International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 9 – No.7, September 2015 – www.ijais.org

Effective Values of Physical Features for Type-2 Diabetic and Non-diabetic Patients Classifying Case Study: Shiraz University of Medical Sciences

S. Vahid Farrahi
M.Sc Student
Shiraz University of Technology,Shiraz, Iran

Mohammad Mehdi Masoumi
M.Sc Student
Shiraz University of Technology, Shiraz, Iran

Marzieh Ahmadzadeh
Assistant Professor
Shiraz University of Technology, Shiraz, Iran

Pezhman Shafikhani
Diabetes Consultant
Shiraz University of Medical Sciences, Shiraz, Iran

ABSTRACT

Currently, one of the major issues regarding diabetes is its early detection. In this research, the dataset, which has been provided by Shiraz University of Medical Sciences, based on National Type-2 Diabetes Prevention and Control Program in Iran, used for finding an answer to a key question. In this program, there are some standard danger factors based on the physical features, which indicate whether a person is susceptible to diabetes (some people are unaware of their disease) or not, such as Body Mass Index (BMI). The World Health Organization (WHO) mentioned standard values for the danger factors which are based on the physical features, based on those; people who are suspicious for diabetes (considering their physical features) are referred back to the lab to take the Fasting Blood Sugar Test (FBS). The key question is that which values of physical features separate the diabetic and non-diabetic people more accurately, in Iran or more specifically in Shiraz. Classification is one of the data mining techniques, which can classify diabetic and non-diabetic people. Decision tree is a classification technique that can provide generalized rules based on the tree. In this paper, C4.5 decision tree algorithm has been used for rules generation. The extracted rules mention different values of physical features than the standard values.

Keywords

Type-2 Diabetes, Decision Tree, Rule Generation, Classification, Data Mining

1. INTRODUCTION

The increasing growth of medical and human health data before birth to after death makes it more complicated to process and find useful knowledge. Health technology systems record data in databases and analyses them to help better medical decisions. Therefore, different computer analysis method has been proposed for different data. It is proved that data mining techniques show good accuracy regarding these methods[1, 2]. Using data mining in analyzing medical data reduces the costs and increases the disease detection accuracy.

In this era, improving the countries' economy and social status has controlled and eliminated many infectious diseases. On the other hand, with the life style changes and life industrialization, non-communicable diseases have become serious threats for human health. Diabetes is one of the four non-communicable disease groups[3]. Currently, in Iran, from every twenty individuals, one has diabetes and half this number is unaware of their disease[3].

Diabetes is a disease that disturbs insulin generation or consumption in the body and Fasting Blood Sugar (FBS) is the primary way to detect it. Moreover, it is possible to detect it in time using screening parameters. Some characteristics that indicate the symptoms are physical features, the history of diabetes in the family, and eventually performing a FBS test.

Diabetes is divided into type-1, type-2, and pregnancy diabetes. Type-1 is common in children and teenagers and under the influence of genetic factors; pancreas is usually unable to produce insulin. In type-2, which 90% of the patients suffer from, despite producing insulin, the body is not capable to consume it. Finally, the third type, pregnancy diabetes, is usually transformed into the type-2 diabetes. Diabetes is not only considered a disease, but an intricate network of dangerous environmental and genetic factors with different pathophysiology. Type-2 diabetes is the most common type of diabetes. The focus of this paper is on type-2 diabetes.

Diabetes is costly and incapacitating due to characteristics like engendering or associating with cardiovascular, kidney, and diseases, as well as disabilities and side effect. Its side effects are divided into early and delayed groups[4]. Therefore, currently, the major issue regarding diabetes is its early detection or inability to detect it. This inability can mainly be attributed to not selecting appropriate patterns by doctors, lack of appropriately using standard features, or human errors.

