Epidemiology and Health System Journal
doi:10.34172/ehsj.26113 2024 Summer;11(3):128-136 http://ehsj.skums.ac.ir

Original Article

Exploring Risk Factors of Type 2 Diabetes Mellitus Using Decision Tree and Random Forest Models: Baseline Data From Kharameh Cohort Study

Maryam Jalali[1], Hamid Reza Niazkar[2], Masoumeh Ghoddusi Johari[2*], Amir Hossein Saem[3], Abbas Rezaianzadeh[1]

[1]Colorectal Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
[2]Breast Diseases Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
[3]School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran

Abstract

Background and aims: Identifying subjects that are at risk of type 2 diabetes mellitus (T2DM) and predicting the associated risk factors are highly important. Thus, this study aimed to explore the risk factors and find the prediction model for T2DM using decision trees (DTs) and random forest (RF) models.

Methods: This cross-sectional study is a part of the Kharameh Cohort Study. Kharameh Cohort is a part of the Fars Cohort, which started in 2014 with 10 663 people aged 40–70. In this study, the risk factors of T2DM were explored using two data mining methods. Accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) were applied to evaluate the models. The data were statistically analyzed using R software.

Results: The DT modeling showed that age, triglycerides (TG), physical activity, systolic blood pressure, low-density lipoproteins (LDL), and body mass index (BMI) were the most associated factors in D2MT, while applying RF revealed that fasting blood sugar, cholesterol, creatinine, TG, gamma-glutamyl transferase physical activity, BMI, and LDL were the most effective on T2DM. The RF model was superior to the DT based on the applied criteria. Sensitivity, specificity, accuracy, and AUC for the RF were 73.4, 70.10, 73.5, and 79.1. These findings for the DT were 63.8, 69.7, 62.8, and 66.8, respectively.

Conclusion: Based on the inferences, a strong association was found between several risk factors and the risk of T2DM. Therefore, predictive analytics using the RF model can be applied to identify the risk factors of other chronic diseases.

Keywords: Diabetes mellitus, Data mining, Risk factors

***Corresponding Author:**
Masoumeh Ghoddusi Johari,
Email: m.ghoddusi94@yahoo.com

Received: December 6, 2023
Accepted: August 19, 2024
ePublished: November 12, 2024

Introduction

Type 2 diabetes mellitus (T2DM) is a common non-communicable disease, and its prevalence is increasing worldwide due to lifestyle changes. It represents approximately 95% of diabetes cases in the majority of populations. According to reported estimates, the number of adult diabetics will rise from 422 to 624 million worldwide by 2040, 54% more than the estimated numbers in 2010.[1] In addition, less than 1% of females and an even smaller number of males have the opportunity to achieve global objectives in preventing the increase in diabetes incidence by 2025.[2]

Geographically, diabetes is spread differently around the world. According to reports, the highest prevalence of diabetes was observed in India, China, and the United States of America. The general pattern of diabetes in the world shows that the prevalence of this disease is higher in developing countries as well as low socio-cultural groups; thus, it is estimated that more than 75% of the total diabetics will be in developing countries by 2025.[3]

Based on the prediction of the World Health Organization, the prevalence of T2DM in Iran will be 8.6% (5 125 000 patients) in 2025 and will reach 9.2 million diabetics in 2030.[4]

Many factors are associated with T2DM, which can be classified into non-modifiable and modifiable groups. Metabolic syndromes such as high levels of triglyceride (TG), cholesterol, high blood pressure (BP), abdominal obesity, low levels of high-density lipoprotein (HDL), and smoking are considered risk factors for high-risk or pre-diabetic individuals.[4] Hence, identifying subjects at risk of T2DM and predicting the associated risk factors are highly essential.

Classical methods for determining risk factors applied in studies include Fisher's linear discriminant analysis and logistic regression. However, these traditional models cannot perform well in many variables, high-dimensional data, nonlinear relationships, outliers, and missing data.

These situations can be handled using data mining methods such as decision trees (DTs), random forests (RFs), neural networks, and support vector machines.[5] Numerous studies have consistently affirmed the superior accuracy and lower error rates of data mining methods compared to traditional classification models.[6] Komi et al explored the early prediction of diabetes via five methods of data mining, and the artificial neural network had the highest accuracy.[7]

In another recent study, data mining algorithms for predicting diabetes were compared, and the results revealed that the DT had the best accuracy.[8]

The mechanism of data mining methods is to extract hidden factors and patterns from a large amount of data, and it is applied in medical studies to explore the associated risk factors of T2DM and other situations.[9] However, to the best of our knowledge, a relatively small number of researchers have utilized data mining to construct prediction models incorporating multiple risk factors. Accordingly, this study seeks to explore the risk factors and find the prediction model for T2DM in the population of Kharameh Cohort Study in the south of Iran using two data mining methods, DT and RF. Accuracy, sensitivity, and specificity have been applied to evaluate the models' performance.

Materials and Methods

Study Design and Population

This analytical cross-sectional study is part of the Kharameh Cohort Study, a branch of Prospective Epidemiological Studies in Iran. The Persian Cohort Study is one of the most significant research projects of the region; its aims and design have been published before.[10] Kharameh is located in the south of Fars province in Iran, with a population of 61 580 people. Kharameh Cohort is a part of the Fars cohort, which was initiated in 2014 with 10,663 individuals ages 40–70 years. The primary aim was to find the prevalence and risk factors of non-communicable diseases. All of the Kharameh population was entered into the study through census.[11]

Initially, the participants completed a written consent form. Then, they received a standardized questionnaire that was used to gather information about their demographic characteristics, including age, gender, body mass index (BMI), marital status, level of education, place of residence, occupational status, social and economic status, and family history of chronic diseases. In addition, data on behavioral factors, such as smoking, alcohol consumption, hookah use, drug use, and physical activity, were obtained through interviews. For the laboratory tests, the participants were requested to fast for 12 hours before blood sampling. Further, their weight and height were measured using a Seca scale and a standard measuring tape, respectively. Furthermore, blood glucose, HDL, and cholesterol levels were estimated using the Mindray brand tool and Pars test kit.[10]

Exclusion Criteria

Unwillingness to participate in the study, mental disorders (intellectual disability), and total daily energy intake (kcal) out of mean ± 3SD were considered as overreport data and the exclusion criteria in this study.[12] Finally, 10 439 subjects (1587 DMT2 and 8852 non-diabetes) were included in our study.

