ACTA SCIENTIFIC NUTRITIONAL HEALTH

Volume 3 Issue 9 September 2019

Research Article

Determining Risk Factors for Type 2 Diabetes and a Controlling Model with Data Mining Approach in Women: A Case Study on Patients in Diabetes Clinic in Lolagar Hospital in Tehran

Fatemeh Banazadeh[1]*, Mohammad Ebrahim Mohammad Pourzarandi[2] and Zahra Banazadeh[3]

[1]*Vice Chancellor for Research and Technology, Iran University of Medical Sciences, Tehran, Iran*

[2]*Department of Industrial Management, PHD Associated Professor, Central Branch, Islamic Azad University, Tehran, Iran*

[3]*Internist Manager of Diabetes Clinic in Iran University of Medical Sciences, Iran*

***Corresponding Author:** Fatemeh Banazadeh, Vice Chancellor for Research and Technology, Iran University of Medical Sciences, Tehran, Iran.

Received: July 29, 2019; **Published:** August 13, 2019

DOI: 10.31080/ASNH.2019.03.0410

Abstract

The importance of predictor variables in every approach is obtained by sensitivity analysis. Accordingly, fasting blood sugar and 2-hour blood sugar (postprandial glucose) have been identified as two important variables in women. Glomerular filtration rate and mean arterial blood pressure in women are other variables having been identified in this research as predictors of the development of diabetes in women. In general, the results indicate that the variables of waist circumference and height are equally important in the incidence of diabetes, even ore than fasting glucose. In order to build predictive models, 6 main and conventional methods in data mining (decision trees with C5.0 algorithms, CART, QUEST, MLP algorithm neural network, support vector machine (SVM), Bayesian simple model (Naïve Bayes)) and a conventional model in epidemiological studies (logistic regression) have been used. The results showed that Naïve Bayes model is not sensitive to unbalanced data in both men and women so that its sensitivity in women with unbalanced data to total variables is 78%. Although other methods represented high characteristics of unbalanced data, they have very low sensitivity. Their average sensitivity to unbalanced data with total variables is 34.8% in women. The performance of all 6 methods of classification is comparable to balanced data; they are even better than logistic regression. In this combination, the best performance belongs to decision trees with QUEST algorithm (20 variables in women). Generally, data mining can be used in epidemiological studies for different purposes; the decision tree methods determine non-linear relationships among variables by creating a tree structure and they can help to identify risk factors in certain subgroups by creating a threshold or decision boundary.

Keywords: Diabetes; Data Mining; Decision Tree; Sensitivity Analysis; Naïve Bayes Model; Quest Algorithm; Neural Networks; Support Vector Machine.

Introduction

With the increasing speed of globalization, the tendency towards unhealthy diets, obesity, sedentary lifestyle, and bad habits, the prevalence of chronic non-communicable diseases has been increasing around the world [1]. Diabetes is common as one of the chronic non-communicable diseases almost everywhere in the world. According to the World Health Organization in 2011, 336 million people have diabetes around the world [2]; this number is rising for various reasons. It is estimated to reach 552 million in 2030 [3]. It is expected that this number increase to 9.3% in 2030. Epidemiologic studies from 1993 to 1997 in Iran represent the prevalence of diabetes among the population over 30 years to 7.67% in women in Tehran [4]. Prospective study by Lolagar Hospital diabetic clinic shows that the overall prevalence of diabetes has increased from 12.5% to 17.5% in women in less than 3 years, which is equal to approximately one per cent per annum growth. Diabetes imposes costs to health care systems so that global cost of diabetes in Iran in the year 2010 was about 1.5 billion dollars [5]. Moreover, diabetes has no cure and it does not seem to be possible to cure it in the near future. Like most diseases, the best treatment for this disease is to prevent it. Due to the increasing volume of information in the field of medicine and health in recent years, the use of different methods of data mining has become common. Development of predictor models for various diseases is one of the most important data mining applications in health [6]. Studies having been conducted in the field of diabetes with data mining methods include blood glucose prediction in diabetes therapy [7] selecting important variables to predict the development of diabetes [8], evaluation of risk factors for diabetes [9], genetic study of diabetes [10-12], evaluation of diseases associated with diabetes [13], and choosing the appropriate drug for the treatment of diabetes [14]. This research employs different data such as clinical information

in patient records, disease registries, data in insurance organizations, case-control studies, and data from cohort studies [15]. Consequently, this research aims to determine risk factors for type 2 diabetes in people over 20 years by data mining methods using data from Lolagar Hospital diabetic clinic in Tehran.

