Article

Not peer-reviewed version

Machine Learning Based Approach for Predicting Diabetes Employing Socio-demographic Characteristics

Md. Ashikur Rahman , Lway Faisal Abdulrazak , Md. Mamun Ali , Imran Mahmud , Kawsar Ahmed * , Francis M. Bui

Posted Date: 3 October 2023

doi: 10.20944/preprints202310.0109.v1

Keywords: Diabetes; Socio demographic characteristics; machine learning; polydipsia; sudden weight loss.

Article

Machine Learning Based Approach for Predicting Diabetes Employing Socio-Demographic Characteristics

Md. Ashikur Rahman [1], Lway Faisal Abdulrazak [2], Md. Mamun Ali [1,3,4] Imran Mahmud [1], Kawsar Ahmed [4,5,6,*], and Francis M. Bui [3,5]

[1] Department of Software Engineering, Daffodil International University, Daffodil Smart City (DSC), Birulia, Savar, Dhaka-1216, Bangladesh; ashikur35-562@diu.edu.bd, imranmahmud@daffodilvarsity.edu.bd

[2] Department of Computer Science, Cihan University Sulaimaniya, Sulaimaniya 46001, Kurdistan Region, Iraq; lway.faisal@sulicihan.edu.krd

[3] Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada; m.ali@usask.ca

[4] Health Informatics Research Lab, Department of Computer Science and Engineering, Daffodil International University, Daffodil Smart City, Birulia, Dhaka-1216, Bangladesh; mamun35-274@diu.edu.bd

[5] Department of Electrical and Computer Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada; francis.bui@usask.ca

[6] Group of Bio-photomatiχ, Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail 1902, Bangladesh; kawsar.ict@mbstu.ac.bd

[*] Correspondence: kawsar.ict@mbstu.ac.bd, k.ahmed@usask.ca and k.ahmed.bd@ieee.org (Kawsar Ahmed)

[†] These authors contributed equally to this work.

Abstract: Diabetes is one of the fatal diseases that play a vital role in the growth of other diseases in the human body. Controlling and curing diabetes in its early stages is the most significant technique to avoid its effects of diabetes. However, lack of awareness and expensive clinical tests are the primary reasons to skip clinical diagnosis and take preventive methods in lower-income countries like Bangladesh, Pakistan, and India. From this perspective, the study aims to build an automated machine learning (ML) model, which will predict diabetes at an early stage using socio-demographic characteristics rather than clinical attributes. Because clinical features are not always known to all people from lower-income countries. To find the best fit supervised ML classifier of the model, we applied six classification algorithms and found that RF outperformed with an accuracy of 99.36%. In addition, the most significant risk factors were found based on the SHAP value by all the applied classifiers. The study reveals that polyuria, polydipsia, and delayed healing are the most significant risk factors for growing diabetes. The findings indicate that the proposed model is highly capable of predicting diabetes in the early stages.

Keywords: diabetes; socio demographic characteristics; machine learning; polydipsia; sudden weight loss.

1. Introduction

Diabetes is one of the diseases that people are most afraid of nowadays. Every country around the globe, whether developed or underdeveloped, is affected by the diabetes epidemic. These days, it affects the entire country and is a hardship for all the nations, especially for emerging nations like Bangladesh, India, and Pakistan. People with little awareness of medical conditions are at greater risk. The World Health Organization (WHO) report says that from 1980 to 2014 about 314 million diabetes patients increased worldwide [1]. Besides, according to this research, diabetes spreads more quickly in developing nations than in high-income ones [1]. From 2000 to 2019 diabetes deaths among certain ages, people have increased by almost 3%. Diabetes and kidney disease caused almost 2 million deaths worldwide in 2019 [1].

Diabetes is a chronic condition brought on by either insufficient insulin production by the pancreas or inefficient insulin utilization by the body. There are mainly two types of diabetes type 1 and type 2. Type 1 diabetes is a condition in which the body stands deficient in insulin production and demands daily insulin injections. There are a few symptoms of type 1 diabetes; those are excessive urine excretion (polyuria), excessive thirst (polydipsia), extreme hunger (polyphagia), sudden weight loss, and vision shifts (visual_blurring). The inefficient usage of insulin causes type 2 diabetes. Also, the symptoms of type 2 diabetes are as same as type 1 [1]. Another sort of diabetes is called Diabetes mellitus. There are also some symptoms of diabetes mellitus those are excessive urine excretion (polyuria), excessive thirst (polydipsia), and extreme hunger (polyphagia) [33]. In addition, Gestational diabetes is a special kind of diabetes. Pregnancy-related hyperglycemia refers to levels of blood sugar that are above average but less than those associated with diabetes. Chronic type 2 diabetes danger is elevated for such mothers and presumably for their offspring as well. Prenatal testing instead of observed symptoms is employed to identify gestational diabetes [1]. Diabetes mellitus seems to be a circumstance in which the body produces insufficient insulin or fails to utilize it properly, leading to excessively high blood sugar amounts.

Diabetes is the source of many other deadly diseases. It is a highly potential disease to harm the heart, blood vessels, eyes, kidneys, and nerves. So, it is urgent to predict diabetes among patients. Otherwise, it can cause other diseases in our bodies. We can prevent this dangerous disease by following some lifestyle rules. Even though following the rules of lifestyle, people have a risk to affect diabetes. If we can predict diabetes at an early stage, it is possible to control it. Changing some lifestyle and obeying the doctor's suggestion, the patient gets relief from this disease. So it is said that predicting diabetes at an early stage is crucial to prevent the mortality rate of this disease. Every year, all the countries around the globe spend a large number of funds on diabetes. A report from the American Diabetes Association (ADA) expresses that the whole world spent $245 billion in 2012, and the amount of money increased by $82 billion in the next five years. In 2017 it had been $327 billion [2]. If we predict diabetes disease at an early stage, it also can prevent a lot of spending money.

In recent years, numerous studies have looked to predict diabetes through several Machine Learning (ML) models. Khanam, Jobeda Jamal, and Simon Y. Foo. (2021) have shown a Neural Network (NN) (NN)-based model with an accuracy of 88.6%. In their study, they utilized a dataset obtained from the Pima Indian Diabetes (PID) dataset. Although they built a model with NN, they did not show the impact of features on this model in their research. And the accuracy of the model has also not been good enough [3]. Islam, M. M. et al. (2020) performed data mining techniques and found Random Forest (RF) gave the best results, with 97.4% accuracy on 10-Flod cross-validation and 99% accuracy on the train-test split. They have used a dataset collected through oral interviews from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh. They have shown good accuracy, but their dataset was unbalanced. They did not use any data balancing techniques [4]. Krishnamoorthi, Raja, et al. (2022) have built a framework for diabetes prediction called the intelligent diabetes mellitus prediction framework (IDMPF) with an accuracy of 83%. They have employed a dataset that is also the Pima Indian Diabetes (PID) dataset. The result of their model is still not good enough, and there has been scope to improve this result [5].

