Available online at www.dergipark.org.tr/en

INTERNATIONAL ADVANCED RESEARCHES and ENGINEERING JOURNAL

Journal homepage: www.dergipark.org.tr/en/pub/iarej

e-ISSN: 2618-575X

International Open Access

Volume 07
Issue 01

April, 2023

Research Article

Early stage diabetes prediction using decision tree-based ensemble learning model

Ozge Sen Kaya [a], ***Sinem Bozkurt Keser*** [b,*] ***and Kemal Keskin*** [b]

[a]*Eskişehir Osmangazi University, Faculty of Engineering, Department of Computer Engineering, 26040, Eskişehir/Turkey*
[b]*Eskişehir Osmangazi University, Faculty of Engineering, Department of Electrical and Electronics Engineering, 26040, Eskişehir/Turkey*

ARTICLE INFO

Article history:
Received 14 October 2022
Accepted 09 April 2023
Published 15 April 2023

Keywords:
Bagging
Decision tree
Diabetes mellitus
Extra trees
Random forest

ABSTRACT

Diabetes is a lifelong disease that has undesirable effects on various organs, such as long-term organ damage, functional disorder, and finally failure of the organ. Diabetes must be treated under the supervision of a doctor. Diabetes is known as a disease that can be seen in many people today and is becoming widespread due to life conditions. If a person with diabetes does not receive any treatment at an early stage, the patient's body can react with serious complications. In addition to the medical methods used in the diagnosis of diabetes, this disease can be detected by an artificial intelligence approach. This research aims to establish the most influential variable among the many variables causing diabetes and to design a model that will predict diabetes to help doctors analyze the disease with selected machine learning methods. In this study, Decision Tree, Bagging with Decision Tree, Random Forest and Extra Tree algorithms were used for the proposed model and the highest accuracy values were obtained with the Extra Trees algorithm with 99.2%.

1. Introduction

Diabetes mellitus refers to a metabolic disease reasoned by persistent hyperglycaemia. Deficiency of insulin secretion or irregularities in insulin activity or both of them is the primary basis for the hyperglycaemia [1]. Chronic diabetes mellitus (DM) constitutes undesirable effects on various organs such as long-term organ damage, function disorder, and finally failure of the organ. Specifically, eyes, nerves, heart, kidneys, and blood vessels are the most affected ones [2]. DM can be classified into three main aetiological types: type 1, type 2, and other specific types. These types differ from each other considering defects, disorders, or processes along with diabetes mellitus. The pancreas contains β-cells, which produce and release insulin in response to blood glucose levels. Type 1 diabetes, namely, insulin-dependent diabetes, arise from autoimmune extermination of the β-cells of the pancreas. The rate of disruption of β-cells is quite variable according to age, sex, genetic factors, lifestyle and diet, such that it proceeds faster in some people (primarily children), and sluggish in others (especially adults). In the latter phase of this type of diabetes, insufficient (almost negligible) insulin is secreted. Patients need insulin therapy throughout their lifetime. Only 5-10 percent of diabetics have this form of diabetes. Type 2 diabetes involves people with insulin resistance and a lack of insulin production. Released Insulin is not enough to compensate for insulin resistance. However, the patients do not require insulin therapy to survive. This form of diabetes accounts for 90-95 percent of diabetics and has mostly been seen in obese people because obesity has a destructive effect on insulin levels. Increased neogenesis was the mechanism through which the relative volume of β-cells was higher in obese than lean nondiabetic individuals [2]. Other types of diabetes mellitus have rarely been seen when compared to type 1 and type 2 diabetes and occur under specific conditions such as genetic defects and pregnancy.

The number of diabetics worldwide has reached 422 million (approximately 8.5 percent of the world population) in 2014 and it has doubled since 1980 [3]. The research on the prevalence of diabetes mellitus for future projections shows that there will be 693 million people with diabetes by 2045 [4]. The growing rate of prevalence is really fast, especially in impoverished countries. When one considers that half of the diabetics are undiagnosed, it

* Corresponding author. Tel.: *+90-222-239-3750*.
E-mail addresses: ozgeesenn1995@gmail.com *(O. Sen Kaya)*, sbozkurt@ogu.edu.tr *(S. Bozkurt Keser)*, kkeskin@ogu.edu.tr. *(K. Keskin)*
ORCID 0000-0002-4713-7536 *(O. Sen Kaya)*, 0000-0002-8013-6922 *(S. Bozkurt Keser)*, 0000-0002-3969-2396 *(K. Keskin)*
DOI: 10.35860/iarej.1188039

becomes difficult to cope with the disease and the death rate is increased. Diagnosing diabetes in time with precision is of great importance. Therefore, it can prevent diabetes-related deaths, improve the quality of life, and reduces the economic burden caused by the disease. Machine learning is a research approach that focuses on providing computers with the ability to recognize complex patterns and extrapolate knowledge from. It has frequently been used in recent years to diagnose or help diagnose diseases. In the context of rapidly increasing diabetics, the classification of patients according to current groups will allow them to start treatment on time and will eliminate some of the problems caused by delays.

