Decision Tree Discovery for the Diagnosis of Type II Diabetes

Asma A. AlJarullah
Department of Information Systems, College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
Asma.aljarullah@gmail.com

***Abstract*—The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in database and generate clear and understandable description of patterns. In this study, decision tree method was used to predict patients with developing diabetes. The dataset used is the Pima Indians Diabetes Data Set, which collects the information of patients with and without developing diabetes. The study goes through two phases. The first phase is data pre-processing including attribute identification and selection, handling missing values, and numerical discretization. The second phase is a diabetes prediction model construction using the decision tree method. Weka software was used throughout all the phases of this study.**

Keywords- Decision Tree; diabetes; data mining; diagnosis

I. INTRODUCTION

Healthcare information systems tend to capture data in databases for research and analysis in order to assist in making medical decisions. As a result, medical information systems in hospitals and medical institutions become larger and larger and the process of extracting useful information becomes more difficult. Traditional manual data analysis has become inefficient and methods for efficient computer based analysis are essential. To this aim, many approaches to computerized data analysis have been considered and examined. Data mining represents a significant advance in the type of analytical tools currently available. It has been shown to be a valid, sensitive, and reliable method to discover patterns and relationships. It has been proven that the benefits of introducing data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to reduce human resources [1][2].

Data Mining is a technique employed to extract non-trivial information from large datasets. There are literally thousands of medical databases throughout the world, each providing valuable information. Clinicians and medical researchers should be able to examine (data mining) and seek information that might allow them to more accurately diagnosis disease, and also assist medical science in the development of rational approaches to palliative/curative medicine.

In recent times, the number of people suffering from diabetes is increasing day by day. It is a disease in which body does not produce insulin or use it properly. This increase the risks of developing, kidney disease, blindness, nerve damage, blood vessel damage and contribute to heart disease [3]. There are two types of diabetes; Type-1 diabetes - also called insulin dependent and type-2 diabetes which is with relative insulin deficiency. Patients with type 2 diabetes do not require insulin cure to remain alive, although up to 20% are treated with insulin to control blood glucose levels [4]. To diagnose diabetes disease at an early stage is quite a challenging task due to complex inter dependence on various factors. There is a critical need to develop medical diagnostic decision support systems which can aid medical practitioners in the diagnostic process. This study deals about the classification/prediction of Type II diabetes.

The dataset used in this study is "The Pima Indians Diabetes Data Set" which was taken from the UCI Machine Learning Repository [5]. The original owner of this data set is the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of this dataset from larger database. In particular, all patients selected are females at least 21 years old of Pima Indian heritage.

Weka software package was used throughout this study. Weka software is a collection of machine learning algorithms for data mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. Weka system is open source software issued under GNU General Public License, where it can be modified by anybody for use [6].

II. RELATED WORK

Many researches have been conducted in the field of diabetes diagnosis systems. A diabetes expert system have been proposed in [7] which adopts a forward, backward and forward-backward chaining inference mechanism and an

uncertainty handling method which can quickly and efficiently, based on the patient's symptoms, judge the possibility of illness, its severity, and its potential complications. Authors in [8] have conducted data analysis on the PIMA dataset and constructed a decision tree prediction model using RapidMiner, they handled the missing values in the data set by removing columns that have very large number of missing values and records that have zero values in any one of the attributes even with attributes that may actually have the value of zero, and this was handled in this study as what will be seen in later sections of this paper. In [2] authors have constructed an artificial neural network model for diagnosis of diabetes, they used certain combination of preprocessing techniques to handle the missing values and compared the results of accuracy of the model for each technique, however the method of handling missing values presented in this paper wasn't employed in that study. Authors in [9] have constructed association rules for classification of type -2 diabetic patients. They generated 10 association rules to identify whether the patient goes on to develop diabetes or not.

III. METHODOLOGY

This study consists of two stages, data pre-processing and decision tree construction. The data pre-processing phase aims at preparing the dataset for the second phase. The second phase includes using one of the decision tree algorithms to construct a decision tree model for the prediction of patients with developing diabetes.

A. Data Pre-processing

Data gathering methods are often loosely controlled, resulting in some errors in data such as abnormal values, impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. The phrase "Garbage In, Garbage Out" is particularly applicable to data mining and machine learning projects. Thus, the representation and quality of data is first and foremost before running an analysis [18]. It has been estimated that data preparation alone accounts for 60% of all the time and effort expanded in the entire data mining process [10]. Many data preprocessing techniques are given in [11]. In this study, the techniques used are attribute identification and selection, handling missing values, and numerical discretization.

