Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE) Vol. 3, Issue 2 (2016) ISSN 2349-1469 EISSN 2349-1477

Diognosis of Diabetes with the Method of Data Mining

Seher Arslankaya[1] and Güneş Çalık[2]

***Abstract*—**The field in which information most rapidly changes in the sense of content and structure is the health sector. In order to offer health services rapidly, accurately, and with a high quality, health professionals should reach the most accurate and updated information and use this information by benefiting from the decision support systems. A great deal of information entered into the hospital information system is saved in the databases. This saved information both offers patients a service understanding of a higher quality and contains valuable information for medical studies. It may be hard to draw a significant conclusion by considering all this information and by analyzing all of it simultaneously. At this point, data mining helps information to occur significantly. The use of Data Mining as a decision support tool when health professionals take the most optimal decisions will be helpful. Data mining is used in many fields in the health sector. In this study, data mining was used for the early diagnosis of Diabetes or Diabetes Mellitus with its name in the literature. The blood values and patient complaints affecting the disorder were evaluated, and at what rates these factors caused the disease was investigated by employing a few of the data mining techniques. Classification was made in the WEKA program with the C4.5, K-Nearest Neighbor and NaiveBayes Algorithms as well as the Multi-Layer Perceptron, and the results obtained were compared and evaluated. At the end of the study, a decision tree was generated with the optimum classification method and software was developed in order to be used for the early diagnosis of newly arriving patients.

***Keywords*—** Data Mining, Diagnosis of Diabetes, WEKA.

I. Introduction

Diabetes or Diabetes Mellitus with its name in the literature is a chronic metabolism disease in which an organism fails to benefit adequately from carbohydrates (CH), fat, and proteins due to insulin deficiency or the defects under the influence of insulin and which continually requires medical care [1].

Before doing blood tests, such complaints as xerostomia, frequent urination, drinking water excessively, foot burning and numbness, arthralgia, loss of weight, and asthenia help to diagnose diabetes. A patient who carries these criteria only cannot definitely be diagnosed with diabetes. In order to make the diagnosis, the blood tests must also be at the disease values besides the complaints of the patient.

The Fasting Blood Sugar, the Oral Glucose Tolerance Test, the Random Plasma Glucose and HbA1c are measured when diagnosing diabetes.

Regarding the Fasting Blood Sugar result, the range 100-125 mg/dl is considered prediabetes, whereas the values equal to and greater than 126 mg/dl are diagnosed as diabetes. The Oral Glucose Tolerance Test result below 140 mg/dl is considered normal. The range 140-199 mg/dl is interpreted as prediabetes, and the values equal to and greater than 200 mg/dl are interpreted as diabetes. If the random blood sugar measurement at any time of the day is equal to or greater than 200 mg/dl, it is diabetes according to the result of the random plasma Glucose measurement. The blood glucose level in the last 3 months is understood according to the result of the HbA1c blood analysis test. The diagnosis of diabetes is reached when values equal to and greater than 6.5% are measured [2].

The hospital information systems contain plenty of recorded patient information of vital importance. Kept in databases, this information is made more significant by using the method of data mining and plays an efficient role in the receiving of services of a higher quality by specialists, the hospital management, and patients. In this study, it was aimed to diagnose diabetes early by employing the methods of data mining and the classification algorithms so as to act early in the diagnosis of the disease, to save time in diagnosis, and for a decrease in the costs of tests.

II. Implementation

Some 369 patient data were studied for the diagnosis of diabetes. The results of the laboratory blood analyses for the diagnosis of diabetes were selected from these data and used. By evaluating the fasting blood sugar, postprandial blood sugar, and HbA1c blood test results as well as patients' ages and complaints of foot burning and numbness, arthralgia, asthenia, the desire to drink an excessive amount of water, pregnancy, a sudden increase in sugar, and frequent urination, a classification was made so as to determine the criteria of priority for the diagnosis of diabetes.

Version 3.8 of the WEKA program was used as the data mining tool for early diagnosis by finding the factors of priority which affected diabetes. At the stage of evaluation of the data by means of the WEKA program, the sample group was first of all divided into two groups as "patients with diabetes" and "healthy". To make better inferences from the data group, the 53% share of the data including the sample was determined as the training set to find out the algorithm and the rules, while the remaining 47% share was determined as the

[1] Sakarya University, Engineering Faculty, Industrial Engineering Department, Turkey

[2] Sakarya University, Engineering Faculty, Industrial Engineering Department, Turkey

http://dx.doi.org/10.15242/IJCCIE.U0716007

test set to examine the accuracy rates of the rules, algorithms, and patterns determined in the training set.