Input Parameters

The parameters entered in our analysis are the ones that were gathered in the Kharameh Cohort Study, and their association was investigated in previous studies.[13-15] They are listed as follows:

- Demographic characteristics: Age, gender, marital status, education, and cigarette smoking habit;
- Anthropometric data: BMI (BMI and its formula is: weight/height^2), weight (kg), height (m^2), waist circumference (w), hip circumference (h), and waist to hip ratio (w/h);
- Glucose levels: Fasting blood sugar;
- Kidney function: Specific gravity (SG), creatinine (Cr), and blood urea nitrogen;
- Liver tests: Serum glutamic-oxaloacetic transaminase, alanine transaminase (ALT), alkaline phosphatase, and gamma-glutamyl transferase;
- BP: Systolic and diastolic BP (SBP and DBP);
- Lipid profile: Cholesterol, HDL, low-density lipoproteins (LDL), and TG;
- Total energy (kcal);
- Physical activity;
- Family history of T2DM: First- and second-degree family history of T2DM (FHD1 and FHD2)
- First-degree family history of hypertension (FHH1).

Diabetes

Blood samples were taken from the participants. DM was defined as fasting blood glucose ≥ 126 mg/dL, a 2-hour value in an oral glucose tolerance test ≥ 200 mg/dL, or taking antidiabetic medication.

Statistical Analysis

Two data mining methods were applied, including DTs and RFs. Each of them is briefly explained in the following paragraphs.

Decision Tree Method

DTs are a subset of supervised machine learning based on splitting the data relative to a specific parameter to explore a predictive model based on the features or classification of the subjects. Classification (categorical outcomes) and regression (continuous outcomes) trees are the two common types of DTs. A DT has two parts, including a node (root and internal nodes) and a leaf (end nodes or the target). In a DT, each node represents an attribute, each link or branch shows a decision, and each leaf demonstrates an outcome.[16] A tree results from a successful data division based on one of the variables.

To form the trees, DT algorithms apply splitting criteria at the internal nodes to minimize the internal nodes' impurity. The node impurity is an index for measuring the homogeneity of the class labels in each node and leaf. The node will be split if the impurity is reduced; otherwise, it will be presented as a leaf. If the impurity is reduced, two branches will be formed, and two new nodes will appear accordingly. The splitting criteria give a rate to each predictor variable. Therefore, those with the best rate will remain in the model.

The classification and regression tree is a tree algorithm variation that can deal with classification and regression issues.[17] In this algorithm, nodes are divided into sub-nodes according to the threshold of an attribute. The root node is considered the training set, and splitting is performed by considering the appropriate attribute and threshold value. This process stops when the maximum possible number of leaves is achieved. A common and applied criterion for measuring the degree of non-homogeneity in the DT method is the Gini index, which is a function that shows the goodness of splitting and helps find the best splitter and pure DT. The Gini impurity value ranges from 0 to 1. The formula of the Gini index at a node D having m classes is:

$$\text{Gini (D)} = 1 - \sum_{i=1}^{m} p_i^2 .$$

where p_i is the probability of belonging the observation in D to class Ci and is estimated by |Ci, D |/|D|, and m represents the possible classes.[17]

Random Forest Method

RF, or random decision forest, is an ensemble learning method applied in classification and regression and works based on constructing a mass of DTs.[18] This method reduces overfitting by combining multiple overfitted DTs to form an ensemble learning algorithm. Every DT in this method has its decision result. By applying the result of voting for each tree in the forest, the sample category for testing is derived based on the rule of the minority following the majority. Finally, the category with the highest vote in all DTs would be the result.[18]

To achieve these final rules, classification trees are derived from bootstrap sampling.[18] The RF has some significant characteristics, including handling missing data, balancing errors in the case of imbalanced data, and estimating the variable's importance.[18]

The main tree training parameters are the number of generated trees (ntree), the number of predictors in each tree (ntry), and the number of observations in a leaf node (node size). The values considered for these parameters in this study are ntree = 500, ntry = 29, and node size = 5. It is reported that RF is not too sensitive to the value of parameters, and the default values usually yield appropriate output. In addition to greater accuracy than other supervised learning methods, it offers variable importance to all input variables. Variable importance represents each variable's contribution to improving classification.[18]

DTs are simple to understand, providing a clear visual to guide decision-making. However, this simplicity has severe disadvantages, including overfitting, errors due to bias, and variance. RFs reduce the variance observed in decisions. Similar to RFs, gradient boosting is a set of DTs. The two main differences are how trees are built and how the results are combined. If the parameters are carefully tuned, gradient boosting can perform better than RFs. However, gradient boosting may only be a good choice if you have a little noise, as it can result in overfitting. They also tend to be harder to tune than RFs. RFs and gradient boosting each excel in different areas. RFs perform well for multi-class object detection and bioinformatics, which tend to have much statistical noise. Gradient boosting performs well when you have unbalanced data, such as when performing a real-time risk assessment.[19]

Both DTs and RFs are effective in handling collinearity in the input features. DTs are inherently robust to collinearity, while RFs enhance this robustness by combining multiple DTsDTRFDT.[17]