1) Attribute Identification and Selection

The dataset consists of 9 attributes as shown in Table 1. The dataset was collected from 768 female. All of them were randomly selected such that they are at least 21 years old of Pima Indian heritage. Important attributes that are very related with diabetes were recorded for each female like the number of times pregnant and age in years. Other important attributes like Plasma glucose concentration measured using a two-hour oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-Hour serum insulin, body mass index, and diabetes pedigree function. Figure 1 shows a sample of the dataset.

According to [12], all of the given attributes have some effect on developing diabetes, thus all of them were gathered from the dataset and used for further cleaning in the following steps.

TABLE I
THE PIMA INDIAN DIABETES DATASET ATTRIBUTES

Attribute	Description	Type
Pregnant	A record of the number of times the patient pregnant	Numeric
Plasma-Glucose	Plasma glucose concentration measured using a two-hour oral glucose tolerance test (mm Hg)	Numeric
Diastolic BP	Diastolic blood pressure	Numeric
Triceps SFT	Triceps skin fold thickness (mm)	Numeric
Serum–Insulin	Two-hour serum insulin (muU/ ml)	Numeric
BMI	Body mass index(weight Kg/height in $(mm)^2$	Numeric
DPF	Diabetes pedigree function	Numeric
Age	Age of the patient (years)	Numeric
Class	Diabetes on set within five years	Nominal

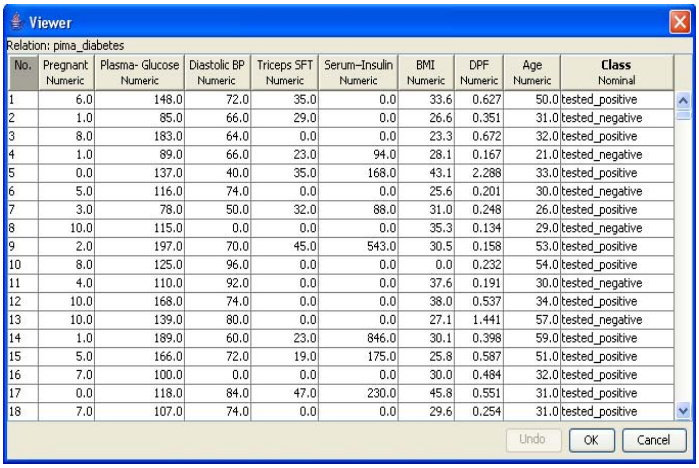

Figure 1. Sample of the Pima Indian diabetes dataset

2) Handling Missing Values

By analyzing the Pima Indian Diabetes dataset, it has been found that there are some missing values that have been filled in with 0's. The numbers of missing values for each attribute are as follows:

- Pregnant : 110
- Plasma- Glucose : 5
- Diastolic BP : 35
- Triceps SFT : 227
- Serum–Insulin : 374
- BMI : 11
- DPF : 0

- Age : 0
- Class : 0

There are four ways for handling missing values. The first and easiest way to deal with missing values is simply delete all the cases with missing values for the variable under consideration. This technique however may lead to the loss of potentially valuable information about patients whose values are missing. The second approach is to replace all missing values with the mean. But this technique might bias the database. The third technique is to replace all the missing values with zeros. However, this technique may lead to poor classification. The fourth technique K-nearest neighbor method replaces missing values in data with the corresponding value from the nearest-neighbor column. The nearest-neighbor column is the closest column in Euclidean distance. However, this technique also might bias the dataset [13].

In this study, the missing values have been handled in the following way:

- For the zero values of the attribute Pregnant, the zero values are left assuming that they are the real values.
- For the zero values of the attributes Plasma- Glucose, Diastolic BP, and BMI, the whole instance was removed, since the number of missing values in these attributes is very small.
- Finally, since the attributes Triceps SFT and Serum–Insulin have very large number of missing values (227 zero values for Triceps SFT and 374 zero values Serum–Insulin), so the instances cannot be removed, instead the attribute themselves were removed in order to keep as much accurate information as possible [9].

After handling the missing values, only 724 instances are remain out of 768 with 5 attributes: plasma-glucose, diastolic BP, BMI, DPF and age.