This paper is aimed at finding the effective physical features values in Shiraz University of Medical Sciences (SUMS) dataset, which help to classify diabetic and non-diabetic patients in advance, without FBS test. In other words, one of the most primary questions is that which values of physical features separate the diabetic patients form non-diabetic people more precisely. This paper focused on finding a response for this critical question and find more exact values using a data mining technique. More precise values means that the diabetic patients and non-diabetic people can be separated more accurately, based on the physical features. In addition, the focus of this paper is on type-2 diabetes.

2. RELATED WORKS

Early detection of diabetes greatly helps to control and delay the extension of this disease and its side effects for the patient. So far, many methods are proposed and analyzed in different studies and research. In most studies in the data mining and machine intelligence domain, a dataset namely, Pima Indian Diabetes dataset, is used to detect diabetes, which consists of 8 features and 768 female individuals at least 21 years old[5]. The features include age, the number of pregnancies, and Body Mass Index (BMI).

In[6], Apriori association rule mining algorithm is employed. This algorithm works with tabular data. Therefore, the proposed method first discretizes numeric values into low, average, and high states and then executes the algorithm. Some of the classifiers that can work with numeric features demonstrate better overall accuracy after discretizing the numerical variables at the preprocessing stage.

In[7], patients are classified by discretizing continuous variables and running the decision tree algorithm. It is showed that the discretized data set has better accuracy in detection of patients.

In[8], the authors proposed a hybrid learning approach for prediction of type-2 diabetic patients. The proposed method used simple K-means for validating chosen class label of given data. Finally, C4.5 decision tree algorithm has been used for data classification.

The notion of this research is predicting non-diabetic people based on physical measures. In addition, this paper by focusing on non-diabetic people detection is trying to separate diabetic patients from non-diabetic people based on physical features more precisely. Therefore, generalized rules of decision tree are used to predict non-diabetic people.

3. THE MEASURES OF THE NATIONAL TYPE-2 DIABETES PREVENTION AND CONTROL PROGRAM

According to the standards of World Health Organization (WHO), the danger measures and factors of detecting type-2 diabetes is collected and shown in Figure 1[3]. Based on these factors, Iran Ministry of Health and Medical Education started a program namely, National Type-2 Diabetes Prevention and Control Program[3], to collect the data of people in the organizations, universities etc. This program aimed at detecting diabetic patients that are unaware of their disease, in advance or people who are susceptible to type-2 diabetes. Simply, one of the major objectives of this national program is detecting diabetic patients in advance in order to control their disease, especially for those who are unaware of their disease.

People with ages higher than 30 years who have one of the following conditions are suggested to take the FBS test:

- Being overweight or obese with the Body Mass Index (BMI) larger than 25.
- A ratio of around the waist to around the hips larger or equal to 0.9.
- Blood pressure higher than 140.90mmHg (in two blood pressure measurements)
- Family history of diabetes in immediate family members.

Considering Figure 1, the process of diabetes screening based on National Type-2 Diabetes Prevention and Control Program is as follows. First, the blood pressure is measured in two sessions. Height, weight, around the waist and hips are then measured. The ratio of around the waist to around the hips and BMI are derived from these values.

Screening volunteers that have at least one of the danger factors, which mentioned above, are referred back to the lab to take the FBS test. For people with FBS higher or equal to 126, the FBS test is repeated for the second time in one to four weeks. If the results of the second test are also higher or equal to 126, these individuals are recognized as diabetes patients. Otherwise, the individuals do not have diabetes. In other words, if the result of the first FBS test is lower than 100, there is no need for repetition, and the individuals are not diabetic.

Moreover, if the result of the second test is also less than 100, the individual is not diabetic. Of course, there are individuals who voluntarily repeat the FBS test despite having results lower than 100 for the first FBS test.

