Ann Med Res

Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

A comparison of multivariate statistical methods to detect risk factors for type 2 diabetes mellitus

Ipek Balikci Cicek[a,*], Saim Yologlu[a], Ibrahim Sahin[b]

[a] *Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye*
[b] *Inonu University, Faculty of Medicine, Department of Internal Medicine, Division of Endocrinology and Metabolic Diseases, Malatya, Türkiye*

ARTICLE INFO

Keywords:
Artificial neural networks
Logistic regression analysis
Decision trees
Type 2 diabetes mellitus
Risk factors

Received: Sep 09, 2022
Accepted: Nov 09, 2022
Available Online: 27.02.2023

DOI:
10.5455/annalsmedres.2022.09.276

Abstract

Aim: The goal of this study is to compare the performances of Logistic Regression (LR), Artificial Neural Networks (ANN) and Decision Tree models, which are machine learning classification methods, in the diagnosis of type 2 Diabetes Mellitus (DM) and to determine the most successful method. It is also the examination of risk factors affecting type 2 DM using these models.

Materials and Methods: The study's data was collected from patients who visited the Diabetes and Thyroid polyclinic at the Inonu University Faculty of Medicine Turgut Ozal Medical Center, Department of Internal Medicine. The k-Nearest Neighbor algorithm, which is one of the missing value assignment methods, was used to eliminate the problems related to missing values. Sensitivity, accuracy, precision, specificity, AUC F1-score, and classification error were used as performance evaluation criteria. Evolutionary algorithm parameter optimization method was used to optimize the parameters of the ANN model. Missing value assignment, modeling and parameter optimization were done with Rapidminer Studio Free version 8.1.

Results: Among the three methods applied in the diagnosis of type 2 DM, the ANN gave the best classification performance. The accuracy, sensitivity, selectivity, precision, F1-score, AUC and classification error values obtained from this method are respectively; 98.94%, 100%, 97.73%, 98.04%, 99.01%, 0.978 and 1.06. For the ANN method, the importance values of the gender, long-term drug use, family history, concomitant disease, cortisone use, stress factor, high blood pressure, smoking, high cholesterol, heart disease, exercise status, carbohydrate use, alcohol consumption, vegetable use, meat use, age, weight, height, starting age, daily bread consumption, LDL, HDL, Total Cholesterol, Triglyceride, Fasting blood sugar the importance values of independent variables are respectively; 0.017, 0.009, 0.013, 0.017, 0.008, 0.016, 0.008, 0.006, 0.053, 0.024, 0.023, 0.040, 0.007, 0.020, 0.007, 0.046, 0.083, 0.049, 0.024, 0.066, 0.084, 0.083, 0.020, 0.031, 0.244.

Conclusion: According to the performance criteria obtained from the three classification models used to predict type 2 DM; it has been found that the best classification performance belongs to the ANN model. According to the ANN method, the three most important risk factors that may cause type 2 DM were found to be fasting blood glucose, LDL, and HDL, respectively.

Introduction

Diabetes is a condition characterized by a lack, ineffectiveness, or inadequate synthesis of the insulin hormone in the body, as well as chronic consequences that disturb carbohydrate metabolism and elevate blood glucose levels [1]. Diabetes is classified into two types: type 1 and type 2. Type 1 diabetes develops when the immune system mistakenly attacks pancreatic beta cells, resulting in insufficient or nonexistent insulin production. Type 2 diabetes develops when the body does not produce enough insulin or when insulin resistance develops [2]. Diabetes, which presents as symptoms such as intense thirst, hunger, and frequent urination, creates severe issues in the patient if not addressed. If prompt steps are not taken and blood sugar levels are not regulated, it has a harmful influence on the veins. Sugar's harmful effects can permanently damage numerous organs and tissues, including the eyes, kidneys, nerve endings, heart, brain, and leg arteries. As a result, early identification of diabetes is crucial for averting

*Corresponding author:
Email address: ipek.balikci@inonu.edu.tr (Ipek Balikci Cicek)

serious harm [3]. Machine learning algorithms for early detection of diabetes are becoming increasingly popular. Machine learning is a system that investigates the development and use of algorithms that can learn and predict data [4].