In this study, the methods of classification through the C4.5, K-Nearest Neighbor, Multi-Layer Perceptron and NaiveBayes Algorithms were employed and the most accurate classification technique to determine the factors affecting diabetes was investigated.

The data group used in this study, where data mining was applied in the field of health, is shown in Table 1.

TABLE I
DISTRIBUTION OF THE DATA TO BE USED FOR DATA MINING

Class	Number of People	Training set (Number of People)	Test Set (Number of People)
Healthy	48	23	25
Patients with Diabetes	321	150	171
Total	369	173	196

A. Generating the Decision Tree by means of the C4.5 Algorithm

An opportunity for generating decision trees on the databases including numerical values was provided by the C4.5 algorithm. Some challenges may be seen concerning dividing numerical qualities into certain ranges. However, various methods are available to compute the optimum t threshold value. Quality values are put in order and become $\{V_1, V_2, \dots , V_n\}$. The set of quality values is divided into two parts, and the mid-point of range $[V_i, V_{i+1}]$ may be considered the threshold value [3]:

$$t_i = (V_i + V_{i+1}) / 2 \tag{1}$$

A decision tree was generated by achieving 82.65% accuracy according to the C4.5 algorithm. The irregularity matrix according to the C4.5 algorithm is shown in Table 2.

TABLE II
THE IRREGULARITY MATRIX

Class	Total Number of Patients	a	b
a=Patients with Diabetes	150	139	11
b=Healthy	23	19	4

According to the irregularity matrix, 139 of 150 test data with diabetes patient values were matched as patients and 11 of them as healthy. In addition, of 23 healthy data, 4 people were matched as "healthy" but 19 people as "patients with diabetes". The elaborated table of accuracy for 150 people according to the irregularity matrix is shown in Table 3.

TABLE III
THE TABLE OF ACCURACY FOR C4.5

Class	Accuracy	Precision	Sensitivity	F-Measure
a=Patients with Diabetes	88.43%	0.88	0.927	0.903
b=Healthy		0.267	0.174	0.211

The resulting decision tree for the diagnosis of diabetes by using the available data through the C4.5 Algorithm is shown in Figure 1.

According to Figure 1, the presence of a sudden increase in sugar is an essential factor to directly make the diagnosis of diabetes. If the patient has not gone to a doctor before, the Fasting Blood Sugar greater than 164 mg/dl may be described as an important result in the diagnosis of diabetes. The blood analysis result of the Fasting Blood Sugar as 86 mg/dl in the case of pregnancy is considered adequate for diagnosis, and a postprandial blood sugar value greater than 180 mg/dl among the people over the age of 31 and the HbA1c test result as 7.5% among the people over the age of 57 are interconnected cases for the diagnosis of diabetes. Briefly, the Fasting Blood Sugar, Postprandial blood sugar, and HbA1c test results as well as age, pregnancy, and the complaints of arthralgia, asthenia, and frequent urination may be considered the criteria for the diagnosis of diabetes.

B. Data Modeling for the NaiveBayes Algorithm

In general, it is used to make classifications and estimations in cases of ambiguity. Its most important disadvantages are the failure to model the relationship among the variables and the assumption that the variables are completely independent of each other [4].

Let $X=\{x_1, x_{2,} .., x_n\}$ be a data set with unknown class membership and $C=\{c_1, c_{2,} .., c_n\}$ be n classes on the data set. According to the Bayes' theorem, probability $P(c_j \mid X)$ is computed as in Equations (2) and (3) by conditioning Set C on X [5].

$$P(x) = \sum_{i=1}^{n} P(x_i|C_j).P(C_j) \tag{2}$$

$$P(x) = \frac{P(x_i|C_j).P(C_j)}{P(x_i)} \tag{3}$$

According to the NaiveBayes Algorithm, matching occurred with 88.43% accuracy. The irregularity table for the NaiveBayes Algorithm is elaborated in Table 4

TABLE IV
THE IRREGULARITY MATRIX ACCORDING TO THE NAIVEBAYES ALGORITHM

Class	Total Number of Patients	a	B
a=Patients with Diabetes	150	139	11
b=Healthy	23	9	14