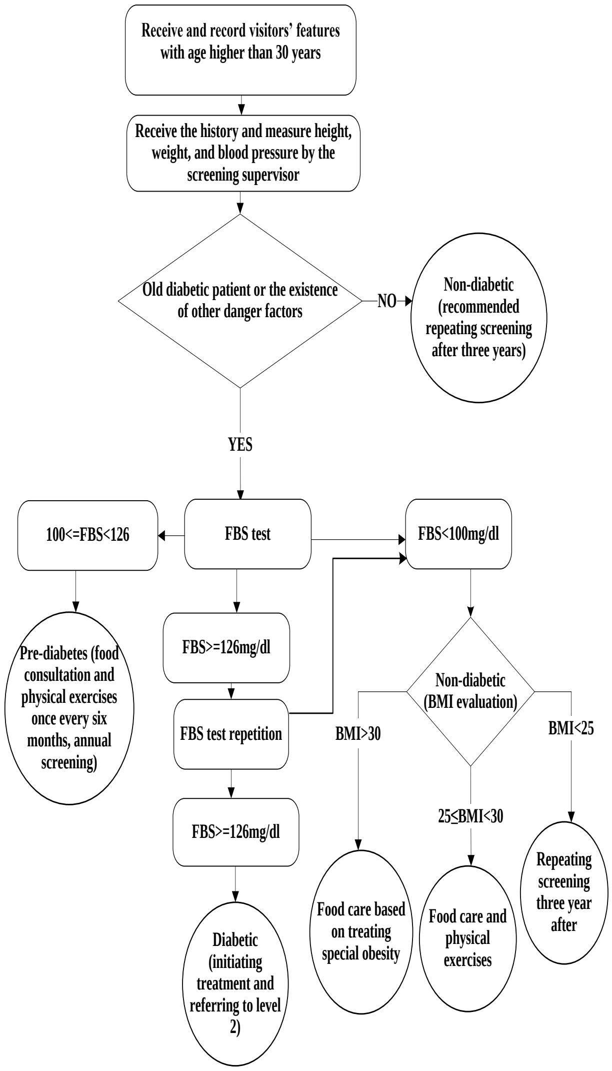
Receive and record visitors' features with age higher than 30 years
Receive the history and measure height, weight, and blood pressure by the screening supervisor
Old diabetic patient or the existence of other danger factors
NO
Non-diabetic (recommended repeating screening after three years)
YES
100<=FBS<126
FBS test
FBS<100mg/dl
Pre-diabetes (food consultation and physical exercises once every six months, annual screening)
FBS>=126mg/dl
Non-diabetic (BMI evaluation)
FBS test repetition
BMI>30
BMI<25
25≤BMI<30
FBS>=126mg/dl
Repeating screening three year after
Food care based on treating special obesity
Food care and physical exercises
Diabetic (initiating treatment and referring to level 2)

Fig. 1: The diabetes screening process of a patient[3].

4. C4.5 DECISION TREE ALGORITHM

Each record is defined as *(x_i, y)*, where x_i is the set of features, $x_{i\in n} = \{A_1, A_2, ..., A_n\}$ and *y* is the class variable with values $y = \{C_1, C_2, ..., C_n\}$. Therefore, classification is a training task using an objective function *f* to predict the value of the class variable from the values in set *y* for each feature set x_i [9].

Decision tree algorithms require an evaluation measure that discretizes the continuous attribute values of training records. Moreover, it should show which value of the discretized features are better splitter and when to stop the split. In the decision tree, each leaf represents the prediction result and the class variable. A set of conditions should be satisfied from the root to the leaf to perform the prediction for the corresponding record. This set of conditions can be considered as the antecedent of a rule and the result as the consequent of that rule. In other words, any path from the root to the leaf represents a rule. For instance, $A\ AND\ B \rightarrow C_1$ is a rule that can be extracted from the tree.