Logistic regression is a method for determining the cause and effect relationship between the independent variables when the dependent variable is observed in categorical, double, triple, and multiple categories. It is a regression approach that obtains the predicted values of the dependent variable as probabilities based on the independent variables. The mathematical connection between the dependent variable and the independent variable or variables is examined using simple and multiple regression analyses [5]. The effects of independent variables on the dependent variable are acquired as probabilities using logistic regression analysis, and risk factors are evaluated as probabilities. In the medical applications of logistic regression models, independent variables are risk variables or variables that determine whether a disease will occur or not. Detection of these variables has an important place in early diagnosis and in the fight against the factors that cause the disease. In summary, logistic regression is a regression approach that aids in assignment and classification by expressing the expected value of the dependent variable as a probability based on the independent variables [6].

Artificial Neural Network (ANN) are computer systems developed with the aim of directly performing the features of the human brain, such as learning, generating, creating and discovering new information, without any assistance [7]. Due to the learning ability of ANN, their adaptability to different problems, the need for less information after learning, their ability to generalize, and their ability to solve difficult mathematical models very quickly; It has been successfully applied in different areas such as learning, association, classification, modeling and prediction, generalization, feature detection and optimization. ANN creates their own experiences with the information they get from the examples and then they make similar decisions on similar issues [8]. ANN is composed of artificial cells that are hierarchically interconnected and capable of working in parallel. The most fundamental function of an artificial neural network is to determine an output set that corresponds to a single input set. To do this, the network is trained (learned) using instances of the relevant event and acquires the capacity to generalize. With this generalization, output sets that correspond to similaroccurrences are identified. It can be seen as a decision-making tool and calculation method that can be used very effectively, especially in cases where there is no information about events but examples are available [9].

Decision trees, one of the ways for making predictions, are a popular and effective technique for information discovery and data mining. Decision trees are a hierarchical and organized method of expressing data rules. Decision trees are a visual modeling technique that simplifies the presentation of a large amount of information about the problem facing the decision maker and sorts the decision alternatives and probabilistic circumstances in a certain order. In this context, decision trees can be viewed as a hierarchical model that incorporates decisions and outcomes. Due to its straightforward graphical layout and guidelines, it is utilized in numerous fields [10]. Among the categorization models in data mining, the model with predictive value is the decision tree model. Decision trees ask questions from the first stage to the final decision alternatives and develop their structure based on the responses they receive; rules (if-then rules) can be written using this tree structure [11].

The aim of this study is to compare the performances of LRA, ANN and Decision Trees models of different machine learning classification methods in diagnosis of type 2 DM and to determine the most successful method. It is also to examine the risk factors affecting type 2 DM by using these models.

Materials and Methods

Ethics committee approval

Ethical approval was obtained from Malatya Clinical Research Ethics Committee with protocol number 2016/144.

Study design and data

Data from patients with and without type 2 DM who visited Inonu University Faculty of Medicine Turgut Ozal Medical Center Internal Diseases Department Diabetes and Thyroid Outpatient Clinic were utilized in the study. The relevant data set consists of a total of 313 patient records, of which 146 (46.6%) had type 2 DM and 167 (53.4%) did not have type 2 DM. In this context, the variables included in the study are shown in Table 1.

Logistic regression model

Regression methods are used to describe the relationship between a response variable and multiple explanatory variables. The logistic regression model is utilized to determine the mathematical connection between categorical or continuous independent variables and categorical dependent variables. One of the reasons why this model is preferred is that there is no assumption criterion that the data come from a normal distribution, as in linear regression. Logistic regression is used for classification purposes when constructing a model to be used to predict which group patients belong to [12]. In logistic regression, the probability of the situation of interest is estimated when the values of the explanatory variables are given. In the simplest form, the logistic model used to estimate the probability of the state of interest when there is one explanatory variable is $\mathrm{P(y)}= 1/1+ \ e^{\ -(\beta 0+\beta 1\mathrm{X})}$. Here P(y) is the probability of Y, and the linear combination of coefficients is the same as in simple linear regression. In logistic regression, there may be more than one explanatory variable as well as one explanatory variable. In such a case, the logistic model is $\mathrm{P(y)}= 1/1+ \ e^{\ -(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn)}$ [13].

