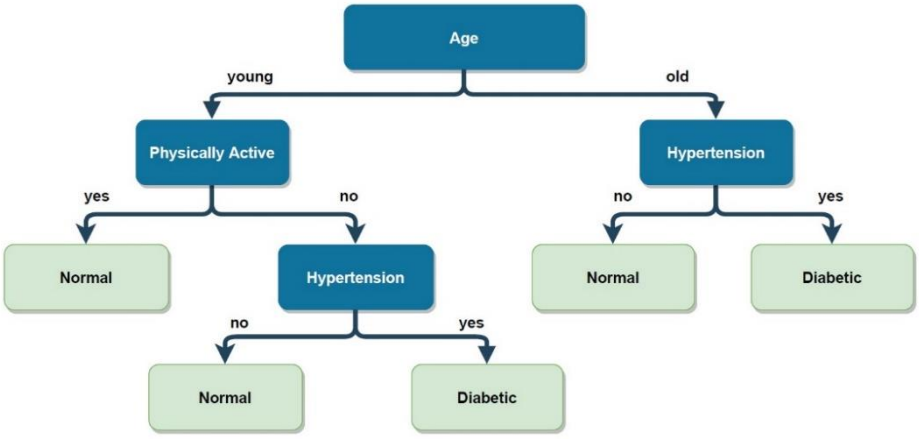

Figure 1*. A simple graph shows a decision tree that predicts whether a person will be diabetic or not*

Although decision trees are flexible and easy to interpret, a single tree is usually prone to overfitting and is unlikely to generalize well. We can fight overfitting by limiting the tree’s depth, but it will drive the decision tree to underfit. Therefore, we use several decision trees combined instead of a single decision tree. This

helps us make predictions that are generalized well. There are two main ways to combine multiple decision trees together, bagging and boosting. Bagging builds many different decision trees in parallel, then adds their outputs to form the final output. The best example of bagging is random forests. Boosting, on the other hand, sequentially builds weak learners in an adaptative way, which means each model in the sequence is fitted to correct the observations in the dataset which were badly handled by the previous model, then combines them to get a strong learner. One of the top boosting algorithms is gradient boosting [17]. Gradient boosting is the ensemble model built by a weighted sum of weak learners, Equation (1)

$s_L(.) = \sum_{l=1}^{L} c_l \times w_l(.)$ where c_l's are coefficients and w_l's are weak learners. (1)

Due to the difficulty of getting to this form of the optimal model, an iterative approach has been taken. The sequential optimization process in gradient boosting is done by casting it to a gradient descent: at each iteration, we fit a weak learner to the opposite of the gradient of the current fitting error with respect to the current ensemble model [17]. Gradient Boosting Decision Tree is a gradient boosting that uses a decision tree as a learner. The top three gradient boosting decision tree algorithms are XGBoost, LightGBM, and CatBoost.

2.3. Artificial Neural Network and Deep Learning

ANNs are one of the main artificial intelligence techniques that were developed by mimicking the working structure of the human brain. Figure 2 shows the similarity between brain structure and ANN structure.

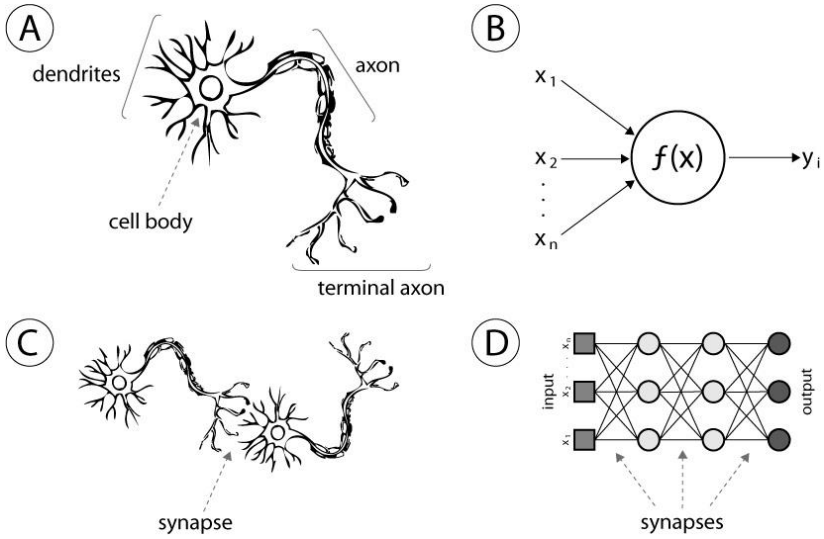

Figure 2. *A graph showing how the ANN are brain inspired*

In general, ANNs are computer programs that can generate and create new information by using previously learned or classified information with the help of neural sensors by imitating the biological neural structure of the human brain. Artificial neural networks are widely used in many fields of application, such as pattern recognition, system identification, robotics, signal processing, nonlinear control areas. In technical terms, the task of the artificial neural network is to produce an output as given in Figure 3 in response to the information provided to it as the input set. To do this, first, the network is trained with specific examples. Then the network generalizes the obtained information. Last, determine the outputs accordingly. Figure 3 shows the structure of artificial neurons.

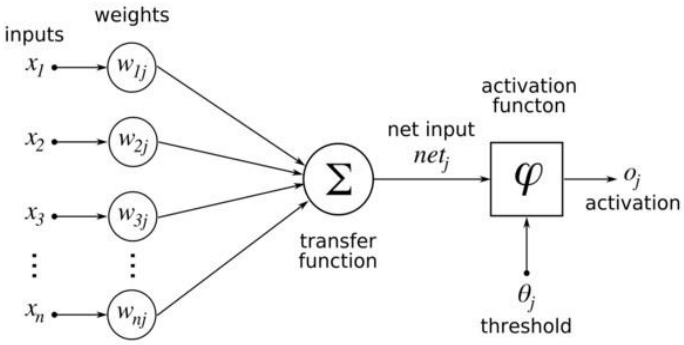

Figure 3. *The structure of the Neuron in ANNs*

ANNs consist of many neurons that are organized into layers – each layer contains several neurons - with links from each layer to the next that links each neuron in the previous layer with each neuron in the next layer. The first layer is called the input layer because it contains the input neurons, and the last layer is the output layer, which contains the output neurons, and the layers in between are called hidden layers. Figure 4 illustrates the structure of the ANNs.