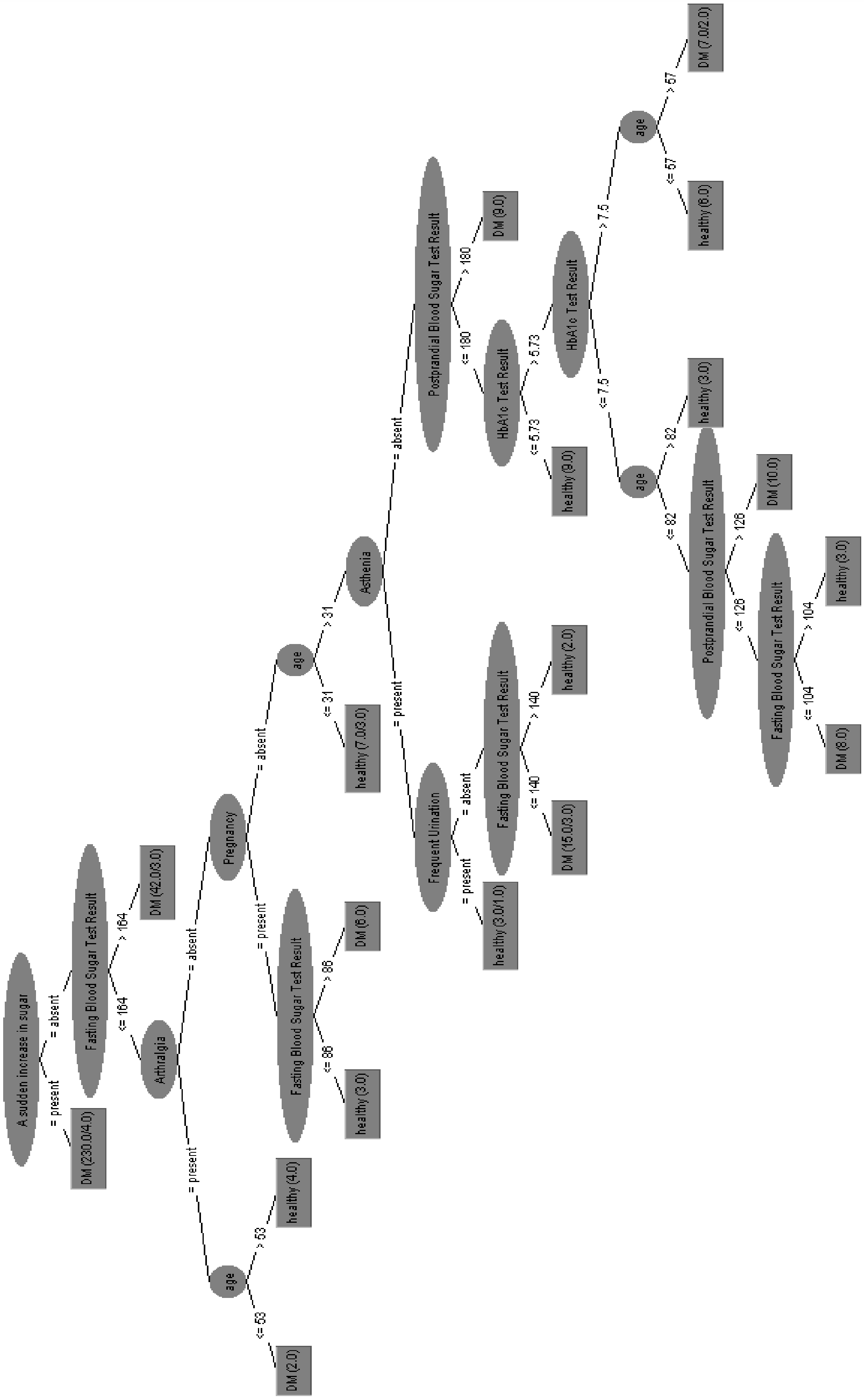


Fig. 1 A Decision Tree according to the C4.5 Algorithm

Of 150 people, 139 patients were accurately classified and 11 people were inaccurately classified by means of the NaiveBayes algorithm according to the irregularity matrix. Of 23 healthy people, 14 were accurately classified, while it was seen that 9 people were inaccurately classified. The table of Accuracy where the accuracy rate of 88.43% as well as precision, sensitivity and F-Measure are shown is elaborated in Table 5.