Islam, Md. Shafiqul et al. (2020) have proposed a model for the prediction of Type 2 diabetes in the future, and they achieved 95.94% accuracy. The collected dataset used in their study was from the San Antonio Heart Study, a widely prescribed investigation. They did a good job, but the number of features was only 11, which is an inefficient number to build and validate a machine-learning model [6]. Hasan, Md Kamrul, et al. (2020) assembled classifiers to propose a model with an AUC of 0.95. In this work, they utilized the Pima Indian Diabetes (PID) dataset [7]. Fazakis, Nikos, et al. (2021) have proposed an ensemble WeightedVotingLRRFs ML model with an AUC of 0.884 for Type 2 diabetes prediction. The collected dataset in their study was from the English Longitudinal Study of Ageing (ELSA) database. In this study, there have not been any feature analysis techniques [8]. Ahmed, Usama, et al. (2022) performed a Machine Learning (ML) based model of the Fused Model for Diabetes Prediction (FMDP) and got an accuracy of 94.87%. For this study, they used a dataset collected from the hospital of Sylhet, Bangladesh. Their study method was well designed, but they have not shown any feature impact on the model and no feature analysis [9]. Maniruzzaman, Md et al. (2020) have introduced a model combining Logistic Regression (LR) feature selection and Random Forest (RF), which gives an accuracy of 94.25% and an AUC of 0.95. The National Health and Nutrition Examination Survey conducted from 2009-2012 has been used in their research. The dataset has only 14 features [10]. Barakat, Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat., (2010) have conducted a study for diabetes Mellitus prediction using a Support Vector Machine (SVM) with an accuracy of 94%, a sensitivity of 93% and specificity of 94%. In this study, they did not introduce any feature analysis techniques [11].

So, in this research, we proposed a Machine Learning (ML)-based model for diabetes prediction at an early stage. In recent years, ML has proven to be a very efficient technique for disease prediction. Nowadays, ML plays an essential role in the biomedical sector to overcome traditional methods of diagnosis, disease prediction, and treatment. Consequently, there is no doubt about using an ML-based prediction model to predict diabetes. Our contributions are mentioned as follows:

- Building an ML model that will predict diabetes using socio-demographic characteristics rather than clinical attributes. Because all people, especially from lower-income countries do not know the clinical features.
- Revealing significant risk factors that indicate diabetes.
- Proposing a best fit clinically usable framework to predict diabetes at an early stage.

2. Materials and Methods

2.1. Dataset Description

The dataset used in this study was collected from Kaggle, an online data repository [12]. There were about 520 observations in this dataset, where 320 observations are diabetes-positive, and others are diabetes negative. The dataset contains 17 features, of which one is the target feature. The value of the target feature is either 0 (Diabetes negative) or 1 (Diabetes positive). The other 16 features are two types such as numeric, and nominal. Details about the datasets are represented in Table 1, including the name, data type, and explanation of each characteristic.

Table 1. Brief explanation of dataset.

Attributes	Data Type	Interpretation
age	numeric	Age of the patient
gender	nominal	Whether the patient male/female
polyuria	nominal	Is not whether the patient had frequent urination
polydipsia	nominal	Determine whether or not the patient had excessive thirst/drinking
sudden_weight_loss	nominal	Whether or not the patient experienced a period of sudden reduced weight
weakness	nominal	Whether the patient experienced a moment of weakness
polyphagia	nominal	Whether or not the patient experienced extreme hunger
genital_thrush	nominal	Whether or not the patient had a yeast infection
visual_blurring	nominal	Whether the patient experienced unclear seeing.
itching	nominal	If the patient had an experienced of itch
irritability	nominal	If or not the patient had an experienced of irritability
delayed_healing	nominal	Whether the patient observed a delay in recovery after being injured
partial_paresis	nominal	If the patient experienced a period of muscle wasting or a group of failing muscles
muscle_stiffness	nominal	Whether or not the patient experienced of muscle stiffness
alopecia	nominal	Patient had hair loss or not
obesity	nominal	Considering his body mass index, determine yet if the patient is obese or not.
class	nominal	Presence of Diabetes (Positive/Negative)

2.2. Data Preprocessing

The preprocessing of data is essential for every ML as well as data mining technique. Because the efficiency of a model mostly depends on data preprocessing. Missing values were handled after

doi:10.20944/preprints202310.0109.v1

obtaining the dataset. However, there were no missing values in this dataset. Then, an encoding technique was employed for the processed dataset. Encoding is a fundamental technique in data preprocessing. If there is any object-type (String) data present in the dataset, these are not used for any ML algorithms. So, it needs to convert the object type (String) data to integer-type data, which is suitable for ML algorithms. In this dataset, there is one feature (gender) that is object type; for that reason, an encoding technique is performed. This research has used the One-Hot Encoding technique. After completing the encoding method, this dataset has been made suitable for ML algorithms. Then, statistical analysis and exploratory data analysis (EDA) was carried out on the processed dataset. The overall research methodology is represented in Figure 1.

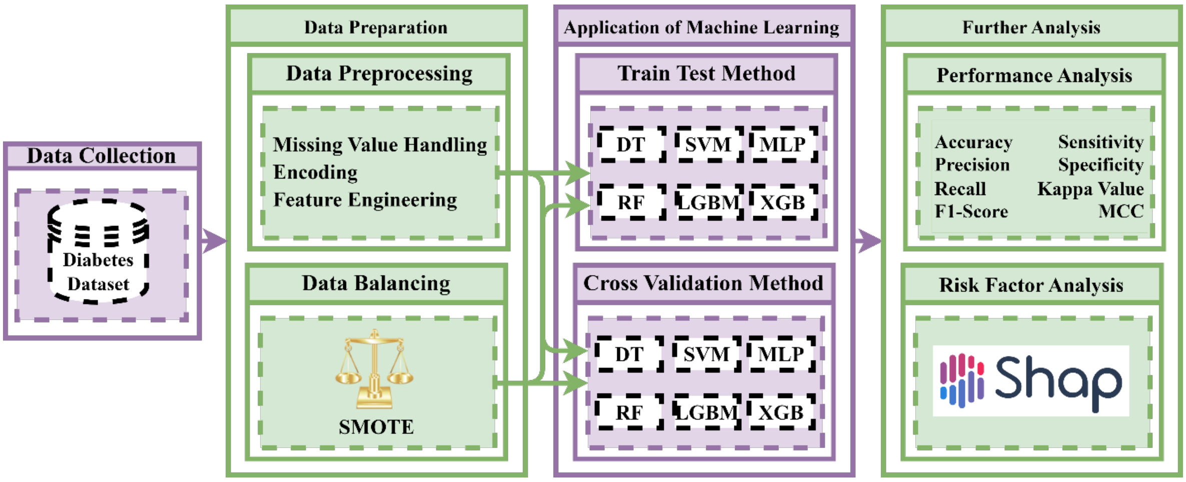

Figure 1. Experimental methodology of the study for building a diabetes prediction model using socio-demographic characteristics by machine learning techniques.