Theoretical foundations

With respect to the research subject and method, concepts and definitions are as follows:

- **Diabetes:** Diabetes is a group of common metabolic disorders with different etiological factors that is accompanied by impairment in metabolism of carbohydrates, protein, and fat due to defects in insulin secretion, reduced insulin action, or both [16,17].
- **Type 2 diabetes (non-insulin):** The most common type of diabetes that occurs in old age gradually. The disease begins typically in people with normal metabolism of carbohydrates, it moves to glucose intolerance, and it ends to diabetes. 10 to 20 years before diagnosis of type 2 diabetes, there is reduction in glucose tolerance along with compensatory increases in insulin [18,19].
- **Data Mining:** The remarkable progress in the acquisition and storage of numerical data has led to the establishment of large databases. In this way, data mining aims to acquire new knowledge from large databases that have not been identified already and it seeks to apply their results in decision-making. Moreover, data mining can generate scientific hypotheses based on experimental data [20].

Research background

In 2006, a study employs data from 4 diabetic patients every day for 5 months to predict the determinants of blood sugar in the next day. Predictor variables in data mining model include fasting plasma glucose, dietary intake, and physical activity. Some parts of the data concerns the incidence of diabetes in women over 21 years from Indians who live in Arizona State in America. The research begins by checking 5000 Indian women in 1965; they had been checked every two years. Diabetes database consists of 768 women; 500 cases had not diabetes and 268 cases had diabetes according to World Health Organization criteria. The database contains 8 variables including the numbers of pregnancy, 2-hour blood sugar, diastolic blood pressure, thickness of subcutaneous fat in the arm, amount of insulin after a 2-hours meal, body mass index, history of diabetes in relatives, and age [21]. The result showed that physical activity level were strongly associated with fasting blood sugar and it is the most important variable for prediction of fasting blood sugar rate [22]. The next study had conducted in Iran in 2013. It compared 2 traditional statistical methods including logistic regression and LDA (Fisher Linear Discriminant Analysis) and 4 methods of data mining including neural networks, SVM, FCM (Fuzzy c-means), and RF in order to present a model for prediction of type 2 diabetes [23]. This research was conducted by employing data from a large cross-sectional study to determine the prevalence of diabetes in Iran. The data includes 6500 samples that had been selected by cluster sampling from the original population. In terms of fasting glucose, the samples were classified in three groups of subjects with diabetes, pre-diabetes and without diabetes. The first two groups were merged into one group (diabetic) and the third group was regarded as an independent group (without diabetes). The results showed that characteristics of each six algorithms is higher than 90% but logistic regression sensitivity, LDA, neural networks, RF, FCM, and SVM are respectively equal to 13.3%, 0.6%, 8%, 8%, 33%, and 82%. The accuracy of all algorithms except FCM was higher than 90%. In this regard, SVM model represented the highest accuracy (985.6%), the highest sensitivity (82%), and the highest characteristics (100%).

Statistical population and statistical sample

Target population is all 3-year old and higher patients referring to Lolagar Hospital diabetic clinic. This region has been selected because this hospital has a vast network of health contacts that have a very important role in calling people to study. Moreover, age and sex distribution in the population of the region is consistent with a total population of Tehran and Iran. In addition, the area is covered by Iran University of Medical Sciences that led to facilitate intersectoral collaboration with the executor of plan [24]. The statistical population contains persons living in the families that are under the coverage of West Tehran Health Center and they have files in Lolagar Hospital. After determining the sample size, a full list of covered households was prepared by referring to the hospital offices. In order to achieve one of the research objectives (evaluating the influence of training on life style improvement), the study population has been divided into three parts geographically. The most distant subjects has been appointed as the intervention group (630 subjects) while subjects with average and close distances have been regarded as controls (n = 1370). Interventions have been conducted in three levels of one, two, and three in three dimensions of increasing physical activity, improving the nutrition, and avoiding tobacco production in a comprehensive approach by investigation department of the clinic, volunteers, and authorities in region, schools, and the society. Inclusion criteria are all patients greater than or equal to 20 years who had diabetes in the beginning of the study. Benchmark for a diabetic person is Fasting blood sugar over 126 milligrams per deciliter, two-hour glucose greater than or equal to 200 milligrams per deciliter, or the use of oral hypoglycemic drugs. After the selection process form 3436 persons who were eligible for inclusion in the first and second phases, 1221 persons had excluded from the study, 243 persons were diagnosed with diabetes, and 1972 persons have not diagnosed with diabetes from baseline to the end of the fourth phase. Therefore, the sample size in this study is 2215 subjects consisting of 243 persons with diabetes and 1972 persons without diabetes. Finally, 664 subjects have been remained in the study from all 5001 participants.

