J Arch Mil Med. 2022 June; 10(2):e118525. https://doi.org/10.5812/jamm-118525.

Published online 2022 June 12. Research Article

A Comparison of Three Research Methods: Logistic Regression, Decision Tree, and Random Forest to Reveal Association of Type 2 Diabetes with Risk Factors and Classify Subjects in a Military Population

Mohammad Sahebhonar[1], Mehrzad Gholampour Dehaki[2], Mohammad Hassan Kazemi-Galougahi[3] and Saeed Soleiman-Meigooni[1, *]

[1]Infectious Diseases Research Center, AJA University of Medical Sciences, Tehran, Iran
[2]Department of Internal Medicine, Faculty of Medicine, AJA University of Medical Sciences, Tehran, Iran
[3]Department of Social Medicine, Faculty of Medicine, AJA University of Medical Sciences, Tehran, Iran

[*] *Corresponding author*: Infectious Diseases Research Center, AJA University of Medical Sciences, Tehran, Iran. Email: dr.saeed.meigooni@gmail.com

Received 2021 August 07; **Accepted** 2022 April 27.

Abstract

Background: Type 2 diabetes mellitus (T2DM) is one of the major non-communicable diseases, causing morbidity and mortality worldwide. There is no study on T2DM status in Iran Army Forces.
Objectives: We aimed to measure the prevalence of T2DM in this population and identify variables associated with T2DM risk in order to classify individuals.
Methods: Data from 3661 Iran Army Ground Forces were employed. Characteristics of the subjects with and without T2DM were compared. We examined the classification ability of logistic regression with two tree-based supervised learning algorithms, decision tree and random forest (RF). The ethical committee of AJA University of Medical Sciences approved this study by the approval code 995685.
Results: The prevalence of T2DM was 3% less than in the general population. Our results showed that the incidence of T2DM increases as subjects become older. The proportions of staff members with T2DM were more than the other military ranks. T2DM is more common in obese and overweight groups. The highest prevalence of T2DM is in the subjects with high levels of lipid profile. The areas below the receiver operating characteristic curve for logistic regression, decision tree, and RF were 73.8%, 77.1%, and 97.1%, respectively.
Conclusions: Age, body mass index, total cholesterol, low-density lipoprotein cholesterol, and triglyceride are associated with T2DM risk. The RF has superior classification performance in comparison with logistic regression and decision tree.

Keywords: Diabetes, Epidemiology, Risk Management, Qualitative Research

1. Background

Diabetes is one of the most chronic health challenges (1). In 2014, the global prevalence of diabetes was 8.5% in the adult population (2). In Iran, it is estimated that about 4.5 to 5.5 million people (about 7% of the general population) have diabetes, which has been increasing during the past decades (3).

Diabetes is one of the top ten causes of death (4). Type 2 diabetes mellitus (T2DM) accounts for more than 90% of all diabetes and is largely preventable (5). Based on the latest report, the diabetes prevalence rate in Iran was 11.4% in adults aged 25 - 70 years (6). Therefore, a large number of adults in Iran have diabetes.

Military personnel is recruited from a relatively healthy population. However, they are not immune to diseases. Diabetes should be less prevalent in these communities due to their particular lifestyle. Some studies have shown that armed conditions have increased the risk of T2DM (7, 8). In order to design prevention interventions and provide better healthcare services, it is necessary to estimate diabetes prevalence and its potential risk factors in military personnel.

There are several class methods that can be used for data with binary outcome variables. Logistic regression is a non-linear parametric predictive model widely used in diabetes studies (7, 9-11). Due to the complex interaction among predictors, there has been an increase in the use of model-free machine learning algorithms. Yu et al. (12) applied a support vector machine (SVM) model to classify subjects with diabetes and pre-diabetes. Khalilia et al. (13) compared SVM, bagging, boosting, and random forest (RF) to predict the risk of several chronic diseases, including diabetes. Casanova et al. (14) examined RF performance relative to logistic regression to classify diabetic retinopathy participants. Uemura et al. (15) investigated unknown factors associated with T2DM using an alternating decision tree.

2. Objectives

To date, very few studies have been done regarding T2DM prevalence in Iran Army Forces. This study was carried out to estimate the prevalence of T2DM in Iran Army Ground Forces and to measure the T2DM rate in the study population subgroups. We hypothesized that a military lifestyle contributes to the risk of T2DM. The next aim was to identify the T2DM risk factors in the population in order to accurately classify patients. In this study, we used three research methods consisting of classic logistic regression and two modern classification algorithms, including decision tree and RF. We discuss the issues of choosing a classification algorithm in relation to the results.

3. Methods

3.1. Study Sample

In this cross-sectional study, we employed a representative sample of data from the Iran Ground Army Forces Health Examination Center for 3661 subjects. Independent demographic and clinical variables included age, military rank (Rank), body mass index (BMI), fast plasma glucose (FPG), total cholesterol (TCL), low-density lipoprotein cholesterol (LDL), and triglyceride (TG) (Table 1).

Table 1. Description of Variables

No.	Symbol	Definition	Unit
1	Age	Age	Year
2	Rank	Army rank	Individual
3	BMI	Body mass index	kg/m^2
4	FPG	Fast plasma glucose	mg/dL
5	TCL	Total cholesterol	mg/dL
6	LDL	Low-density lipoprotein	mg/dL
7	TG	Triglyceride	mg/dL

Subjects were identified as having T2DM if their FPG was greater than 125 mg/dL (16). Military rank was considered as an indicator of socioeconomic status, and was categorized into three groups of staff, conscripts (juniors and non-commissioned officers), and officers. BMI was calculated as weight (kg) divided by the square of height (m^2).

