Avicenna Journal of Medical Biochemistry

doi:10.15171/ajmb.2018.02 2018 June;6(1):3-7 http://ajmb.umsha.ac.ir

Research Article

# Prediction and Diagnosis of Diabetes by Using Data Mining Techniques

**Seyede Somayeh Mirzajani[1,2*], Siamak Salimi[3]**

[1]Research & Technology Deputy, Hamadan University of Medical Sciences, Hamadan, Iran
[2]Masters of Department of Computer Engineering, Malayer Branch, Islamic Azad University, Hamadan, Iran
[3]PhD Student of Bioinformatics, Tehran University, Tehran, Iran

***Corresponding author:**
Seyede Somayeh Mirzajani,
Email:
so_mirzajani@yahoo.com

Received: 7 April 2018
Accepted: 9 June 2018
ePublished: 21 June 2018

## Abstract

**Background:** Diabetes mellitus (DM) is one of the most common diseases in the world. Complications of this disease include nephropathy, cardiac arrest, blindness, and even mutilation of the body. The accurate diagnosis of this condition is very important.
**Objectives:** This study was to identify and provide a model for diagnosis of DM using data mining.
**Methods:** The data used in this study were obtained from 768 women aged 21-83 year old. Nine variables were selected for investigation. The neural network, Basin network, C5.0, and support vector machine models were compared for predicting diabetes and their precision to this end. Clementine 12 software was used to analyze the data.
**Results:** ThThe proposed method for classification of records with the C5.0 algorithm for accuracy data is 80.2% and for accuracy data 87.5%. In comparison with similar studies, it was better to diagnose people with diabetes, while glucose, body mass index and age variables were important in this study.
**Conclusion:** The C5.0 algorithm showed the highest value of accuracy, specificity, and sensitivity compared with other methods studied. Therefore, the C5.0 algorithm probably performs the best classification among other algorithms and is recommended as the best method for diabetes prediction using available data.
**Keywords:** Diabetes mellitus, Bayesian network, Neural network, Decision tree, Support vector machine, Data mining.

## Background

Diabetes mellitus (DM) is one of the most commonly diagnosed diseases in the world. According to the World Federation of Diabetes, more than 370 million people were diagnosed with the disease in 2012 that is projected to reach 439 million in 2030, on the other every year it is added to it (1). Despite the high prevalence of the disease, no effective method to reduce its incidence has yet been offered, although different methods are currently being used to treat and control the disease (1). Diabetes can leads to many complications such as nephropathy, heart disease, blindness, amputation, etc. It has been reported as being the fourth leading cause of death in most human societies. Diabetes can be of two types: type 1 or insulin-dependent diabetes, in which the pancreas cannot secrete insulin; and type 2 or non-insulin dependent diabetes, the pancreas secretes insulin but its absorption in the body is very low (2).

Data mining is a method of analyzing large data and identifying hidden patterns that cannot be accomplished manually (3). The purpose of identifying the cause of disease, diagnosis, and anticipation of diseases is to provide useful information for health experts and professionals; therefore, the complexity of medical information and the availability of data mining algorithms make data mining important to medical information (4).

In data mining, contrary to the statistics, we are not seeking to discover or prove what is already there, but we are looking for a way to predict algorithms that are not already known (5).

Su et al could diagnose diabetes in 89% of patients based on some data mining methods, including artificial neural networks and decision tree, by means of 3-dimensional body photos, in that study, 3-dimensional and 2-dimensional photos were taken from all organs diabetic patients and healthy people (6). Purnami et al using a support vector machine (SVM), accurately detected 93% of type 2 diabetes in 768 people, with blood pressure and insulin level being the most important in predicting the disease (7). Using SVM, Worachartcheewan et al improved the accuracy of the diagnosis one year after development of type 2 diabetes so that they made an accurate diagnosis in 94% of patients (8).

In 2011, in a study in Turkey on the type and amount of different drugs to treat patients with type 2 diabetes, 6 different types of drug combinations were tested on patients, and then a model was designed based on data mining, including the fuzzy neural network and dependency rules. Using this model, the correct drug combination and the correct dosage were achieved for 80% of the patients (9). In another study, researchers were able to accurately predict 73.32% of the neural network algorithm (10).

Considering the above mentioned, this study was aimed to identify and provide an appropriate model for diagnosis of diabetes.