Machine learning is a powerful tool for predicting and diagnosing diabetes. It includes the analysis of big data sets, the identification of patterns, and the generation of predictions based on these patterns utilizing algorithms and statistical models. Machine learning can evaluate medical data associated with diabetes, such as blood glucose levels, blood pressure, and body mass index, to predict a person's risk of getting the disease. Compared to more traditional methods, it can offer accurate predictions. A large amount of data may be analyzed by machine learning algorithms, and these algorithms are capable of identifying small patterns. As a result, individuals may see better results from their diabetes treatment and diagnosis. Machine learning has gained importance in the health sector compared to other methods due to reasons such as easy to use and fast. Machine learning methods are used in disease diagnosis and in different studies by making use of data sets obtained in diseases. These studies have achieved some research of algorithm comparison and model establishing for DM prediction. The data validity and prediction accuracy, however, were not good enough for actual use. To increase accuracy, we must provide a novel prediction model. Therefore, we chose a dataset to test the usability and adaptation of our model. The main objectives of this study are to predict diabetes at an early stage so that patients may begin the right treatments on time, and to discover the correlations among the variables that contribute to diabetes. Finally, this research will help us to discover the best machine learning classifier to predict diabetes.

2. Related Works

Nowadays, the use of machine learning algorithms in medical diagnosis, including type 2 diabetes mellitus (T2DM), is rapidly increasing [1]. The process from expert medical diagnosis to evaluation and decision making is the key factor here [2]. Several classification algorithms are utilized to predict T2DM in the early stage. The compound of ant colony optimization (ACO) and fuzzy logic is proposed in [3] to diagnose diabetics by utilizing the public Pima Indian Diabetes Database (PIDD) which is existing at the University of California, Irvine (UCI) machine learning repository. In the experiments, 84.24% classification accuracy is obtained with the proposed method. Karegowda et. al. [4] compare Neural Networks Back Propagation Networks (BPN), Genetic Algorithms (GA), and DT algorithms to diagnose T2DM. The proposed GA-correlation-based feature selection approach results with 84.71% accuracy in the experiments. The classification technique based on the Gaussian process (GP) has been adopted in linear, polynomial, and radial-based kernel [5]. The performance of the GP-based classification method is compared with LDA, QDA, and NB by utilizing PIDD in the experiments. 81.97% accuracy performance is obtained with the GP-based model which is larger compared to other methods. In [6], NB, SVM, Random Forest (RF), and Simple Classification and Regression Tree (CART) supervised learning algorithms are compared to recommend the best approach based on efficient performance results for the prediction of T2DM. Experimental results of each algorithm used on the PIDD demonstrate that SVM performed best in the prediction of diabetes disease having 79.13% accuracy. Three machine learning classification algorithms namely DT, SVM, and NB are used to construct a model [7]. The experiments are performed on PIDD to evaluate the performance of the algorithms. According to test results, NB outperforms other algorithms to prognosticate the likelihood of T2DM in patients with 76.30% accuracy. To construct an adaptive model with better accuracy, the k-means clustering algorithm is enhanced with Logistic Regression (LR) [8]. In the experiments, 95.42% accuracy is obtained with the proposed algorithm using 10-fold cross-validation. DT, RF, and Artificial Neural Network (ANN) are used to predict T2DM [9]. In the experiments, the dataset is including 14 attributes collected from the hospital in Luzhou, China during the physical examination. The attributes are 14 physical examination indexes such as age, pulse rate, breathe, left systolic pressure (LSP), right systolic pressure (RSP), left diastolic pressure (LDP), right diastolic pressure (RDP), height, weight, physique index, fasting glucose, waistline, low density lipoprotein (LDL), and high density lipoprotein (HDL). Principal Component Analysis (PCA) and minimum redundancy maximum relevance (mRMR) are utilized to decrease the dimension of the dataset. Five-fold cross-validation is used to evaluate the algorithms, and RF performs better with 80.84% accuracy. ANN, RF, and K-means clustering techniques are applied for the estimation of T2DM using PIDD. In experiments, the ANN algorithm outperforms the other models with an accuracy of 75.70%, and by using association rule mining, the results have shown that there is a powerful connection between BMI and glucose with diabetes [10]. Four machine learning classifiers (LR, MLP, SVM, and RF) are evaluated on a dataset including an aggregate of 5319 cases and 36,224 controls. The dataset contains a total of 116 attributes with 18 demographic, 12