2.3. Data Balancing Techniques

Synthetic Minority Oversampling Technique (SMOTE) has been utilized to balance the imbalance dataset. SMOTE is an oversampling approach to balance the imbalanced data. It is one of the most widely used balancing techniques. It is employed to address the imbalance issue. It attempts to balance the number of classes by randomly creating minority class samples and duplicating them. SMOTE introduces unique minority instances by synthesizing existing minority instances. For something like the minority class, linear interpolation is used to create virtual training data. By choosing a random one or several of the k-nearest neighbors for every instance in the minority class, such synthetic training records are constructed. This data is regenerated after the oversampling procedure, and several categorization methods can be used to analyze the data input [29]. It selects instances inside the feature set that are close to each other, draws a line between the instances, and then creates a new instance at a location somewhere along the line.

2.4. Performance Evaluation Metrics

Accuracy and other statistical evaluation metrics were considered to find the best-fit ML model among all the applied classifiers. All the applied supervised ML classifiers were compared among each other based on the criteria used to evaluate their efficiency. In most cases, ML models are assessed using sensitivity, specificity, and accuracy; these are generated by a confusion matrix. Classification accuracy is the ratio of the models correctly classified to all other possible outputs. Accuracy is a suitable metric whenever the target feature categories in the data are pretty equal [13]. Specificity describes the percentage of true negatives estimated to be negatives [14]. A metric called "Sensitivity" shows the proportion of actual positive events that were assumed to be positive [14]. The following equations are used to determine the value of all the statistical evaluation metrics [13,14].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

In addition to these three evaluation metrics, for more precise evaluation, several other evaluation metrics were considered to determine how effectively each algorithm performed. These are Matthew's Correlation Coefficient (MCC), kappa statistics, recall, precision, and f1-measure. Matthews correlation coefficient (MCC), takes the confusion matrix's four parameters into account, as well as a maximum level (near to 1) shows that both classifications are well estimated, though if one category is significantly under (or over) represented [15,16]. A recall is a metric that represents the number of positives that the Machine Learning (ML) algorithms obtained [17]. This score will calculate the harmonic mean of accuracy and recall. The weighted average of accuracy and recall is utilized to evaluate the F1 score [17]. The precision determines the ratio of true positives to all expected positives [17]. The observed and estimated accuracy is analyzed using the Kappa statistic [18]. In the equation, the terms TP, FP, and TN, FN are respectively substituted as True Positive, False Positive, True Negative, and False Negative.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

$$KappaStatistics = \frac{observedaccuracy - expectedaccuracy}{1 - expectedaccuracy} \tag{7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

A 5-fold cross-validation (CV) approach is demonstrated in Figure 2. The dataset is broken down into five groups, four of which take part in model training and one of which evaluates model training after each round. A 5-fold CV is used in this work. To prevent overfitting inside a classification algorithm, cross-validation is the most efficient method.

Train-test-split is a traditional approach to ML algorithms. In train-test-split the dataset is broken into two parts, one part is used to train the model, and the rest of the other part is used for testing the model [34]. In this study, we have used 70% of the total data for training our ML models, and the rest of the 30% of the data we used for testing the model and analyzing the performance of ML models.

2.5. K-Fold Cross-validation and Train Test Split

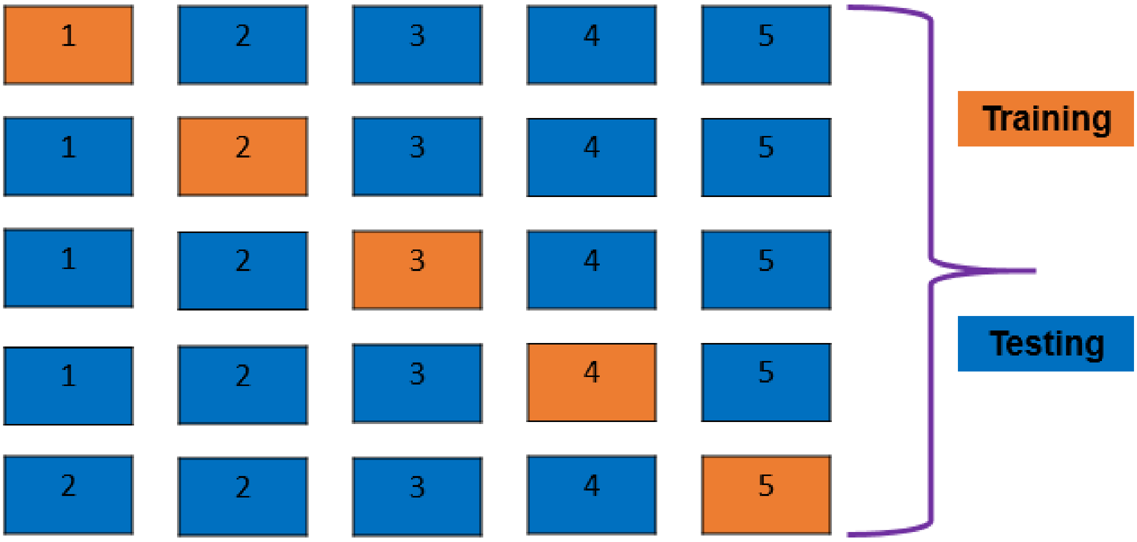

Figure 2. Visual representation of K-Fold CV method in supervised machine learning model training and testing.

2.6. Machine Learning Algorithms

Several supervised and classified Machine Learning (ML) approaches were applied in this research. Here is some suggested supervised machine learning (ML) methods for predicting diabetes.

2.6.1. Decision Tree Classifier

Decision Tree (DT) is a supervised ML algorithm, that can be used in both classification and regression tasks. It consists of a tree-structured classifier, in which each leaf node represents the classification outcome and inside nodes represent the attributes of a dataset [19]. A DT consists of two nodes one is the Decision Node and the other is the Leaf Node. Decision nodes are used for taking actions and include some branches, on the other hand, Leaf nodes show the results of those decisions and do not consist of any additional branches. Because it grows on succeeding branches to form a structure resembling a tree, it is called as a "decision tree" because, like a tree, it starts from the root. The leaves stand in for the options or possibilities. These decision nodes split up the data. In terms of Decision Tree (DT) building, there are two metrics: one is Entropy/ Gini-index, and the other is Information Gain (IG). The two metric calculation equations are shown below.

$$Entropy(S) = \sum_{i=1}^{n} -p_i \log(p_i) \tag{9}$$

$$IG(S, A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{\mid S_v \mid}{\mid S \mid} \times Entropy(S_v) \tag{10}$$

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{11}$$

2.6.2. Random Forest Classifier

Random Forest (RF) is a type of ensemble learning and a supervised ML classifier. An ensemble of DTs, the majority of which were trained using the "bagging" method, is combined to create a forest.