3) Numerical Discretization

Discretization of numerical attributes was made to reduce the complexity of the problem. Another advantage is that some classifiers that can take numeric attributes can achieve improved accuracy if the data is discretized prior to learning. The values of each attribute were grouped or binned into a small number of groups or bins. Discretization of numerical attributes can be either manually done based on user specification or automatically done through Weka software. In [8], the authors discretized the attributes using information from various sources including [12] [14] [15]. In [9], the authors used approximate equal interval binning and also taken advice from medical experts. In this study, the attributes were discretized taking into consideration the discretization mentioned in [8] and [9] with some changes to match the dataset after handling the missing values. Table 2 summarizes the cut-off values along with the names of the bins for each attribute.

TABLE 2
DISCRETIZATION BINS FOR THE PIMA INDIAN DIABETES DATASET ATTRIBUTES

Attribute	Bins	Source
Pregnant	low (0,1), medium (2, 3, 4, 5), high (>6)	[8]
Plasma-Glucose	low (<95), medium (95-140), high (>140)	[12]
Diastolic BP	normal (<80), normal-to-high (80-90), high (>90)	[9] [14]
BMI	low (<24.9), normal (25-29.9), obese (30-34.9), severely-obese (>35)	[9] [15]
DPF	low (<0.5275), high (>0.5275)	Automatically through Weka
Age	range1 (<28.5), range2 (>28.5)	Automatically through Weka

Figure 2 shows a sample of the dataset after applying the previous pre-processing techniques.

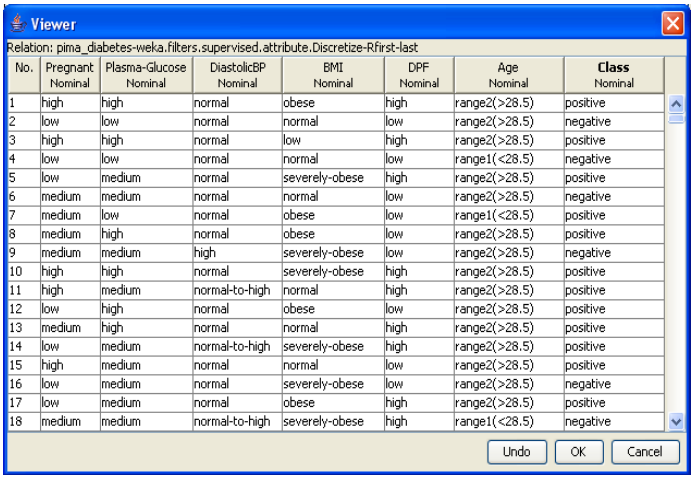

Figure 2. Sample of the Pima Indian diabetes dataset after pre-processing

B. Decision Tree Construction

After the dataset has been prepared, Weka software was used to construct the decision tree. With choice of Weka knowledge explorer (exploratory data analysis) which is an easy to use graphical interface, the dataset file was uploaded, and then all attributes were selected for classification. The dataset was then classified by choosing the J48 algorithm which is a decision tree learner and is the implementation of Quinlan C4.5 in Weka software. C4.5 is a program for machine learning [16].

The test option chosen was 10-fold cross-validation. Cross-validation is a method of estimating the accuracy of a classification and it works as follows:

1. Separate data in to fixed number of partitions (or folds)
2. Select the first fold for testing, whilst the remaining folds are used for training.
3. Perform classification and obtain performance metrics.
4. Select the next partition as testing and use the rest as training data.
5. Repeat classification until each partition has been used as the test set.

6. Calculate an average performance from the individual experiments.

10-fold cross validation was chosen because experiments have shown that this is the best choice to get an accurate estimate [16][17].

1) The Decision Tree Constructed Using Weka Software

By applying Weka's classifier J48 decision tree algorithm on the dataset, the decision tree was generated; it can be seen in Figure 3. Plasma-Glucose was made the root. If it is low, then diabetes result is negative which means that the patient is not developing diabetes. If the Plasma-Glucose is medium, then the age of the patient is tested, if it is less than 28.5, then the diabetes result negative, and if the age is greater than 28.5, then the patient's DPF is tested, if it is low then the diabetes result is negative and so on as indicated by the tree. The numbers between brackets at leaf nodes indicate the number of training instances associated with the leaf node, and the number of instances that were incorrectly classified. Some leaf nodes have only one number, which means that all associated instances were correctly classified.