In large trees, however, rules may not be general and hard to understand for humans. Therefore, irrelevant conditions should be removed from the antecedent to make rules generalized and comprehensible. This can be performed based on the training dataset used to build the decision tree[10]. C4.5 is a decision tree algorithm that is used in the data mining and machine learning domains and uses divide and conquer to build the tree and uses Entropy (Equation 1) to measure the impurity value. In Equation 1, $P(i|j)$ shows the number of class *i* variables in group *t*.

$$(1)\ Entropy(t) = -\sum_{i=0}^{c-1} P(i|j)\log_2 P(i|j)$$

5. RESEARCH STEPS

This study aimed to classify non-diabetic and diabetic patients, based on physical feature values. As mentioned earlier, there are some standard values for the people who suggested taking the FBS test. There is a key question, which this research focused on finding an answer to it. The key question is that which values of physical features classify or separate diabetic and non-diabetic patients more precisely.

Figure 2 shows the process of this study in order to find more accurate values for the physical features in the SUMS' dataset. The research stages can be divided into three levels:

- Preprocessing and specifying the status of the individuals based on the information recorded in the database.
- Running the algorithm and extracting the rules.
- Testing the rules using the SUMS' datasets.

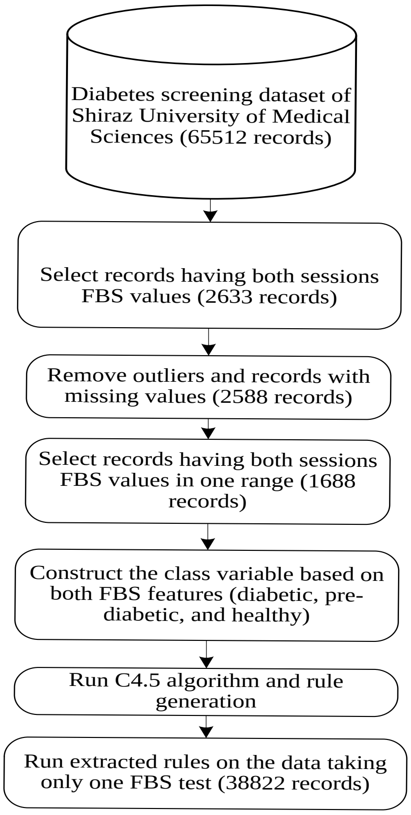
Diabetes screening dataset of Shiraz University of Medical Sciences (65512 records)
Select records having both sessions FBS values (2633 records)
Remove outliers and records with missing values (2588 records)
Select records having both sessions FBS values in one range (1688 records)
Construct the class variable based on both FBS features (diabetic, pre-diabetic, and healthy)
Run C4.5 algorithm and rule generation
Run extracted rules on the data taking only one FBS test (38822 records)

Fig. 2: Research Steps.

6. DATASET

The employed dataset has 65512 records based on the National Type-2 Diabetes Prevention and Control Program prepared by SUMS from the medical centers' diabetes screening volunteers with ages higher than 30 years since late 2009 until 2015.

Each record includes the demographic and physical (body) characteristics of the volunteers. Moreover, variables irrelevant to the purpose of this research are removed. As, the history can also be effective in detecting the disease. Therefore, age, height, weight, around the waist, BMI, around the hips, the ratio of around the waist to around the hips, the diabetes history in immediate family members, history of high blood pressure, and history of tobacco consumption are considered in this research. In addition, the attributes, which show high blood pressure in two sessions, are ignored. Since in most records, blood pressure is measured in one session, which means that it is not a reliable index.

7. DATA PREPROCESSING

Data preprocessing is important to find accurate results without ambiguity. There are many factors causing false results. Human error when entering data, cases where occur rarely in the data (outliers), are some instances of this factors. Therefore, it is necessary to analyze the data before extraction.

7.1 Selecting Records Containing Both First and Second Fasting Blood Sugar Values

Among the collected records, only 2633 records had both FBS test sessions performed and not missing values for these two features. These records were selected for the next stage of the preprocessing.

7.2 Removing Outliers

A number of selected records were identified and removed as outliers due to reasons like user errors in inserting the data. Moreover, records with a few number of occurrences, including people with ages higher than 90 years, height less than 140cm, or weight less than 45kg, were also identified as outliers and removed from the dataset. Records having missing values for features like age, height, weight, etc. were also removed from the dataset. After removing outliers, 2588 records were used in the next preprocessing stage.