The bagging approach's core concept is that by integrating multiple learning methods, the outcome is improved [20]. Based on voting methods, this supervised learning methodology forecasts the outcome. The Random Forest (RF) forecasts that the ultimate prediction will be 1, and vice versa, if the majority of the trees in the forest offer a prediction of 1 [21]. Additionally, Random Forest (RF) is a quantitative approach that applies decision tree classifiers to various resamples of the dataset before using averaging to improve prediction accuracy and prevent overfitting. When bootstrap=True, the max samples option controls the size of the resamples; otherwise, each tree is created using the entire dataset [22]. Random Forest (RF) also uses the same metrics that have been used in Decision Tree (DT) classifiers. Like Entropy, Entropy/ Gini-Index, and Information Gain (IG). The equation of those metrics is already shown above in the Decision Tree (DT) subsection.

2.6.3. Support Vector Machine

Support Vector Machines (SVM) is a group of supervised learning approaches that deal with classification tasks, analysis of regression problems, as well as outliers' identification. Because of their capacity to choose a decision boundary that minimizes the distance from the adjacent data points in all classifications, SVMs differ from other classification algorithms. The decision boundary classifier or the highest margin hyperbolic decision boundary generated by SVMs is referred to as plane and plane. SVM has two types, the first is Simple SVM, and the second is Kernel SVM [23]. In this study, we have used the Kernel SVM. On Kernel SVM, we used the linear kernel SVM. The majority of other kernel functions are slower than linear kernel functions, and there are fewer parameters to optimize. The equations that are used by linear kernel SVM have been described below [23].

$$f(X) = w^T \times X + b \tag{12}$$

Throughout this equation, w stands for the weight matrix that you would like to optimize, X for the data you intend to interpret, and b stands for the predicted linear coefficient from either the training dataset or the test dataset. The above equation establishes the output range of the SVM.

2.6.4. XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a method for ensemble learning. Sometimes, sole reliance on the output of one Machine Learning (ML) model may not have been effective. A technique for systematically combining the prediction skills of several learners is ensemble learning. As a result, a mono framework that incorporates the output of several models is produced [24]. Additionally, the decentralized gradient boosting framework XGBoost was created to be very effective, flexible, and portable. It develops Machine Learning (ML) methods using the Gradient Boosting framework. In order to swiftly and reliably carry out a wide range of data science applications, XGBoost uses concurrent tree boosting [25]. Efficiency and implementation duration were taken into consideration when developing the XGBoost algorithm. In comparison to other boosting algorithms, it works substantially faster. With XGBoost, problems with regression and classification can both be resolved. This method significantly improves the decision tree chain's weight-dependent efficiency. For this work, we have used the default XGBoost algorithm; we do not tune any hyper-parameters of this algorithm.

2.6.5. LightGBM Classifier

The LightGBM/LGBM (Light Gradient Boosting Machine) gradient-boosting approach uses concepts from tree-based modeling. It can manage enormous volumes of data due to the decentralized architecture, ability for parallel learning, and use of GPUs. The speed of LGBM is six times that of XGBoost. A rapid and precise machine learning technique is XGBoost. Conversely, LGBM, which executes more quickly with comparable predictive performance and simply provides additional hyperparameters to modify, is potentially posing a threat. The key performance difference is that

doi:10.20944/preprints202310.0109.v1

whereas LGBM separates the tree vertices one at a time, XGBoost does it one layer at a time [26]. Furthermore, LGBM is a gradient-boosting technique that employs similar tree-based instructional strategies. A different method develops trees parallel to the ground, whereas LGBM produces trees upwardly, or, to put it another way, LGBM produces trees leaf by leaf, whereas another method produces trees level by level. The leaf with the greatest delta erosion will be produced [27]. This study has used the default parameter for LGBM classifiers.

2.6.6. Multi-Layer Perceptron

The least complex Artificial Neural Network is called Multi-Layer Perceptron (MLP). It is a synthesis of various perceptron algorithms. Perceptors are designed to mimic the functions that the human brain performs in an attempt to overcome issues. Such perceptrons are orthogonal in MLP and have a significant degree of connectivity. Effective parallel processing facilitates quicker computing. Frank Rosenblatt developed the perceptron in 1950. That, such as the human brain, is capable of learning complicated tasks. A perceptron structure (Output Unit) is made up of Sensory Unit (Input Unit), Associator Unit (Hidden Unit), and Response Unit [28]. A completely associated input layer and an output layer make up the perceptron. The input and output layers of MLPs are the same, but there could be several hidden layers somewhere in the input layer or output layer. The MLP model is developed continuously. The cost function's partial derivatives are employed to modify each phase's parameters.

2.7. Feature Importance and Model Explanation

The most significant thing in any ML strategic approach is choosing the appropriate method. While taking into account several assessment matrices and scientifically assessing the results, we chose the superior model for the current study. Showing the feature impact of each model on their prediction is also an essential concept in ML approaches. Features' impact plays a vital role in building an effective ML model. The features' impacts show why those features are important in building a specific model and how those features influence the model's prediction side by side. Show features' impact on the model's prediction will significantly affect studies for forecasting in the disciplines of social science and healthcare. SHAP (SHapley Additive exPlanations) plots have been utilized in the research to show the features' impact on the model's prediction. The significance of receiving a specific value for a specific characteristic in comparison to the forecast we would provide if that attribute had a quantitative amount is quantified by SHAP values [30].

3. Result Analysis & Discussion

In this study, six supervised ML algorithms were employed to build a model to predict diabetes using socio-demographic characteristics in the early stages. Before applying machine learning techniques. Exploratory data analysis (EDA) is performed to explore the hidden knowledge of the applied dataset. Then, ML techniques are conducted to build a potential model to predict diabetes and to find out the most significant socio-demographic risk factors related to diabetes. All the findings of the study are represented in this section.