3.2. Descriptive Study

The distribution of the subjects was examined by age groups, rank, BMI, and various levels of TCL, LDL, and TG. Age groups were determined based on age quantiles. Subjects were categorized in three BMI strata as normal with BMI < 25 kg/m^2, overweight with BMI range between 25 and 30 kg/m^2, and obese with BMI $\geq$ 30 kg/m^2. A total cholesterol level of under 200 mg/dL was ideal. A level between 200 to 239 mg/dL was in borderline class, and more than 240 mg/dL was at high-risk. LDL level lower than 100 mg/dL was ideal, between 100 and 129 mg/dL was close to ideal, between 130 and 159 mg/dL was in borderline class, and more than 160 was elevated. The ideal level of TG was lower than 150 mg/dL, and the borderline was between 150 and 199 mg/dL. A level over 200 mg/dL was known to be high. Cholesterol classes were defined based on the U.S. National Institute of Health Guide. Mean age, BMI, TCL, LDL, and TG were calculated and compared using the *t*-test statistics.

3.3. Analytical Study

3.3.1. Statistical Analyses

To explore the effects of risk factors on T2DM, we considered a classic binary multiple logistic regression model as well as two modern supervised machine learning algorithms, including decision tree and RF. We defined dependent variable y = 1 for T2DM and y = 1 for control subjects. For multiple logistic regression, the below equation was used:

$$\pi\ (x) = \frac{1}{1 + exp\left(-\left(\beta_0 + \Sigma_{i=1}^{p}\ \beta_i X_i\right)\right)} \quad (1)$$

Where, π (x) is the probability that y = 1 for a given value of independent variables X_is, β_0 is intercept, and β_is are regression coefficients. To find the most parsimonious model, we used a backward stepwise variable selection. Akaike Information Criterion was applied to assess the importance of each factor on the goodness of fit.

Compared to parametric logistic regression, nonparametric tree-based methods do not require a predefined relationship between dependent and independent variables. A decision tree consists of hierarchical nodes formed by binary recursive partitioning of the data set into one independent variable at a time. Partitioning occurs based on the Gini impurity index. The results are represented graphically as a decision tree (17).

Random forest (18) is an ensemble classifier that composes of many decision trees (ntree), and each tree is constructed of a bootstrap sample of variables (mtry) and observations. Each tree generates a classification. Based on all the trees, the forest selects the classification with the most votes (19). We set $ntree$ to 1000 and run RF for different $mtry$ values to classify T2DM. RF gives variable importance according to the degree of association between a dependent variable and observations.

Data were split into training and testing partitions with a ratio of 70 to 30%. Due to the rare occurrence of T2DM, this data set is class-imbalanced. Using imbalanced data in most classifiers will produce models with high accuracy but low prediction performance for the minority class. To deal with imbalanced data, we used Synthetic Minority Over-sampling Technique (SMOTE) (20) in the training set. SMOTE created artificial data for the minority training set based on randomly chosen samples from the *k* nearest minority class neighbors. To overcome the overfitting problem and increase the predictive performance of models on the testing set, we performed 10-fold cross-validation with three repeats for analyzing the training set.

3.3.2. Algorithm Evaluation

In order to assess classifier performance, we compared the accuracy, sensitivity, and specificity metrics according to confusion matrix (21). The area under the receiver operating characteristic curve (AUC) was computed to evaluate the overall performance of the three classifiers. We performed all calculations and statistical analyses using the R software (22) and the packages caret (21), DMwR (23), MASS, rpart (24), rpart.plot (25), RF (26), and ggplot2 (27).

3.4. Approval Code

The Ethical committee of AJA University of Medical Sciences approved this study by the approval code 995685.

4. Results

Of the 3661 subjects, 517 were excluded due to missing values for one or more variables or measured values being outside the variable's range. The main analysis was performed for 3144 samples.

4.1. Descriptive Results

In the study population, the mean age was 36.1 ± 7 years, the mean BMI was 25.8 ± 3.2 kg/m^2, the mean FPG was 91.2 ± 19.5 mg/dL, the mean TCL was 173.6 ± 34.1 mg/dL, the man LDL was 103.8 ± 29.2 mg/dL, and the mean TG was 131.7 ± 58.5 mg/dL. Data set consisted of 1412 (44.9%) officers, 1121 (35.7%) conscripts, and 611 (19.4%) staff members. Also, 94 (3%) subjects from 3144 samples (3%) were found to have T2DM.

The results showed that T2DM patients had a significantly higher mean age, BMI, FPG, TCL, and TG (Table 2). There was no significant difference in the mean LDL levels. Prevalence of T2DM increased as subjects became older. The ratio of staff members with T2DM was more than the other ranks. T2DM is more common in obese and overweight groups. The highest prevalence of T2DM was in the subjects with high levels of TCL, LDL, and TG (Table 3).

Table 2. Comparison of Mean (SD) of the Variables Between Individuals with or Without Type 2 Diabetes Mellitus (T2DM)

Characteristic	T2DM		P-Value [a]
	Yes	No	
Age	41.9 (6.2)	36.0 (6.9)	< 0.001
BMI	27.2 (3.3)	25.7 (3.1)	< 0.001
FPG	176.0 (45.9)	88.6 (9.9)	< 0.001
TCL	182.1 (43.1)	173.4 (33.7)	< 0.05
LDL	107.0 (33.1)	103.7 (29.1)	ns [b]
TG	159.8 (78.7)	130.8 (57.6)	< 0.001

Abbreviations: BMI, body mass index; FPG, fast plasma glucose; LDL, low-density lipoprotein cholesterol; TCL, total cholesterol; TG, triglyceride.
[a] P-value obtained from *t*-test
[b] Statistically non-significant

Figure 1 depicts the distribution of rank by age and BMI. For officers, conscripts, and staff members, the mean age was 38.9 ± 6.7, 31.1 ± 3.7, and 39.8 ± 7.2 years and the mean BMI was 26.1 ± 3.2, 25.7 ± 3.3, and 26.3 ± 3.9 kg/m^2, respectively.