TABLE V.
THE TABLE OF ACCURACY ACCORDING TO THE NAIVEBAYES ALGORITHM

Class	Accuracy	Precision	Sensitivity	F-Measure
a=Patients with Diabetes	88.43%	0.939	0.927	0.933
b=Healthy		0.560	0.609	0.583

C. The K-Nearest Neighbor Algorithm and Classification Results

The K-Nearest Neighbor Algorithm is based on the logic of classifying the set that the object intended to be classified belongs to in the same set as the one belonging to the largest number of units out of its nearest K unit objects [6].

In general, Euclidean or Manhattan distance equations are used to calculate the distance between points in the KNN algorithm [7]

$$d(i,j) = \sqrt{\sum_{k=1}^{p} \left(x_{ij} - x_{jk}\right)^2} \tag{4}$$

K: Number of the nearest neighbors of the given point

D. Distance

According to the Irregularity matrix with the K-Nearest Neighbor Algorithm, it is seen that of a total of 150 patients with diabetes, 141 people were classified as “Patients with Diabetes” but 9 people as “healthy”. Of a total of 23 healthy people, 18 people were classified as “healthy” and 5 people as “Patients with Diabetes”. Considering these data, 91.90% accuracy was observed with the method of the k-nearest neighbor algorithm. In the k-nearest neighbor algorithm, the value of k was determined as 1. When k=1, the proximity of k data out of the data intended to be classified is considered when making a classification. That is, the results were obtained considering the proximity of 1 neighbor here. The irregularity matrix and accuracy table values of the program output are shown in Table 6.

TABLE VI
THE IRREGULARITY MATRIX FOR THE K-NEAREST NEIGHBOR ALGORITHM

Class	Total Number of Patients	a	B
a=Patients with Diabetes	150	141	9
b=Healthy	23	5	18

The table of Accuracy where the K-Nearest Neighbor Algorithm is evaluated is elaborated in Table 7.

TABLE VII
THE TABLE OF ACCURACY FOR THE K-NEAREST NEIGHBOR ALGORITHM

Class	Accuracy	Precision	Sensitivity	F-Measure
a=Patients with Diabetes	91.90%	0.966	0.94	0.953
b=Healthy		0.667	0.783	0.720

E. Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) concentrates on generating smart codes which imitate the learning mechanism of the human brain by setting up a parallel connected network model. In an MLP model, the system is trained first and the network can compute the outputs as the functional mapper by using the most recently updated network parameters [8].

The Multi-Layer Perceptron is trained by continually being backfed through the standard backpropagation algorithm. MLPs are supervised networks. That’s why an MLP requires the training of the desired response. The MLP learns how the input data will be converted into the desired response data; therefore, the MLP is widely used in the pattern classification. With one hidden layer or two hidden layers, the MLP enables the input and output maps to virtually converge. Many applications of artificial neural networks contain MLPs [9].

When the results of the MLP classification method performed in the WEKA program were interpreted, it was seen that of 150 patients, 144 people were accurately classified, but 6 people were inaccurately classified according to the irregularity matrix. It was also seen that of 23 healthy people, 14 people were inaccurately classified, whereas 9 people were accurately classified. The irregularity matrix according to the Multi-Layer Perceptron Algorithm is elaborated in Table 8.

TABLE VIII
THE IRREGULARITY MATRIX FOR THE MULTI-LAYER PERCEPTRON ALGORITHM

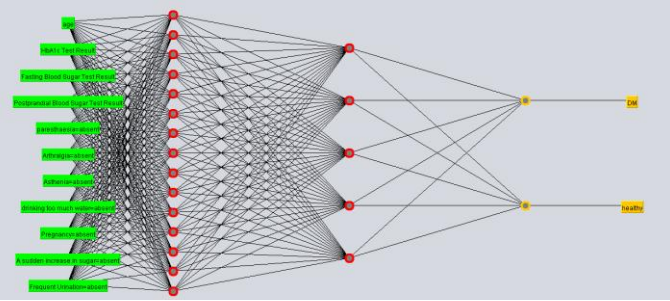
Class	Total Number of Patients	a	b
a=Patients with Diabetes	150	144	6
b=Healthy	23	14	9

According to the table of Accuracy shown in Table 9, it was observed that 349 of 369 people were accurately classified, thereby attaining 88.43% accuracy.