## Methods

This study was done by the CRISP methodology. This includes 5 phases: problem identification, data collection and description, data preparation, data modeling and data evaluation (12).

Understanding the problem: In this phase that addresses the goal of identifying the problem, because of inappropriate diet and physical inactivity of people, the number of people with diabetes is increasing day by day. The goal is to provide a model to predict the likelihood of making a diagnosis of diabetes to achieve rapid and inexpensive diagnosis of the disease.

Data collection and description: Most studies regarding machine learning on diabetes are based on the Indian data collected from the UCI data set (13). This data set contains 9 variables, 8 of which were considered input variables and one served as the response variable. The data set also contains 768 records and two classes, with the first class including 500 healthy women and the second class including 268 women with DM aged 21-81 years (Table 1).

### Data Preparation

When data are collected from the environment, they are examined for potential errors such as incompleteness, noise, etc. Therefore, we need a solution that performs pre-processing of these data. We are preparing to get into the methods used for data mining. This phase includes data cleaning, data integration, data transfer, data reduction.

One of the famous methods used for data extraction and data cleaning is the Scatter diagram (Figure 1), which is available in most data mining software and statistical analyses.

As observed, the structure of each of these points determines a row from one of the cells. The closer cells indicate that this cell is close to our data, and farther ones mean that these data do not adequately agree with the characteristics of the other data. If the number of the farther pillars of a sample is higher than those of other samples, its respective data should be cleaned. If our data differs from a few data sets, we need to integrate all data into one data set; finally, by selecting the feature in the software, we eliminate variables that have a less significant impact on the prediction of the disease.

**Table 1.** Used the Variable in Study

| Type | Data Range | Name |
|---|---|---|
| Number of times pregnant | 0-17 | Numerical |
| Plasma glucose concentration | 0-199 | Numerical |
| Diastolic blood pressure | 0-122 | Numerical |
| Triceps skin fold thickness | 0-99 | Numerical |
| 2-Hour serum insulin | 0-846 | Numerical |
| Body mass index | 0-1.67 | Continual |
| Diabetes pedigree function | 42.2-78 | Continual |
| Age | 21-81 | Numerical |

### Data Modeling

Data mining was used to predict the probability of diabetes from classification. This model is one of the most commonly used methods of machine learning for prediction of medical data (14). In data sorting algorithms, the division of data into two educational and experimental parts is 75 to 25 and using the technique of the method, the fold-10 division of the models are created, the fold-10 method randomly divides the whole data into 10 sections. Therefore, each time, one of the 10 parts is considered experimental data and the other 9 are considered educational data. With the increase in this section, the results may be more favorable yet time-consuming. Finally, we build our model data set and use the experimental data set to investigate the accuracy of the model. The data mining categorization algorithms include decision tree, SVM, Bayesian network and neural networks.

### Decision Tree

The decision tree is one of the most powerful tools for classification and prediction which is capable of generating understandable manifestations of the relationships existing in a data set. The decision structure can also be introduced in the form of mathematical and computational techniques

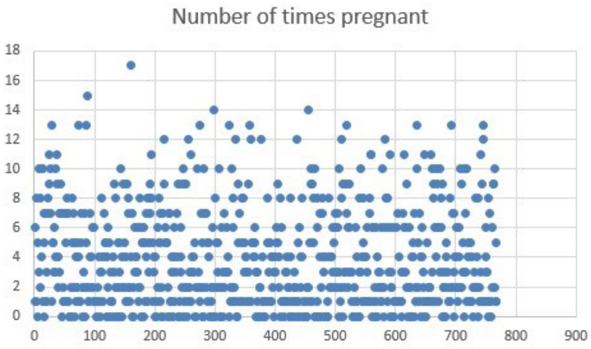

**Figure 1.** Scatter Diagram.

that help describe, categorize and publicize a data set. the decision tree is a unique way of providing a system that facilitates future decisions and makes the system define an appropriate way. The most important feature of decision tree is the ability to break the complex decision-making process into a simpler set of decisions that can easily be interpreted (15).

The decision tree is an explicit description using the decomposition of the algorithm. This tree structure is similar to the flowchart, including the highest node, represented by the root of the tree, branches, representing the outputs of the test, and the leaves, which represent the nodes or the distribution of the categories (16). The rules created by the decision tree are expressed as "then" and "if". C5.0 is a well-known algorithm that is a decision tree. C5.0 is an algorithm for making decision trees.