4.2. Analytical Results

The original training set was imbalanced. All three classification algorithms were highly biased in prediction toward the majority class. We applied SMOTE to undersample the majority class as well as oversample the minority class in the training set. The features of the original training set and training set after conducting SMOTE are compared in Figure 2.

The stepwise logistic regression model selected six variables of age, rank, BMI, TCL, LDL, and TG as risk factors associated with having T2DM (Table 4). Notably, the results showed that with one year increase in age, we expect a 16% increase in the odds of having T2DM. The odds of incidence of T2DM in officers was 68% less than in staff members. One unit increase in BMI raised the odds of having T2DM by 12%. The logistic regression model had a prediction accuracy of 82.7% (95% confidence interval: 80.1%, 85.1%), a sensitivity of 64.3%, and a specificity of 83.3%. In the testing set, 2.9% had

Table 3. Distribution of Individuals Based on Age Group, Military Rank (Rank), Body Mass Index (BMI), Total Cholesterol (TCL), Low-Density Lipoprotein Cholesterol (LDL) and Triglyceride (TG) [a]

Characteristic	T2DM [b]		Total
	Yes	No	
Age (y) [c]			
19 - 31	6 (0.7)	897 (99.3)	903
32 - 34	8 (1.2)	674 (98.8)	682
35 - 42	29 (3.3)	839 (96.7)	868
43 - 57	51 (7.4)	640 (92.6)	691
Rank			
Officer	49 (3.5)	1363 (96.5)	1412
Conscripts	14 (1.2)	1107 (98.8)	1121
Staff	31 (5.1)	580 (94.9)	611
BMI (kg/m^2)			
Normal: < 25	22 (1.6)	1337 (98.4)	1359
Overweight: 25 - 30	51 (3.4)	1437 (96.6)	1488
Obese: ≥ 30	21 (7.1)	276 (92.9)	297
TCL (mg/dL)			
Ideal: < 200	64 (2.6)	2391 (97.4)	2455
Borderline: 200 - 239	21 (3.7)	546 (96.3)	567
High: ≥ 240	9 (7.4)	113 (92.6)	122
LDL (mg/dL)			
Ideal: < 100	41 (2.7)	1453 (97.3)	1494
Close to ideal :100 - 129	32 (3.2)	966 (96.8)	998
Borderline: 130 - 159	14 (2.6)	524 (97.4)	538
High: ≥ 160	7 (6.1)	107 (93.9)	114
TG (mg/dL)			
Ideal: < 150	50 (2.4)	2057 (97.6)	2107
Borderline: 150 - 199	16 (2.6)	594 (97.4)	610
High: ≥ 200	28 (6.6)	399 (93.4)	427

[a] Subjects were identified as having T2DM if their fast plasma glucose was greater than 125 mg/dL.
[b] Values in parentheses show the proportion of T2DM patients in each sub-category.
[c] Age groups were defined based on age quantiles.

T2DM, of whom 1.9% were correctly detected. The logistic regression model had an AUC of 73.8% (95% confidence interval: 64.7%, 82.9%).

The classification decision tree revealed that age and BMI with interactions between them were the most important predictors that affect T2DM (Figure 3). The decision tree yielded cut-off points of 35 years of age and 25 kg/m^2 for BMI. According to the results, the incidence of T2DM was higher in cases aged ≥ 35 and with a BMI ≥ 25. The final value for max depth was 8, which culminated in the highest accuracy through cross-validation. The accuracy of prediction was 85.8% (95% confidence interval: 83.4%, 87.9%) with a sensitivity 67.8% and a specificity 86.4%. The prevalence of T2DM in test data was 2.9%, and the detection rate in test data was 2.0%. The AUC value was 77.1% (95% confidence interval: 68.2%, 85.9%) for T2DM. In terms of the classification decision tree, we achieved slightly better results than those from the multiple logistic regression model.

The RF identified age (100%) as the most important correlated variable with T2DM. As shown in Figure 4, the elimination of age from the model causes the largest decrease in performance of the model. Other variables ranked based on their relative importance to the age. After the age, BMI (51.3%) had the highest importance. In contrast to the for-