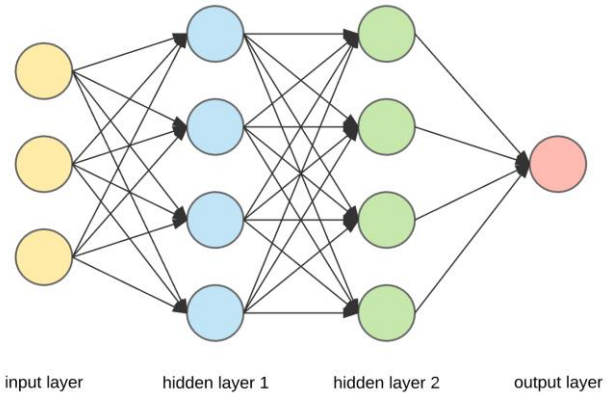

Figure 4. *The basic structure of ANNs*

Deep learning is a term that is used to describe ANNs with many hidden layers. The word *deep* is used to indicate that these networks use a large number of layers, which makes the learning process much deeper than the other networks. Figure 5 shows the difference between simple ANN and Deep Learning Neural Networks.

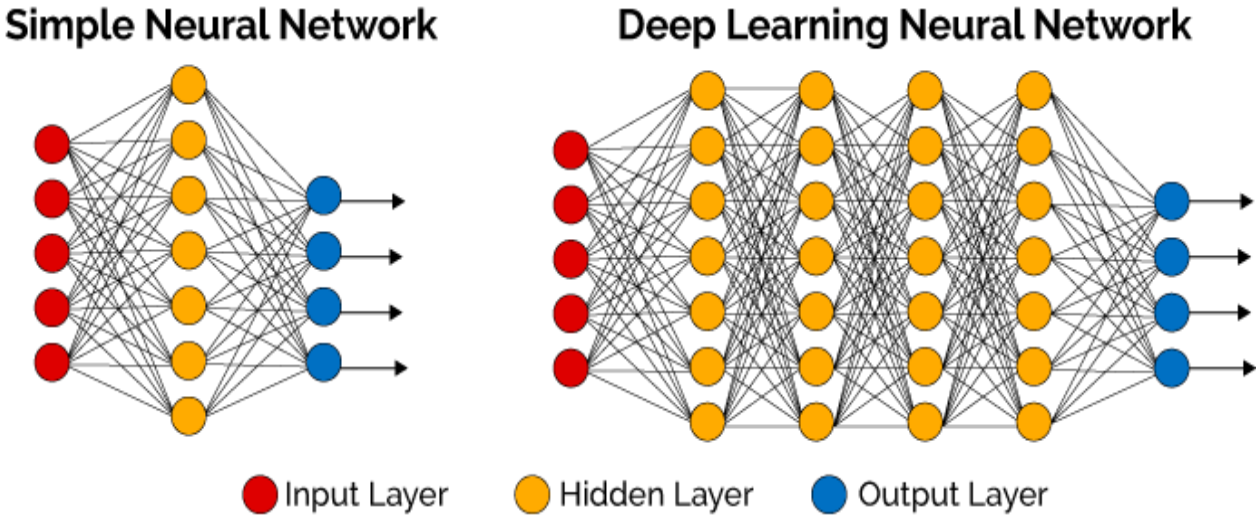

Figure 5. *The difference between simple and deep learning neural networks*

2.4. Imbalanced Datasets

An imbalanced dataset is any dataset with skewed class proportions. In other words, some classes have a larger proportion than the others in the dataset. The classes which have a large proportion of the dataset are called majority classes, and the other smaller proportions are called minority classes [18]. This kind of dataset is widely common in the medical field because we are trying to predict the abnormal of the population (people with the disease) in a normal population (healthy people), which is a small portion of any normal population. To handle such a problem, many techniques have been proposed such as under-sampling, oversampling, weighted-class, and threshold-moving.

Under-sampling: In this method, we reduce the size of the majority class by sampling the same amount of the minority class, as shown in Figure 6. The sampling process could be done randomly or with other techniques like cluster centroids and Tomek links. However, one big disadvantage of this method is that we lose important information from the dataset that can be used to make better predictions [19].

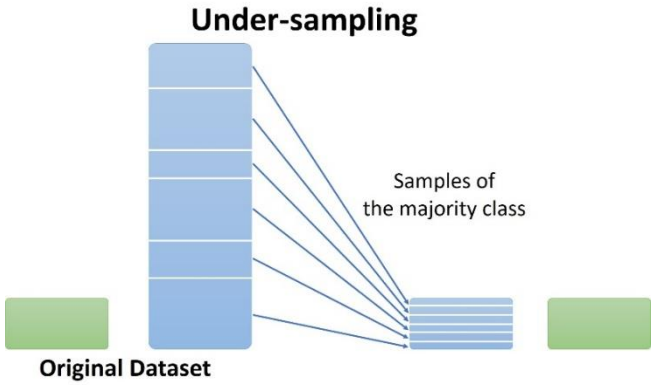

Figure 6. *Under-Sampling*

Over-sampling: Conversely, here we increase the size of the minority class to match the size of the majority class either by duplicating or synthesizing new ones [19], as shown in Figure 7. There are many algorithms for oversampling like ADASYN and SMOTE and its versions. While oversample preserves the information of the majority class, it introduces new synthesized data that is not representative of the real ones, which introduces its own biases to the model. However, oversampling is preferable over under-sampling because it keeps all the original data.

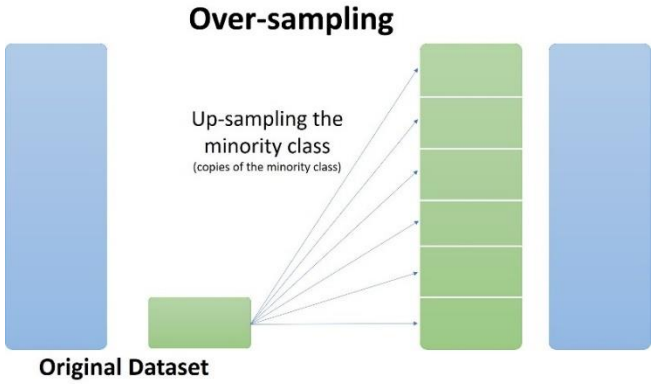

Figure 7. *Over-Sampling*

Combination of over and under sampling methods: Another approach to handle an imbalanced dataset is to use a combination of over-sampling and under-sampling. This method uses SMOTE for oversampling the data first, then cleaning using ENN or Tomek links. This approach has better results than the previous two techniques.