ID 3 algorithm can be used to express the classification, as with a decision tree or set of rules (17). In many applications, the set of rules is preferred because their perception is simpler than that of a decision tree.

## Support Vector Machine

The basis of the categorization of this algorithm is the linear classification of data, and in the division of the data line, the line that it chooses has the greatest interval of confidence. In a learning process involving two classes, the objective is to find a function for classification so that members of the two classes can be identified in the data set (18).

## Bayesian Network

This algorithm assumes a categorization of objects according to the law of Bayes, and assumes input variables independent of each other. It has a very simple structure and, despite its simplicity, has a high predictive accuracy, such that Wu et al reported this algorithm as being the most effective algorithm to anticipate recurrence of breast cancer (19).

## Neural Networks

Artificial neural network is an algorithmic learning method that has been developed from the human brain and is used in statistical fields, artificial intelligence, and classification. The Neural Network consists of several layers called the input layer, the hidden layer(s), and the output layer. The neural networks are divided into 2 types in terms of the connection of the nodes.

1. Antecessor nerve networks: The nodes of each layer are connected only to the next layer (Figure 2).
2. Retrogressive neural networks: The nodes of each layer are connected to the nodes of the next layer or to themselves.

In this study, we used antecessor neural networks or multilayer perception (20).

The output of the artificial neural network algorithm is in the form of a black box. These networks can be used as appropriate methods for generating analytical and estimating models and using different data (20).

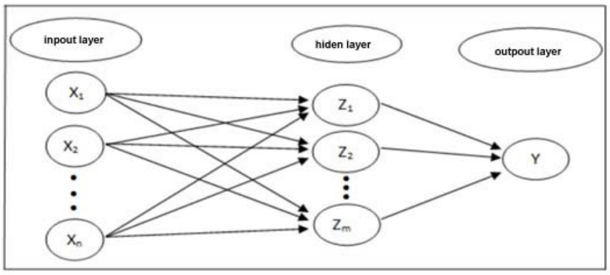

**Figure 2.** View of Neural Network.

Data evaluation: In this study, for the evaluation and analysis of the accuracy, sensitivity and specificity criteria, which are briefly described below, confusion matrix was used:

Specificity: If the answer is negative for a person, in a low percentage of cases, the model will also have a negative result. In other words, if the test is very specific and positive, we can be relatively sure that the person will develop diabetes. It is calculated by equation 1.

$$Specificity = \frac{TN}{FP + TN} \tag{1}$$

Sensitivity: If the answer is positive for a person, he/she will also have a positive result in a low percentage of the cases. In other words, if the test is very sensitive and the answer to it is negative, we can almost be sure that the person will not develop diabetes. It is calculated by equation 2.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

Accuracy: This criterion is defined as the percentage of correct classes and is calculated by equation 3.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

Where:

TN: The real category is negative and the algorithm is correctly detected as negative.

TP: The real category is positive and the algorithm is recognized correctly.

FP: The real category is negative and the algorithm has been mistakenly detected as positive.

FN: The real category is positive and the algorithm is detected by a negative error.

## Results

The artificial neural networks, Bayesian networks, SVM, and C5.0 algorithms were studied in the data set. The precision produced for training and testing data is

**Table 2.** Accuracy Values Calculated foe Models

| Algorithm Type | Algorithm Name | Learning Data | Test Data |
|---|---|---|---|
| Decision tree | C5.0 | 81.01 | 78.11 |
| Neural network | NN | 77.70 | 77.31 |
| Bayesian network | BN | 79.26 | 70.61 |
| Support vector machine | SVM | 78.4 | 75.77 |

**Table 3.** Index Size for Studied Algorithms

| Algorithm Type | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Decision tree | 80.2 | 59.5 | 87.5 |
| Neural network | 77.6 | 55.13 | 84.73 |
| Bayesian network | 77.08 | 54.36 | 85.4 |
| Support vector machine | 72.73 | 66.81 | 82.37 |

according to Table 2. The highest accuracy was obtained using the decision tree using the C5.0 algorithm, so this algorithm is used to predict diabetes.

Table 3 shows the values of the indexes calculated for each of the studied algorithms. For the C5.0 algorithm, by means of confusion matrix, accuracy, sensitivity, and specificity were calculated at 80.2, 59.5, and 87.5, respectively. These values represent that the Tree can produce comprehensive rules to predict the diagnosis of diabetes.