Research methodology

This is a secondary data analysis extracted from cohort study data in Lolagar Hospital diabetic clinic. It is a longitudinal study to estimate the incidence and prevalence of metabolic disorders, determine risk factors important non-communicable diseases, change life style to improve and prevent increasing these diseases [24]. The first phase is cross-sectional study to determine the prevalence of cardiovascular disease from February 1999 to August 2001. The second phase is a perspective study from September 2001 ongoing. Information is collected in the same manner as the first phase (cross-sectional phase) that is repeated every 3 years on the same population. Incidence rate of disease is gathered on an annual basis. Collected data includes medical history, anthropometric measurements, ECG, blood biochemical tests, London School of Hygiene Cardiovascular (Rose) Questionnaire and Fontain stage 2, nutritional status and physical activity, smoking status, medical examination for blood pressure, pulse, thyroid examination, drug history, and clinical examinations. Analysis of the data is a part of the data mining process; in this research, the process is performed based on the standard model (CRISP-DM), which is defined in six steps [25].

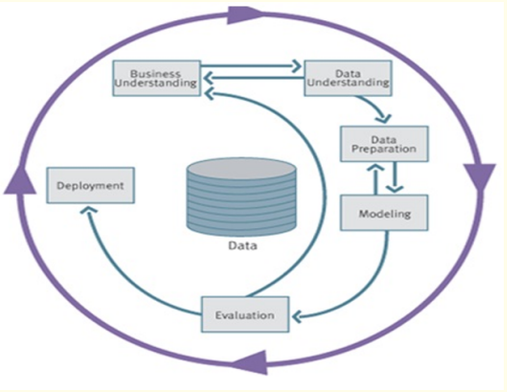

Figure 1: Data mining process based on the standard model of CRISP-DM.

Since this research aims to identify diabetes risk factors in a population without diabetes, the statement is predictive or classification in data mining. There are various approaches to build predictor models in data mining; this study uses neural network, SVM, Naive, Bayes, and decision tree. As the model built by the first three methods can identify only predictive factors and it is not possible to recognize risk factors by this model, three decision tree algorithms are performed to identify the risk factors and their relationship with the use of these algorithms.

Research findings

Diabetes incidence in the population

The rate of diabetes in women by the end of examination is shown in Table 1. According to Table 1, the incidence of diabetes in men is 11%.

Gender	Without diabetes Number (Percent)	Diabetic Number (Percent)	Total Number (Percent)
Female	1114	140 (11%)	1254 (100%)

Table 1. Diabetes incidence rate in the population by gender.

Basic characteristics of the participants in the study

Basic characteristics (quantitative variables) and the number of included women in the examination have been represented in Table 2 after inclusion. According to Table 2, there is a significant difference in both groups of diabetic and non-diabetic persons in terms of quantitative variables. As observed, the average of all variables, except, blood glomerular filtration rate is lower in non-diabetic subjects rather than diabetic subjects.

Findings and comparison of implementing different models on women

In this section, the performance of each model in different combinations is evaluated to determine the combination in which the model has revealed the best performance. Then, the specifications of the models are analyzed in the intended combination. Indexes of accuracy, G-Means, F-Measure, AUC, sensitivity, and characteristic for each model are represented in the following Table 3. According to Table 3, MLP model in the third combination has the highest sensitivity and G-Means. Since the fourth combination (with 20 variables) is very close to the third combination (with 54 variables), MLP model in the fourth combination has the best performance. Table 4 shows that the performance indexes of QUEST model in third and fourth combinations are the same. Considering the number of variables in third combination (54 variables) and fourth combination (20 variables), the fourth combination has the best performance in QUESt model. According to the statements about QUEST model, the fourth combination has the best performance in CART model (Table 5).

Results of Table 6 indicate that NB model with unbalanced data and 20 variables (second combination) and balanced data with 54 variables (third combination) have the best performances. In third combination, the model has the highest sensitivity but the number of variables in this combination is 54; thus, it is not economic to use this model in modeling due to the number of variables. Considering the number of variables and sensitivity as the most important index, the fourth combination of NB has a favorable performance. Results of Table 7 shows that the fourth combination has the best performance in logistic regression. In this combination, balanced data with 20 variables has led to the highest sensitivity, F-measure, AUC, and G-Means.

Findings for comparison of best models in women

In general, the research results show that the fourth combination of each model has the best performance in terms of perfor-