Weighted-Class: This method balances the data by changing the weight that each training example carries when computing the loss during training. Normally, each example and class in our loss function will carry equal weight, i.e., 1.0. However, in imbalanced datasets, we want minority examples to hold more weight when computing the loss [19,20]. The resulting model will have a balanced performance without losing any information from the original dataset or obtaining any irrelevant information. In our opinion, this is the best technique to handle an imbalanced dataset and we will use it during the development of our model.

Threshold-moving: This method can be applied to any soft classifier, a classifier that provides a score to each example indicates the degree of this example is a member of that class, this score can be used as a threshold to generate other classifiers. By varying this threshold, we can accomplish the classification problem with higher accuracy [19].

All of the imbalanced data handling techniques have been tested to find the best-performing one. Before the testing, we had an initial intuition that the weight-class method would perform the best due to its nature, fighting imbalance during the training process and after we performed all the tests, the results conformed to our initial intuition as shown in Figure 8.

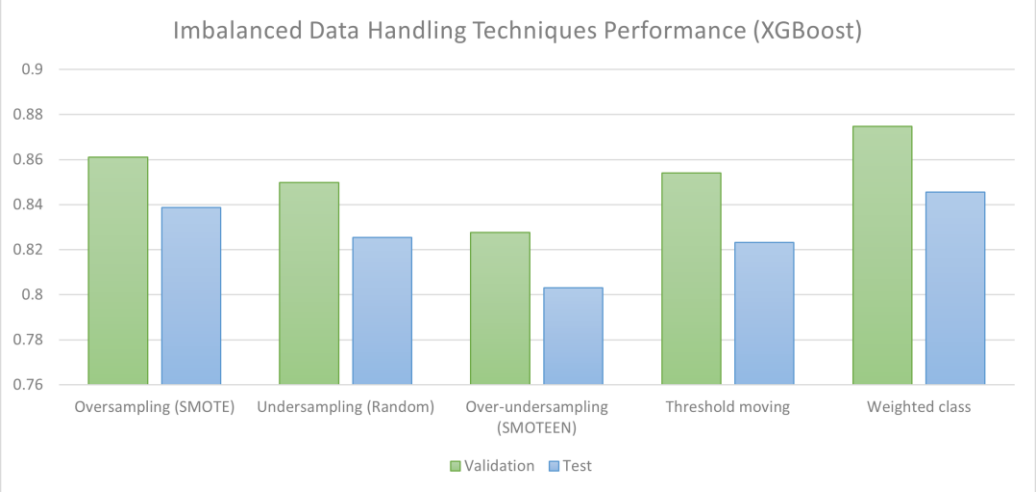

Figure 8. *Imbalanced Data Handling Techniques Performance (XGBoost)*

3. MATERIALS AND METHOD

3.1. Dataset

The dataset that we used has been collected from 6 waves of the National Health and Nutrition Examination Survey: 2007-2008, 2009-2010, 2011-2012, 2013-2014, 2015-2016, and 2017-2018. At the time of writing this thesis, the most recently available complete dataset of NHANES was 2017-2018 [21]. Due to the original dataset having almost 1800 features (variables) most of them are not related to our work, we had to first identify and separate the important features. We chose the main aspects of health behaviors, which are physical activity, dietary, and smoking features to the final dataset. In addition, we chose the hypertension feature because of its relationship to diabetes according to a recent study which has been done on 4.1 million adults and showed that people with high blood pressure have a higher risk of developing type 2 diabetes [22]. Although there are contradictions among the studies that focused on the relationship between alcohol consumption and diabetes, we did include alcohol consumption in the selected features because it is one of the main aspects of health behaviors. Other demographic features have been included too, which are age, gender, race, marital status, education level, annual family income, and the ratio of family income to poverty guidelines. The resulting dataset consists of 30 features and 55939 observations, and all values of the dataset have been coded in numerical form. Table 1 shows the selected features and their descriptions.

Table 1. *A table shows the selected features and their descriptions*

Features (Variables)	Description
SEQN	Sequence number
RIDAGEYR	Age in years
RIAGENDR	Gender
RIDRETH1	Race
DMDMARTL	Marital status
INDFMIN2	Annual family income (in Dollars)
INDFMPIR	A ratio of family income to poverty guidelines
DMDEDUC3	Education level - Children/Youth 6-19
DMDEDUC2	Education level - Adults 20+
PAQ655	Heavy Workout Days Per Week
PAD660	Daily Heavy Workout in minutes
PAQ670	Moderate Workout Days Per Week
PAD675	Daily Moderate Workout in minutes
DR1TCHOL	Cholesterol (mg)
DR1TFIBE	Dietary fiber (gm)
DR1TSODI	Sodium (mg)
DR1TCARB	Carbohydrates (gm)
DR1TTFAT	Total fat (gm)
DR1TSUGR	Total sugars (gm)
DR1TPROT	Protein (gm)
DR1TPOTA	Potassium (mg)
DR1TALCO	Alcohol (gm)
SMQ020	Smoked at least 100 cigarettes in life
SMQ040	Current smoker
URXPREG	Pregnancy test result
BPQ020	Ever told you had high blood pressure
BPQ030	Told had high blood pressure - 2+ times
BPD035	Age told had hypertension
BMXWT	Weight (kg)
BMXBMI	Body Mass Index (kg/m2)
DIQ010	Diabetes status

3.2. Research Flow

The flow of the research methodology is illustrated in Figure 9.