The results were statistically analyzed by R software (version 4.1) using the rpart package for the DT model and the RF package for RF models. In addition, a 10-fold cross-validation was applied for the evaluation of our method. This significantly reduces underfitting and substantially reduces overfitting. K-fold cross-validation helps generalize the machine learning model, which results in better predictions on unknown data.[20] In this validation technique, the data will be randomly separated into ten equal datasets, and the models (RF or DT) will be constructed based on the training dataset. The nine datasets will be used as testing data to confirm the model's effectiveness. This procedure will be repeated ten times, reserving a different one-tenth for testing.[20]

The overall model discrimination and inherent validity of our classification models were assessed using bootstrap (500 replications) optimism-corrected area under the receiver operating characteristic curve (ROC) by applying pROC and boot packages. The ROC curve shows the clinical sensitivity and specificity relationship for every possible cut-off.[21] It is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is different. The ROC curve is created by plotting the true-positive rate against the false-positive rate at various threshold settings. The true-positive rate is also known as sensitivity, recall, or probability of detection. The false-positive rate is also known as the probability of false alarm rate and can be calculated as (1– specificity).[21]

The relationship between the above measurements is as follows:

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Sensitivity = TP / (TP + FN)

Specificity = TN / (FP + TN)

TP, TN, FP, and FN are true positive, true negative,

false positive, and false negative, respectively.[21]

While specificity, sensitivity, and accuracy are valuable metrics, they each have significant limitations to consider when evaluating and comparing diagnostic tests. A more comprehensive assessment typically requires considering multiple performance measures in the appropriate clinical context. These metrics have some limitations, including (1) the trade-off between sensitivity and specificity, (2) dependence on prevalence, (3) the difficulty in comparing tests directly, (4) context dependence, (5) oversimplification of accuracy, and (6) lack of uncertainty representation. As regards the first limitation, increasing sensitivity often leads to a decrease in specificity, and vice versa, and the appropriate balance depends on the clinical context and consequences of false results. Concerning dependence on prevalence, the predictive values of a test (positive and negative) rely on the condition's prevalence in the population, and sensitivity, specificity, and accuracy alone do not provide enough information about a test's clinical utility. Regarding the third limitation, it can be challenging to compare the performance metrics of different diagnostic tests, and values can vary based on the study population, setting, and measurement methods. In terms of context dependence, the appropriate balance between sensitivity, specificity, and accuracy depends on the clinical context and the consequences of false results. As regards the fifth limitation, accuracy provides a single summary statistic that can oversimplify the complex tradeoffs involved in test performance and does not distinguish between false positives and false negatives. Concerning lack of uncertainty representation, these metrics do not directly convey the degree of uncertainty or reliability associated with a test's results.[22]

The AUC is generally a measure of the usefulness of a diagnostic test. Hence, a greater area implies a more accurate test. In addition, the accuracy comparison of two or more tests is possible by comparing each test's AUC using DeLong's test. The higher value of the AUC stands for better performance of the classifier method.[23]

Results

Tables 1 and 2 represent means ± standard deviations (SD)/number (%) for the anthropometrics and clinical data of 10 439 subjects, including 1587 DMT2 and 8852 non-diabetics, along with the *P* values for comparing these factors. According to Table 1, the mean ± SD of age was 56.11 ± 7.86 for diabetics, and 23.6% of those with a family history of diabetes had DMT2. About 10% of males and 19.2% of females were DMT2 cases. There were significant differences ($P<0.05$) between diabetic and non-diabetic subjects in terms of age, BMI, gender, FHH1, FHD1, FHD2, education level, occupational status, marital status, alcohol consumption, and smoking status. Based on the results (Table 2), all clinical parameters significantly differed between diabetics and non-diabetics.

Two models (DT and RF) were assessed in this study. The data were divided into a training dataset (70%) and a test dataset (30%). Both models were built on a training dataset (7250 records). A testing dataset (3189 records) was used to evaluate the models.

After applying the RF model, BMI, FHD1, cholesterol, TG, Cr, physical activity, age, LDL, and ALT were the most important and influential factors in T2DM (Figure 1). The extracted rules through this model are presented in Table 3.

DT models revealed that age, TG, BMI, SBP, FHD1, physical activity, LDL, Cr, and DBP were the most associated factors in DMT2. The classification and regression tree is displayed in Figure 2, and the extracted rules are provided in Table 3.

Table 1. Comparison of Participants' Characteristics (Number (%) for Categorical and Mean ± SD for Continuous) Between Diabetics and Non-diabetics

Characteristics		Non-diabetics	Diabetics	*P* Value
		Mean ± SD/ Number (%)	Mean ± SD/ Number (%)	
Age (y)		51.36 ± 8.15	56.11 ± 7.86	**< 0.001**
BMI (kg/m²)		25.91 ± 4.48	27.15 ± 4.14	**< 0.001**
Weight		69.15 ± 12.33	69.38.06 ± 12.16	0.49
Height		162.98 ± 9.31	163.28 ± 9.16	0.22
W		95.52 ± 12.10	95.83 ± 11.88	0.35
H		100.93 ± 8.37	100.87 ± 8.27	0.96
W/h		0.95 ± 0.07	0.95 ± 0.07	0.1
FHH1	Yes	4651 (82.6)	978 (17.4)	**< 0.001**
	No	4413 (87.7)	621 (12.3)	
FHD1	Yes	2959 (76.4)	914 (23.6)	**< 0.001**
	No	6104 (89.9)	685 (10.1)	
FHD2	Yes	1759 (79.2)	462 (20.8)	**< 0.001**
	No	7305 (86.5)	1137 (13.5)	
Gender	Male	4016 (90)	446 (10)	**< 0.001**
	Female	4778 (80.8)	1134 (19.2)	
Education level	Low	7419 (83.8)	1432 (16.2)	**< 0.001**
	Moderate	865 (90)	96 (10)	
	High	510 (90.7)	52 (9.3)	
Occupational status	Employed	4760 (90.3)	509 (9.7)	**< 0.001**
	Unemployed	4034 (79)	1071 (21)	
Marital status	Single	160 (93)	12 (7)	**< 0.001**
	Married	7921 (86)	1292 (14)	
	Widow	623 (69.8)	296 (30.2)	
	Divorced	56 (96.6)	2 (3.4)	
	Temporary Marriage	34 (87.2)	5 (12.8)	
Alcohol consumption	Yes	290 (93.9)	19 (6.1)	**< 0.001**
	No	8504 (84.5)	1561 (15.5)	
Smoking status	Yes	2312 (91.1)	227 (8.9)	**< 0.001**
	No	6482 (82.7)	1353 (17.3)	