Variable	Total women (1254 persons)		Without Diabetes (1114 persons)		Diabetic (140 persons)		P Value*
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	
Age	39.6	12.3	38.7	12.1	47.1	11.7	0.001
Total length of stay in Tehran (years)	33.1	13.2	32.3	13.0	39.1	13.4	0.001
Two-hour glucose (milligrams per deciliter)	10.8	27.5	105.1	24.5	137.9	31.9	0.001
Blood glomerular filtration rate	63.3	10.8	63.8	10.9	59.4	9.9	0.001
Fasting blood sugar (milligrams per deciliter)	89.0	9.6	87.6	8.4	99.7	11.4	0.001
Wrist circumference (cm)	15.9	1.0	15.9	0.99	16.5	1.0	0.001
Body mass index (kilograms per meter squared)	27.3	4.7	26.9	4.5	30.5	4.9	0.001
Waist-to-hip ratio	0.83	0.1	0.82	0.1	0.88	0.1	0.001
Waist-to-Height Ratio	0.5	0.1	0.54	0.1	0.61	0.1	0.001
Ratio of triglycerides to HDL	3.7	2.9	3.5	2.7	5.2	3.7	0.001
Ratio of total cholesterol to HDL	4.8	1.6	4.7	1.5	5.5	1.7	0.001
Pulse pressure (mmHg)	39.2	12.1	38.6	11.5	44.6	14.5	0.001
Average arterial pressure (mmHg)	89.2	11.7	88.3	11.3	96.6	12.6	0.001

Table 2: Basic characteristics (quantitative variables) in both diabetic and non-diabetic women.

* P value <0.05 for the examination of both diabetic and non-diabetic groups based on t-test.

	First Combination	Second Combination	Third Combination	Fourth Combination	n
Sensitivity	41%	42%	78%	77%	67%
Characteristic	97%	97%	80%	80%	79%
Accuracy	91%	91%	80%	80%	0.78
F-Means	0.50	0.49	0.45	0.45	0.39
G-Means	0.63	0.64	0.79	0.78	0.73
Area under curve (AUC)	0.85	0.85	0.85	0.86	0.83

Table 3: The performance of mlp model in different data combinations and variable of women.

	First Combination	Second Combination	Third Combination	Fourth Combination	Fifth Combination
Sensitivity	29%	29%	78%	78%	73%
Characteristic	98%	98%	78%	78%	78%
Accuracy	90%	90%	78%	78%	78%
F-Means	40%	0.40	0.43	0.43	0.42
G-Means	0.53	0.53	0.78	0.78	0.75
Area under curve (AUC)	0.67	0.62	0.81	0.81	0.81

Table 4: The performance of quest model in different data combinations and variable of women.

	First Combination	Second Combination	Third Combination	Fourth Combination	Fifth Combination
Sensitivity	33%	33%	70%	70%	66%
Characteristic	96%	96%	79%	79%	78%
Accuracy	90%	90%	78%	78%	77%
F-Means	0.41	0.41	0.41	0.41	0.38
G-Means	0.56	0.56	0.74	0.74	0.72
Area under curve (AUC)	0.65	0.65	0.81	0.81	0.81

Table 5: The performance of cart model in different data combinations and variable of women.

	First Combination	Second Combination	Third Combination	Fourth Combination	Fifth Combination
Sensitivity	78%	78%	81%	79%	76%
Characteristic	76%	77%	72%	72%	72%
Accuracy	76%	77%	73%	73%	73%
F-Means	0.41	0.42	0.39	0.39	0.3
G-Means	0.77	0.77	0.76	0.75	0.74
Area under curve (AUC)	0.83	0.83	0.83	0.83	0.82

Table 6: The performance of naive model in different data combinations and variable of women.

	First Combination	Second Combination	Third Combination	Fourth Combination	Fifth Combination
Sensitivity	34%	38%	71%	74%	71%
Characteristic	97%	98%	79%	79%	78%
Accuracy	90%	91%	78%	78%	77%
F-Means	0.42	0.48	0.41	0.43	0.40
G-Means	0.57	0.61	0.75	0.76	0.74
Area under curve (AUC)	0.84	0.86	0.83	0.86	0.84

Table 7: The performance of logistic model in different data combinations and variable of women.

mance indexes such as sensitivity, G-Means. Comparing the sensitivity of models in their best performance represent that Naïve model has the highest and CART model has the lowest sensitivity.

As shown in Figure 2, the highest sensitivtiy belong to NB model, MLP model, and QUEST mode, respectively. Considering the balance between sensitivity and characteristic, the best models are respectively QUEST model and MLP model. 19 variables have been used in the fifth combination of training data (20 variables without 2-hour blood sugar). In terms of sensitivity and G-Means, the results indicate that the performances of Nb and CART algorithms have respectively the highest and the lowest sensitivity. If the balance between sensitivity and characteristic has been more important, QUESt model is the best model (Figure 3).

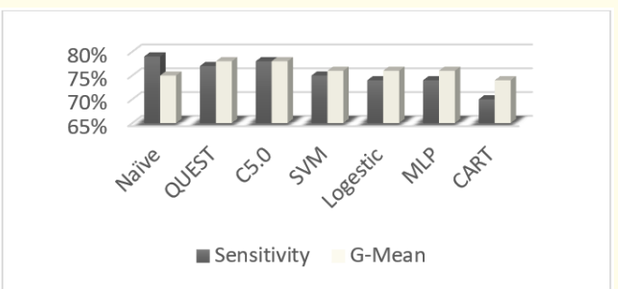

Figure 2: Comparing sensitivity and G-Means indexes in different models with the fourth combination. (women).