2) Confusion Matrix and Accuracy of the Decision Tree

From the selected attributes and 724 instances of the dataset, 566 instances were correctly classified representing 78.1768 % of total records and 158 instances were incorrectly classified representing 21.8232 % of total records. The confusion matrix was as follows:

a	b	← classified as
417	57	a = negative
101	149	b = positive

The entries in the confusion matrix have the following meaning in the context of this study; Letters a, b are instances which are actually non-diabetic (negative), diabetic (positive) respectively which have then been classified as any of the two classes, a= non-diabetic (negative), or b= diabetic (positive). The false-positive value is the number of instances that were misclassified as positive (diabetic) and equals 101, whereas the false-negative value is the number of instances that were misclassified as negative (non-diabetic) and equals 57. Since the misclassification tends to be false-positive by about 64% of the misclassified instances, this means that much of the accuracy of the resulting decision tree was lost because of the false positive error, which is of no risk compared with false negative error.

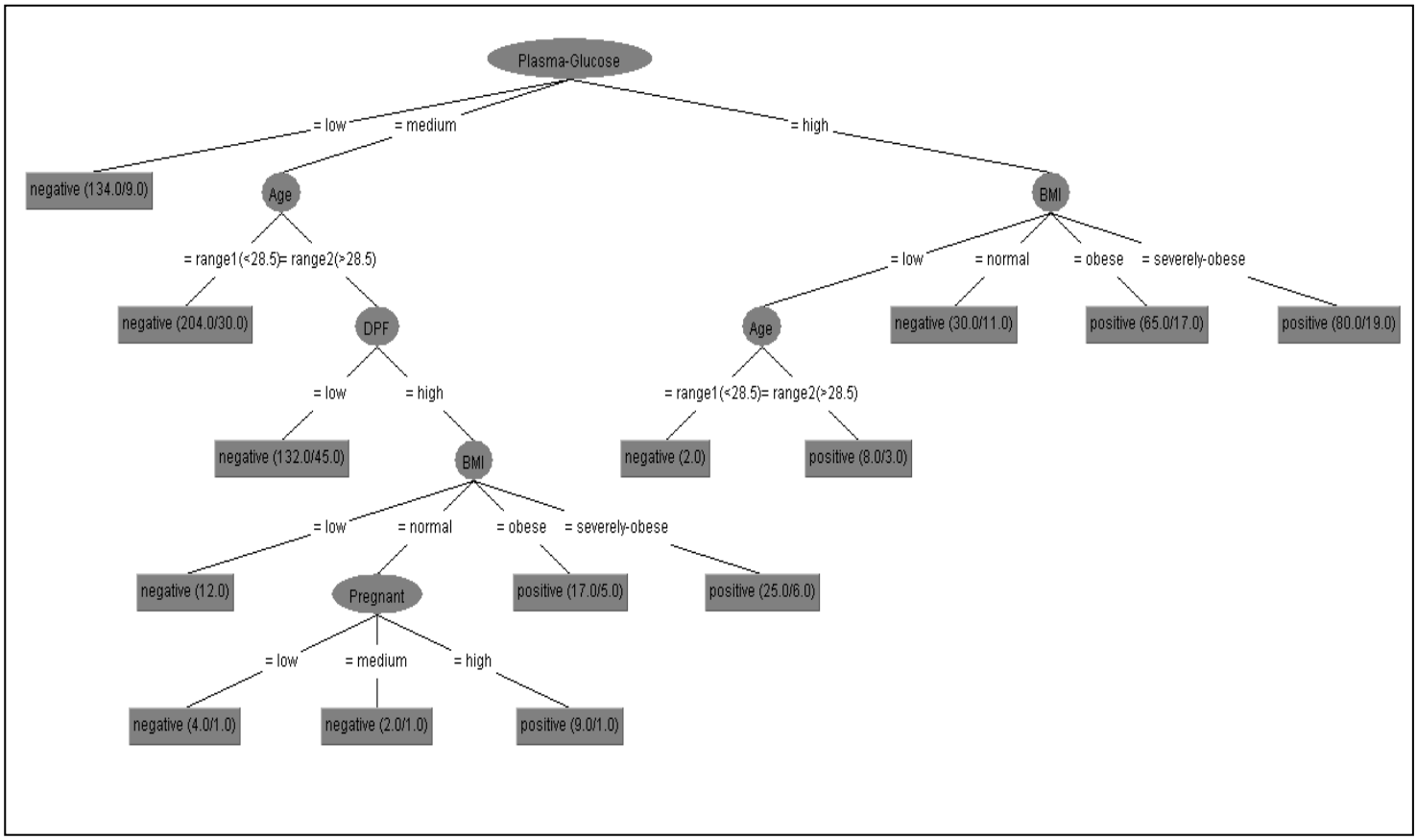

Figure 3. The J48 decision tree for diabetes diagnosis using the Pima Indian diabetes dataset

IV. Discussion and Cnclusion

The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in database and generate clear and understandable description of patterns.