Figure 9. *The flow of methodology*

3.2. Dataset Preprocessing

One of the most important steps for building a machine learning model is data preprocessing. Most of the available datasets have missing and/or incorrect values. To build an efficient machine learning model, we had to clean the dataset before feeding it to the model for the training process. Cleaning data is a term used to describe the process of handling the missing and incorrect values in the dataset either by filling the missing data with various techniques or by removing them. Our dataset contains observations of people with ages ranging from less than 1 to 80 years old. Our study focused on people aged 18 years and above because it is more common in adults. After removing all the observations of people with an age less than 18 years old, the dataset size became 35485 observations. Moreover, our dataset includes a lot of variables with missing dataset. Table 2 shows the number of missing values for each variable:

Table 2. *A table shows the number of missing values in each variable and its ratio*

Features (Variables)	Number of missing values	Ratio of missing values
RIDAGEYR	0	0.00%
RIAGENDR	0	0.00%
RIDRETH1	0	0.00%
DMDMARTL	1769	4.99%
INDFMIN2	691	1.95%
INDFMPIR	3564	10.04%
DMDEDUC2	1769	4.99%
DMDEDUC3	33716	95.01%
PAQ670	21467	60.50%
PAQ655	27442	77.33%
PAD660	27461	77.39%
PAD675	21500	60.59%
DR1TCHOL	3062	8.63%
DR1TFIBE	3062	8.63%
DR1TSODI	3062	8.63%
DR1TCARB	3062	8.63%
DR1TTFAT	3062	8.63%
DR1TSUGR	3062	8.63%
DR1TPROT	3062	8.63%
DR1TPOTA	3062	8.63%
DR1TALCO	3062	8.63%
SMQ020	886	2.50%
SMQ040	20545	57.90%
URXPREG	28205	79.48%
BPQ020	0	0.00%
BPQ030	23118	65.15%
BPD035	23157	65.26%
BMXWT	789	2.22%
BMXBMI	847	2.39%
DIQ010	0	0.00%
EDUCATION	46	0.13 %

The next step was to fill in the missing values and to do this we had to go through all the features of the dataset with missing values and, according to Table 2, our dataset has five features without any missing values, thus, we skipped those features and focused on the other 25 features. Various techniques were used to fill in the missing values, such techniques include statistical methods, logic based on the understanding of the dataset, and machine learning. The latter is the most commonly used technique. We used logic and fact-based filling methods to fill in some features like pregnancy, marital status, and education level, and statistical methods like median to fill other features like workout minutes in a week. As for other missing values, we had to drop them because we couldn't handle them with any technique.

3.3. The Final Dataset

The resulting dataset after preprocessing the original dataset has 14682 observations and 21 features alongside the target variable. Five of the features have numerical values and the other 16 have categorical values (represented as numbers). Table 3 shows each feature of the final dataset and its type.

Table 3. *A table shows each feature with its type and number of categories in the preprocessed dataset*

Features	Type	Number of Categories
Age	Numerical	-----
BMI	Numerical	-----
Workout mins in a week	Numerical	-----
Alcohol	Numerical	-----
Income Ratio to Poverty	Numerical	-----
Gender	Categorical	2
Hypertension	Categorical	2
Annual Family Income	Categorical	12
Cholesterol	Categorical	4
Dietary fiber	Categorical	4
Sodium	Categorical	5
Sugar	Categorical	6
Total fat	Categorical	5
Protein	Categorical	4
Carbohydrate	Categorical	6
Potassium	Categorical	6
Race	Categorical	5
Marital status	Categorical	7
Education level	Categorical	5
Smoking Status	Categorical	4
Pregnant	Categorical	4
Diabetes (Target variable)	Categorical	2

3.4. Building Model

Before building the model, we determined which algorithms we will work with to obtain our model. We focused our work on recent algorithms, namely: Gradient boosting algorithms and Artificial neural networks (ANN). We selected the top three gradient boosting algorithms: XGBoost, LightGBM, and CatBoost, alongside ANN and Deep learning. We opt for mentioning the process of building the model to show our work in more detail manner and what choices we made to achieve such results, moreover, we

wanted to make it easier for other researchers to follow up our work to improve upon it. The following steps explain the process.

A. Imbalanced target variable: Imbalanced target variable: Almost every medical dataset is imbalanced due to the nature of the population which include normal people and those with the disease (abnormal). The target variable (Diabetes) has two different values: diabetic, and normal. 73.63% of the values in this variable represent the normal participants and 22.4% represent the participants with diabetes as shown in Figure 10. With a huge difference like this in the percentage of the target's values, the acquired model will be extremely biased and insufficient. To overcome this problem, many solutions were proposed including under-sampling the majority class, oversampling the minority class, weighted-class training, and threshold-moving. We used the weighted-class method to overcome dataset imbalance. We chose this method because it preserves the dataset from any changes and has performed the best in our case. This method assigns weights to the classes based on their presence in the output variable during the training process.

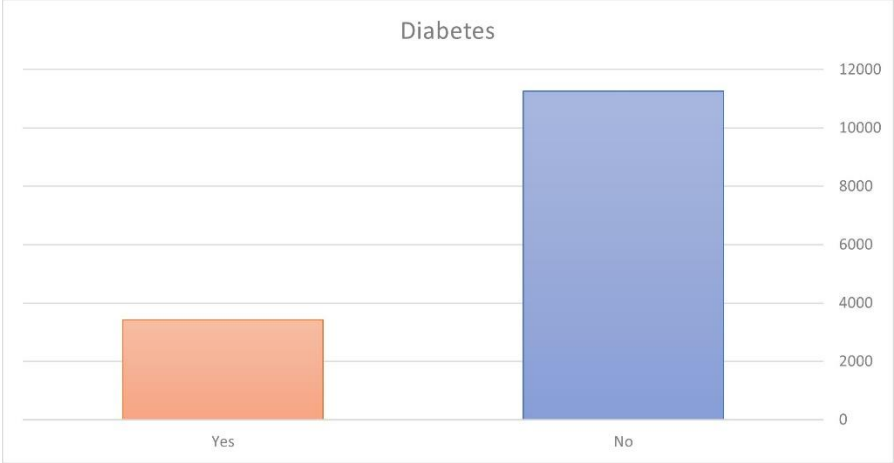

Figure 10. *A graph shows the number of values for each outcome*

B. Preparing the dataset: Before feeding the data to the model, we had to split the data into three sets: one for training and the second to validate the accuracy of the model through the training process (development), and the last one to evaluate the performance of the trained model with data that was never tested before. We use the data from 2007 to 2016 for training and validation, 80% of the data for training, 20% for validation, and the data from 2017-2018 for testing the model.