Note. *Bold figures represent a statistically significant association ($P<0.05$). Independent sample *t* tests and chi-square tests were used to compare quantitative and qualitative factors between groups. SD: Standard deviation; BMI: Body mass index; FHH1: First-degree family history of hypertension; FHD1: First-degree family history of T2DM; FHD2: Second-degree family history of T2DM.

Table 2. Comparison of Mean ± SD for Clinical Parameters Between Diabetics and Non-diabetics

Characteristics	Non-diabetics	Diabetics	*P* Value*
SBP (mm Hg)	114.17 ± 17.50	119.74 ± 19.09	**< 0.001**
DBP (mm Hg)	72.01 ± 10.32	73.34 ± 10.41	**< 0.001**
Cholesterol (mg/dL)	187.85 ± 41.22	181.94 ± 43.98	**< 0.001**
HDL-cholesterol (mg/dL)	47.98 ± 12.61	47.08 ± 12.62	**0.01**
LDL cholesterol (mg/dL)	114.75 ± 34.75	104.59 ± 36.75	**< 0.001**
TG (mg/dL)	126.39 ± 77.03	152.36 ± 93.76	**< 0.001**
FBS	91.83 ± 16.82	142.88 ± 61.19	**< 0.001**
SG	1.02 ± 0.006	1.02 ± 0.006	**0.005**
BUN	13.93 ± 3.98	14.52 ± 4.89	**< 0.001**
Cr	0.98 ± 0.23	0.99 ± 0.23	**< 0.001**
SGOT	22.37 ± 9.34	21.45 ± 11.02	**< 0.001**
SGPT	24.91 ± 15.34	26.67 ± 18.79	**< 0.001**
ALP	208.83 ± 65.75	223.19 ± 71.96	**< 0.001**
GGT (IU/L)	24.85 ± 21.07	30.52 ± 28.01	**< 0.001**
Energy	2498.46 ± 820.60	2151.93 ± 675.19	**< 0.001**
Physical activity	38.80 ± 6.30	36.87 ± 4.94	**< 0.001**

Note. *Bold figures represent statistically significant association ($P < 0.05$). Independent sample *t* test was used for comparison. SBP: Systolic blood pressure; DBP: Diastolic blood pressure; HDL: High-density lipoprotein; LDL: Low-density lipoprotein; TG: Triglyceride; FBS: Fasting blood sugar; SG: Specific gravity; BUN: Blood urea nitrogen; Cr: Creatinine; SGOT: Serum glutamic-oxaloacetic transaminase; SGPT: Serum glutamate pyruvate transaminase; ALP: Alkaline phosphatase; GGT: Gamma-glutamyl transferase.

Sensitivity, specificity, accuracy, and the 95% confidence intervals for the DT and RF models are presented in Table 4. In addition, AUC is reported to compare the two models. The models were evaluated using a confusion matrix on a test dataset (Table 5). The RF model had an accuracy of 86.96%. Of the 2657 non-diabetic individuals in testing datasets, 2439 were correctly classified using the RF, with a specificity of 91.8%. For the 532 diabetic patients in the testing dataset, the RF could correctly classify 334 individuals, with a sensitivity of 62.78%.

The DT model had an accuracy of 81.84%. Of the 2657 non-diabetic individuals in test datasets, 2300 were classified correctly using the DT model, with a specificity of 86.56%. For the 532 diabetic patients in the testing dataset, the DT model could correctly categorize 310 individuals, with a sensitivity of 58.27%.

Discussion

T2DM is a problem that usually cannot be diagnosed before its complications[24]; thus, finding and evaluating prediction models based on risk factors are important issues.[25] In our study, 27 variables were significantly associated with diabetes incidence in univariate analysis. DT and RF models were applied to investigate factors associated with T2DM. These methods are among the machine learning approaches that are preferred to the classical methods for many reasons, such as the ability to handle nonlinear associations, multilevel interactions between variables, simple interpretation,

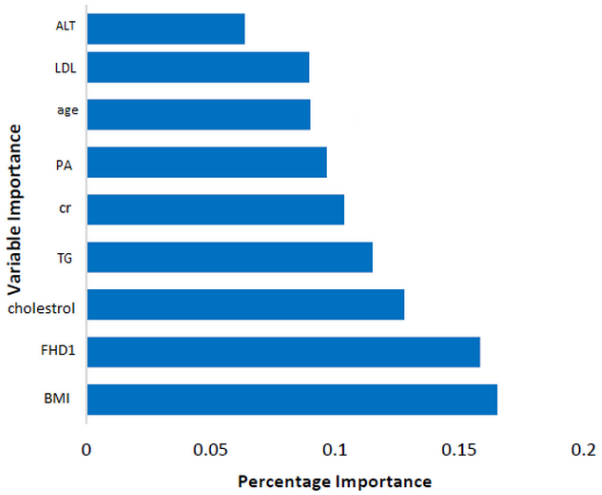

Figure 1. The Importance of Input Variables in the Random Forest Model

and the generation of rules.[26] Additionally, they are more applicable in public health program settings.[27] DTs are practically effortless to use models[28] and robust statistical methods for classification and prediction that have many applications in medical studies.[29] In RF modeling, numerous classification trees are assembled by selecting random training datasets and random sets of variables. Finally, to provide a prediction for each observation, the result of each tree is combined. Because of the modeling process in an RF, it is usually more accurate than a single DT model.[28]

The findings of this study highlight critical risk factors associated with T2DM through the application of DT and RF modeling techniques. Both models identified a range of factors with notable overlaps, providing a comprehensive understanding of the determinants of T2DM.