Artificial neural networks

Artificial intelligence (AI) is the ability of a computer or computer-controlled machine to execute activities linked to higher mental processes, such as reasoning, making sense, generalizing, and learning from prior experiences, which are widely regarded to be human traits [14]. In classification problems, ANN is generally used as one of the

Table 1. The variables investigated in the research.

Variables	Variable Type	Explanation	Variable Role
Type 2 DM	Qualitative	Yes/No	Dependent
Sex	Qualitative	Female/Male	Independent
Family History	Qualitative	Yes/No	Independent
Long-term drug use	Qualitative	Yes/No	Independent
Cortisone Use	Qualitative	Yes/No	Independent
Concomitant Disease	Qualitative	Yes/No	Independent
Hypertension	Qualitative	Yes/No	Independent
Stress Factor	Qualitative	A lot/ Few/None	Independent
Heart disease	Qualitative	Yes/No	Independent
High cholesterol	Qualitative	Yes/No	Independent
Smoking	Qualitative	Drinking/Not drinking	Independent
Alcohol consumption	Qualitative	Yes/No	Independent
Exercise status	Qualitative	Regular / Irregular / Occasional	Independent
Carbohydrate Use	Qualitative	Yes/No	Independent
Vegetable Use	Qualitative	Yes/No	Independent
Meat use	Qualitative	Yes/No	Independent
Age	Qualitative	Natural number	Independent
Weight	Qualitative	Natural number	Independent
Length	Qualitative	Natural number	Independent
Age of onset	Qualitative	Natural number	Independent
Daily bread consumption	Qualitative	Positive Real Number	Independent
HDL	Qualitative	Positive Real Number	Independent
LDL	Qualitative	Positive Real Number	Independent
Triglyceride	Qualitative	Positive Real Number	Independent
Total Cholesterol	Qualitative	Positive Real Number	Independent
Fasting Blood Glucose	Qualitative	Positive Real Number	Independent

Table 2. Distribution table of Type 2 DM variable.

Type 2 DM			
Yes		No	
Count	Percentage	Count	Percentage
146	46.6	167	53.4

Table 3. Number of missing values of variables.

Variables	Daily bread consumption	HDL	LDL	Triglyceride	High Cholesterol	Fasting Blood Glucose
Missing Values	1	1	1	1	2	22

AI techniques. Artificial Neural Networks are approaches that try to create a new system by imitating the functioning of the human brain [7]. ANN structure is created based on the structure of biological nerve cells in our brain. In ANNs, just like our brains, there are mechanisms for learning and decision-making according to the information learned [15]. Artificial nerve cells are called process elements [10].

Inputs denoted G_1, G_2, G_3, ..., G_N are known as inputs of an ANN. A_1, A_2, A_3,....A_N are defined as Weights and show the effect of the information coming into the artificial neuron. The Aggregation Function Function (NET), on the other hand, calculates the net information coming to a neuron. Various functions are used to find this net value, but the most used is the expression that finds the total weight shown in Equation. Here G_i the input value, A_i the weight of this input value, and $\text{NET}=\Sigma_i^n G_i A_i$ the total value of the function. The Activation Function (FNET) in the artificial neuron determines the output value to be produced by calculating the net inputs to the cell.

Decision tree

Decision trees are the most popular induction method widely used for classification purposes. It consists of four parts, the internal node, the root node, the leaf node, and the branches [16]. In this structure, the internal nodes represent the split criterion on the explanatory variables, the leaf nodes represent a class label, and the root node represents the initial variable of the tree. The branches connect the nodes. The value contained in the leaf nodes is the decision expression. When it is desired to find the value of a feature, it is progressed on the tree until the result is reached [17].