medical, and 86 dental attributes of 40,519 patients. In the experiments, RF was superior to other predictive models providing overall accuracy (94.14%) [11]. Leverage F-Score Feature Selection and Fuzzy SVM are used to predict T2DM using PIDD [12]. The fuzzy SVM algorithm gives 89.02% promising accuracy for predicting patients with T2DM. Juliet and Bhavadharani have been discussed the role of Naïve Bayes (NB), Decision Tree (DT), K-Star, LR, Support Vector Machine (SVM) methods for classifying Type 2 Diabetes Mellitus by using PIDD [13]. In the experiments, Logistic Regression provides the best accuracy with 77.73%. The missing values and class imbalance problems of PIDD are handled with NB, and the Adaptive synthetic sampling method (ADASYN), respectively. Then, an RF algorithm is performed to diagnose T2DM. In the experiments, RF outperforms NB, SVM, and DT with 87.10% accuracy using 10-fold cross-validation [14]. In [15], K-Nearest Neighbor (KNN) algorithm is developed by removing noise, decreasing the dimension, and weighting distance with k-means clustering (KMC), principal component analysis (PCA), and autoencoder (AE), respectively. In the experiments, KMC improves the accuracy of KNN from 81.6% to 86.7%, combining KMC and PCA improves the KNN accuracy to be 90.9%, combining KMC and AE enhances the KNN to gives an accuracy of 97.8%, KMC, PCA, and Weighted KNN (WKNN) increases the accuracy to be 94.5%, and finally, the combination of KMC, AE, and WKNN achieves the best accuracy of 98.3%. Since the attributes in PIDD have a high non-linearity; AE gives higher accuracies than PCA. LR, KNN, DT, RF, SVM, NB, ANN, and Gradient Boosting (GB) are compared using clinical data obtained from the Dryad Digital Repository [16]. The clinical data contains age, gender, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), smoking and drinking status, family history of diabetes, alanine aminotransferase (ALT), fasting plasma glucose (FPG), total cholesterol (TC), low density lipoprotein (LDL), high density lipoprotein cholesterol (HDL-C), triglyceride (TG), year of follow up. GB outperforms other algorithms with 95.50% accuracy.

Table 1. A summary of research work for diagnoses of T2DM using machine learning algorithms (NA: Not Available)

Authors	Methodology	Dataset	Tool	Best Results
Ganji and Abadeh, 2011[3]	ACO and fuzzy logic	PIDD	Weka 3	84.24% accuracy
Karegowda et. al., 2011 [4]	GA-correlation based feature selection	PIDD	Python	84.71% accuracy
Maniruzzaman et. al., 2017 [5]	GP-based classification	PIDD	NA	81.97% accuracy
Mir et. al., 2018 [6]	NB, SVM, RF, and Simple CART	PIDD	Weka 3.82	79.13% accuracy with SVM
Sisodia et. al., 2018[7]	DT, SVM, and NB	PIDD	Weka	76.30% accuracy with NB
Wu et. al., 2018 [8]	K-means clustering algorithm is enhanced with LR	PIDD	Weka	95.42% accuracy
Zou et. al., 2018 [9]	DT, RF, and ANN	Dataset is including 14 attributes collected from the hospital in Luzhou, China	Weka, Java, Matlab	80.84% accuracy with RF
Alam et. al., 2019 [10]	ANN, RF, and K-means clustering	PIDD	NA	75.70% accuracy with ANN
Hegde et. al., 2019 [11]	LR, MLP, SVM, and RF	Dataset is retrieved from Marshfield Clinic Health System's data-warehouse	Weka	94.14% accuracy with RF
Lukmanto et. al., 2019 [12]	Fuzzy SVM	PIDD	NA	89.02% accuracy
Juliet and Bhavadharani, 2019 [13]	NB, DT, K-Star, LR, SVM	PIDD	NA	77.73% accuracy with LR
Wang et. al., 2019 [14]	NB, ADASYN, RF, DMP_MI	PIDD	Python	87.10% accuracy
Khairunnisa et. al., 2020 [15]	KMC, PCA, WKNN, AE	PIDD	NA	98.3% accuracy
Tarokh and Darabi, 2020 [16]	LR,NN,DT,RF,SVM,NB,GB	The dataset is retrieved from 32 health care centers in 11 provinces in China	Python	95.50% accuracy with GB
Gupta et.al.,2020 [17]	MLP, GP, LDA, QDA, SGD, RRC, SVM, KNN, DT, NB, LR, RF, ELM, RBF	PIDD, DCA	NA	94.59% accuracy for DCA, and 79.22% for PIDD

For the diagnosis of T2DM, two real-world datasets diabetic clinical dataset (DCA) and PIDD are evaluated by using 15 different classifiers (MLP, GP, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Statistical Gradient Descent (SGD), Ridge Regression Classifier (RRC), SVM, KNN, DT, NB, LR, RF, and Extreme Learning Machine (ELM) for multiquadric, Radial Basis Function (RBF), sigmoid activation function with 10-fold cross-validation [17]. The DCA dataset is collected from a medical expert in the Indian state of Assam between 2017 and 2018. The experimental results show that LR yields better than other techniques (94.59% accuracy for DCA, and 79.22% for PIDD). Table includes several machine learning algorithms that different researchers have used to make comparisons along with dataset, tool, and outcome.