The DT and RF models identified several common factors, including TG, physical activity, BMI, and LDL cholesterol. The significance of these factors aligns with existing literature, which emphasizes their role in the pathophysiology of T2DM. For instance, elevated TG and LDL levels are associated with insulin resistance and metabolic syndrome, which are vital contributors to the development of diabetes.[30,31] Physical activity is also a well-established protective factor, as it enhances insulin sensitivity and aids in weight management.[32] While there was an overlap, the distinct outputs of each model provided further insights. The DT model highlighted age, SBP, FHD1, Cr, and DBP as significant factors. Conversely, the RF model emphasized cholesterol and ALT levels. This divergence may reflect the strengths of each modeling approach—DTs offer clear visualizations of decision paths. Simultaneously, RFs can capture complex interactions and nonlinear relationships among variables.[33] Numerous previous studies support the association between various risk factors and the development of diabetes. For example, in a study exploring the association between liver enzymes and the risk of T2DM, serum ALT concentrations were independently

Table 3. The Extracted Decision Rules Obtained From RF and DT Models

Model	Class	
	The Person With Diabetes (Probability)	The Person Without Diabetes (Probability)
RF Model		
r1: BMI > 23.5 & FHD1 = no & cholesterol < 253	-	181.236 (77%)
r2 BMI > 23.5 & FHD1 = no & cholesterol > 253 & TG < 199	-	50.80 (62.5%)
r3: BMI > 23.5 & FHD1 = yes & cholesterol < 253 & TG < 199	386.750 (51.5%)	-
r4: BMI > 23.5 & FHD1 = yes & cholesterol > 253 & TG > 199	119.199 (60%)	-
R5: BMI < 23.5 & FHD1 = no & Cr < 1.7 & TG < 199	-	589.755 (78%)
R6: BMI < 23.5 & FHD1 = yes & physical activity < 30 & age > 45	20.38 (53%)	-
R7: BMI < 23.5 & FHD1 = yes & TG > 220 & physical activity > 30 & age < 45	-	37.47 (79%)
R8: BMI < 23.5 & physical activity < 30 & age > 45 & LDL < 110	-	27.32 (0.84%)
R9: BMI < 23.5 & physical activity < 30 & age > 45 & LDL > 110 & FHD1 = 1	7.11 (0.64%)	-
R10: BMI < 23.5 & physical activity < 30 & age > 55 & Cr > 1.7 & ALT > 37	10.14 (0.71%)	
R11: BMI < 23.5 & physical activity < 30 & age > 55 & Cr < 1.7 & ALT > 37	-	9.17 (53%)
R12: BMI < 23.5 & physical activity > 30 & age < 55 & ALT < 37 FHD1 = no	-	8.9 (88.8%)
R13: BMI < 23.5 & physical activity > 30 & age < 55 & ALT < 37 FHD1 = yes	-	8.10 (80%)
DT model		
d1: Age < 49 & BMI < 24.8	-	3604.3921 (92%)
d2: Age < 49 & BMI > 24.8 & physical activity > 33	-	53.243 (22%)
d3: Age < 49 & BMI > 24.8 & physical activity < 33 & Cr < 1.3	-	89.198 (45%)
d4: Age < 49 & BMI > 24.8 & physical activity < 33 & Cr > 1.3 & FHD = yes	45.79 (57%)	-
d5: Age < 49 & BMI > 24.8 & physical activity < 33 & Cr > 1.3 & FHD = no	100.198 (51%)	-
d6: Age > 49 & TG > 173 & SBS > 12 & FHD = yes & DBP > 83	9.14 (65%)	-
d7: Age > 49 & TG > 173 & SBS > 12 & FHD = yes & DBP < 83	9.17 (53%)	-
d8: Age > 49 & TG > 173 & SBS > 12 & FHD = no	-	56.149(38%)
d9: Age > 49 & TG > 173 & SBS < 12	-	34.93 (37%)
d9: Age > 49 & TG < 173 & FHD1 = yes & LDL > 110	12.21 (57%)	-
d10: Age > 49 & TG < 173 & FHD1 = yes & LDL < 110	-	31.85 (36%)
d11: Age > 49 & TG < 173 & FHD1 = no	-	9.36 (23%)

Note. T2DM: Type 2 diabetes mellitus; RF: Random forest; DT: Decision tree; BMI: Body mass index; FHD1: First-degree family history of T2DM; FHD2: Second-degree family history of T2DM; TG: Triglyceride; LDL: Low-density lipoproteins; ALT: Alanine transaminase; Cr: Creatinine.

associated with T2DM in both genders.[34] According to the findings of Moon et al, Cr was considered a risk factor for T2DM.[35] BMI and age emerged as the most critical risk factors in RF and DT models, respectively, which is consistent with the findings of other studies identifying obesity and older age as primary predictors of T2DM.[36,37] As populations continue to age and obesity rates rise globally, understanding the implications of these factors is crucial for public health strategies aimed at diabetes prevention. The results of this study conform to those of previous research that utilized DT modeling to explore diabetes risk factors. Some studies have consistently reported BMI, age, and family history as significant predictors of T2DM.[38,39]

Sensitivity and specificity values were used to compare the two mentioned models. The sensitivity and specificity are significant indices that are used for model validity.[40] The two values were higher for the RF model, which is consistent with the results of other studies.[41,42] AUC is an even better performance criterion for predictive evaluation than accuracy.[43] In our study, the AUC of the RF model for testing the dataset was significantly higher than that of the DT, which matches the findings of another study.[5]

Two essential strengths of our study were modeling a large sample with a relatively large number of variables and rendering a new vision in using data mining methods for exploring potential associated risk factors of T2DM.