C. Training the model: In the training process of the model, we tested the four selected algorithms. Each algorithm was trained with different hyperparameters to tune it and obtain the best hyperparameters. The training process was achieved as shown in Figure 11:

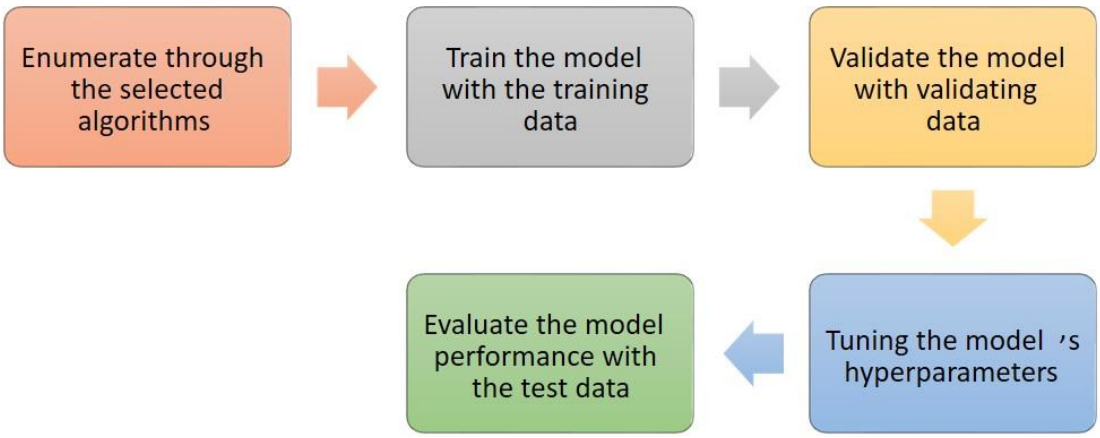

Figure 11. *Flowchart of the steps of the training process*

D. Selection of the best model: In the training process, as explained above, we had four algorithms to try and build the best model. Therefore, in order to select the best model, we first needed to train each model with the training dataset, then find the model's best hyperparameters with the validation dataset and finally evaluate the model with the test dataset. After completing the training process, we found the best model which has been built with the XGBoost algorithm with the histogram method. The Cross-Validation score of the model was 0.864, and the overall accuracy was 87.7% with the validation dataset and 84.96% with the test dataset.

4. RESULTS AND DISCUSSIONS

4.1. Results

For the evaluation process, besides the overall accuracy and Cross-validation (10-fold), we used various metrics namely Sensitivity, Specificity, Precision, Recall, F1-Score, ROC-AUC-Score, the False Positives, the False Negatives, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and MCC. The Confusion Matrix is used to calculate most of these metrics. The confusion matrix is a performance measurement for the classification problems in machine learning where output is a limited number of classes. If the output classes are two, the confusion matrix is a table with 4 different combinations of predicted and actual values. Figure 12 explains the confusion matrix for two classes.

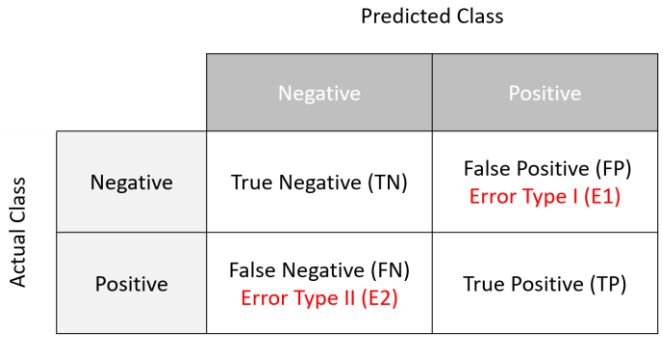

Figure 12. *A graph illustrates the Confusion Matrix output*

- Sensitivity / Recall is a measure that shows the proportion of actual positives that was identified correctly. In our study, Sensitivity / Recall is basically the model accuracy for the Yes outcome. Sensitivity / Recall can be defined mathematically as follows [23], (Equation (2))

$$Sensitivity/Recall = \frac{TP}{TP+FN}. \quad (2)$$

- Specificity is a measure that shows the proportion of actual negatives that was identified correctly. In our study, Specificity is basically the model accuracy for the No outcome. Specificity can be defined mathematically as follows [23,24], (Equation (3))

$$Specificity = \frac{TN}{TN+FP}. \quad (3)$$

- Precision / PPV is a measure that shows the proportion of positive identifications that was actually correct. Also, Precision / PPV can be defined mathematically as follows [25], (Equation (4))

$$Precision/PPV = \frac{TP}{TP+FP} \ . \quad (4)$$

- NPV is a measure that shows the proportion of negative identifications that was actually correct. Also, NPV can be defined mathematically as follows [26], (Equation (5))

$$NPV = \frac{TN}{TN+FN}. \quad (5)$$

- F1-Score is defined as a weighted average of the precision and the recall and can be defined mathematically as follows [27], (Equation (6))

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall}. \quad (6)$$

- ROC-AUC-Score: The area under the curve (AUC) statistic is an empirical measure of classification performance based on the area under a ROC curve. It computes the area under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores [28].

- Matthew's correlation coefficient (MCC): is an evaluation metric that is used in classification problems and its main advantage that it uses all of the confusion matrix elements and measure the correlation between the true and the predicted class, and can be defined mathematically as follows [29], (Equation (7))

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}. \quad (7)$$

All of the selected algorithms have been evaluated against the above-mentioned metrics and the results are shown in Table 4.