Ensemble learning (EL) algorithms are applied to improve the performance of the classification algorithm and solve the problems of unstable classification. The aim of the EL is to aggregate multiple versions of base classifiers to construct the final prediction. In the field of medical diagnosis, various studies are published to enhance the performance of the prediction model by using EL algorithms. NB, KNN, and DT are utilized with EL algorithms such as bagging and boosting to predict T2DM patients using 10-fold cross-validation [18]. In the experiments, the dataset is collected from 27 Primary Care Units (PCU) in Sawanpracharak Regional Hospital between 2011 and 2013. The dataset contains a total of 48,763 instances with 20,743 diabetes and 28,020 non-diabetes, and 15 input attributes and 1 output attribute. 95.31% accuracy is obtained from bagging with a DT classifier that is superior to other algorithms. To improve the performance of the classification algorithm, voting with KNN, SVM, LR, and stacking with RF, AdaBoost (AB), LR are applied using PIDD [19]. Best accuracy results are obtained as 80.08% with stacking in the experiments. The results show that ensemble classifier models performed better than the basic classifiers alone. The ensembling of two boosting classifiers such as AB and XB gives the best combination for predicting T2DM [20]. By using PIDD, 95.00% Area Under Curve (AUC) values were obtained that are better than other algorithms in the experiments. Besides, it is proven that by applying preprocessing steps such as outlier rejection filling missing values the quality of the PIDD can be improved. LR regularised generalized linear model (Glmnet) with Least Absolute Shrinkage and Selection Operator (Lasso) regression (L1), Random Forests (RF), eXtreme Gradient Boosting (XGBoost) with tree booster using regression tree as a base classifier and Light Gradient Boosting Machine (LightGBM) with L1 loss regression are evaluated in early prediction of T2DM [21]. It is found that LightGBM results in much more stable results compared to other algorithms. In a study, an EL algorithm with base classifiers Linear Discriminant Analysis (LDA), SVM, and RF is applied on National Health and Nutrition Examination Survey (NHANES) database to predict T2DM patients. In the experiments, NHANES database including 8057 instances and 12 attributes, 74.50% best accuracy is obtained with EL algorithm with LDA. Several ensemble learning techniques have been proposed in the literature for the diagnosis of T2DM summarized in Table 2.

3. Methodology

The flow chart of the proposed approach is shown in Figure 1. The initial step in the proposed approach is to gather publicly available data sets on diabetic symptoms. In the second stage, the data preprocessing is performed. Furthermore, in the third stage, data is split into training dataset (80%) and test dataset (20%). The training dataset is utilized to train the models (Decision Tree (DT), Bagging with Decision Tree, Random Forest (RF), and Extra Trees) and testing dataset is used to test the models in terms of various performance metrics (accuracy, precision, recall, and F1 score,). After comparing the models in terms of accuracy, the efficient model with the highest accuracy for the diagnose of the diabetes is determined. Finally, the outcomes of the patients are predicted according to this model.

3.1 Dataset

The dataset used in this study is utilized to predict early diabetes mellitus from the open-source machine learning repository UCI. The dataset was obtained by questionnaire from 520 patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh, and were confirmed by doctors. Dataset consists of 17 features such as Age, Sex, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital Thrush, Visual Blurring, Itching, Irritability, Delayed Healing, Partial Paresis, Muscle Stiffness, Alopecia, Obesity, and Class. These features are shown in detail in Table 3.

In this dataset, the first parameter is the age parameter as seen in the Table 3. The diabetes prevalence worldwide has increased from %4.7 to %8.5 of the population [22]. According to the IDF Diabetes Atlas estimates, 1 in 11 adults has diabetes in 2015. These adults are 16 years of age and above 16 years of age. The second parameter is sex. Studies show that middle-aged men have a higher prevalence of diabetes than women of similar age, and conversely, the prevalence is higher in older women than men. Other parameters are important parameters in the diagnosis of diabetes: polyuria, excessive urination; polydipsia, excessive thirst; sudden weight loss, weakness, polyphagia which is associated with hunger, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis or muscular stiffness, alopecia which is point baldness due to hair loss, obesity which is excessive or abnormal fat mass which results to health risks.

Table 2. A summary of research work for diagnoses of T2DM using ensemble learning algorithms (NA: Not Available)

Authors	Methodology	Dataset	Tool	Best Results
Nai-arun et. al.,2014 [18]	NB, KNN, DT, Bagging and boosting	Dataset is collected from 27 Primary Care Units (PCU) in Sawanpracharak Regional Hospital during 2011-2013	NA	%95.31 accuracy with bagging and base classifier DT
Patil et. al., 2019 [19]	KNN, SVM, LR, RF, AB, LR	PIDD	Python	80.08% accuracy
Hasan et. al., 2020 [20]	KNN, DT, RF, AB, NB, XB	PIDD	Python	95.00 % AUC
Kopitar et. al., 2020 [21]	LR,RF,XGBoost, LightGBM	The dataset is collected from the NHANES database including 8057 instances and 12 attributes	NA	74.50% accuracy with EL algorithm with LDA

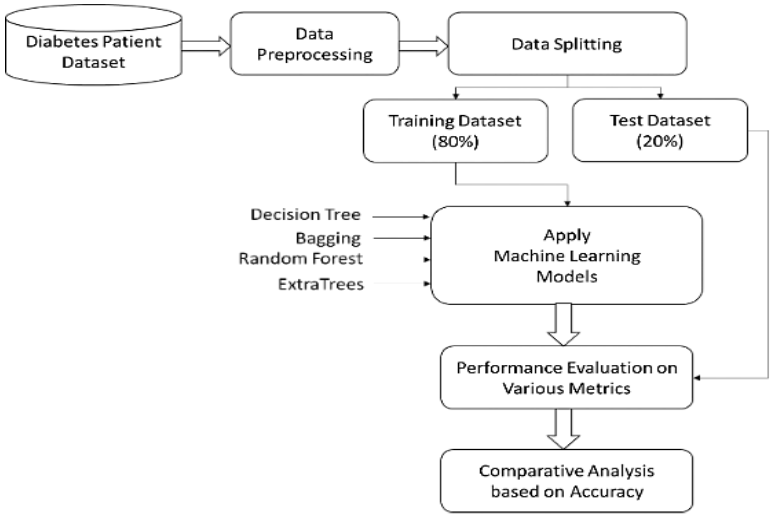

Figure 1. The proposed model

3.2 Data Preprocessing

Preprocessing was done to analyze the features in the dataset and to make the classification more efficient. Although there are no null values in the dataset, there are categorical values. The categorical values were converted to numerical values. There are 38.46% (1) and 61.54% negative (0) classes in the dataset. The scatter plots of each feature over the classes are given in Figure 2.