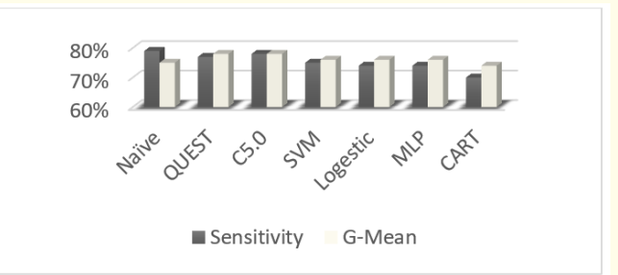

Figure 3: Comparing indexes of sensitivity and G-Means in different models with fifth combination.

Predictor variables based on best performance of the various models in women

Since all models used in the fourth combination have the best performance, the importance of predictor variables have been determined based on each model, and they are shown in Table 8. NB model has not represented any specific model for prediction; thus, it is not entered in Table 8. As the fourth combination has shown the best performance and it has employed 20 variables, it is concluded that the 20 variables are important factors to predict diabetes in women based on NB model. According to Table 8, 10 of the 20 variables having been used for building different models are regarded as important factors. Some variables are observed in only one model while others may be observed in all models or most models. Important coefficient of the variables and their rankings are different in different models.

Rank	MLP	SVM	C5.0	QUEST	CART	Logistic Regression
1	Fasting blood sugar 0.16	Fasting blood sugar 0.29	Fasting blood sugar 0.55	Waist-to-Height 0.45	Fasting blood sugar 0.43	Fasting blood sugar 0.29
2	2-hour blood sugar 0.11	Waist-to-Height 0.17	Waist-to-Height 0.18	Fasting blood sugar 0.37	2-hour blood sugar 0.17	2-hour blood sugar 0.20
3	Triglyceride to HDL 0.1	2-hour blood sugar 0.17	Average arterial pressure 0.15	2-hour blood sugar 0.11	Waist-to-Height 0.15	BMI 0.11
4	BMI 0.07	Triglyceride to HDL 0.06	2-hour blood sugar 0.12	cholesterol to HDL 0.01	BMI 0.1	Waist-to-Height 0.1
5	Average arterial pressure 0.07	Average arterial pressure 0.05		Average arterial pressure 0.01	Wrist circumference 0.09	Average arterial pressure 0.07
6	Waist-to-Height 0.07	BMI 0.05		Glomerular filtration of blood 0.01	Duration of stay in the city 0.01	Waist-to-hip 0.06
7	cholesterol to HDL 0.07	Duration of stay in the city 0.04		Triglyceride to HDL 0.01	Family history of diabetes 0.01	Duration of stay in the city 0.04
8	Waist-to-hip 0.06	education 0.04		Duration of stay in the city 0.01	cholesterol to HDL 0.01	Triglyceride to HDL 0.04
9	Glomerular filtration of blood 0.06	Waist-to-hip 0.03		BMI 0.01	use of enzyme inhibitors 0.01	education 0.04
10	Duration of stay in the city 0.05	Taking aspirin 0.03		use of enzyme inhibitors 0.01	Triglyceride to HDL 0.01	Family history of diabetes 0.03

Table 8: Importance of predictor variables based on the best performance of models in women (fourth combination).

Table 8 represents variable importance in each model. In this combination, fasting blood sugar is regarded as the first important variable in the five models. In QUEST model, the first important variable is waist-to-height ratio. 2-hour blood sugar is the second important variable in three models. In general, 4 variables are observed in all models including fasting blood glucose, 2-hour blood glucose, waist-to-height ratio and mean arterial pressure. In MLP and QUEST models, which are among the best models in the fourth combination, 9 variable of the 10 first variables are common in both models. Waist-to-hip variable in MLP model and the use of enzyme inhibitor are seen in QUEST model.