3.1. Exploratory Data Analysis (EDA) result

Figure 3 depicts the results of exploratory data analysis of the diabetes dataset for all the features. In Figure 3, N refers to the negative, whereas P refers to the positive. In addition to that F, and M refer to females and males, respectively. According to Figure 3, females are more affected by diabetes compared to males. The figure also shows that patients with polyuria, polydipsia, or sudden weight loss are more likely to have diabetes. Of the patients who do not have polyuria, polydipsia, and sudden weight loss, around 70% do not have diabetes. For polyphagia, irritability, partial paresis, and obesity syndromes, patients have a high diabetes risk. Among those who do not have polyphagia, irritability, partial paresis, or obesity, about half have diabetes. According to Figure 3, patients, more than 30 years old, are highly vulnerable to the effect of diabetes.

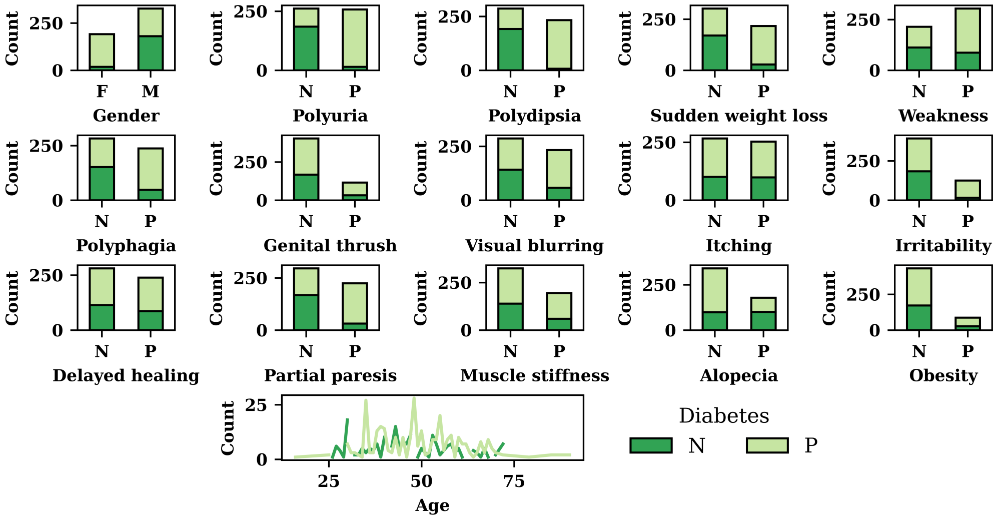

Figure 3. Exploratory Data Analysis Result

3.2. Performance Evaluation of ML Models

Six machine learning models such as MLP, SVM, DT, LGBM, XGB, and RF were applied and their performances were compared among each other to find the best-fit model to predict diabetes in the early stage. The results of the ML models are represented in the following sections.

At first, the imbalanced dataset was trained using the train test split method, where 70% of the dataset was utilized to train the model, and 30% of the dataset, was employed for testing the built models. The result of the train test split method on the imbalanced dataset is represented in Table 2. According to Table 2, the lowest performance is generated by SVM and MLP classifiers. RF has the highest accuracy score of 98.44% among the six ML algorithms. Furthermore, RF also gives the maximum scores for the rest of the performance metrics: precision, recall, f1-score, sensitivity, specificity, kappa-statistics, and MCC value, which are respectively 0.9800, 0.9899, 0.9849, 0.9785, 0.9899, 0.9687, and 0.9687.

Table 2. Performance evaluation on imbalance dataset for train test split method.

Algorithm	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity	Kappa Statistics	MCC
SVM	92.19%	0.9117	0.9394	0.9254	0.9032	0.9394	0.8434	0.8438
MLP	93.23%	0.9388	0.9293	0.934	0.9355	0.9293	0.8645	0.8645
LGBM	94.27%	0.9782	0.9091	0.9424	0.9785	0.9091	0.8855	0.8879
XGB	96.35%	0.9791	0.9495	0.9641	0.9785	0.9495	0.9271	0.9275
DT	97.39%	0.9896	0.9596	0.9743	0.9892	0.9596	0.9479	0.9484
RF	98.44%	0.98	0.9899	0.9849	0.9785	0.9899	0.9687	0.9687

Table 3 shows the results for different performance metrics on the imbalanced dataset. RF has the highest accuracy score of 98.44% among the six ML algorithms. Furthermore, RF also gives the maximum scores for the rest of the performance metrics: precision, recall, f1-score, sensitivity, specificity, kappa-statistics, and MCC value, which are respectively 0.9800, 0.9899, 0.9849, 0.9785, 0.9899, 0.9687, and 0.9687.

Table 3. Performance evaluation on balance dataset for train test split method.

Algorithm	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity	Kappa Statistics	MCC
MLP	93.59%	0.9423	0.9608	0.9514	0.8889	0.9608	0.8571	0.8575
SVM	93.59%	0.951	0.951	0.951	0.9074	0.951	0.8584	0.8584
DT	94.87%	0.9796	0.9412	0.96	0.9629	0.9412	0.8886	0.89
LGBM	98.08%	1	0.9706	0.9851	1	0.9706	0.958	0.9589
XGB	98.72%	1	0.9804	0.9901	1	0.9804	0.9719	0.9723
RF	99.37%	1	0.9902	0.9951	1	0.9902	0.9859	0.986

Table 4 represents detailed information about the ML approaches for 5-fold CV results on the balanced dataset. The maximum CV accuracy is 94.87% for RF classifiers. DT shows the highest precision value of 0.9784, and RF gives the highest recall and f1-scores of 0.9608 and 0.9608. At the same time, DT also gains the maximum sensitivity value of 0.9629. The maximum specificity, kappa-statistics, and MCC values given through RF are 0.9608, 0.8867, and 0.8867, respectively.

Table 4. Performance evaluation on balance dataset for 5-fold CV method.

Algorithm	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity	Kappa statistics	MCC
DT	91.61%	0.9784	0.8921	0.9333	0.9629	0.8921	0.8228	0.8291
RF	94.87%	0.9608	0.9608	0.9608	0.9259	0.9608	0.8867	0.8867
SVM	89.74%	0.8981	0.9509	0.9238	0.7963	0.9509	0.7673	0.7703
XGB	92.95%	0.9691	0.9216	0.9447	0.9444	0.9216	0.8475	0.8496
LGBM	92.95%	0.9691	0.9216	0.9447	0.9444	0.9216	0.8475	0.8496
MLP	92.95%	0.9505	0.9412	0.9458	0.9074	0.9412	0.8449	0.845

5-flood CV results on the balanced dataset for the ML approaches have been described in Table 5. According to Table 5, RF and LGBM have the highest maximum accuracy of 93.23%. Besides, the RF and LGBM show the highest precision value of 0.9574. The highest recall and specificity value is 0.9091 which is generated by DT, RF, XGBoost, and LGBM classifiers. Both RF and LGBM show a maximum sensitivity score of 0.9570. RF and LGBM have shown maximum values of f1-score, kappa-statistics, and MCC of 0.9326, 0.8647, and 0.8658, respectively.