7.3 Class Variable Construction

As mentioned earlier, if the results of first FBS test show that the individual is not diabetic, the results can be considered. Else, if the first FBS test shows that the individual is diabetic the FBS test must be repeated in order to decide whether the individual is diabetic or not. Therefore, the class variable constructed based on the records that have both values of FBS tests. This is an evidence, which mentions that an individual is diabetic definitely, or not.

Considering Section 3, only individuals whose results of both FBS test sessions indicated diabetic, pre-diabetes, or non-diabetic can be used in the model. Both FBS tests of these records have FBS values in the same numeric range; in other words, the results of both tests indicated pre-diabetic, diabetic, or non-diabetic. There are records, whose the first FBS test result indicated the person is non-diabetic, but the second test showed diabetes or vice-versa. Although, for those who their first FBS test show that they are non-diabetic it not necessary to repeat the test. These records were removed assuming possible errors in performing the FBS test or failing to adhere the recommendations before taking the test. Consequently, 1688 records remained and used as the train set and test sets.

Based on the standards of WHO and the instructions of the national program, the FBS measure is defined as a numeric value in three ranges. In this study, people who have FBS value less than 100 for "both FBS tests" are labeled as "non-diabetic" and people with values larger or equal to 100 and smaller than 126 are labeled "pre-diabetes", i.e. in risk of developing diabetes, and the rest, with values larger than 125 are labeled as "diabetic".

8. C4.5 EXECUTION AND REPITATION

After preprocessing and selecting 1688 records from the SUMS' dataset and the class variable construction, C4.5 algorithm was executed on these records. 70% of the records were selected as the training set and the rest as the test data. The algorithm was executed 10 times with different seeds for the random number generator. The four best rules for non-diabetic individuals were selected and presented in Figure 3.

9. MODEL EVALUATION AND ITS ACCURACY

The data used by the model as the training set are randomly selected. In other words, 70% of the records are randomly selected as the training set. Therefore, the accuracy of the algorithm is different by selecting different records as the training set. In fact, it is impossible to achieve an appropriate estimation of the accuracy of the entire population only by a limited population and limited number of samples. One solution is using confidence interval[11].

Equation 2 shows that the probability that accuracy is between the confidence interval c_1 and c_2 is $1-\alpha$. If random quantity $X = \{x_1, x_2, \dots, x_n\}$ has a distribution with mean μ and standard deviation σ, the sample means $\overline{X}$, that is obtained by random sampling with size n, has a distribution with mean μ and standard deviation σ/√n, which becomes a normal distribution by increasing n (Equation 3). Therefore, Equation 4, can be used to achieve the confidence interval of accuracy, where $\overline{X}$ is the mean accuracy of repeating the model by changing the seed of random number generator and s is the standard deviation of accuracy in 10 repetitions.

(2) $probability\{c_1 \le \mu \le c_2\} = 1 - \alpha$

(3) $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

(4) $\overline{X} - t_{[1-\frac{\alpha}{2};n-1]} \frac{s}{\sqrt{n}}, \overline{X} + t_{[1-\frac{\alpha}{2};n-1]} \frac{s}{\sqrt{n}}$

It can be assumed that numbers have a normal distribution. The seed of the random number generator is changed in each repetition and numbers are independent. Therefore, it is true to say that each repetition is performed independently and the obtained numbers have a normal distribution. According to Equation 4, the model's accuracy with 95% confidence is between 79.92 ± 2.79.

10. GENERALIZED RULES OF THE TREE

Four rules were extracted from C4.5 decision tree to predict non-diabetic individuals more accurately. The values of physical features in these rules are different from the standard values. These rules are for the people with age lower or equal to 47 years old. In addition, the rules are shown in Figure 3.