## Ranking the Importance of Variables

In the C5.0 algorithm model, the order of importance of the variables used to predict the response variable is shown in Figure 3.

According to Figure 3, the variables plasma glucose concentration, age, parity, diabetes pedigree function, and body mass index are most important for predicting diabetes.

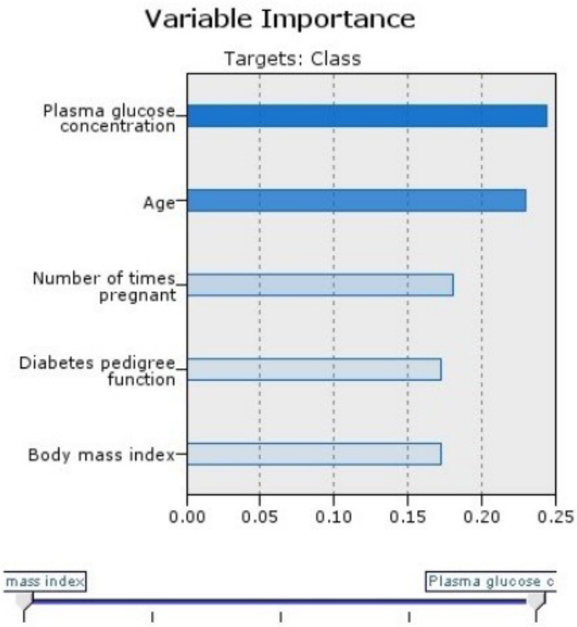

**Figure 3.** Importance of Variables.

## Discussion

In this study, using data mining algorithms, we sought to draw a model to predict the risk of diabetes by using C5.0 decision tree algorithms, neural networks, SVM, and Bayesian networks. Among the models produced, the C5.0 model has the highest accuracy to predict development of diabetes.

Gao et al (17) created a system of data processing for type 2 diabetes by combining C4.5 algorithms. Huang et al (21) conducted a study to identify the major factors influencing diabetes control, by using Feature Selection in the patient management system. Han et al (22) applied the Rapid Miner software using the ID3 Decision Tree algorithm to diabetic patients database. Anbananthen et al (23) used artificial neural network and decision tree developed by using the C4.5 algorithm for detecting individuals with diabetes based on age-related and blood pressure characteristics. Fang (24) clustered the data of patients with diabetes using different techniques. The features that are important in these models are age, family history, and weight. The accuracy of the model is based on 80% clustering.

The results show that blood glucose concentration and increased age are two major contributor to diabetes. Comparison of previous research findings on data mining and predicting diabetes clearly shows the model presented in this study has a high accuracy.

Increasing the accuracy of identifying people with diabetes depends on the magnitude of the database, so in future studies, by using larger databases, the number of records and the accuracy of the algorithm can be increased. Data mining can also be used to reduce the processing time of feature selection algorithms in order to reduce the number of variables. It is possible to use this method to develop decision-making systems of medicine to help diagnose diabetes in healthcare centers.

## Conclusion

In this study, a systematic effort was made to identify and review machine learning and data mining approaches for diabetes. Diabetes is rapidly emerging as one of the greatest global health issues of the 21st century.

Data mining is a valuable asset to deal with the abundant clinical data collected from patients and generated from the research and management of diabetes, so that researchers and clinicians can be assisted in providing better health care for the patients affected by this disease of the modern society.

The results showed that the C5.0 algorithm has the highest accuracy, specificity, and sensitivity compared to other methods studied. Therefore, the C5.0 algorithm has the best performance among other algorithms and is introduced as the most effective method to treat diabetes using available data types.

By comparison with other methods, it seems that:

- Speed - C5.0 is significantly faster than the other networks (neural networks, SVM, and Bayesian);
- Memory usage - C5.0 is more memory efficient.
- Smaller decision trees - C5.0 yields similar results to C4.5 with considerably smaller Decision Trees;
- Support for boosting - Boosting improves the trees and makes them more accurate; and
- Weighting - C5.0 allows to weight different cases and misclassification types.

From biological perspective, blood glucose concentration could be an effective biomarker for diagnosis and recent research has been aimed to present genomic elements.

## Conflict of Interest Disclosures

The authors declare no potential conflicts of interest relevant to this article.