Table 9 represents 4 rule for prediction of non-diabetic cases and 3 rules for prediction of diabetic cases; they are extracted from QUEST decision tree. Based on the first rule, incidence of diabetes will be met with the possibility of 88% if fasting blood sugar is less than or equal to 94 and waist-to-height ratio is less than or equal to 0.55. Based on the second rule, diabetes will not occur with the possibility of 72% if fasting blood sugar is less than or equal to 94, waist-to-height ratio is between 0.55 to 0.66, and 2-hour blood sugar is less than or equal to 134. According to third rule, diabetes will not occur with the possibility of 57% if fasting blood sugar is less than or equal to 94, waist-to-height ratio is more than 0.66, and 2-hour blood sugar is less than or equal to 125.5. On the word of fourth rule, diabetes will not occur with the possibility of 72% if fasting blood sugar is more than 94 and waist-to-height ratio is less than 0.52. In line with Rule 5, diabetes will occur with the possibility of 69% if fasting blood sugar is less than or equal to 94, waist-to-height ratio is between 0.55 to 0.66, and 2-hour blood sugar is more than 134. As stated by Rule 6, diabetes will occur with the possibility of 75% if fasting blood sugar is less than or equal to 94, waist-to-height ratio is more than 0.66, and 2-hour blood sugar is more than 125.5. As said by Rule 7, diabetes will occur with the possibility of 81% if fasting blood sugar is less than or equal to 94 and waist-to-height ratio is more than 0.52.

Predictor variables based on the fifth combination in women

In the fifth combination, the variable of 2-hour blood sugar has been eliminated from the 20 variables used in the fourth combination. Model parameters are the same parameters that were used in the fourth combination. 10 important variables in each model are shown in Table 10. Table 10 shows the importance of variables in each model in the fifth combination. As observed, fasting blood sugar is represented as the first variable to predict diabetes in all

Rules For 0
Rule 1 If fbs1 <= 93.85 and wtohei <= 0.55 and then 0 (824;0.877) Rule 2 If fbs1 <= 93.85 and wtohei <= 0.66 and fbs1 > 88.003 and wtohei > 0.55 and bs2hr1 <= 134.42 then 0(505; 0.717) Rule 3 If fbs1 <= 93.85 and wtohei > 0.66 and bs2hr1 <= 125.47 then 0 (125; 0.568) Rule 4 If fbs1 > 93.85 and wtohei <= 0.52 then 0 (138; 0.739) **Rules For 1** Rule 5 If fbs1 <= 93.85 and wtohei <=0.66 and wtohei > 0.55 and bs2hr1 > 134.42 then 1 (136; 0.691) Rule 6 If fbs1 <= 93.85 and wtohei > 0.66 and bs2hr1 > 125.47 then 1 (52; 0.75) Rule 7 If fbs1 > 93.85 and wtohei < 0.52 then 1 (1, 381; 0.811)

Table 9: Extracted rules from quest decision tree model in the fourth combination (women).

Rules for 0: rules for non-diabetic

Rules for 1: Rules for diabetic

Fbs1: fasting blood sugar

Bs2hr1: 2-hour blood sugar

Wtohei: Waist-to-Height Ratio

Ageyr: Age

Rank	MLP	SVM	C5.0	QUEST	CART	Logistic Regression
1	Fasting blood sugar 0.17	Fasting blood sugar 0.4	Fasting blood sugar 0.57	Fasting blood sugar 0.44	Fasting blood sugar 0.43	Fasting blood sugar 0.29
2	Glomerular filtration of blood 0.1	Waist-to-Height 0.15	Waist-to-Height 0.25	Fasting blood sugar 0.44	BMI 0.18	BMI 0.18
3	Waist-to-Height 0.09	Average arterial pressure 0.07	Average arterial pressure 0.18	Average arterial pressure 0.06	Waist-to-Height 0.15	Waist-to-hip 0.12
4	Triglyceride to HDL 0.08	BMI 0.07		Pulse pressure 0.01	Wrist 0.09	Average arterial pressure 0.08
5	cholesterol to HDL 0.07	Waist-to-hip 0.07		cholesterol to HDL 0.01	Average arterial pressure 0.07	Triglyceride to HDL 0.06
6	BMI 0.07	Triglyceride to HDL 0.06		Triglyceride to HDL 0.01	Age 0.06	Duration of stay in the city 0.05
7	Average arterial pressure 0.07	Duration of stay in the city 0.05		Duration of stay in the city 0.01	Pulse pressure 0.04	Family history of diabetes 0.05
8	Duration of stay in the city 0.06	education 0.04		BMI 0.01	Waist-to-hip 0.04	education 0.05
9	Waist-to-hip	Family history of diabetes		Waist-to-hip 0.01	Family history of diabetes	Taking aspirin 0.01
10	A history of heart attack or stroke in male relatives	Taking aspirin 0.01			Glomerular filtration of blood 010.01	A history of heart attack or stroke in male relatives

Table 10: Importance of predictor variables based on the models used in fifth combination (women).

models. Waist-to-height ratio and BMI are seen in 5 models and mean arterial pressure is seen in all models. Duration of stay in the city, triglycerides to HDL, and taking aspirin are seen in two models. Family history of diabetes is observed in 3 models.