The purpose of the Decision Tree is to divide the training set into subgroups over the independent variables, starting from the root node. Different division criteria are used for different algorithms when subdividing. The decision tree classifier consists of two steps. In the first stage, the tree is created and the tree created in the second stage is pruned to prevent overfitting. Parametric statistical assumptions

Table 4. Descriptive statistics table for quantitative independent variables in the study.

Variables	Type 2 DM: Yes n=146		Type 2 DM: No n=167		p
	Mean±SD	Median (Min-Max)	Mean±SD	Median (Min-Max)	
Age	54.5±13.2	57.0 (19.0-80.0)	44.2±12.8	43.0 (19.0-76.0)	<0.001*
Weight	81.2±15.7	79.0 (50.0-135.0)	75.6±14.8	75.0 (45.0-169.0)	0.002*
Length	166.2±8.9	165.0 (148.0-190.0)	162.0±21.3	165.0 (1.0-190.0)	0.056*
Age Of Onset	45.8±13.3	47.0 (8.0-78.0)	38.0±13.8	38.0 (0.0-72.0)	<0.001*
Daily Bread Consumption	0.96±0.62	1.00 (0.00-3.00)	1.09±0.74	1.00 (0.00-3.00)	0.182*
HDL	44.1±11.5	42.5 (23.0-90.4)	48.1±11.0	46.1 (27.0-93.3)	<0.001*
LDL	117.1±36.5	114.3 (22.0-217.0)	121.8±38.1	119.6 (9.1-241.6)	0.269**
Triglyceride	173.1±94.7	152.0 (42.0-664.0)	130.7±75.8	108.0 (34.0-426.0)	<0.001*
High Cholesterol	196.2±42.5	194.0 (72.0-344.0)	196.6±44.4	192.0 (101.0-367.0)	0.887*
Fasting Blood Glucose	191.8±99.2	161.0 (67.0-538.0)	95.4±15.7	94.0 (69.0-199.0)	<0.001*

* Mann Whitney U test, **:Independent samples t-test, SD: Standard Deviation. Min: minimum, Max: maximum.

are not made in decision tree methods. Insights can be presented on terminal nodes with a few logical if-then conditions. There are no implicit assumptions in a normal data distribution or linear relationships between variables and response variable. Decision tree methods can reveal relationships and express them as a few decision rules that more computationally intensive methods overlook [18].

Data analysis

Quantitative data is represented by mean ± standard deviation and median (minimum-maximum), whereas qualitative data is represented by number (percentage). The Kolmogorov-Smirnov test was used to assess conformity to normal distribution. In terms of independent variables, whether a statistically significant difference exists between the "no" and "yes" groups, which are the categories of the dependent variable (Type 2 DM), and whether there is a relationship, Pearson chi-square test, Continuity Correction test, Independent samples t-test, Mann-Whitney U test, and Fisher's Exact test. It was investigated using the chi-square test, with $p<0.05$ considered statistically significant. For all analyses, the IBM SPSS Statistics 26.0 package program was used.

Modeling process

The k-Nearest Neighbor algorithm (kNN), which is one of the missing value assignment methods, was used to eliminate the problems related to missing values. While assigning in the kNN-based algorithm, other observation values that are similar to the missing observation values (the similarity criterion is generally Euclidean distance are chosen) are taken into account. LR, ANN and decision tree one of the ML methods, was used in the modeling. While applying these methods, first of all, the data set was divided into 70% as training data set and 30% as test data set. Sensitivity, accuracy, precision, specificity, AUC, F1-score, and classification error were used as performance evaluation criteria. In addition, variable importances were calculated, which gives information about how much the input variables explain to the output variable. Parameter optimization is one of the important factors that can affect the prediction performance of models. In order to obtain more accurate results from the data set and to increase the power of the performance output of the model, the parameters of the model should be adjusted very well. Evolutionary algorithm parameter optimization method was used to optimize the parameters of the ANN model. Missing value assignment, modeling and parameter optimization were done with Rapidminer Studio Free version 8.1. [19].

Results

The data set used in the study consists of 25 independent variables and 1 dependent variable. The distribution table of the dependent variable, type 2 DM, is shown in Table 2.

Variables with missing values among 26 variables are shown in Table 3.