Rule1: if (158<height≤172) & (around the waist≤87) & (the ratio of around the waist to around the hips≤0.87) & (age≤47) → Non-diabetes

Rule2: if (height≤154) & (around the waist≤87) & (the ratio of around the waist to around the hips≤0.87) & (has family history of diabetes) & (age≤47) → Non-diabetes

Rule3: if (82<around the waist≤87) & (the ratio of around the waist to around the hips≤0.87) & (has family history of diabetes) & (age≤47) → Non-diabetes

Rule4: if (BMI≤34.77) & (around the waist>96) & (the ratio of around the waist to around the hips≤0.87) & (age≤47) →Non-diabetes

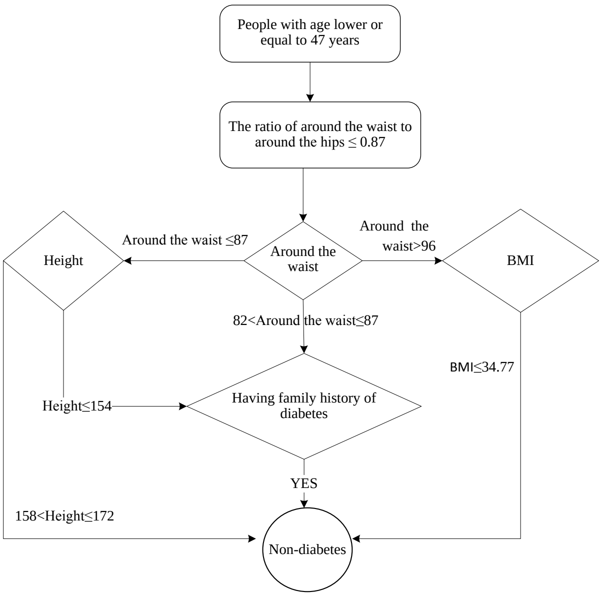

Fig. 3: Decision Tree based rules for non-diabetic individuals

11. RESULTS AND RULES TESTING

38882 records were selected as the test data to evaluate the rules. This test set is completely different form the test set that has been used in model evaluation. Approximately, all of these records only have the value of one FBS test. This means that they have not repeat the FBS test for the second time. Therefore, these data have not been used in the model construction. These records have been used in order to test the rules. For those who have taken the FBS test only once, if the results show that they are non-diabetic, the results are reliable, else they should repeat the FBS test. For those who their first FBS test show that they are diabetic, it is necessary to repeat the test for the second time. So, if the extracted rules separate the individuals that have taken the FBS test only once and the test show that they are not diabetic, the rules classified these individuals correctly.

Figure 4, presents the individuals satisfying the rules extracted in Section 10 and have at least one of the danger factors, including BMI larger or equal to 25, the ratio of around the waist to around the hips larger than or equal to 0.9, or history of diabetes in immediate family members. Accordingly, these people are required to perform the first FBS test. This test is only repeated if the result of the first session test is larger than 100 to determine the definite state of the individual.

For example, the number of individuals that satisfy the Rule1 is 1443. Because all of them have at least on of the physical danger factors, they have to take the first FBS test. The result of their first FBS test show that 1266 are non-diabetic based on the first FBS result. Since their first FBS test was less than 100. On the other hand, the number of diabetic patients, which satisfy Rule1, is 157. Since they have FBS values larger or equal to 100. However, the first FBS test cannot be reliable in order to mention that they are diabetic. For these individuals the test should be repeated for more reliability and one test cannot be definitive.

Moreover, considering Rule1, despite having FBS higher than 100, results of the repeated test for 20 people showed that they are non-diabetic. They repeated the test for the second time, so they are definitely non-diabetic, since the second FBS test shows this. These people may have failed to satisfy to the essential conditions before the test or results have been ensued due to test error.