Table 5. Performance evaluation on imbalance dataset for 5-fold CV method.

Algorithm	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity	Kappa Statistics	MCC
DT	91.67%	0.9278	0.9091	0.9184	0.9247	0.9091	0.8333	0.8334
RF	93.23%	0.9574	0.9091	0.9326	0.957	0.9091	0.8647	0.8658
SVM	87.50%	0.8947	0.8586	0.8763	0.8925	0.8586	0.7501	0.7507
XGB	92.71%	0.9474	0.9091	0.9278	0.9462	0.9091	0.8542	0.8549
LGBM	93.23%	0.9574	0.9091	0.9326	0.957	0.9091	0.8647	0.8658
MLP	91.14%	0.9271	0.8989	0.9128	0.9247	0.8989	0.8229	0.8233

Figure 4 shows the ROC-curve and Precision-Recall (PR) curve for six different ML techniques that have been applied in this study. The results for the balanced dataset are shown in Figures 4(A) and 4(B). On the other hand, Figures 4(C) and 4(D) show the results of the imbalanced dataset. For a balanced dataset, the highest AUC score is 1.00 for RF, XGBoost, and LGBM, as shown in the figure. At the same time, RF, XGBoost, and LGBM show the highest AUCPR value, which is also 1.00. On the contrary, the maximum AUC score is 0.999, as shown by RF. Respectably, RF also shows the highest AUCPR value of 0.999 for an imbalanced dataset.

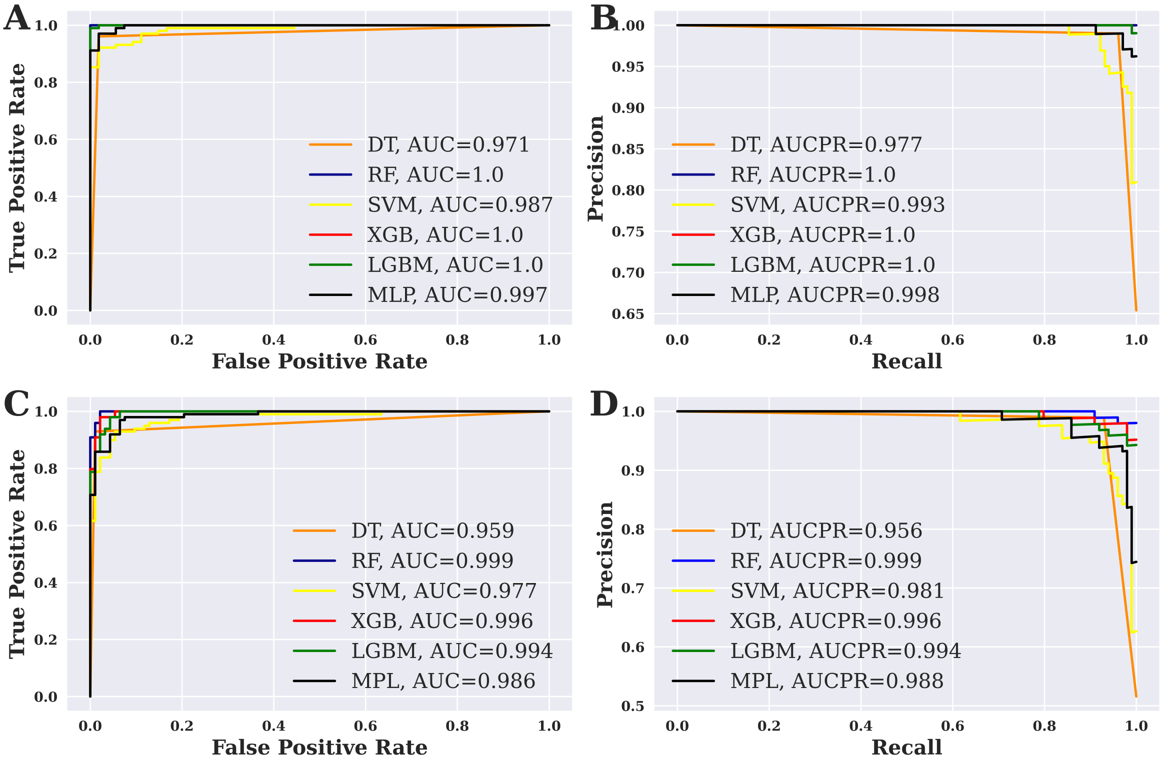

Figure 4. ROC Curve and PR Curve Analysis

3.3. Overall Performance Evaluation for ML Methods

The results of performance metrics for six ML approaches have been shown in Figure 5. In Figure 5, we compare the results of train-test-split and cross-validation techniques for the balanced dataset. It is shown that the results of train-test-split are mostly higher than those of cross-validation. But in a few cases, the results of cross-validation are increasing. For DT, precision and sensitivity show the same results for both train-test-split and cross-validation. SVM yields the same recall and specificity results for both techniques. Furthermore, the MLP shows few exceptions; in the MLP, the results of precision and sensitivity for cross-validation are higher than the train-test-split's result.

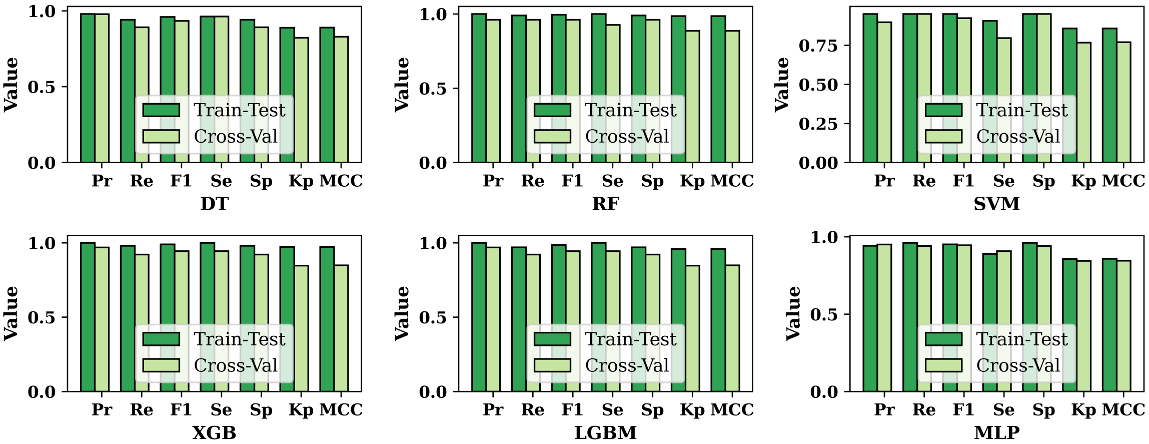

Figure 5. Compare the results of different performance metrics based on train-test-split and cross-validation for the balanced dataset.