The value assignment to the observations with missing values was made using the kNN algorithm. After assigning the missing values to the observations, the descriptive statistics table of the independent variables is shown in Table 4.

The distribution table of the qualitative independent variables in the study is shown in Table 5.

The classification performances of the three models we used before and after applying the parameter optimization method to the ANN model are given in Table 6 according to the determined performance criteria.

According to Table 6, classification performance values of the LR method for test data for accuracy, sensitivity, specificity, precision, F1-score, AUC, classification error criteria, respectively; 82.26%, 84.85%, 79.31%, 82.35%, 83.58%, 0.874, 17.74. For the same classification performance values of the ANN method, 75.53%, 78.00%, 72.73%, 76.47%, 77.23%, 0.855, 24.47, respectively. The same classification performance values of the Decision Trees method were found to be 85.48%, 90.91%, 79.31%, 83.33%, 86.96%, 0.878, 14.52, respectively.

According to Table 7, classification performance values of the LR method for test data accuracy, sensitivity, specificity, precision, F1-score, AUC, classification error criteria, respectively; 82.26%, 84.85%, 79.31%, 82.35%,

Table 5. Distribution table for qualitative independent variables in the study.

Variables	Categories	Type 2 DM: Yes n=146	Type 2 DM: No n=167	p
		n (%)	n (%)	
Sex	Female	70 (47.9)	124 (74.3)	<0.001*
	Male	76 (52.1)	43 (25.7)	
Family history	Yes	80 (54.8)	72 (43.1)	0.039*
	No	66 (45.2)	95 (56.9)	
Long-term drug use	Yes	37 (25.3)	37 (22.2)	0.508*
	No	109 (74.7)	130 (77.8)	
Cortisone use	Yes	4 (2.7)	3 (1.8)	0.709***
	No	142 (97.3)	164 (98.2)	
Concomitant disease	Yes	97 (66.4)	65 (38.9)	<0.001*
	No	49 (33.6)	102 (61.1)	
Hypertension	Yes	65 (44.5)	35 (21.0)	<0.001*
	No	81 (55.5)	132 (79.0)	
Stress factor	None	27 (18.5)	39 (23.4)	0.573*
	Few	31 (21.2)	34 (20.4)	
	A lot	88 (60.3)	94 (56.3)	
Heart disease	Yes	51 (13.2)	22 (13.2)	<0.001*
	No	95 (86.8)	145 (86.8)	
High cholesterol	Yes	54 (37.0)	18 (10.8)	<0.001*
	No	92 (63.0)	149 (89.2)	
Smoking	Smoking	34 (23.3)	40 (24.0)	0.890*
	No Smoking	112 (76.7)	127 (76.0)	
Alcohol consumption	Yes	4 (2.7)	2 (1.2)	0.423***
	No	142 (97.3)	165 (98.8)	
Exercise status	Regular	23 (15.8)	20 (12.0)	0.164***
	Irregular	98 (67.1)	128 (76.6)	
	Occasional	25 (17.1)	19 (11.4)	
Carbohydrate use	Yes	94 (64.4)	89 (53.3)	0.047*
	No	52 (35.6)	78 (46.7)	
Vegetable use	Yes	124 (84.9)	141 (84.4)	1.0**
	No	22 (15.1)	26 (15.6)	
Meat use	Yes	95 (65.1)	105 (62.9)	0.687*
	No	51 (34.9)	62 (37.1)	

*: Pearson chi-square test, **: Continuity Correction test, ***: Fisher's Exact test.

Table 6. Classification performance for test data before parameter optimization method is applied to ANN model from LR, ANN and Decision Tree methods.

Methods	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1-score(%)	AUC	Classification Error
LR	82.26	84.85	79.31	82.35	83.58	0.874	17.74
ANN	75.53	78	72.73	76.47	77.23	0.855	24.47
Decision Tree	85.48	90.91	79.31	83.33	86.96	0.878	14.52

83.58%, 0.874, 17.74 and also, for the same classification performance values of the ANN method, %98.94, %100, %97.73, %98.04, %99.01, 0.978, 1.06. The same classification performance values of the Decision Trees method were found to be 85.48%, 90.91%, 79.31%, 83.33%, 86.96%, 0.878, 14.52.