These scatter plots represent the distribution of diabetes disease according to the parameters. The data in the dataset according to the parameters were visualized with these plots. When the dataset examined, it is seen that there is an unbalanced distribution problem. Due to the unbalanced distribution problem in the dataset, The SMOTE approach was used. The SMOTE method is regarded as the most popular and frequently the most effective sampling technique [23]. Numerous imbalanced dataset issues have been addressed using the technique created in 2002. This technique differs from other approaches in that it creates artificial instances based on the k nearest neighbors of the instances under examination, as opposed to replicating the minority class data.

Table 3. Feature Details

Features Name	Features Type	Data Type	Possible Value
Age	Predictive	Integer	16-90
Sex	Predictive	Object	Male, Female
Polyuria	Predictive	Object	Yes, No
Sudden Weight Loss	Predictive	Object	Yes, No
Weakness	Predictive	Object	Yes, No
Polyphagia	Predictive	Object	Yes, No
Genital Thrush	Predictive	Object	Yes, No
Visual Blurring	Predictive	Object	Yes, No
Itching	Predictive	Object	Yes, No
Irritability	Predictive	Object	Yes, No
Delayed Healing	Predictive	Object	Yes, No
Partial Paresis	Predictive	Object	Yes, No
Muscle Stiffness	Predictive	Object	Yes, No
Alopecia	Predictive	Object	Yes, No
Obesity	Predictive	Object	Yes, No
Class	Responsive	Object	Positive, Negative