Figure 6 compares the accuracy of the six ML approaches for both datasets. It has been shown that the balanced dataset's accuracy is always higher than the imbalanced dataset for all the classifiers that have been used in this research. Among the six ML algorithms, RF shows the highest accuracy for both the dataset and the techniques of train-test-split and cross-validation.

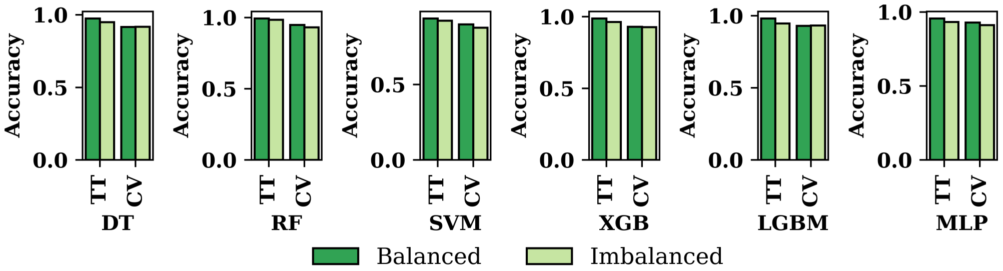

Figure 6. Accuracy of the six applied classifiers for balanced and imbalanced dataset based on Train-Test-split and Cross-Validation.

3.4. Risk Factor Analysis and Model Explanation Based on SHAP Value

This study aims to emphasize the features' impact on predicting diabetes for different ML techniques. We have been utilizing the SHAP Summery plot to carry out and show the feature's impact on the model. Using SHAP summary plot features are categorized in terms of how they affect the forecast. It takes into account the absolute SHAP value; hence, it matters if the feature affects the prediction either positively or negatively [31]. Feature's impact on model prediction utilizing the SHAP Summery plot for six ML algorithms has been shown in Figure 7.

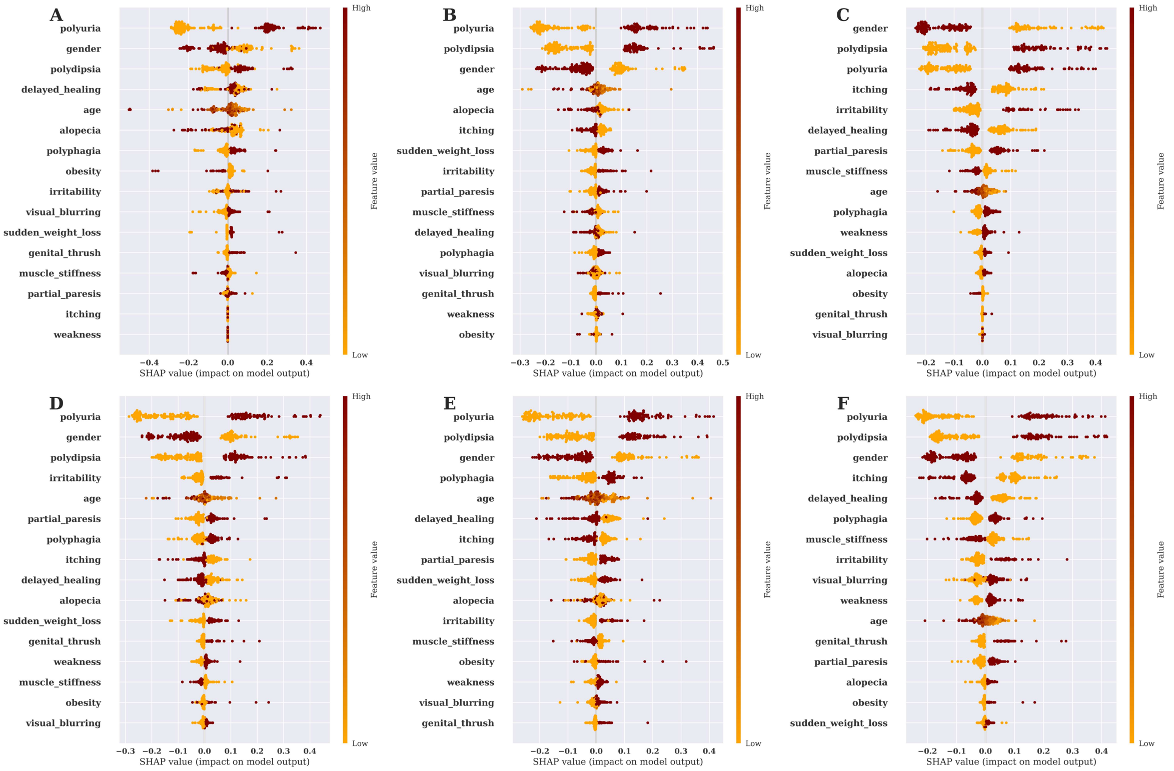

Figure 7. SHAP summary plot for features-impact on model prediction. (A) Feature's importance on model prediction by DT, (B) Feature's importance on model prediction by RF, (C) Feature's importance on model prediction by SVM, (D) Feature's importance on model prediction by XGBOOST, (E) Feature's importance on model prediction by LGBM, (F) Feature's importance on model prediction by MLP.

3.5. Discussion

Researchers have conducted a lot of research on diabetes prediction, but there is still room for improvement in diabetes prediction research. For predicting diabetes in this work, we employ a socio-demographic diabetes dataset. After collecting the dataset, we preprocessed it to make it suitable for further analysis. We have applied six supervised ML algorithms DT, RF, SVM, XGBOOST, LGBM, and MLP to predict diabetes. After applying the ML approaches, we assessed the results of the applied ML approaches utilizing different performance metrics like accuracy, precision, recall, f1-measure, sensitivity, specificity, kappa-statistics, and MCC. Among the applied ML algorithms, RF shows the highest result with 99.37% accuracy; 1.00 precision; 0.9902 recall; f1-score is 0.9951; sensitivity is 1.00; 0.9902 specificities; kappa-statistics and MCC 0.9859 and 0.9860 respectively for the train-test-split techniques, which effectively predicts diabetes. The same socio-demographic diabetes dataset that we analyze for diabetes prediction has been also analyzed by Islam, M. M., et al., (2020) and has been shown to have the highest result of 99.00% accuracy; for the RF approaches [4]. And Ahmed, Usama, et al., (2022) also used the same dataset and got 94.87% accuracy; 0.9552 sensitivity; 0.9438 specificities; f1-score is 0.9412 [9]. The impact of features on the model has an essential role in the ML field for any disease prediction. So, in this work, we also show the features-impact on model prediction of the six ML algorithms by utilizing the SHAP summary plot, which is graphically expressed in Figure 8.