TABLE IX
THE TABLE OF ACCURACY FOR THE MULTI-LAYER PERCEPTRON ALGORITHM

Class	Accuracy	Precision	Sensitivity	F-Measure
a=Patients with Diabetes	88.43%	0.911	0.960	0.935
b=Healthy		0.600	0.391	0.474

2 hidden layers were used on the neural network that occurred in the multi-layer perceptron algorithm. The classification was made, with the 1st hidden layer consisting of 15 neural networks and the 2nd hidden layer comprising 5 neural networks. The structure of the neural network is elaborated in Figure 2.


age
HbA1c Test Result
Fasting Blood Sugar Test Result
Postprandial Blood Sugar Test Result
paresthaesia=absent
Arthralgia=absent
Asthenia=absent
drinking too much water=absent
Pregnancy=absent
A sudden increase in sugar=absent
Frequent Urination=absent
DM
healthy

Learning took place via the variables of age, HbA1c, fasting blood sugar, postprandial blood sugar, foot burning and numbness, arthralgia, asthenia, drinking water excessively, pregnancy, a sudden increase in sugar, and frequent urination, and they were classified as patients with diabetes and healthy people with an accuracy rate of 88.43%.

III. Conclusion

In this study, the test results and patient complaints of some 14,000 patients were examined with doctor's support for the diagnosis of diabetes, thereby determining some 12 variables which triggered diabetes. The unnecessary patient data were eliminated in line with the specified criteria, and the data were reduced to 369 people with the doctor's observation. With the new data obtained, a classification was made by means of the C4.5 algorithm and the factors directly and indirectly affecting diabetes were examined with the decision tree displayed. Classifications were made with the K-Nearest Neighbor, Multi-Layer Perceptron, and NaiveBayes Algorithms besides the C4.5 algorithm, and the most accurate classification method for the diagnosis of diabetes was investigated. The F-measures were examined so as to test the accuracy, sensitivity, and precision of the investigated classification methods, and it was intended to reach the most accurate classification technique. It was aimed to prevent cost and loss of time when diagnosing in line with the revealed decision tree.

It was seen that the K-Nearest Neighbor Algorithm was the optimum classification method for the diagnosis of diabetes among all classification methods examined in Table 10.

Table X:
Classification of All Algorithms

Algorithms	Precision	Sensitivity	F-Measure	Percentage of Accuracy
C4.5 Algorithm	0.798	0.827	0.811	82.65%
K-Nearest Neighbor Algorithm	**0.926**	**0.919**	**0.922**	**91.90%**
MLP Algorithm	0.870	0.884	0.874	88.43%
NaiveBayes Algorithm	0.889	0.884	0.886	88.43%

With the methods employed, the blood test results of the patients and patients' complaints were evaluated; the patients were enabled to save time; it was ensured that doing unnecessary blood tests was prevented; and whether an individual going to the hospital with different complaints had diabetes or not was also detected beforehand. Thanks to this, it was concluded that individuals who had a high quality of life and were not drug-dependent could be reintroduced to life by preventing the progress, and even occurrence, of the disease by taking measures early for diabetes.

References

[1] TEMD *Diabetes Mellitus ve Komplikasyonlarının Tanı, Tedavi ve İzlem Kılavuzu*, 2015

[2] 2016, Erişim tarihi: 03.02.2016 http://bilheal.bilkent.edu.tr/aykonu/ay2013/diabet/diabet.htm

[3] Özkan, Y., (2008). *Veri Madenciliği Yöntemleri.* İstanbul: Papatya Yayıncılık.

[4] Wang, J., (2006). Encyclopedia of Data Warehousing and Mining. Information Science Reference, Volume: 49, ss: 140.

[5] Olmuş, H. ve Erbaş, S. O., 2003, "*Bayes Ağlarda Koşullu Bağımsızlıkların İncelenmesi üzerine bir Çalışma*", TÜİK İstatistik Araştırma Dergisi, Cilt 2, Sayı 1, s. 89-103

[6] Koyuncugil, A.S. , Özgülbaş N. (2009) Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları

[7] Tan, P.N., Steinbach, M., Kumar, V., 2006 , *Introduction to Data Minining, Addison-Wesley*

[8] Göktepe, A.B., Agar, E. and Lav, A.H., (2004). *Comparison of Multilayer Perceptron and Adaptive Neuro-Fuzzy System on Backcalculating the Mechanical Properties of Flexible Pavements*. ARI The Bulletin of the Istanbul Technical University, Volume: 54(3), ss: 65-77.

[9] Kökver, Y. , Barişçi, N. , Çiftçi, A. , Ekmekçi, Y., *Hipertansiyona Etki Eden Faktörlerin Veri Madenciliği Yöntemleriyle İncelenmesi*, Journal of New World Sciences Academy, 2014

http://dx.doi.org/10.15242/IJCCIE.U0716007