Conclusion

The findings related to the implementation of the MLP neural network model in women showed that the model has very low sensitivity and very high characteristics with unbalanced data. This model was 41% sensitive in women with regard to the first combination that used unbalanced data and all variables. The results showed that this method has significantly increased the sensitivity of MLP model in the group of women. Moreover, the results of SVM indicate that the model has very low sensitivity to unbalanced data in the group of women so that the model was 37% sensitive in women with regard to the first combination that used unbalanced data and all variables. By balancing the data and all the variables, the model sensitivity in women increased to 72%. In the final model (fourth combination), SVM sensitivity is 74% in women. Comparing the performance of this model with logistic regression analysis shows that the sensitivity of this model in women is the same in both models. According to the research, 3 decision tree algorithms (C5.0, QUEST, CART) have very low sensitivity and very high characteristic to unbalanced data. The sensitivity in women for the three models has been 71%, 78%, and 70%, respectively. Balancing data increases sensitivity of the three models significantly. This research implies that decision tree algorithms will have better performances for prediction of positive class by balancing data. Sensitivity of these three algorithms in women is respectively 75%, 77%, and 70%. Comparing the sensitivity of these three models with sensitivity of logistic regression analysis (fourth combination) shows that the sensitivity of these algorithms in women is 7%, 9%, and 2%, respectively, more than logistic model. Accordingly, all tree algorithms in the final model (fourth combination) for women are more than 70% but their characteristic (between 72% and 79%) is less than MLP and SVM models. The results also showed that decision tree algorithms are appropriate models for data in the current study with a sensitivity of 70%. In terms of Naïve Bayes model performance, results showed that the algorithm had the highest sensitivity with unbalanced data and total variables (the first combination) in women (about 78%). Results of the most important predictors of diabetes in women are that 20 variables entered into the modeling based on the fourth combination. In total, 15 variables in 6 models of diabetes were identified as the most important predictor variables. According to the results, fasting blood sugar is the first important variable in all models and 2-hour blood sugar is among the first three variables in terms of importance. Waist-to-height has been identified as the first 6 variables in all models. BMI has been identified in most cases and family history has been identified in two models. Ratio of triglycerides to HDL, ratio of cholesterol to HDL, education, and waist-to-hip are also variables having been found as ten first variables in some models. It is noted that the importance of importance of waist circumference to height is more than fasting sugar and 2-hour sugar in QUEST model that is the best model. In the fifth combination, which 2-hour glucose was removed from the list of input variables of the model, fasting blood sugar was identified as the most important factor; then, waist to height and average arterial pressure are important variables in most cases. Other variables that have been identified in fourth combination have the same behavior in the fifth combination. New variables are observed in fifth combination compared to fourth combination. The new variables are history of heart attack or stroke in male relatives and pulse pressure; they are observed in 2 models and the variable of age have been observed in one model. Results of most important risk factors for diabetes in women are as follows. In order to examine the relationship between variables, tree structure and rules have been derived from QUEST model, which is the best model in fourth combination, According to the rules of the model, a fasting blood sugar of 94 is crucial for the occurrence or non-occurrence of diabetes so that diabetes will not occur with a possibility of 88% in a subject with fasting blood sugar of 94 if the waist-to-height is less than 0.55. The possibility is not related to 2-hour glucose levels. However, 2-hour glucose level becomes important by increasing the ratio waist to height so that the level of 2-hour glucose must be reduced as long as the waist to height rate increases in order to decrease the possibility of diabetes. Thus, in fasting glucose levels of less than 94, waist to height and to 2-hour glucose level are the determinants of of diabetes. If the amount of fasting glucose reaches up to 94, waist-to-Height more than 0.52 is most important risk factor for diabetes. In this manner, if waist to height ratio is more than 0.52 in fasting glucose level more than 94, chance of diabetes will reaches to 81% while it reaches to 26% in less than 0.52. If the amount of 2-hour glucose level is disregarded, in fasting glucose less than 94, average arterial pressure decreases to less than 98.9 with increasing waist-to-height ratio to more than 0.6; hence, the risk of diabetes will be reduced. Otherwise, the risk of diabetes will increase from 41% to 66% in average arterial pressure more than 98.9. In general, the principles indicate that waist-to-height importance in the incidence of diabetes is equal to or more than fasting blood sugar in women, because women will not be target of diabetes with a possibility of 74% by keeping waist-to-height ratio of less than 0.52; this possibility is independent of fasting blood sugar and 2-hour blood sugar. The model was less sensitive by increasing the number of tree layers; then, the relationships among variables have been uncovered. Other factors including cholesterol to HDL ratio (more than 5.3), duration of stay in the city more than 43 years, triglyceride to HDL ratio (more than 13.5), and glomerular filtration between 58.5 and 65.7 will increase the risk of diabetes in certain sub-groups. Therefore, they are risk factors. The subgroups are explained fully in the results. The results of risk factors for women indicate that fasting blood sugar, 2-hour blood sugar, waist to height ratio, and average arterial pressure are predictors of risk for diabetes in any case. For example, diabetes has a preventable nature in women more than men because the risk of dia-

betes in men will be 56% with a fasting blood sugar of more than 95 even if their waist-to-height ratio decreases to less than 0.44. Nevertheless, diabetes will not occur 74% with a decrease in this ratio independent of fasting blood sugar or 2-hour blood sugar. In women, waist-to-height ratio should be between 0.5 and 0.6 to reduce the risk of diabetes even if fasting blood sugar reaches to less than 88. By increasing fasting blood sugar to less than 88 in men, the chance for not developing diabetes will increase to 86%. Accordingly, the amount of waste-to-height ratio is more important for women rather than men.