Therefore, in this work, there are some limitations. First of all, the number of instances in this dataset is only 520, which is enough to build an ML-based prediction model but not good enough. As a result, we should collect more data in the future. The attributes of the diabetes dataset are only socio-demographic, but the socio-demographic data on diabetes are not sufficient to accurately predict diabetes. For that reason, we should collect clinical data in the future and merge them together to build an effective diabetes prediction model. Also, in the future, this study should be focused on utilizing more effective ML approaches to build an effective prediction model and develop an end-user website for diabetes prediction.

4. Conclusions

Diabetes is now one of the most alarming diseases in the world. Data mining and ML are now being used to predict diabetes alongside traditional clinical tests. Inspired by this, the study aimed to build an automated model to predict diabetes at an early stage. To fulfill the objective, six ML approaches were applied and compared to their performances to find the best-fit classifier that will predict diabetes based on a socio-demographic attribute in an early stage with significant accuracy. It has been observed that the best-fit classifier is RF with an accuracy of 99.37%. This study also aimed to find the feature's impact on model prediction, and that has been successfully done in this study. The proposed method will be further developed with more state-of-the-art technology with more data in the future. The findings will be more beneficial for researchers who have an interest in diabetes disease research based on ML techniques and will also be helpful for physicians to diagnose diabetes at very early stages.

Author Contributions: Md. Ashikur Rahman, Md. Mamun Ali, Imran Mahmud, Francis M. Bui, and Kawsar Ahmed: Provided the concept and performed the experiments; Wrote the paper; Analyzed and interpreted the data. Md. Ashikur Rahman, and Md. Mamun Ali: Interpreted the data; Md. Ashikur Rahman, and Md. Mamun Ali: Handling the manuscript and Analyzing data. Kawsar Ahmed and Md. Mamun Ali: Edited and reviewed the manuscript. Lway Faisal Abdulrazak: Funding the project, Francis M. Bui, Kawsar Ahmed, and Md. Mamun Ali: Designed the experiments and supervised the whole project.

Funding: This research work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Data Availability Statement: The corresponding author can provide the data that were utilized to support the study upon request.

Conflicts of Interest: No conflicts of interest are disclosed by the authors.

doi:10.20944/preprints202310.0109.v1

References

1. Roglic, Gojka. "WHO Global report on diabetes: A summary." International Journal of Noncommunicable Diseases 1.1 (2016): 3.
2. Balfe, Myles, et al. "What's distressing about having type 1 diabetes? A qualitative study of young adults' perspectives." BMC endocrine disorders 13 (2013): 1-14.
3. Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." ICT Express 7.4 (2021): 432-439.
4. Islam, M. M., et al. "Likelihood prediction of diabetes at early stage using data mining techniques." Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.
5. Krishnamoorthi, Raja, et al. "A novel diabetes healthcare disease prediction framework using machine learning techniques." Journal of Healthcare Engineering 2022 (2022).
6. Islam, Md Shafiqul, et al. "Advanced techniques for predicting the future progression of type 2 diabetes." IEEE Access 8 (2020): 120537-120547.
7. Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.
8. Fazakis, Nikos, et al. "Machine learning tools for long-term type 2 diabetes risk prediction." IEEE Access 9 (2021): 103737-103757.
9. Ahmed, Usama, et al. "Prediction of diabetes empowered with fused machine learning." IEEE Access 10 (2022): 8529-8538.
10. Maniruzzaman, Md, et al. "Classification and prediction of diabetes disease using machine learning paradigm." Health information science and systems 8.1 (2020): 1-14.
11. Barakat, Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat. "Intelligible support vector machines for diagnosis of diabetes mellitus." IEEE transactions on information technology in biomedicine 14.4 (2010): 1114-1120.
12. Dataset: Available online: https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification (accessed on 17 November 2022)
13. Sanni, Rachana R., and H. S. Guruprasad. "Analysis of performance metrics of heart failured patients using Python and machine learning algorithms." Global transitions proceedings 2.2 (2021): 233-237.
14. Silva, Fabrício R., et al. "Sensitivity and specificity of machine learning classifiers for glaucoma diagnosis using Spectral Domain OCT and standard automated perimetry." Arquivos brasileiros de oftalmologia 76 (2013): 170-174.
15. Chicco, Davide, Niklas Tötsch, and Giuseppe Jurman. "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation." BioData mining 14.1 (2021): 1-22.
16. Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC genomics 21.1 (2020): 1-13.
17. Erickson, Bradley J., and Felipe Kitamura. "Magician's corner: 9. Performance metrics for machine learning models." Radiology: Artificial Intelligence 3.3 (2021): e200126.
18. Mohamed, Amr E. "Comparative study of four supervised machine learning techniques for classification." International Journal of Applied 7.2 (2017): 1-15.
19. Priyam, Anuja, et al. "Comparative analysis of decision tree classification algorithms." International Journal of current engineering and technology 3.2 (2013): 334-337.
20. Azar, Ahmad Taher, et al. "A random forest classifier for lymph diseases." Computer methods and programs in biomedicine 113.2 (2014): 465-473.
21. Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." Shanghai archives of psychiatry 27.2 (2015): 130.
22. Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.
23. Zhang, Yongli. "Support vector machine classification algorithm and its application." Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3. Springer Berlin Heidelberg, 2012.
24. Ramraj, Santhanam, et al. "Experimenting XGBoost algorithm for prediction and classification of different datasets." International Journal of Control Theory and Applications 9.40 (2016): 651-662.

25. XGBoost Documentation : Available online: https://xgboost.readthedocs.io/en/stable/ (accessed on 24 December 2022)
26. Rufo, Derara Duba, et al. "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)." Diagnostics 11.9 (2021): 1714.
27. Abdurrahman, Muhammad Hafizh, Budhi Irawan, and Casi Setianingsih. "A review of light gradient boosting machine method for hate speech classification on twitter." 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE). IEEE, 2020.
28. Desai, Meha, and Manan Shah. "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)." Clinical eHealth 4 (2021): 1-11.
29. Maulidevi, Nur Ulfa, and Kridanto Surendro. "SMOTE-LOF for noise identification in imbalanced data classification." Journal of King Saud University-Computer and Information Sciences 34.6 (2022): 3413-3423.
30. Marcílio, Wilson E., and Danilo M. Eler. "From explanations to feature selection: assessing SHAP values as feature selection mechanism." 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI). Ieee, 2020.
31. Bowen, Dillon, and Lyle Ungar. "Generalized SHAP: Generating multiple types of explanations in machine learning." arXiv preprint arXiv:2006.07155 (2020).
32. Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.
33. Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.
34. Tan, Jimin, et al. "A critical look at the current train/test split in machine learning." arXiv preprint arXiv:2106.04525 (2021).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.