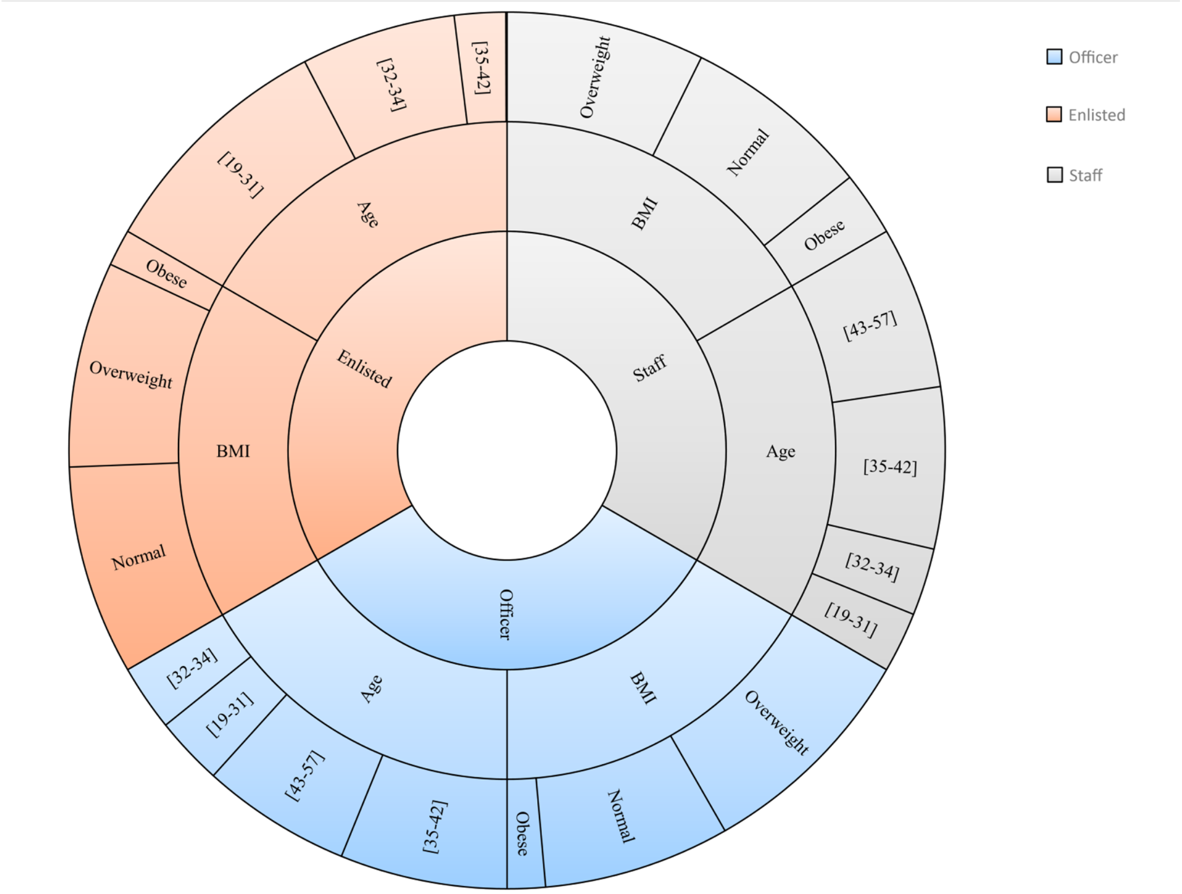

Figure 1. Distribution of rank by age and body mass index (BMI). Normal < 25 kg/m^2, overweight = BMI ≥ 25 and < 30, and obese = BMI ≥ 30

Table 4. Multiple Logistic Regression Analysis for Type 2 Diabetes Mellitus

Characteristic	OR (95% CI)	P-Value [a]
Age	1.16 (1.14 - 1.17)	< 0.001
Rank		
Staff	1.00 (reference)	
Officer	0.63 (0.53 - 0.76)	< 0.001
Conscripts	0.87 (0.66 - 1.14)	ns [b]
BMI	1.12 (1.09 -1.16)	< 0.001
TCL	1.01 (1.00 - 1.02)	< 0.01
LDL	0.99 (0.98 - 1.00)	< 0.01
TG	1.00 (1.00 - 1.00)	< 0.01

Abbreviations: OR, odds ratio; BMI, body mass index; LDL, low-density lipoprotein cholesterol; TCL, total cholesterol; TG, triglyceride.

[a] P-value obtained from multiple logistic regression analysis

[b] Statistically non-significant

mer models, the results of RF showed that LDL (29.7%), TCL (28.8%), and TG (25.8%) were associated with T2DM. The RF had the highest prediction accuracy of 94.4% (95% confidence interval: 92.7%, 95.8%), the sensitivity of 100%, and specificity of 94.2%. All cases in the testing set were detected. The RF yielded the best AUC of 97.1% (95% confidence interval: 96.4%, 97.9%). These results indicated that RF outperformed decision tree and multiple logistic regression.

5. Discussion

The findings in this study showed that, as we believed, the incidence of T2DM is much lower in the study population than the prevalence of T2DM in the general population. Military personnel is chosen according to pre-employment medical tests. In addition, the military lifestyle demands particular conditions, including regular physical activity, more mobility, a healthier dietary program, and periodic medical examination.