This study aimed at the discovery of a decision tree model for the diagnosis of type 2 diabetes. The dataset used was the Pima Indian diabetes dataset. Pre-processing was used to improve the quality of data. The techniques of pre-processing applied were attributes identification and selection, handling missing values, and numerical discretization. Next, Weka's J48 decision tree classifier was applied to the modified dataset to construct the decision tree model. The accuracy of the resulting model was 78.1768%.

Considering the Pima Indian diabetes dataset, there might be other risk factors that the data collections did not consider. According to [12] other important factors include gestational diabetes, family history, metabolic syndrome, smoking, inactive lifestyles, certain dietary patterns etc. The proper prediction model would need more data gathering. This can be achieved by collecting diabetes datasets from multiple sources, generating a model from each dataset, then combining the results of the generated models in order to find a best prediction model. And the most important thing before all of this can be done; the datasets themselves must be available.

References

[1] Marjan Khajehei, Faried Etemady, "Data Mining and Medical Research Studies," cimsim, pp.119-122, 2010 Second International Conference on Computational Intelligence, Modelling and Simulation, 2010

[2] Jayalakshmi, T.; Santhakumaran, A.; , "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks," Data Storage and Data Engineering (DSDE), 2010 International Conference on , vol., no., pp.159-163, 9-10 Feb. 2010

[3] E.I.Mohamed, R.Linderm, G.Perriello,N.Daniele, S.J.Poppl, & A.DeLorenzo. "Predicting type 2 diabetes using an electronic nose-based artificial neural network analysis," Diabetes nutrition & metabolism, 15(4),215–221.202.

[4] J.C.Pickup, G. Williams, (Eds.), Textbook of diabetes, Blackwell Science, Oxford.

[5] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[6] WEKA, by university of Waikato, http://www.cs.waikato.ac.nz/ml/weka/

[7] Jiang Ming-Yan; Chen Zhi-Jian; , "Diabetes expert system," Intelligent Processing Systems, 1997. ICIPS '97. 1997 IEEE International Conference on , vol.2, no., pp.1076-1077 vol.2, 28-31 Oct 1997

[8] Jianchao Han; Rodriguez, J.C.; Beheshti, M.; , "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner," Future Generation Communication and Networking, 2008. FGCN '08. Second International Conference on , vol.3, no., pp.96-99, 13-15 Dec. 2008

[9] Patil, B.M.; Joshi, R.C.; Toshniwal, D.; , "Association Rule for Classification of Type-2 Diabetic Patients," Machine Learning and Computing (ICMLC), 2010 Second International Conference on , vol., no., pp.330-334, 9-11 Feb. 2010

[10] Pyle, D., 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, California.

[11] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Preprocessing for Supervised Leaning", International Journal of Computer Science, 2006, Vol 1 N. 2, pp 111-117.

[12] Seibel, J. A. (2007) Diabetes Guide, WebMD, http://diabetes.webmd.com/guide/oral-glucose-tolerance-test.

[13] Jayalskshmi, T.; Santhakumaran, A.; , "Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks," Machine Learning and Computing (ICMLC), 2010 Second International Conference on , vol., no., pp.109-112, 9-11 Feb. 2010

[14] Stein, D. W. (2006) Hypertension / High Blood Pressure Guide, WebMD, http://www.webmd.com/ hypetensiondiagnosing- high-blood-pressure

[15] Zelman, K. M. (2008), How Accurate is Body Mass Index, or BMI? WebMD, http://www.webmd.com/diet/features/how-accurate-body-mass-index-bmi

[16] I. H. Witten and E. Frank, "Data mining," Practical Machine Learning Tools and Techniques, 2000.

[17] Andrew Roberts, "AI32 —Guide toWeka", 2005.

[18] Data Pre-processing - Wikipedia, the free encyclopedia, <http://en.wikipedia.org/wiki/Data_Pre-processing>