## Acknowledgements

This article was derived from a master's thesis on computer engineering approved by Islamic Azad University of Malayer.

## References

1. Janahmadi Z, Nekooeian AA, Mozafari M. Hydroalcoholic extract of Allium eriophyllum leaves attenuates cardiac impairment in rats with simultaneous type 2 diabetes and renal hypertension. Res Pharm Sci. 2015;10(2):125-33.
2. Nazarzadeh M, Bidel Z, Sanjari Moghaddam A. Meta-analysis of diabetes mellitus and risk of hip fractures: small-study effect. Osteoporos Int. 2016;27(1):229-30. doi: 10.1007/s00198-015-3358-9.
3. El-Sappagh S, Elmogy M, Riad AM. A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis. Artif Intell Med. 2015;65(3):179-208. doi: 10.1016/j.artmed.2015.08.003.
4. Jurian N, Ashoori M. Predicting the effectiveness of preeclampsia medications based on dose and method of drug consumption using data mining. The Iranian Journal of Obstetrics, Gynecology and Infertility. 2014;17(123):13-22. doi: 10.22038/ijogi.2014.3588.
5. Amereh M. Survey data mining algorithms and compare them on a case study. Ecommerce. 2014;71:40-2. [Persian].
6. Su CT, Yang CH, Hsu KH, Chiu WK. Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data. Comput Math Appl. 2006;51(6-7):1075-92. doi: 10.1016/j.camwa.2005.08.034.
7. Purnami SW, Embong A, Zain JM, Rahayu SP. A new smooth support vector machine and its applications in diabetes disease diagnosis. J Comput Sci. 2009;5(12):1003-8. doi: 10.3844/jcssp.2009.1003.1008.
8. Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. ScientificWorldJournal. 2015;2015:581501. doi: 10.1155/2015/581501.
9. Gulcin Yıldırım E, Karahoca A, Ucar T. Dosage planning for diabetes patients using data mining methods. Procedia Comput Sci. 2011;3:1374-80. doi: 10.1016/j.procs.2011.01.018.
10. Jaafar SFB, Ali DM. Diabetes mellitus forecast using artificial neural network (ANN). Asian Conference on Sensors and the International Conference on New Techniques in Pharmaceutical and Biomedical Research; 2005. doi: 10.1109/ASENSE.2005.1564523.
11. Baron-Epel O, Heymann AD, Friedman N, Kaplan G. Development of an unsupportive social interaction scale for patients with diabetes. Patient Prefer Adherence. 2015;9:1033-41. doi: 10.2147/ppa.s83403.
12. Shearer C. The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing. 2000;5(4):13-22.
13. Asuncion A, Newman D. UCI machine learning repository. 2007. Available from: https://archive.ics.uci.edu/ml/index.php.
14. Hische M, Larhlimi A, Schwarz F, Fischer-Rosinsky A, Bobbert T, Assmann A, et al. A distinct metabolic signature predicts development of fasting plasma glucose. J Clin Bioinforma. 2012;2:3. doi: 10.1186/2043-9113-2-3.
15. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013;29(2):93-9. doi: 10.1016/j.kjms.2012.08.016.
16. Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2006.
17. Gao J, Luo SL, Jia HB, Zhang TM, Han YW. Type 2 diabetes data processing with EM and C4. 5 algorithm. IEEE/ICME International Conference on Complex Medical Engineering; 2007:371-7. doi: 10.1109/ICCME.2007.4381759.
18. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10(5):988-99. doi: 10.1109/72.788640.
19. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst. 2008;14(1):1-37. doi: 10.1007/s10115-007-0114-2.
20. Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Heidelberg: Springer; 2001.
21. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med. 2007;41(3):251-62. doi: 10.1016/j.artmed.2007.07.002.
22. Han J, Rodriguez JC, Beheshti M. Diabetes data analysis and prediction model discovery using rapidminer. Second International Conference on Future Generation Communication and Networking; 2008; 96-9. doi: 10.1109/FGCN.2008.226.
23. Anbananthen KSM, Sainarayanan G, Chekima A, Teo J. Artificial neural network tree approach in data mining. Malaysian Journal of Computer Science. 2007;20(1): 51-62.
24. Fang X. Are you becoming a diabetic? A data mining approach. Sixth International Conference on Fuzzy Systems and Knowledge Discovery; 2009. doi: 10.1109/FSKD.2009.807.