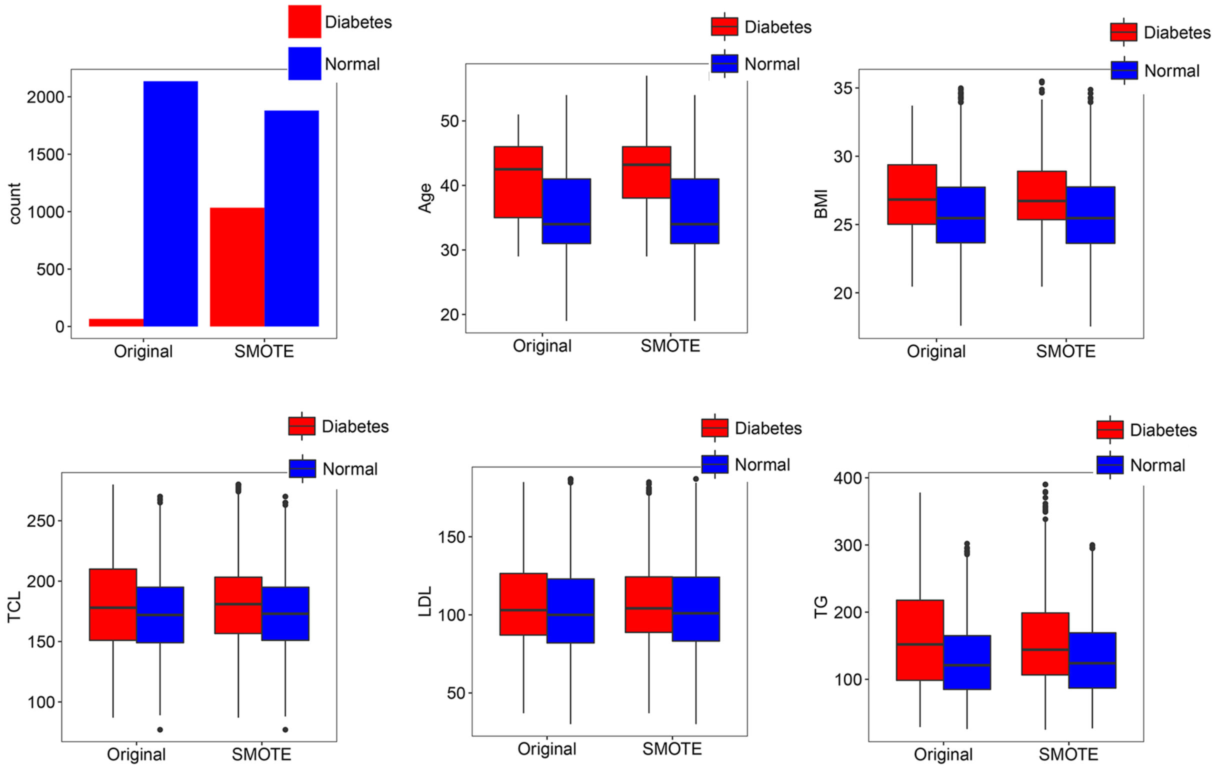

Figure 2. Comparison of the original training set and training set after applying Synthetic Minority Over-sampling Technique (SMOTE) for the number of individuals in each category: Age, body mass index (BMI), total cholesterol (TCL), low-density lipoprotein cholesterol (LDL), and triglyceride (TG). Subjects were identified as having type 2 diabetes mellitus if their fast plasma glucose level was greater than 125 mg/dL.

Previous studies have demonstrated increased T2DM risk associated with physical inactivity (28-30). As the results showed, the mean age and BMI were almost similar between officers and staff members. Therefore, the higher T2DM incidence in staff may confirm physical inactivity and sedentary behaviors in this group. Some studies have reported a stressful lifestyle as a risk factor for T2DM (7, 31). We used rank as a marker for socioeconomic status. However, the lower prevalence of T2DM in conscripts is probably more related to their age and BMI circumstances than to their life status.

The risk of T2DM was categorized into modifiable and non-modifiable factors (32). Among the variables in this study, only age was non-modifiable. Obesity is a well-established risk factor for T2DM (33). The incidence of overweight or obesity in the study population was around 56.8%, and about 2.3% of T2DM individuals were overweight or obese. Kuwahara et al. (11) suggested that preventing weight gain plays an important role in the reduction of T2DM risk. Diabetic dyslipidemia is an abnormal change in lipid profile as a consequence of T2DM (34). Previous studies have illustrated how insulin resistance in T2DM patients causes high TG levels and decreased HDL cholesterol levels (34-36). T2DM individuals have elevated LDL cholesterol levels, but they may not have higher LDL levels (37). Our findings of lipid profiles in T2DM patients are consistent with the reports of the aforementioned studies.

In this study, we evaluated three classification methods, and all of them suffered from the class-imbalanced problem. The use of SMOTE sampling technique was useful to resolve the imbalanced data problem.

Among classification methods, logistic regression has been extensively used in scientific research to measure the association between dependent and independent variables. Logistic regression is a parametric model and works based on a pre-determined set of variables. Therefore, its classification performance depends on the given model. Due to the intricate relationship among underlying features, this method may not have enough power to accurately classify subjects (38).

By contrast, the decision tree is a non-parametric method mainly developed to classify the population rather to test the significance of variables on outcome (39). However, the major drawback of the decision tree is moderate-