Bibliography

1. D Maher and J Sekajugo. "Research on health transition in Africa: time for action". *Health Research Policy and Systems* 9.1 (2011): 10-20.
2. C Mathers., *et al.* "The global burden of disease: 2004 update". World Health Organization (2008).
3. DR Whiting., *et al.* "IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030". *Diabetes Research and Clinical Practice* 94.3 (2011): 311-321.
4. L Navayi., *et al.* "The prevalence of diabetes and IGT in Islam Shahr and Compare GTT screening result to detect impaired glucose tolerance". *Journal of Research in Medical* 21.1 (1997): 85-97.
5. P Zhang., *et al.* "Global healthcare expenditure on diabetes for 2010 and 2030". *Diabetes Research and Clinical Practice* 87.3 (2010): 293-301.
6. R Bellazzi and B Zupan. "Predictive data mining in clinical medicine: current issues and guidelines". *International Journal of Medical Informatics* 77.2 (2008): 81-97.
7. S Fong., *et al.* "Evaluation of stream mining classifiers for real-time clinical decision support system: a case study of blood glucose prediction in diabetes therapy". *Biomed Research International* 19.2 (2013): 15-25.
8. Y Huang., *et al.* "Feature selection and classification model construction on type 2 diabetic patients' data". *Artificial Intelligence in Medicine* 41.3 (2007): 251-262.
9. XH Meng., *et al.* "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors". *The Kaohsiung Journal of Medical Sciences* 29.2 (2013): 93-99.
10. D Rebholz-Schuhmann., *et al.* "A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources". *Drug Discovery Today* 19.7 (2014): 882-889.
11. C Brown., *et al.* "Searching QTL by gene expression: analysis of diabesity". *BMC Genetics* 6.1 (2005): 14-18.
12. U Covani., *et al.* "Relationship between human periodontitis and type 2 diabetes at a genomic level: a data-mining study". *Journal of Periodontology* 80.8 (2009): 1265-1273.
13. AM Shin., *et al.* "Diagnostic analysis of patients with essential hypertension using association rule mining". *Healthcare Informatics Research* 16.2 (2010): 77-81.
14. H Liu., *et al.* "An efficacy driven approach for medication recommendation in type 2 diabetes treatment using data mining techniques". *Studies in Health Technology and Informatics* 192.3 (2012): 1071-1071.
15. M Marinov., *et al.* "Data-mining technologies for diabetes: a systematic review". *Journal of Diabetes Science and Technology* 5.6 (2011): 1549-1556.
16. World Health Organization, Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation (2011).
17. WHO. Consultation, Definition, diagnosis and classification of diabetes mellitus and its complications (1999).
18. S Wild., *et al.* "Global prevalence of diabetes estimates for the year 2000 and projections for 2030". *Diabetes Care* 27.5 (2004): 1047-1053.
19. F Azizi. "Diabetes mellitus in the Islamic Republic of Iran". *IDF Bulletin* 41.4 (1996): 38-39.
20. Yoo., *et al.* "Data mining in healthcare and biomedicine: a survey of the literature". *Journal of Medical Systems* 36.4 (2012): 2431-2448.
21. W C Knowler., *et al.* "Diabetes incidence in Pima Indians: contributions of obesity and parental diabetes". *American Journal of Epidemiology* 113.2 (1981): 144-156.
22. M Yamaguchi., *et al.* "Prediction of blood glucose level of type 1 diabetics using response surface methodology and data mining". *Medical and Biological Engineering and Computing* 44.6 (2006): 451-457.
23. L Tapak., *et al.* "Real-data comparison of data mining methods in prediction of diabetes in Iran". *Healthcare Informatics Research* 19.3 (2013): 177-185.
24. F Azizi., *et al.* "Cardiovascular risk factors in an Iranian urban population: Tehran lipid and glucose study (phase 1)". *Sozial-Und Präventivmedizin* 47.6 (2002): 408-426.
25. P Chapman., *et al.* "Wirth, CRISP-DM 1.0 Step-by-step data mining guide (2000).

Volume 3 Issue 9 September 2019