Table 4. *Values of each evaluation metric for all of the tested algorithms*

	XGBoost		**XGBoost-hist**		**LightGBM**		**CatBoost**		**ANN**	
	Valid	Test	Valid	Test	Valid	Test	Valid	Test	Valid	Test
Accuracy	0.875	0.846	0.877	0.850	0.873	0.846	0.861	0.836	0.853	0.829
Sensitivity	0.840	0.834	0.842	0.840	0.842	0.835	0.835	0.837	0.806	0.804
Specificity	0.885	0.850	0.887	0.853	0.882	0.850	0.869	0.835	0.867	0.839
Precision	0.686	0.679	0.690	0.685	0.680	0.678	0.656	0.658	0.643	0.654
NPV	0.949	0.931	0.950	0.934	0.950	0.931	0.946	0.931	0.937	0.918
F1-Sore	0.755	0.748	0.759	0.755	0.753	0.749	0.735	0.737	0.715	0.721
AUC	0.863	0.842	0.865	0.847	0.862	0.843	0.852	0.836	0.836	0.821
FN (E2)	91	104	90	100	90	103	94	102	111	123
FP (E1)	220	247	216	242	226	248	250	272	255	266
MCC	0.885	0.850	0.887	0.853	0.882	0.850	0.869	0.835	0.867	0.839
CV 10-fold	0.864		0.864		0.860		0.855		0.828	

Table 4 shows that the XGBoost-hist model has the best overall model which outperformed the other models across all metrics. The default XGBoost model came second with scores that are very close to the XGBoost-hist. Similarly, the LightGBM model had the third-best scores across all metrics with insignificant differences between its scores and XGBoost's Scores. As for the CatBoost model, while its performance was quite well it lagged behind the other gradient boosting algorithms. Finally, the artificial neural networks model has the lowest scores among all of the tested algorithms. While ANN is well known for its state-of-the-art performance across many machine learning tasks, it was not the best-performed algorithm for our task. Moreover, we noticed a pattern that the sensitivity/recall score is a little bit less than the specificity score and that is expected since we have an imbalanced dataset.

Although sensitivity/recall and precision/PPV scores are not the best we got, in our case, it is better to increase recall and NPV over precision and PPV. Precision and recall are important metrics, but we had to find a balance between them because if one is increased, the other will decrease. In order to find the best

balance between precision and recall, we used the F1-Score. One other important metric is the number of False-Positives (E1) and the False-Negatives (E2). In our case, we opted for decreasing the more costly number of False-Negatives (E2). In this study, the False Positive (E1) is a low-risk error because if the model predicts that the patient has diabetes but in reality, he is diabetes-free will not be a problem since the patient can easily get tested for the disease and verify the absence of the disease. But if the model predicts that the patient is diabetes-free but in reality, he has diabetes this will be a high-risk error, because the patient will think that he doesn't have diabetes and may not get tested for the disease which leads the diabetes complications to get worse.

Figure 13 shows the overall accuracy results of each algorithm, this metric value is between 0 and 1 a closer value to 1 means a better model, we can notice that the XGBoost algorithm with the histogram method has the closest value to 1 among the tested algorithms.

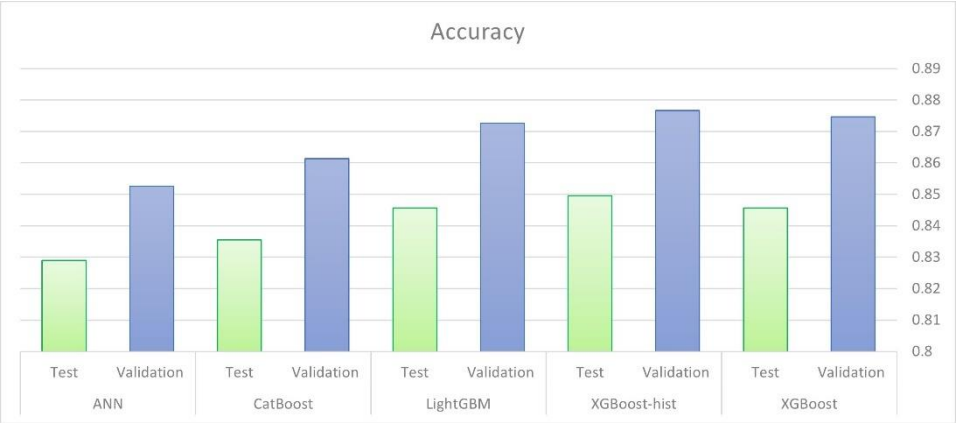

Figure 13. *The overall accuracy results*

Figure 14 shows the cross-validation (10-fold) results of each algorithm, this metric value is the same as the overall accuracy and here XGBoost algorithm with and without the histogram method has the closest value to 1 among the tested algorithms.

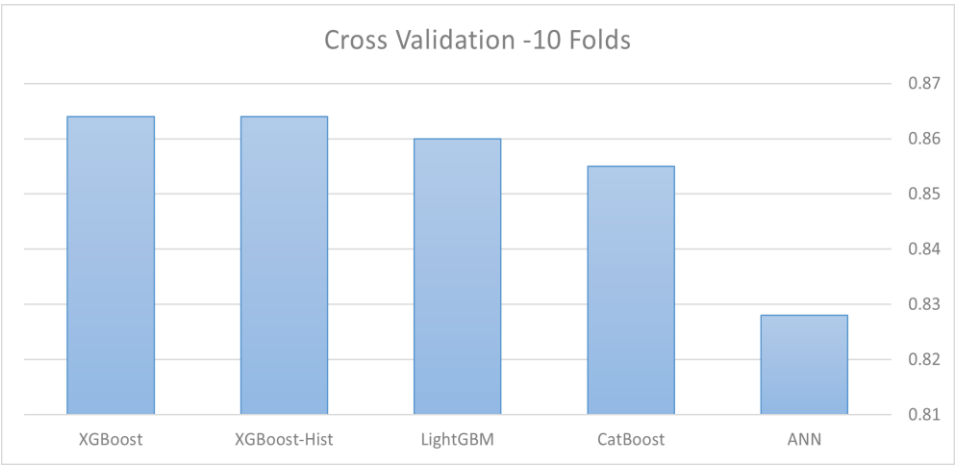

Figure 14. *The Cross-Validation (10-fold) results*

Figure 15 shows the sensitivity or recall results of each algorithm, and like the previous two, this metric value is between 0 and 1 a closer value to 1 means a better model. Here both the XGBoost algorithm with the histogram method and LightGBM algorithm have the closest value to 1 among the tested algorithms.