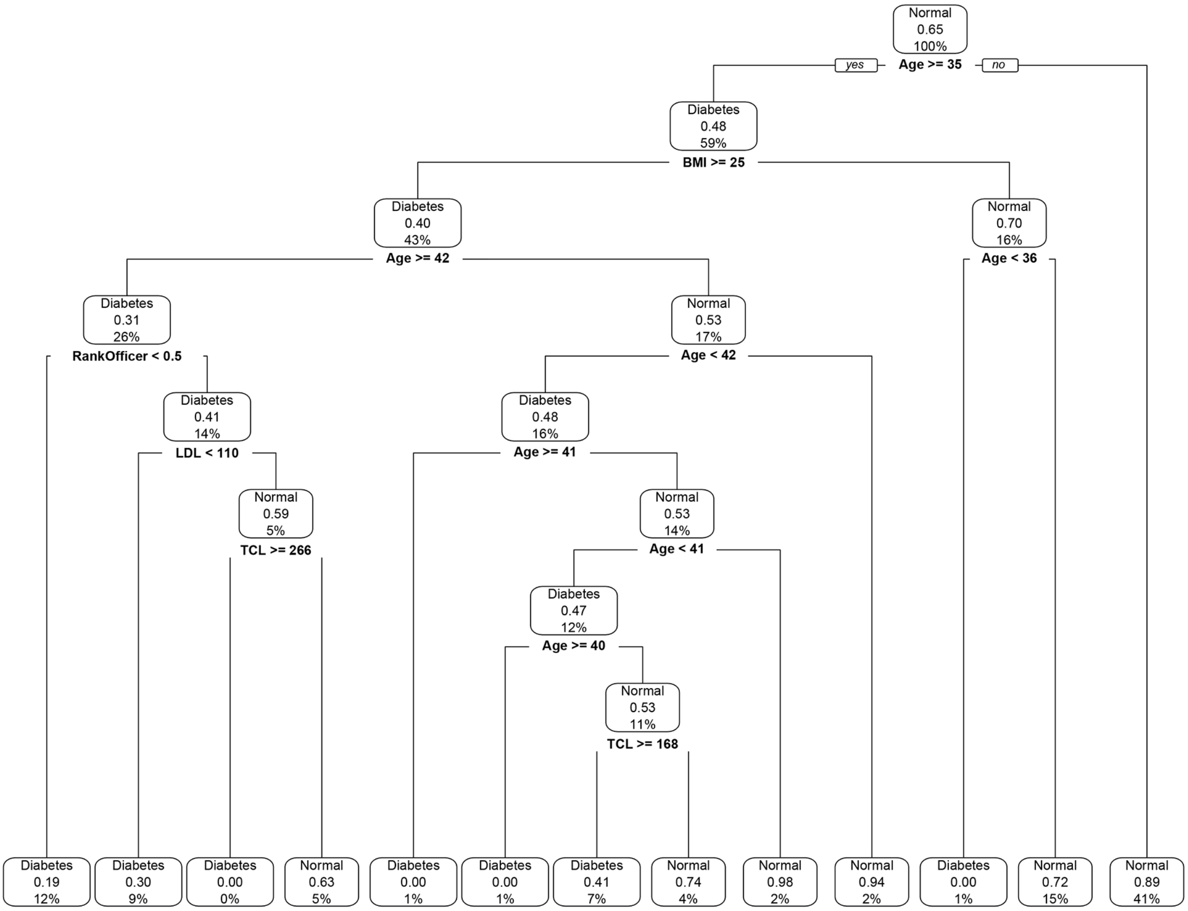

Figure 3. The classification decision tree of demographic and biological risk factors for type 2 diabetes mellitus. Information in each class model includes: Label, the probability of a fitted class, i.e. the correct classification rate at the node, and the percentage of observations that fall in the node. Subjects were identified as having type 2 diabetes mellitus if their fast plasma glucose level was greater than 125 mg/dL. BMI, body mass index; FPG, fast plasma glucose; LDL, low-density lipoprotein cholesterol; TCL, total cholesterol

to-high variance, which is an important cause of decision tree weak performance (40).

RF is also a model-free classification technique, working based on an ensemble of decision trees. The salient feature of RF is low variance due to randomness grown of many trees (41). Consequently, RF is less prone to overfitting and is better in generalization. In addition, RF provides a measure of variable importance, which is more informative than choosing a group of variables that their combination is predictive. Khalilia et al. (13) showed that RF has superior performance compared to SVM, bagging, and boosting methods in disease prediction. Casanova et al. (14) pointed out that the accuracy of RF in classification of diabetic retinopathy participants was much higher than the accuracy of logistic regression.

Typically, the performance of statistical models is assessed using predictive accuracy. However, study of diseases requires a relatively high rate of correct classification of patients. Our results confirm that RF is more powerful in finding complex relations among risk factors. Specifically, with regard to sensitivity and specificity, the RF more correctly classified cases.

Footnotes

Authors' Contribution: Mohammad Sahebhonar contributed to collecting and analyzing the data and writing the primary and final draft of the manuscript. Mehrzad Gholampour Dehaki contributed to the literature review and wrote the primary draft of the manuscript. Mohammad Hassan Kazemi-Galougahi contributed to the literature review and wrote the primary draft of the manuscript. Saeed Soleiman-Meigooni contributed to the literature review, writing the primary and final draft of the manuscript.