While we created an accurate predictive model for T2DM using data mining techniques and identified potential risk factors, it is noted that not all criteria for predictive models are typically met. Critical criteria for predictive models involve incorporating all clinically relevant factors, the logical coherence for clinicians who will use it, and the necessity for the model to be tested on independent samples.[44] Furthermore, limiting our sample to a single city in Iran poses another constraint on our study. In other words, our model would be more representative if the data were collected from various regions across the country.

Identifying these risk factors carries significant

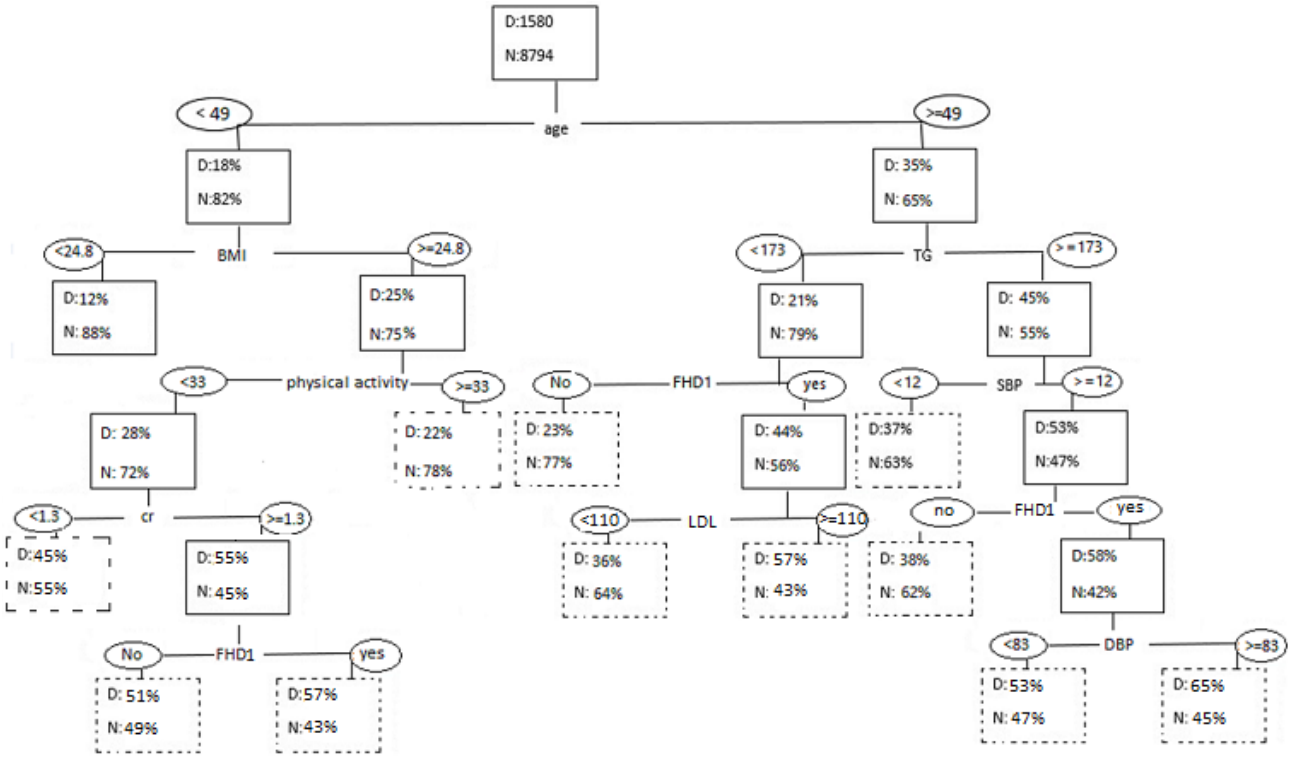

Figure 2. The Decision Tree Model for the Training Data

Table 4. Sensitivity, Specificity, Accuracy, AUC (%), and the 95% CIs of the Models

Model	Sensitivity% (CI)	Specificity% (CI)	Accuracy% (CI)	AUC (CI)	*P* Value*
DT	58.27 (53.95, 62.5)	86.56 (85.21, 87.84)	81.84 (80.46, 83.17)	79.1 (76.8, 80.8)	0.002
RF	62.78 (58.52, 66.9)	91.80 (90.69, 92.81)	86.96 (85.74, 88.11)	66.8 (63.4, 68.21)	

Note. DT: Decision tree; RF: Random forest; CI: Confidence interval; AUC: Area under the receiver operating characteristic curve. *DeLong's test for comparing the AUC.

Table 5. Confusion Matrix of Testing Dataset for the RF and DT Models

Model	Actual Outcome	Predicted Outcome	
		Diabetics (n)	Non-diabetics (n)
RF Model	Diabetics	334	198
	Non-diabetics	218	2439
DT Model	Diabetics	310	222
	Non-diabetics	357	2300

implications for clinical practice and public health initiatives. Targeted interventions focusing on lifestyle modifications, such as increased physical activity and weight management, could help mitigate the risk of T2DM. Furthermore, public health campaigns to raise awareness about maintaining healthy cholesterol and TG levels may be beneficial.