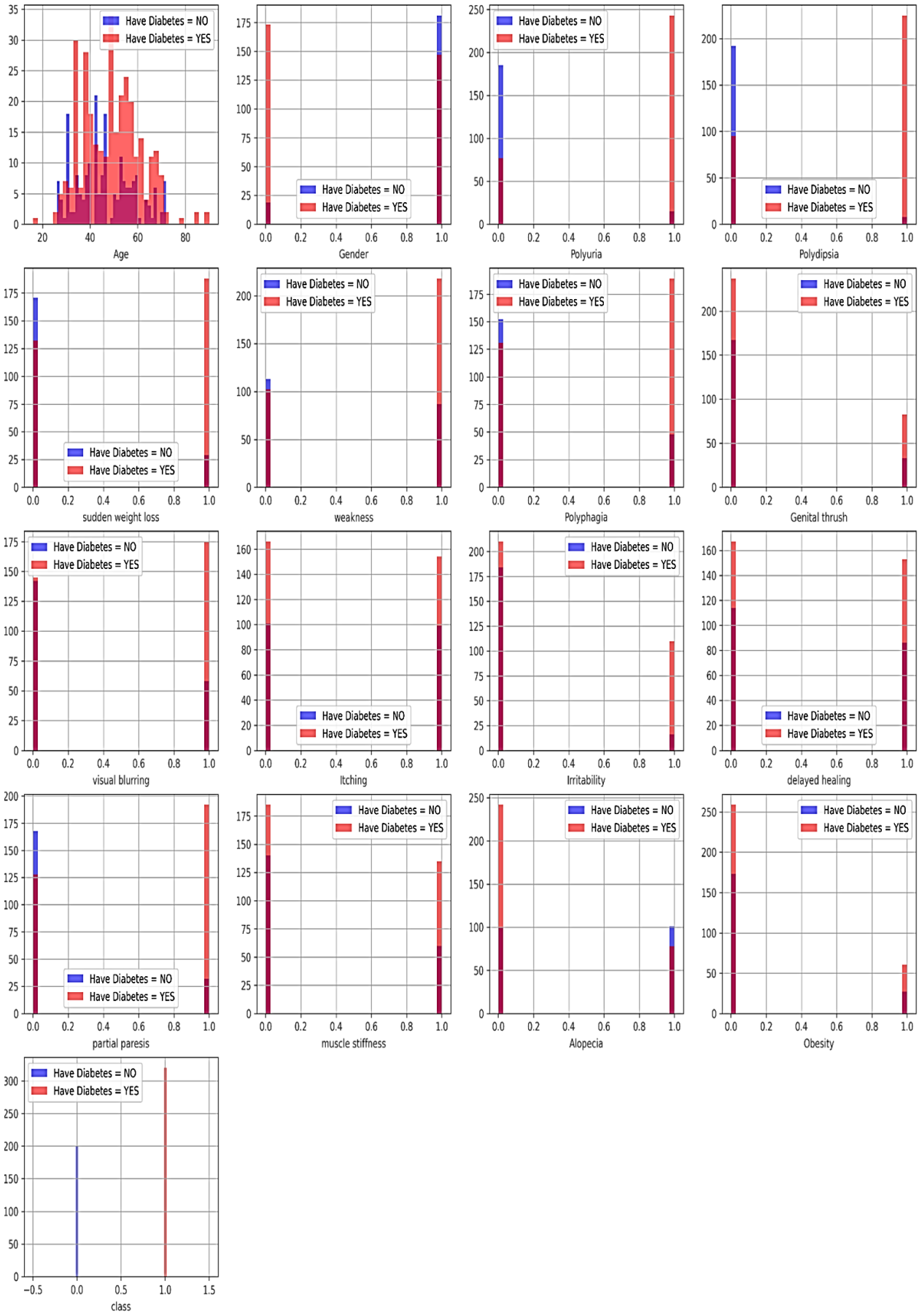

Figure 2. The scatter plots of each feature

Following is a brief summary of how the algorithm operates:

1. The k-closest neighbors of each observation pertaining to the minority class are searched,
2. The difference between the observation of the minority class and the observation with the k nearest neighbors (K-Nearest Neighbors) is taken,
3. The difference determined in Step 2 is multiplied by a random integer (α) is selected at random from (0,1),
4. Using the Equation (1), a new synthetic observation is produced:

$$x_{yeni} = xi + (xj - xi) * \alpha \quad (1)$$

5. Repeat steps 1-4 to obtain the desired number of synthetic observations.

3.3 Machine Learning Algorithms

3.3.1 Decision Tree

The first algorithm used in prediction of the diabetes mellitus is the decision tree algorithm [24]. The basic idea in the decision tree algorithm, which is a powerful classification algorithm among machine learning algorithms, is based on repeatedly dividing the input data into groups with the help of a classification algorithm. As a result of this division, nodes are formed and these nodes are labeled. The division continues in depth until all nodes of the group have the same class label. The advantage of this algorithm is that it is used very frequently and it is an algorithm that can be easily interpreted and created easily. Decision trees, which are one of the supervised learning methods, consist of roots, nodes, branches and leaves like a tree. There are some criteria used in feature selection, that is, in determining which node will be selected. These criteria are; information gain, gain ratio and gini index.

3.3.2 Bagging with Decision Tree

The bagging method is one of the ensemble learning methods derived by the award-winning statistician Breiman in 1996. In bagging, the dataset is distributed across several bootstrap copies. Each replica is plotted with replacement, regardless of the original dataset; on average, each copy contains 63.2% of the original data. The process is accomplished by repeatedly running the weak learner on various bootstraps. At each iteration, the classifier learned from the weak learner is combined into the strong composite classifier to achieve higher accuracy than any single component classifier can do alone [25]. In this study, a bagging algorithm is utilized to transform weak learners into strong learners.

	Predicted Class Label	
Real Class Label	TP	TN
	FP	FN

3.3.3 Random Forest

Random Forest, another most commonly used machine learning technique, is a combination of the Bagging method developed by Breiman in 1996 and the Random Subspace method developed by Kim Ho [26]. This technique combines multiple classifiers to improve performance. In different subsets of the given dataset, the Random Forest classifier consists of decision trees. Each decision tree output is averaged to improve the prediction accuracy in the given dataset. Random Forest takes predictions from multiple trees, calculates the maximum number of predictions, and predicts the final output. In the Random Forest algorithm, random data points are first selected from the training data set. A subset of the decision tree associated with this selected point is developed. The first two steps are applied again by determining the number of the decision tree to be created. Finally, at the given data points, each decision tree is estimated and new data points are assigned to the category that wins the majority vote.

3.3.4 Extra Trees

The Extra Tree algorithm, which consists of a collection of decision trees, is referred to collections of other decision tree algorithm collections as bootstrapping (bagging) and random forest [27]. The training dataset is used to create a large number of unpruned decision trees as part of the Extra Trees technique. When utilizing regression or classification, estimates are performed by averaging the decision tree estimation or by employing majority vote. The Extra Trees approach fits each decision tree on the whole training dataset, in contrast to bagging and random forest, which only use a bootstrap sample of the training dataset to construct each decision tree. With regard to calculation time, the Extra Trees approach is faster.

3.4 Performance Evaluation

In machine learning, the confusion matrix is used to show the relationship between real class labels and predicted class labels. It is a visualization tool used to show the accuracy values obtained as a result of the classification process. It is used to represent true positive (TP), correct negative (TN), false positive (FP), false negative (FN).

• True Positive (TP): It indicates that the patient has diabetes.

• True Negative (TN): It indicates that the patient does not have diabetes.

• False Positive (FP): It indicates that a person who does not have diabetes has been misdiagnosed with diabetes.

• False Negative (FN): It indicates that a person with diabetes is misdiagnosed as not having diabetes.