According to the LR, ANN and Decision Trees methods, the importance values that can be used to determine which independent variable is more effective on the dependent variable are given in Table 8.

Table 7. Classification performance for test data after applying parameter optimization method to ANN model from LR, ANN and Decision Trees methods.

Methods	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1-score(%)	AUC	Classification Error
LR	82.26	84.85	79.31	82.35	83.58	0.874	17.74
ANN	98.94	100	97.73	98.04	99.01	0.978	1.06
Decision Tree	85.48	90.91	79.31	83.33	86.96	0.878	14.52

Table 8. The importance values of independent variables in Type 2 DM.

Variables	LR	ANN	Decision Tree
Sex	0.782	0.017	0.176
Family History	0.482	0.013	0.044
Long-term drug use	0.16	0.009	-
Cortisone Use	0.232	0.008	-
Concomitant Disease	0.742	0.017	0.097
Hypertension	0.33	0.008	-
Stress Factor	0.031	0.016	-
Heart disease	0.003	0.024	-
High cholesterol	0.992	0.053	-
Smoking	0.186	0.006	-
Alcohol consumption	0.172	0.007	-
Exercise status	0.089	0.023	-
Carbohydrate Use	0.709	0.04	-
Vegetable Use	0.22	0.02	-
Meat Use	0.497	0.007	-
Age	0.287	0.046	0.117
Weight	0.458	0.083	0.34
Lenght	0.155	0.049	-
Age of onset	0.974	0.024	0.018
Daily bread consumption	0.752	0.066	0.065
HDL	2.395	0.083	-
LDL	4.897	0.084	-
Trigliserid	2.618	0.031	-
Total cholesterol	6.588	0.02	-
Fasting Blood Glucose	5.108	0.244	0.297

According to Table 8, for the LR method, gender, long-term drug use, family history, cortisone use, concomitant disease, high blood pressure, stress factor, heart disease, high cholesterol, smoking, alcohol consumption, exercise status, carbohydrate use, vegetable use, meat use, age, weight, height, starting age, daily bread consumption, HDL, LDL, Triglyceride, Total Cholesterol, Fasting blood glucose the importance values of independent variables, respectively; 0.782, 0.482, 0.160, 0.232, 0.742, 0.330, 0.031, 0.003, 0.992, 0.186, 0.172, 0.089, 0.709, 0.220, 0.497, 0.287, 0.458, 0.155, 0.974, 0.752, 2.395, 4.897, 2.618, 6.588, 5.108; For the ANN method, the importance values of the same variables are respectively; 0.017, 0.013, 0.009, 0.008, 0.017, 0.008, 0.016, 0.024, 0.053, 0.006, 0.007, 0.023, 0.040, 0.020, 0.007, 0.046, 0.083, 0.049, 0.024, 0.066, 0.083, 0.084, 0.031, 0.020, 0.244; for the decision tree method, the importance values of the same variables are respectively; It was found as 0.176, 0.044, 0.097, 0.117, 0.340, 0.018, 0.065, 0.297.

Discussion

Diabetes is an important disease that significantly affects the quality of life of human beings, and its incidence is increasing in the world and in Turkey. DM can cause organ loss and death. According to the World Health Organization's most recent statistics, there are 422 million people with diabetes worldwide, most of whom live in low- and middle-income countries. Diabetes is also directly responsible for 1.6 million fatalities annually [20]. As a result, diabetes is acknowledged as one of the main causes of mortality worldwide, and both its prevalence and number of cases are sharply rising. In the face of this negative picture, countries set a target to stop the increase in diabetes globally by 2025 and decided to cooperate [1].

Early diagnosis and follow-up are of great importance in order to prevent diabetes or to minimize its damage. With the technological developments, it becomes possible to diagnose many diseases with the help of artificial intelligence and learning techniques. In this way, the diagnosis of diseases and the reporting of related examinations are completed in a shorter time, and as a result, the time spent by the patients in the health institution is reduced [21].