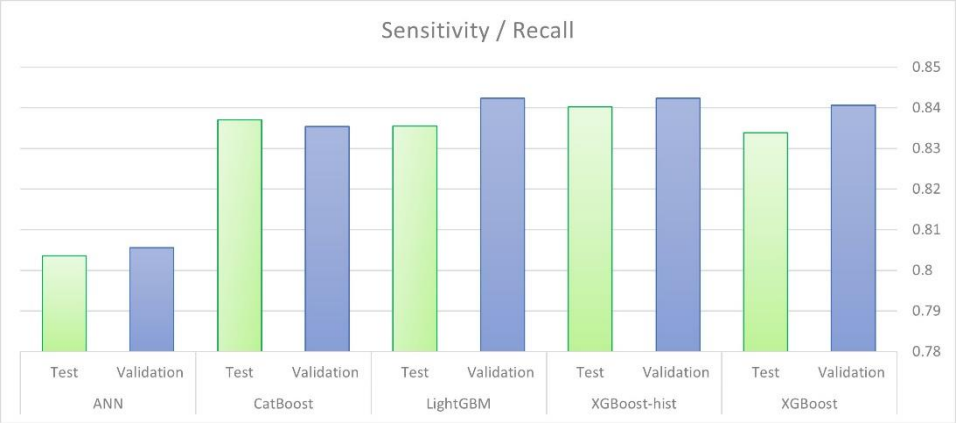

Figure 15. *The sensitivity/recall results*

Figure 16 shows the specificity results of each algorithm, this metric value also is between 0 and 1 a closer value to 1 means a better model, we can notice that the XGBoost algorithm with the histogram method has the closest value to 1 among the tested algorithms.

Figure 16. *The specificity results*

Figure 17 shows the precision or PPV results of each algorithm, this metric value is range from 0 to 1 too and a closer value to 1 means a better model, and as the pattern continues that the XGBoost algorithm with the histogram method has the closest value to 1 among the tested algorithms.

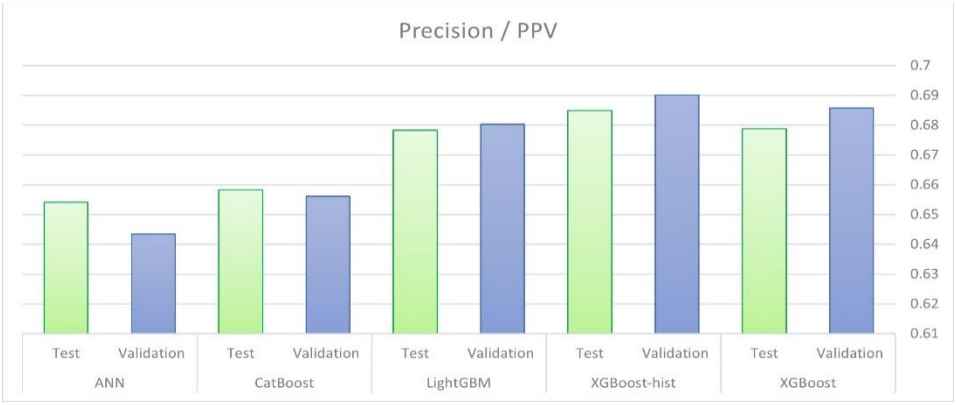

Figure 17. *The precision / PPV results*

Figure 18 shows the NPV results of each algorithm, like the previous metrics, this metric value is between 0 and 1 a closer the value to 1 means better model, and the same pattern continue with the XGBoost algorithm with the histogram method has the closest value to 1 among the tested algorithms.

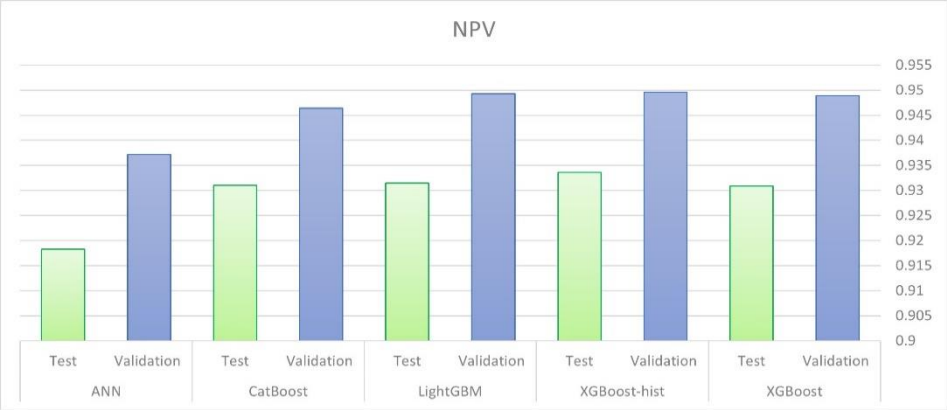

Figure 18. *The NPV results*

Figure 19 shows the f1-score results of each algorithm, this metric value is ranging from 0 to 1 also, and a closer value to 1 means a better model, and as expected from the previous figures the XGBoost algorithm with the histogram method has the closest value to 1 among the tested algorithms.

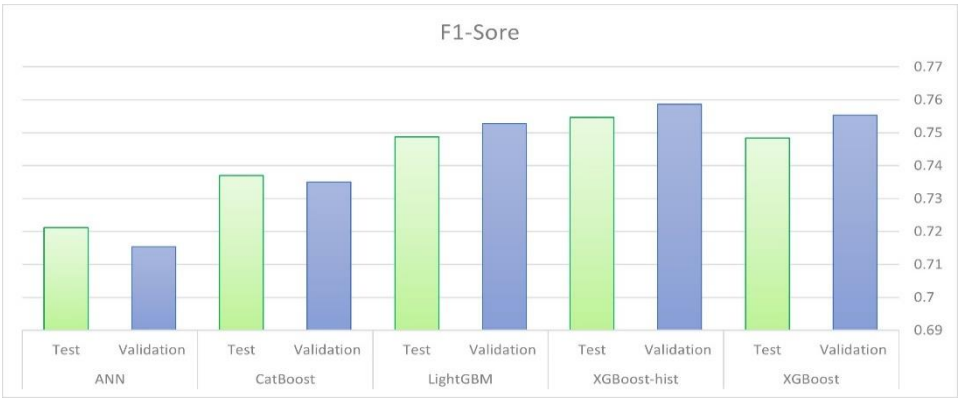

Figure 19. *The f1-score results*

4.2. Discussion

Our classification task is one of the most complicated tasks in the machine learning field. The goal of our study is to find a connection between diabetes prevalence and health behaviors alongside some demographic characters only, which can be obtained easily without requiring any lab results. This is what differentiates our work from the others as mentioned in the related works in the introduction section. Although, there are studies that support this relationship they include either a few variables or too many variables that include pharmacological interventions. Furthermore, the human level of accuracy cannot be measured which results in an unknown value for the Bayes error, the smallest error that can be achieved. Another complication is the dataset, although it has a reliable source it has a lot of missing values along with an unbalanced outcome. The original dataset has more than 55 thousand observations, but after removing all the observations of the participants aged less than 18 years old it drops down to only 35 thousand. The final dataset has 14682 observations with a ratio of 3.28 to 1 for normal, for every 3.28 observations for diabetes -free comes 1 observation for with diabetes. With these drawbacks, the process of building the model became more challenging.