Table 4. **C**omparison of algorithms using 10-CV

	ACC	PREC	RC	FS
DT	%97.35	%97.79	%96.87	%97.33
Bagging with Decision Tree	%97.50	%98.41	%96.56	%97.48
RF	%98.59	%98.75	%98.44	%98.59
Extra Trees	**%98.91**	**%98.75**	**%99.06**	**%98.91**

* ACC: Accuracy, PREC: Precision, RC: Recall, FS: F1-Score

Table 5. Confusion matrix results of classification algorithms

DT

Real-Class Label	Predicted Class Label: **0**	**1**
0	61	1
1	1	65

Bagging with Decision Tree

Real-Class bel	Predicted Class Label: **0**	**1**
0	61	1
1	2	64

RF

Real-Class abel	Predicted Class Label: **0**	**1**
0	62	0
1	2	64

Extra Trees

Real-Class Label	Predicted Class Label: **0**	**1**
0	62	0
1	1	65

Table 6. Overall performance comparison of algorithms

	ACC (Training)	ACC (Test)	PREC (Test)	RC (Test)	FS (Test)
DT	%100.00	%98.44	%98.00	%98.00	%98.00
Bagging with Decision Tree	%99.61	%97.66	%97.00	%98.00	%98.00
RF	%100.00	%98.44	%97.00	%100.00	%98.00
Extra Trees	**%100.00**	**%99.22**	**%98.00**	**%100.00**	**%99.00**

* ACC: Accuracy, PREC: Precision, RC: Recall, FS: F1-Score

Table 7. Evaluation of related studies

References	The Best Algorithm	The Best Result	The Difference with Our Work
Zou et. al., 2018	Random Forest	%80.84	+18.36%
Nai-arun et. al.,2014	Bagging and base classifier DT	%95.31	+3.89%
Hegde et. al., 2019	Random Forest	%94.14	+5.06%
Başer et.al, 2021[28]	Random Forest	%84,78	+14.42%

Using these values in the confusion matrix given in the table, the following metrics used to determine the performance of the classification algorithm are calculated.

Measures	Definitions	Formula
Accuracy	Refers to the ratio of the number of correctly classified samples to the total number of samples	$\frac{TP+TN}{\text{Total no of samples}}$
Precision (P)	Classifiers correctness/accuracy is measured by Precision	$\frac{TP}{TP+FP}$
Recall (R)	Refers to the power of the classification algorithm to correctly predict positive samples.	$\frac{TP}{TP+FN}$
F1-Score	It is the geometric mean of the sensitivity and recall criteria and allows the values of both criteria to be considered together.	$\frac{2*(P*R)}{P+R}$

4. Test Results

In this section, we discuss the experimental test and evaluate the machine learning algorithms to predict diabetes mellitus. These tests aim to estimate whether the patient has diabetes mellitus or not. In addition, the sckit-learn library, which is a popular library containing functions of algorithms developed in the field of machine learning and data mining developed with Python programming language, has been utilized. The tests were performed on Windows 10 operating system, using Python version 3.8.10. All tests experiments

were performed on a computer with 12GB of RAM (Intel ® Core™ i7-4700HQ CPU @ 2.40Ghz 2.40GHz).

The problem in this study was considered as a classification problem and Decision Tree (DT), Bagging with DT, Random Forest (RF), and Extra Trees algorithms were used in the classification stage. Firstly, 10-CV was utilized in order to demonstrate the performance of the algorithms. The results of 10-CV are given in Table 4.

According to the proposed system, the dataset was split into 80% train and 20% test as given in Figure 1. Then, the confusion matrix obtained as a result of the tests performed with machine learning algorithms is shown in Table 5. In confusion matrices given Table 5, 0 represents cases without diabetes mellitus and 1 represents cases with diabetes mellitus. It is seen that the Extra Trees algorithm is more successful than other algorithms that predict diabetes mellitus when Table 5 is examined. The performance measures calculated using the confusion matrix are given in Table 6.

When the results given in Table 6 are examined, it is seen that the Extra Trees algorithm performs the best prediction with an accuracy of 99.22%. The DF algorithm has an accuracy of 98.44%, the Bagging algorithm has an accuracy of 97.66% and the RF algorithm has an accuracy of %98.44. We observed that decision tree-based classifiers give good results for this dataset. Because the features in the dataset are in the logic of a decision mechanism.

For data with unbalanced class distribution, it would be more accurate to use the f1-score performance metric, where precision and recall metrics are considered together, instead of accuracy. Accordingly, considering the values in Table 5, it is seen that better f1-score values are obtained by applying the Extra Trees algorithm. When the algorithm is examined in terms of processing times, it is seen that there is not much difference between them. In Table 7, we compare our work with the literature in order to demonstrate the efficiency of the proposed approach.

5. Conclusions

Diabetes is one of the diseases that can have undesirable effects on various organs and last throughout life. Diabetes, which is a disease that should be treated under the supervision of a doctor, can lead to serious complications in the body if not treated. In addition to the medical methods used in the diagnosis of diabetes, this disease can be detected by artificial intelligence approached. With these approaches, experts are assisted in the diagnosis of the disease. This paper aims to predict the diabetes. First of all, a data preprocessing was done on the dataset obtained from UCI. SMOTE algorithm was used during this preprocessing. Then, different machine learning algorithms were used to predict diabetes. In this paper, using Decision Tree, Bagging and Decision Tree, Random Forest and Extra Trees algorithms, the most successful accuracy rate was obtained from the Extra Trees algorithm with 99.22%. In future studies, it is aimed to detect diabetes by using deep learning on the dataset and machine learning methods not used in this study. In future studies, it is aimed to realize an artificial intelligence-based decision support system that will significantly help experts with big data technologies, deep learning and transfer learning approaches in the diagnosis of diabetes.