Future research should continue to explore the interactions between these risk factors and their cumulative effects on diabetes risk. Longitudinal studies could provide deeper insights into how these factors influence T2DM development.

Conclusion

In this study, some models were proposed for determining the risk factors of T2DM using two data mining methods that did not need laboratory tests. The application of DT and RF models has elucidated vital risk factors associated with T2DM, emphasizing the importance of BMI, age, TG, physical activity, and LDL. The findings of this study, in conjunction with those from related research, can be utilized to improve the diagnostic processes for T2DM and mitigate the complications arising from delayed diagnosis. This integration of knowledge can significantly enhance clinical practices and patient outcomes in managing T2DM.

Acknowledgements

The authors are grateful to the officers and data management staff of the Kharameh cohort. This article is part of the Persian cohort study in Kharameh.

Authors' Contribution

Conceptualization: Mayam Jalali.
Data curation: Masoumeh Ghoddusi Johari, Abbas Rezaianzadeh, Hamid Reza Niazkar, and Amir Hossein Saem.
Formal analysis: Mayam Jalali.
Investigation: Mayam Jalali, Masoumeh Ghoddusi Johari, and Abbas Rezaianzadeh.
Methodology: Maryam Jalali, Masoumeh Ghoddusi Johari, Abbas Rezaianzadeh, Hamid Reza Niazkar, and Amir Hossein Saem.
Project administration: Masoumeh Ghoddusi Johari and Abbas Rezaianzadeh.

Resources: Masoumeh Ghoddusi Johari and Abbas Rezaianzadeh.
Software: Mayam Jalali.
Supervision: Masoumeh Ghoddusi Johari and Abbas Rezaianzadeh.
Validation: Masoumeh Ghoddusi Johari and Abbas Rezaianzadeh.
Visualization: Mayam Jalali.
Writing–original Draft: Mayam Jalali.
Writing–review & editing: Maryam Jalali, Masoumeh Ghoddusi Johari, and Abbas Rezaianzadeh.

Competing Interests
The authors report no conflict of interests.

Consent for Publication
Not applicable.

Data Availability Statement
The datasets used and analyzed during the current study are available by emailing the data owner.

Ethical Approval
The Ethics Committee of Shiraz University of Medical Sciences approved the study (ethical code: IR.SUMS.REC.1402.298). The confidentiality of their data was emphasized.

Funding
None.

References
1. Stene LC, Tuomilehto J. Epidemiology of type 1 diabetes. In: Holt RI, Flyvbjerg A, eds. Textbook of Diabetes. John Wiley & Sons; 2024. p. 41-54. doi: 10.1002/9781119697473.ch4.
2. Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: early detection should be focused. Health Sci Rep. 2024;7(3):e2004. doi: 10.1002/hsr2.2004.
3. Liu J, Bai R, Chai Z, Cooper ME, Zimmet PZ, Zhang L. Low- and middle-income countries demonstrate rapid growth of type 2 diabetes: an analysis based on Global Burden of Disease 1990-2019 data. Diabetologia. 2022;65(8):1339-52. doi: 10.1007/s00125-022-05713-6.
4. Mohammadi G, Rostamian Motlagh F, Hassanabadi S, Babakhanian M, Abounoori M, Darbani M, et al. Type 2 diabetes trends in Semnan province of Iran and forecasting until 2025: a time series modeling study in 2015-2020. Middle East J Rehabil Health Stud. 2022;9(2):e119455. doi: 10.5812/mejrh-119455.
5. Pham BT, Prakash I. Evaluation and comparison of LogitBoost ensemble, Fisher's linear discriminant analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. Geocarto Int. 2019;34(3):316-33. doi: 10.1080/10106049.2017.1404141.
6. Ghavidel A, Pazos P. Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review. J Cancer Surviv. 2023. doi: 10.1007/s11764-023-01465-3.
7. Komi M, Jun L, Yongxin Z, Xianguo Z. Application of data mining methods in diabetes prediction. In: 2017 2nd International Conference on Image, Vision and Computing (ICIVC). Chengdu: IEEE; 2017. p. 1006-10. doi: 10.1109/icivc.2017.7984706.
8. Azrar A, Ali Y, Awais M, Zaheer K. Data mining models comparison for diabetes prediction. Int J Adv Comput Sci Appl. 2018;9(8):320-3.
9. Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A. Investigating health-related features and their impact on the prediction of diabetes using machine learning. Appl Sci. 2021;11(3):1173. doi: 10.3390/app11031173.
10. Poustchi H, Eghtesad S, Kamangar F, Etemadi A, Keshtkar AA, Hekmatdoost A, et al. Prospective epidemiological research studies in Iran (the PERSIAN Cohort Study): rationale, objectives, and design. Am J Epidemiol. 2018;187(4):647-55. doi: 10.1093/aje/kwx314.
11. Jalali M, Keshani P, Ghoddusi Johari M, Rezaeianzadeh R, Hosseini SV, Rezaianzadeh A. The association between Index of Nutritional Quality (INQ) and obesity: baseline data of Kharameh cohort. Biomed Res Int. 2022;2022:8321596. doi: 10.1155/2022/8321596.
12. Rezazadeh A, Rashidkhani B. The association of general and central obesity with major dietary patterns of adult women living in Tehran, Iran. J Nutr Sci Vitaminol (Tokyo). 2010;56(2):132-8. doi: 10.3177/jnsv.56.132.
13. do Vale Moreira NC, Hussain A, Bhowmik B, Mdala I, Siddiquee T, Fernandes VO, et al. Prevalence of metabolic syndrome by different definitions, and its association with type 2 diabetes, pre-diabetes, and cardiovascular disease risk in Brazil. Diabetes Metab Syndr. 2020;14(5):1217-24. doi: 10.1016/j.dsx.2020.05.043.
14. Jitraknatee J, Ruengorn C, Nochaiwong S. Prevalence and risk factors of chronic kidney disease among type 2 diabetes patients: a cross-sectional study in primary care practice. Sci Rep. 2020;10(1):6205. doi: 10.1038/s41598-020-63443-4.
15. Zhang Y, Yang J, Ye J, Guo Q, Wang W, Sun Y, et al. Separate and combined associations of physical activity and obesity with lipid-related indices in non-diabetic and diabetic patients. Lipids Health Dis. 2019;18(1):49. doi: 10.1186/s12944-019-0987-6.
16. Patel HH, Prajapati P. Study and analysis of decision tree based classification algorithms. Int J Comput Sci Eng. 2018;6(10):74-8.
17. Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh A. A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. J Res Health Sci. 2018;18(2):e00412.
18. Liu Y, Wang Y, Zhang J. New machine learning algorithm: random forest. In: Liu B, Ma M, Chang J, eds. Information Computing and Applications. ICICA 2012. Vol 7473. Berlin, Heidelberg: Springer; 2012. p. 246-52. doi: 10.1007/978-3-642-34062-8_32.
19. Golden CE, Rothrock MJ Jr, Mishra A. Comparison between random forest and gradient boosting machine methods for predicting *Listeria* spp. prevalence in the environment of pastured poultry farms. Food Res Int. 2019;122:47-55. doi: 10.1016/j.foodres.2019.03.062.
20. Pal K, Patel BV. Data classification with K-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). Erode, India: IEEE; 2020. p. 83-7. doi: 10.1109/iccmc48092.2020.iccmc-00016.
21. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol. 2022;75(1):25-36. doi: 10.4097/kja.21209.
22. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ. 2013;185(11):E537-44. doi: 10.1503/cmaj.121286.
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45.
24. Chang W, Liu Y, Xiao Y, Yuan X, Xu X, Zhang S, et al. A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics (Basel). 2019;9(4):178. doi: 10.3390/diagnostics9040178.
25. Serbis A, Giapros V, Kotanidou EP, Galli-Tsinopoulou A, Siomou E. Diagnosis, treatment and prevention of type