For these reasons, in this study, it is aimed to determine the most successful method in the diagnosis of type 2 diabetes by comparing the performances of LRA, ANN, Decision Tree models from different machine learning methods and to express the importance values of risk factors affecting type 2 diabetes.

Classification with machine learning for early diagnosis of diabetes gives results with high accuracy. Many different machine learning algorithms are used to make classifications. Considering the studies in the literature for the diagnosis of different diabetes diseases; In a study conducted with the data set obtained from 250 patients aged between 25-78 years, Bayesian algorithm was used and an accuracy rate of 88.8% was achieved [22]. In another study, the J48 model was used to classify diabetic treatment plans such as insulin, drug therapy and diet, and a success value of 70.8% was obtained [23]. In one study, complications, genetic background and environment were examined for the prediction of diabetes. Various machine learning algorithms have been used, Support vector machine (SVM) has emerged as the most successful and widely used algorithm [24]. In another study, diabetes prediction was made using 7 different features such as glucose, age, blood pressure and body mass index. SVM, LR and ANN were used as machine learning algorithms for the analyses. Considering the classification performances, the best results were obtained with SVM [25]. In another study, k-NN, Naive Bayes and Random Forest (RO) methods were used in the diagnosis of diabetes. The accuracies of these methods were found to be 66.19%, 72.66% and 73.72%, respectively [26]. In a different study, several machine learning methods were used for diabetes diagnosis. The highest accuracy success was achieved with LR, Linear Discriminant Analysis and AdaBoost models %96, %94 and 93% [27]. In another similar study, Deep Learning and SVM methods were used on the data set created with 500 diabetic patients and 268 healthy people. The deep learning model performed better and reached a success value of 77.474% [28].

In this study; According to the findings of three classification models used to predict patients with and without type 2 DM; it has been determined that the best classification performance belongs to the Artificial Neural Networks model. In the ANN model; accuracy, sensitivity, selectivity, precision, F1-score, AUC, classification error were 98.94%, 100%, 97.73%, 98.04%, 99.01%, 0.978 and 1.06, respectively.

In addition, the factors associated with type 2 DM were determined by ANN, LRA, Decision Trees, and the effects of the variables on the disease were estimated. According to the ANN model, which gave the best performance result, the most important factor that could cause type 2 DM was obtained as fasting blood glucose. Clinically, increased fasting blood glucose, that is, a fasting blood value of 126 mg/dl or higher, has an effect on type 2 DM [29]. In the classification of type 2 DM, LDL value was determined as the second most important factor that could affect this disease, while HDL value, which is one of the blood lipids, was determined as the third important risk factor. The main factor influencing the relationship between cholesterol and coronary heart diseases is the LDL cholesterol value, and type 2 DM is also an important determinant of risk. Additionally, people with low HDL cholesterol levels are at risk for type 2 DM [30]. Another important factor was determined to be the weight variable. Those with a Body Mass Index (BMI) above 25 kg/m^2 and especially those with a large belly circumference are at risk for type 2 DM [31].

Disclosure

This study was presented as an oral presentation in XX. National and III. International Biostatistics Congress 26-29 October 2018 - Gaziantep.

Ethics approval

Ethical approval was obtained from Malatya Clinical Research Ethics Committee with protocol number 2016/144.