Declaration

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author(s) also declared that this article is original, was prepared in accordance with international publication and research ethics, and ethical committee permission or any special permission is not required.

Author Contributions

O. Sen Kaya performed the literature research, analysis of the methodology, preparation of algorithm figures, examination of the test results, writing the article. S. Bozkurt Keser performed the literature research, design and implementation of the research, analysis of the results, organizing and writing the article. K. Keskin contributed to organizing, writing, and proofreading the article.

References

1. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I., Machine learning, and data mining methods in diabetes research. Computational and structural biotechnology journal, 2017. **15**: p. 104-116.
2. Choubey, D.K., Paul, S., and Bhattacharjee, J., Soft computing approaches for diabetes disease diagnosis: a survey. International Journal of Applied Engineering Research, 2014. **9**(21): p. 11715-11726.
3. Ganji, M.F. and Abadeh, M.S., A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications, 2011. **38**(12): p. 14650-14659.
4. Karegowda, A.G., Manjunath, A., and Jayaram, M., Application of genetic algorithm optimized neural network connection weights for medical diagnosis of Pima Indians diabetes. International Journal on Soft Computing, 2011. **2**(2): p. 15-23.
5. Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., and Suri, J. S., Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer methods and programs in biomedicine, 2017. **152**: p. 23-34.
6. Mir, A. and Dhage, S.N., Diabetes disease prediction using machine learning on big data of healthcare. in 2018 fourth international conference on computing communication control and automation (ICCUBEA). 2018. IEEE.
7. Sisodia, D. and Sisodia, D. S., Prediction of diabetes using classification algorithms. Procedia computer science, 2018. **132**: p. 1578-1585.

8. Wu, H., Yang, S., Huang, Z., He, J., and Wang, X., Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 2018. **10**: p. 100-107.
9. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H., Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 2018. **9**: p. 515.
10. Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., and Abbas, Z., A model for early prediction of diabetes. Informatics in Medicine Unlocked, 2019. **16**: p. 100204.
11. Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., and Acharya, A., Development of non-invasive diabetes risk prediction models as decision support tools designed for application in the dental clinical environment. Informatics in medicine unlocked, 2019. **17**: p. 100254.
12. Lukmanto, R. B., Nugroho, A., and Akbar, H., Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. Procedia Computer Science, 2019. **157**: p. 46-54.
13. Juliet, M.P.L. and T. Bhavadharani, An improved prediction model for type 2 diabetes mellitus disease using clustering and classification algorithms. International Research Journal of Engineering and Technology (IRJET), **6**(2): p. 1179-1186.
14. Wang, Q., Cao, W., Guo, J., Ren, J., Cheng, Y., and Davis, D. N., DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data With missing values. IEEE Access, 2019. **7**: p. 102232-102238.
15. Khairunnisa, S., Suyanto, S., and Yunanto, P. E. Removing Noise, Reducing dimension, and Weighting Distance to Enhance k-Nearest Neighbors for Diabetes Classification. in 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). 2020. IEEE.
16. Tarokh, M.J., Type 2 Diabetes Prediction Using Machine Learning Algorithms. Jorjani Biomedicine Journal, 2020. **8**(3): p. 4-18.
17. Gupta, D., Choudhury, A., Gupta, U., Singh, P., and Prasad, M., Computational approach to clinical diagnosis of diabetes disease: a comparative study. Multimedia Tools and Applications, 2021: p. 1-26.
18. Nai-Arun, N., and Sittidech, P., Ensemble learning model for diabetes classification. in Advanced Materials Research. 2014. Trans Tech Publ.
19. Patil, M. K., Sawarkar, S. D., and Narwane, M. S. Narwane, Designing a Model to Detect Diabetes using Machine Learning. Int. J. Eng. Res. Technol, **8**(11), p: 333-340
20. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., and Hasan, M., Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 2020. **8**: p. 76516-76531.
21. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., and Stiglic, G., Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports, 2020. **10**(1): p. 1-12.
22. Gamara, R. P. C., Bandala, A. A., Loresco, P. J. M., and Vicerra, R. R. P., Early stage diabetes likelihood prediction using artificial neural networks. in 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). 2020, IEEE.
23. Hu, F., Li, H., A novel boundary oversampling algorithm based on neighborhood rough set model: NRSBoundary-SMOTE. Mathematical Problems in Engineering, 2013.
24. Quinlan, J. R., Induction of decision trees, Machine Learning, **1**, p: 81-106, 1986.
25. Perveen, S., Shahbaz, M., Guergachi, A., and Keshavjee, K., Performance analysis of data mining classification techniques to predict diabetes. ScienceDirect, 2016. **82**: 115-121.
26. Breiman, L., 2001. Random forests. Machine Learning, **45**(1): p. 5-32, 2001.
27. Geurts, P., Ernst, D., and Wehenkel, L., Extremely Randomized Trees, Machine Learning, **63**(1), p. 3-42, 2006.
28. Başer, B. Ö., Yangın, M., and Sarıdaş, E. S., Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması. Journal of Natural & Applied Sciences, **25**(1), 2021.