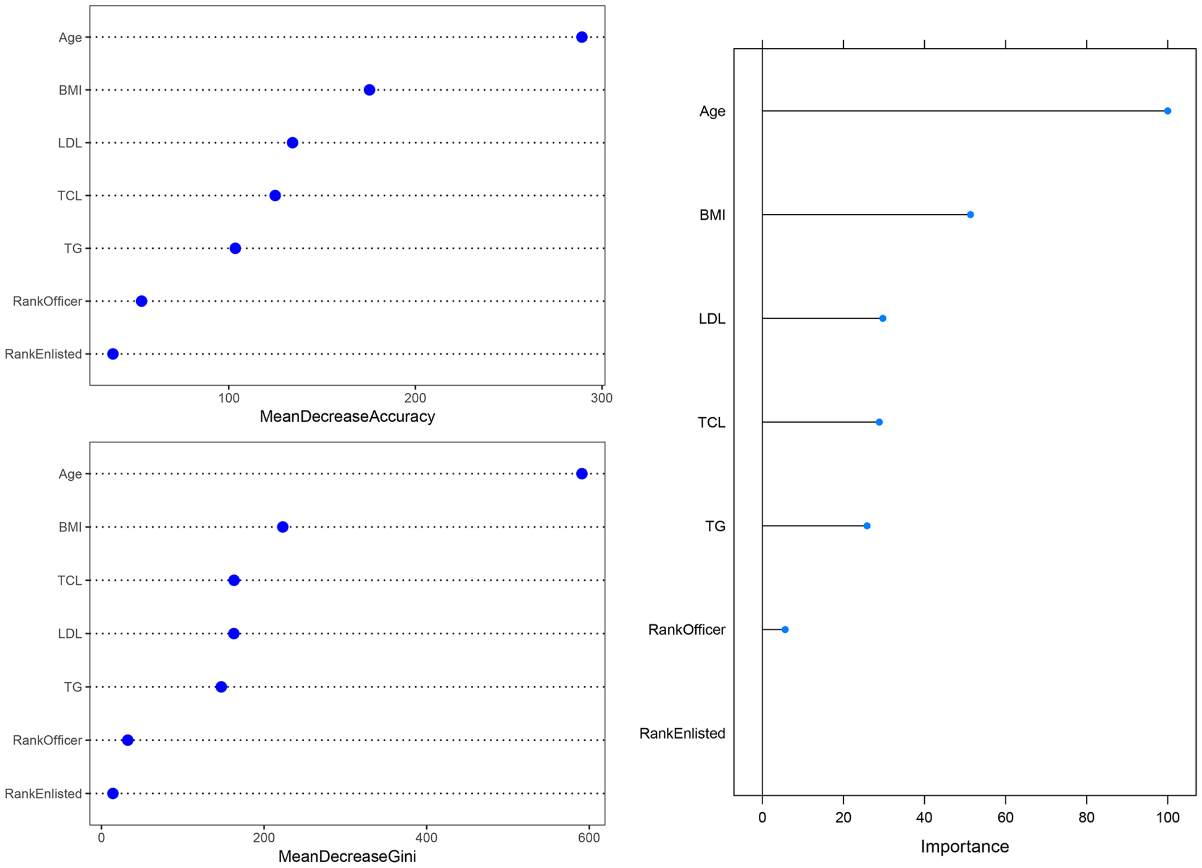

Figure 4. The variable importance in random forest. The upper left figure shows variable importance based on a mean decrease in accuracy, the lower left figure shows variable importance based on a decrease in Gini Index, and the right figure shows overall variable importance. BMI, body mass index; LDL, low-density lipoprotein cholesterol; TCL, total cholesterol; TG, triglyceride.

Conflict of Interests: The authors declared that they do not have any conflict of interest.

Data Reproducibility: The data presented in this study are openly available in one of the repositories or will be available on request from the corresponding author by this journal representative at any time during submission or after publication. Otherwise, all consequences of possible withdrawal or future retraction will be with the corresponding author.

Ethical Approval: The Ethical committee of AJA University of medical sciences approved this study by the approval code 995685 at 2016-06-06.

Funding/Support: This study was supported by the AJA University of Medical Sciences.

References

1. N. C. D. Risk Factor Collaboration. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet*. 2016;**387**(10027):1513–30. [PubMed ID: 27061677]. [PubMed Central ID: PMC5081106]. https://doi.org/10.1016/S0140-6736(16)00618-8.
2. World Health Organization. *Global report on diabetes*. Geneva, Switzerland: World Health Organization; 2016.
3. Ministry of Health and Medical Education. *[7% of the country's population has diabetes / Controlling and preventing diabetes is one of the four goals of global control by 2025]*. Tehran, Iran: Ministry of Health and Medical Education; 2016. Persian. Available from: http://iec.behdasht.gov.ir/index.aspx?fkeyid=&siteid=143&pageid=52371&newsview=130819.
4. World Health Organization. *Fact sheet; the top 10 causes of death*. Geneva, Switzerland: World Health Organization; 2017.
5. American Diabetes Association. 2. Classification and Diagnosis of Diabetes. *Diabetes Care*. 2016;**39 Suppl 1**:S13–22. [PubMed ID: 26696675]. https://doi.org/10.2337/dc16-S005.
6. Esteghamati A, Etemad K, Koohpayehzadeh J, Abbasi M, Meysamie A, Noshad S, et al. Trends in the prevalence of diabetes and impaired fasting glucose in association with obesity in Iran: 2005-2011. *Diabetes Res Clin Pract*. 2014;**103**(2):319–27. [PubMed ID: 24447808]. https://doi.org/10.1016/j.diabres.2013.12.034.
7. Boyko EJ, Jacobson IG, Smith B, Ryan MA, Hooper TI, Amoroso PJ, et al. Risk of diabetes in U.S. military service members in relation to combat deployment and mental health. *Diabetes Care*.

2010;**33**(8):1771–7. [PubMed ID: 20484134]. [PubMed Central ID: PMC2909060]. https://doi.org/10.2337/dc10-0296.