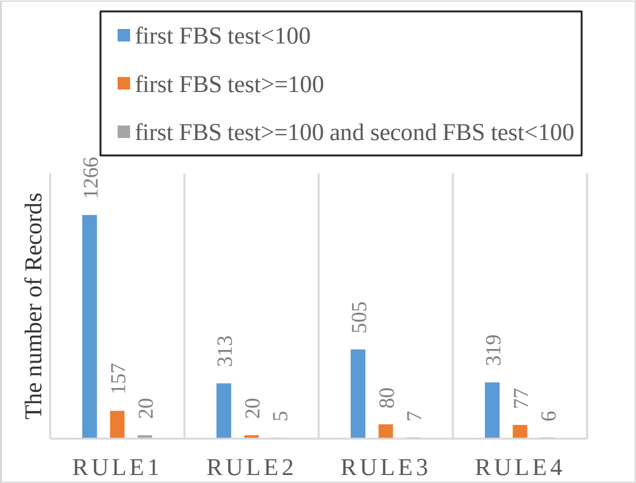

Fig. 4: The number of individuals satisfying the four extracted rules

In Table 1, the individuals that have taken the FBS test only once and results show they are diabetic, have been considered as error of the rule, although it cannot be mentioned that they are diabetic definitely until they repeat the test for the second time.

Table 1: Accuracy of the Extracted Rules

Rules	Accuracy of detecting Non-diabetic (%)	Detection error (%)
Rule1	89.11	10.89
Rule2	94.08	5.92
Rule3	86.48	13.52
Rule4	80.84	19.16
Total	87.96	12.04

12. CONCLUSION AND FUTURE WORKS

This study, focused on classifying the diabetic and non-diabetic patients based on their physical features. C4.5 decision tree algorithm has been used to extract some rules for separating the non-diabetic from diabetic individuals. There are some standard danger factors, which individuals referred back to the lab if they have at least one of them. These danger factors such as BMI have standard values which means if an individual has the BMI larger than 25, is referred back to the lab for taking the FBS test. This research found some rules that mention different values for physical features for the people who susceptible to diabetes.

Since this plan was implemented in a geographical region of Shiraz, it is suggested to conduct the research in different areas. Moreover, data mining and analysis to discover the relationship between BMI and diabetes and specifying the accurate value of this feature to distinguish non-diabetic and diabetic individuals more accurately is considered as the future work. Moreover, studying the data to predict diabetic patients according to the features not taken into account.

13. ACKNOWLEDGMENTS

We are grateful of the health administration of Shiraz University of Medical Sciences for the collection and cooperation in representing the data.

14. REFERENCES

[1] M. Khajehei, and F. Etemady, "Data Mining and Medical Research Studies." pp. 119-122.

[2] K. Rajesh, and V. Sangeetha, "Application of data mining methods and techniques for diabetes diagnosis," *International Journal of Engineering and Innovative Technology (IJEIT),* vol. 2, no. 3, 2012.

[3] M. Alavinia, M. Ghotbi, A. Mahdavi Hezareh, A. Kermanchi, A. Nasli, and S. Yarahmadi, "Regulations of the Type-2 Diabetes National Prevention and Control Program," Ministry of Health and Medical Education, 2012.(Language in Persian)

[4] "Iran Diabets Society," http://ids.org.ir/index.php/fa/explore/layouts/diabets123.(Language in Persian)

[5] M. Lichman. "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml.

[6] M. Patil, R. Joshi, and D. Toshniwal, "Association rule for classification of type-2 diabetic patients." pp. 330-334.

[7] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes." pp. 303-307.

[8] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients," *Expert systems with applications*, vol. 37, no. 12, pp. 8102-8108, 2010.

[9] T. Pang-Ning, M. Steinbach, and V. Kumar, "Introduction to data mining." p. 74.

[10] J. R. Quinlan, *C4. 5: programs for machine learning*: Elsevier, 2014.

[11] R. Jain, D. Menasce, L. W. Dowdy, V. A. Almeida, C. U. Smith, and L. G. Williams, "The Art of Computer Systems Performance Analysis: Techniques," 2010.