5. CONCLUSION AND FUTURE WORK

The use of machine learning in the medical field appears to be promising and has astonishing achievements. With the help of machine learning tools and algorithms, we were able to build a model that can predict with a Cross-Validation (10-fold) score of 0.864 and an average accuracy of 86%. The robustness, model performance with unseen data, of the model is nearly 85%. Also, the model could predict the outcome with

a relatively low number of E2 and a high score of NPV. With results like these, our model can be classified as an acceptable model and can be deployed as an assistive tool for the diagnosis of type 2 diabetes. With a dataset of higher quality, our model performance will improve significantly. Finally, our model has been built with data that has been collected from the American population, so it may not perform as well with data from other regions because it may have a different distribution than the American population. However, our model can be tuned to perform better with a particular population with data from that population.

As for the future work, since a high-quality dataset leads to a high-performance model, the next step is to gather more data with higher quality, which means more observations, less or no missing values, more balanced outcomes, and from different parts of the world to help the model to generalize well. Also, other variables could be added like sleep patterns and hours, and water consumption. The new dataset then needs to be preprocessed in an optimal way. Next, we rebuild the model with the new dataset. These changes will result in a model with higher performance close to state-of-the-art performance. Then, the achieved model could be used for the diagnosing of type 2 diabetes with little human supervision.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] https://www.healthline.com/health/diabetes. Access Date: 15.09.2019

[2] https://www.diabetes.org/a1c/diagnosis. Access Date: 15.09.2019

[3] Kasl, S. V., Cobb, S., "Health Behavior, Illness Behavior and Sick Role behavior", Archives of Environmental Health: An International Journal, 12(2): 246-266, (1966).

[4] Rogers, R.W., Prentice-Dunn, S., and Gochman, D.S., "Handbook of health behavior research 1: personal and social determinants", New York, NY, US: Plenum Press, Xxviii, 505: 113–132, (1997).

[5] Feldman, A. L., Long, G. H., Johansson, I., Weinehall, L., Fhärm, E., Wennberg, P., Rolandsson, O., "Change in lifestyle behaviors and diabetes risk: evidence from a population-based cohort study with 10 year follow-up", International Journal of Behavioral Nutrition and Physical Activity, 14: 39, (2017).

[6] Gillies, C. L., Abrams, K. R., Lambert, P. C., Cooper, N. J., Sutton, A. J., Hsu, R. T., Khunti, K., "Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis", BMJ, 334: 299, (2007).

[7] Wareham, N. J., "Mind the gap: efficacy versus effectiveness of lifestyle interventions to prevent diabetes", The Lancet Diabetes & Endocrinology, 3: 160–161, (2015).

[8] Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., Nathan, D. M., "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin", The New England journal of medicine, 346(6): 393-403, (2002).

[9] Inzucchi, S. E., "Diagnosis of diabetes", New England Journal of Medicine, 367(6): 542-550, (2012).

[10] Jaleel, J. A., Salim, S., Aswin R. B., "Computer aided detection of skin cancer", in 2013 International Conference on Circuits, Power and Computing Technologies (ICCPCT), 1137-1142, (2013).

[11] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H., "Predicting diabetes mellitus with machine learning techniques", Frontiers in genetics, 9: 515, (2018).

[12] Juneja, A., Juneja, S., Kaur, S., Kumar, V., "Predicting Diabetes Mellitus with Machine Learning Techniques Using Multi-Criteria Decision Making", International Journal of Information Retrieval Research (IJIRR), 11(2): 38-52, (2021).

[13] Muhammad, L. J., Algehyne, E. A., Usman, S. S., "Predictive supervised machine learning models for diabetes mellitus", SN Computer Science, 1(5): 1-10, 5, (2020).

[14] Tigga, N. P., Garg, S., "Prediction of type 2 diabetes using machine learning classification methods", Procedia Computer Science, 167: 706-716, (2020).

[15] https://deepai.org/machine-learning-glossary-and-terms/machine-learning. Access Date: 25.10.2019

[16] https://www.expert.ai/blog/machine-learning-definition/. Access Date: 25.10.2019

[17] https://towardsdatascience.com/ensemble-methods-baggingboosting-and-stacking-c9214a10a205. Access Date: 26.10.2019

[18] https://developers.google.com/machine-learning/data-prep/construct/samplingsplitting/imbalanced-data. Access Date: 24.01.2020

[19] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., Herrera, F., "Learning from imbalanced data sets", 11: 82-117, Berlin: Springer, (2018).

[20] https://towardsdatascience.com/handling-imbalanced-datasets-in-deeplearning-f48407a0e758. Access Date: 22.10.2019

[21] https://www.cdc.gov/nchs/nhanes/index.htm. Access Date: 18.12.2020

[22] Knowles, J. W., Reaven, G., "Usual blood pressure and new-onset diabetes risk: evidence from 4.1 million adults and a meta-analysis", Journal of the American College of Cardiology, 67(13): 1656-1657, (2016).

[23] https://link.springer.com/referenceworkentry/10.1007%2F978-0387-30164-8_752. Access Date: 18.12.2020

[24] https://link.springer.com/referenceworkentry/10.1007%2F978-0387-30164-8_770. Access Date: 18.12.2020

[25] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_645. Access Date: 18.12.2020

[26] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_582. Access Date: 18.12.2020

[27] https://scikitlearn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score. Access Date: 18.12.2020

[28] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_28. Access Date: 18.12.2020

[29] Chicco, D., Tötsch, N., Jurman, G., "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation", BioData Mining, 14, 13, (2021).