8. Boyko EJ, Seelig AD, Jacobson IG, Hooper TI, Smith B, Smith TC, et al. Sleep characteristics, mental health, and diabetes risk: a prospective study of U.S. military service members in the Millennium Cohort Study. *Diabetes Care*. 2013;**36**(10):3154–61. [PubMed ID: 23835691]. [PubMed Central ID: PMC3781550]. https://doi.org/10.2337/DC13-0042.
9. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet*. 2012;**380**(9836):37–43. [PubMed ID: 22579043]. https://doi.org/10.1016/S0140-6736(12)60240-2.
10. Zou W, Ni L, Lu Q, Zou C, Zhao M, Xu X, et al. Diabetes Onset at 31-45 Years of Age is Associated with an Increased Risk of Diabetic Retinopathy in Type 2 Diabetes. *Sci Rep*. 2016;**6**:38113. [PubMed ID: 27897261]. [PubMed Central ID: PMC5126680]. https://doi.org/10.1038/srep38113.
11. Kuwahara K, Honda T, Nakagawa T, Yamamoto S, Hayashi T, Mizoue T. Body mass index trajectory patterns and changes in visceral fat and glucose metabolism before the onset of type 2 diabetes. *Sci Rep*. 2017;**7**:43521. [PubMed ID: 28266592]. [PubMed Central ID: PMC5339907]. https://doi.org/10.1038/srep43521.
12. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak*. 2010;**10**:16. [PubMed ID: 20307319]. [PubMed Central ID: PMC2850872]. https://doi.org/10.1186/1472-6947-10-16.
13. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak*. 2011;**11**:51. [PubMed ID: 21801360]. [PubMed Central ID: PMC3163175]. https://doi.org/10.1186/1472-6947-11-51.
14. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of random forests methods to diabetic retinopathy classification analyses. *PLoS One*. 2014;**9**(6). e98587. [PubMed ID: 24940623]. [PubMed Central ID: PMC4062420]. https://doi.org/10.1371/journal.pone.0098587.
15. Uemura H, Ghaibeh AA, Katsuura-Kamano S, Yamaguchi M, Bahari T, Ishizu M, et al. Systemic inflammation and family history in relation to the prevalence of type 2 diabetes based on an alternating decision tree. *Sci Rep*. 2017;**7**:45502. [PubMed ID: 28361994]. [PubMed Central ID: PMC5374531]. https://doi.org/10.1038/srep45502.
16. Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet Med*. 1998;**15**(7):539–53. [PubMed ID: 9686693]. https://doi.org/10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S.
17. Loh WY. Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*. 2011;**1**(1):14–23. https://doi.org/10.1002/widm.8.
18. Breiman L. Random Forests. *Machine Learning*. 2001;**45**(1):5–32. https://doi.org/10.1023/a:1010933404324.
19. Breiman L, Cutler A. *Random Forests - classification description*. Berkeley, California, USA: University of California, Berkeley; 2007. Available from: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;**16**:321–57. https://doi.org/10.1613/jair.953.
21. Kuhn M. Building Predictive Models inRUsing thecaretPackage. *J Stat Softw*. 2008;**28**(5). https://doi.org/10.18637/jss.v028.i05.
22. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Core Team; 2015.
23. Torgo L. *Functions and data for 'Data Mining with R' R package*. R package; 2013.
24. Therneau T, Atkinson B, Ripley B. *Recursive Partitioning and Regression Trees*. 2017. Available from: https://cran.r-project.org/web/packages/rpart/index.html.
25. Milborrow S. *Plot rpart Models: An Enhanced Version of plot.rpart*. 2017. Available from: https://cran.r-project.org/web/packages/rpart.plot/index.html.
26. Liaw A, Wiener M. *Classification and regression based on a forest of trees using random inputs*. 2015. Available from: https://cran.r-project.org/web/packages/randomForest/index.html.
27. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York, USA; 2009. https://doi.org/10.1007/978-0-387-98141-3.
28. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med*. 2002;**346**(6):393–403. [PubMed ID: 11832527]. [PubMed Central ID: PMC1370926]. https://doi.org/10.1056/NEJMoa012512.
29. Lee IM, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet*. 2012;**380**(9838):219–29. [PubMed ID: 22818936]. [PubMed Central ID: PMC3645500]. https://doi.org/10.1016/S0140-6736(12)61031-9.
30. Aune D, Norat T, Leitzmann M, Tonstad S, Vatten LJ. Physical activity and the risk of type 2 diabetes: a systematic review and dose-response meta-analysis. *Eur J Epidemiol*. 2015;**30**(7):529–42. [PubMed ID: 26092138]. https://doi.org/10.1007/s10654-015-0056-z.
31. Nyberg ST, Fransson EI, Heikkila K, Ahola K, Alfredsson L, Bjorner JB, et al. Job strain as a risk factor for type 2 diabetes: a pooled analysis of 124,808 men and women. *Diabetes Care*. 2014;**37**(8):2268–75. [PubMed ID: 25061139]. [PubMed Central ID: PMC4113178]. https://doi.org/10.2337/dc13-2936.
32. Chen L, Magliano DJ, Zimmet PZ. The worldwide epidemiology of type 2 diabetes mellitus–present and future perspectives. *Nat Rev Endocrinol*. 2011;**8**(4):228–36. [PubMed ID: 22064493]. https://doi.org/10.1038/nrendo.2011.183.
33. Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, et al. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA*. 2003;**289**(1):76–9. [PubMed ID: 12503980]. https://doi.org/10.1001/jama.289.1.76.
34. Wu L, Parhofer KG. Diabetic dyslipidemia. *Metabolism*. 2014;**63**(12):1469–79. [PubMed ID: 25242435]. https://doi.org/10.1016/j.metabol.2014.08.010.
35. Chehade JM, Gladysz M, Mooradian AD. Dyslipidemia in type 2 diabetes: prevalence, pathophysiology, and management. *Drugs*. 2013;**73**(4):327–39. [PubMed ID: 23479408]. https://doi.org/10.1007/s40265-013-0023-5.
36. Schofield JD, Liu Y, Rao-Balakrishna P, Malik RA, Soran H. Diabetes Dyslipidemia. *Diabetes Ther*. 2016;**7**(2):203–19. [PubMed ID: 27056202]. [PubMed Central ID: PMC4900977]. https://doi.org/10.1007/s13300-016-0167-x.
37. Cowie CC, Howard BV, Harris MI. Serum lipoproteins in African Americans and whites with non-insulin-dependent diabetes in the US population. *Circulation*. 1994;**90**(3):1185–93. [PubMed ID: 8087927]. https://doi.org/10.1161/01.cir.90.3.1185.
38. Garcia-Magarinos M, Lopez-de-Ullibarri I, Cao R, Salas A. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. *Ann Hum Genet*. 2009;**73**(Pt 3):360–9. [PubMed ID: 19291098]. https://doi.org/10.1111/j.1469-1809.2009.00511.x.
39. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med*. 2003;**26**(3):172–81. [PubMed ID: 14644693]. https://doi.org/10.1207/S15324796ABM2603_02.
40. Dietterich T, Kong EB. *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. State College, PA 16801, United States: Citeseer; 1995.
41. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics*. 2012;**99**(6):323–9. [PubMed ID: 22546560]. [PubMed Central ID: PMC3387489]. https://doi.org/10.1016/j.ygeno.2012.04.003.