References

1. Başer BÖ, Yangın M, Sarıdaş ES. Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi. 2021;25(1):112-20.
2. Ozougwu J, Obimba K, Belonwu C, Unakalamba C. The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus. J Physiol Pathophysiol. 2013;4(4):46-57.
3. Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW. Global trends in diabetes complications: a review of current evidence. Diabetologia. 2019;62(1):3-16.
4. Cihan P, Coşkun H. Diyabet Tahmini için Makine Öğrenmesi Modellerinin Performans Karşılaştırılması Performance Comparison of Machine Learning Models for Diabetes Prediction.
5. Neslihan İ, Aşır G. Lojistik Regresyon Analizi Yardımıyla Denekte Menopoz Evresine Geçişe İlişkin Bir Sınıflandırma Modelinin Elde Edilmesi. Selçuk Üniversitesi Fen Fakültesi Fen Dergisi. 2005;1(25):19-28.
6. Bircan H. Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. Kocaeli Üniversitesi Sosyal Bilimler Dergisi. 2004(8):185-208.
7. Öztemel E. Yapay Sinir Ağları, Papatya Yayıncılık Eğitim Bilgisayar Sis. San ve Tic AŞ, İstanbul. 2006.
8. Haykin S. Neural Networks, a comprehensive foundation, Prentice-Hall Inc. Upper Saddle River, New Jersey. 1999;7458:161-75.
9. Altaş D, Gülpınar V. Karar Ağaçları ve Yapay Sinir Ağlarının Sınıflandırma Performanslarının Karşılaştırılması. Trakya Üniversitesi Sosyal Bilimler Dergisi. 2012.
10. Murthy SK. Automatic construction of decision trees from data: A multi-disciplinary survey. Data mining and knowledge discovery. 1998;2(4):345-89.
11. Maimon OZ, Rokach L. Data mining with decision trees: theory and applications: World scientific; 2014.
12. Hosmer D. Lemeshow S. Applied logistic regression. USA: John Wiley and Sons; 2000.
13. Sanz J, Paternain D, Galar M, Fernandez J, Reyero D, Belzunegui T. A new survival status prediction system for severe trauma patients based on a multiple classifier system. Computer methods and programs in biomedicine. 2017;142:1-8.
14. Nabiyev VV, Zeka Y. Problemler. Yöntemler, Algoritmalar, Seçkin Yayıncılık. 2005:83-6.
15. Haykin S. Neural networks and learning machines, 3/E: Pearson Education India; 2009.
16. Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Computing and Applications. 2013;23(7):2387-403.
17. Sathyadevan S, Nair RR. Comparative analysis of decision tree algorithms: ID3, C4. 5 and random forest. Computational intelligence in data mining-volume 1: Springer; 2015. p. 549-62.
18. Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H, et al. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. Computer methods and programs in biomedicine. 2017;141:105-9.
19. Rapidminer DA. RapidMiner 4.1 User Guide. Dortmund; 2008.
20. Bilgin G. Makine öğrenmesi algoritmaları kullanarak erken dönemde diyabet hastalığı riskinin araştırılması. Journal of Intelligent Systems: Theory and Applications. 2021;4(1):55-64.
21. Asif M. The prevention and control the type-2 diabetes by changing lifestyle and dietary pattern. Journal of education and health promotion. 2014;3.
22. Sapon MA, Ismail K, Zainudin S, editors. Prediction of diabetes by using artificial neural network. Proceedings of the 2011 International Conference on Circuits, System and Simulation, Singapore; 2011.
23. Ahmed TM. Developing a predicted model for diabetes type 2 treatment plans by using data mining. Journal of Theoretical and Applied Information Technology. 2016;90(2):181.
24. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017;15:104-16.
25. Joshi TN, Chawan P. Diabetes prediction using machine learning techniques. Ijera. 2018;8(1):9-13.
26. Al Helal M, Chowdhury AI, Islam A, Ahmed E, Mahmud MS, Hossain S, editors. An optimization approach to improve classification performance in cancer and diabetes prediction. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE); 2019: IEEE.
27. Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019;165:292-9.
28. Thaiyalnayaki K. Classification of diabetes using deep learning and svm techniques. International Journal of Current Research and Review. 2021;13(01):146.
29. Eroğlu N. Diyabetin Komplikasyonlarından Korunmak için Tanı, Tedavi ve İzlem. İzmir Katip Çelebi Üniversitesi Sağlık Bilimleri Fakültesi Dergisi.4(1):31-3.
30. Ergin E, Akın S, Kazan S, Erdem M, Tekçe M, Aliustaoğlu M. Diyabetik Hastalarda Lipid Profili: Farkındalık ve Tedavideki Başarı Oranlarımız. Kartal Eğitim ve Araştırma Hastanesi Tıp Dergisi. 2013;24(3).

31. Karslıoğlu DH. Obezite, Tip 2 Diyabet ve Beslenme. Klinik Tıp Bilimleri.7(3):36-43.