2 diabetes mellitus in children and adolescents. World J Diabetes. 2021;12(4):344-65. doi: 10.4239/wjd.v12.i4.344.
26. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013;29(2):93-9. doi: 10.1016/j.kjms.2012.08.016.
27. Scott IA. Machine learning and evidence-based medicine. Ann Intern Med. 2018;169(1):44-6. doi: 10.7326/m18-0115.
28. Mittal S, Hasija Y. Applications of deep learning in healthcare and biomedicine. In: Dash S, Acharya BR, Mittal M, Abraham A, Kelemen A, eds. Deep Learning Techniques for Biomedical and Health Informatics. Vol 68. Cham: Springer; 2020. p. 57-77. doi: 10.1007/978-3-030-33966-1_4.
29. Speiser JL, Durkalski VL, Lee WM. Random forest classification of etiologies for an orphan disease. Stat Med. 2015;34(5):887-99. doi: 10.1002/sim.6351.
30. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130-5. doi: 10.11919/j.issn.1002-0829.215044.
31. Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature. 2006;444(7121):840-6. doi: 10.1038/nature05482.
32. Grundy SM. Metabolic syndrome: connecting and reconciling cardiovascular and diabetes worlds. J Am Coll Cardiol. 2006;47(6):1093-100. doi: 10.1016/j.jacc.2005.11.046.
33. Colberg SR, Sigal RJ, Yardley JE, Riddell MC, Dunstan DW, Dempsey PC, et al. Physical activity/exercise and diabetes: a position statement of the American Diabetes Association. Diabetes Care. 2016;39(11):2065-79. doi: 10.2337/dc16-1728.
34. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi: 10.1023/a:1010933404324.
35. Ahn HR, Shin MH, Nam HS, Park KS, Lee YH, Jeong SK, et al. The association between liver enzymes and risk of type 2 diabetes: the Namwon study. Diabetol Metab Syndr. 2014;6(1):14. doi: 10.1186/1758-5996-6-14.
36. Moon JS, Lee JE, Yoon JS. Variation in serum creatinine level is correlated to risk of type 2 diabetes. Endocrinol Metab (Seoul). 2013;28(3):207-13. doi: 10.3803/EnM.2013.28.3.207.
37. Kahn SE, Cooper ME, Del Prato S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. Lancet. 2014;383(9922):1068-83. doi: 10.1016/s0140-6736(13)62154-6.
38. Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, et al. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med. 2001;345(11):790-7. doi: 10.1056/NEJMoa010492.
39. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract. 2019;157:107843. doi: 10.1016/j.diabres.2019.107843.
40. Alromaihi D, Bastawrous M, Bastawrous D. Factors affecting glycemic control among patients with type 2 diabetes in Bahrain. Bahrain Med Bull. 2019;41(3):146-9.
41. Ho WH, Lee KT, Chen HY, Ho TW, Chiu HC. Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network. PLoS One. 2012;7(1):e29179. doi: 10.1371/journal.pone.0029179.
42. Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. BMJ. 2012;345:e5900. doi: 10.1136/bmj.e5900.
43. Wang CJ, Li YQ, Wang L, Li LL, Guo YR, Zhang LY, et al. Development and evaluation of a simple and effective prediction approach for identifying those at high risk of dyslipidemia in rural adult residents. PLoS One. 2012;7(8):e43834. doi: 10.1371/journal.pone.0043834.
44. Jaskowiak PA, Costa IG, Campello RJ. The area under the ROC curve as a measure of clustering quality. Data Min Knowl Discov. 2022;36(3):1219-45. doi: 10.1007/s10618-022-00829-0.