Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 28, S. 1265-1268, Kasım 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 28, pp. 1265-1268, November 2021

Research Article

Classification of Type 2 Diabetes Using Machine Learning Techniques

Ziynet Pamuk[1], Ceren Kaya[2*]

[1] Zonguldak Bulent Ecevit University, Faculty of Engineering, Department of Biomedical Engineering, Zonguldak, Turkey, (ORCID: 0000-0003-3792-2183), ziynet.pamuk@beun.edu.tr

[2*] Zonguldak Bulent Ecevit University, Faculty of Engineering, Department of Biomedical Engineering, Zonguldak, Turkey, (ORCID: 0000-0002-1970-2833), ceren.kaya@beun.edu.tr, crnkaya@hotmail.com

(1st International Conference on Applied Engineering and Natural Sciences ICAENS 2021, November 1-3, 2021)

(**DOI:** 10.31590/ejosat.1014878)

ATIF/REFERENCE: Pamuk, Z., & Kaya, C. (2021). Classification of Type 2 Diabetes Using Machine Learning Techniques. *European Journal of Science and Technology*, (28), 1265-1268.

Abstract

Diabetes is a lifelong chronic disease defined by disorders in protein, fat and carbohydrate metabolism as a result of complete or partial deficiency of insulin hormone secreted from the pancreas. This disease is caused by the absence or deficiency of insulin hormone in the body. Normal metabolism also breaks down in the intestines to convert nutrients into glucose. Then, when this glucose passes through the intestines into the blood, the level of sugar in the blood rises. In healthy people, glucose in the blood is transported to cells with the help of insulin hormone, which is secreted from the pancreas. Because sugar can not be transported to the cell if there is a deficiency or impaired effect of insulin hormone in the body, glucose increases in the blood and develops an increase in blood sugar (hyperglycemia), called diabetes. Early diagnosis of diseases that will occur in insulin, which is vital for the human body, is of great importance. The aim of this study is to use machine learning techniques to diagnose Type 2 diabetes using medical laboratory data. As machine learning techniques, J48, Random Forest, Random Tree and IBk algorithms in the WEKA programme were used. In this study, 400 patient data were investigated. 6 laboratory tests such as age, gender, glucose, HbA1C, HGB and urine were selected as input data. All four algorithms used were successfully trained. The highest accuracy value was found 96.97% in Random Forest algorithm, with recall and F-measure values of 98.47% and 96.24%, respectively.

Keywords: Type 2 Diabetes, WEKA, Machine Learning, J48, Random Forest, Random Tree, IBk Algorithms.

Tip 2 Diyabetin Makine Öğrenmesi Teknikleriyle Sınıflandırılması

Öz

Diyabet, pankreastan salgılanan insülin hormonunun tam veya kısmi eksikliği sonucu protein, yağ ve karbonhidrat metabolizmasındaki bozukluklarla tanımlanan, ömür boyu süren kronik bir hastalıktır. Bu hastalığa vücutta insülin hormonunun yokluğu veya eksikliği neden olmaktadır. Normal metabolizma ayrıca besinleri glikoza dönüştürmek için bağırsaklarda parçalanır. Daha sonra bu glikoz bağırsaklardan kana geçtiğinde kandaki şeker seviyesi yükselir. Sağlıklı insanlarda kandaki glikoz, pankreastan salgılanan insülin hormonu yardımıyla hücrelere taşınır. Vücutta insülin hormonunun eksikliği veya etkisinin bozulması durumunda şeker hücreye taşınamadığından, kanda glikoz yükselir ve diyabet adı verilen kan şekerinde yükselme (hiperglisemi) gelişir. İnsan vücudu için hayati önem taşıyan insülinde oluşacak hastalıkların erken teşhisi büyük önem taşımaktadır. Bu çalışmanın amacı, tıbbi laboratuvar verilerini kullanarak Tip 2 diyabeti teşhis etmek için makine öğrenmesi tekniklerini kullanmaktır. Makine öğrenmesi teknikleri olarak WEKA programında yer alan J48, Rastgele Orman, Rastgele Ağaç ve IBk algoritmaları kullanılmıştır. Bu çalışmada 400 hasta verisi incelenmiştir. Girdi verisi olarak yaş, cinsiyet, glikoz, HbA1C, HGB ve idrar gibi 6 laboratuvar testi seçilmiştir. Kullanılan dört algoritmanın tamamı başarıyla eğitildi. En yüksek doğruluk değeri %96.97 oranında Rastgele Orman algoritmasında bulunurken, duyarlılık ve F-ölçüsü değerleri sırasıyla %98.47 ve %96.24 olarak bulunmuştur.

Anahtar Kelimeler: Tip 2 Diyabet, WEKA, Makine Öğrenmesi, J48, Rastgele Orman, Rastgele Ağaç, IBk Algoritmaları.

[*] Corresponding Author: ceren.kaya@beun.edu.tr , crnkaya@hotmail.com

1. Introduction

Diabetes is a disease that develops as a result of the deficiency, ineffectiveness or insufficient production of insulin hormone in the body, as well as the chronic complications that disrupt the carbohydrate metabolism and increase the glucose level in the blood. Diabetes, which is seen with symptoms such as intense thirst, intense hunger, and frequent urination, causes many complications in the patient unless treated. If timely measures are not taken and blood sugar is not controlled, it has a negative effect especially on the veins. The toxic effects of sugar can cause permanent damage to many organs and tissues such as eyes, kidneys, nerve endings, heart, brain, and leg vessels (Özlüer Başer et al., 2021).

According to the current data of World Health Organization (WHO), approximately 422 million people in the world, mostly seen in low and middle-income countries, have diabetes, and the cause of 1.6 million deaths each year is directly related to diabetes. For this reason, diabetes is recognized as one of the leading causes of death in the world, and both the number of cases and its prevalence are increasing dramatically (Özlüer Başer et al., 2021).

Diabetes is classified into 4 groups as Type 1, Type 2, Gestational diabetes (GDM) and other specific types. Type 1 diabetes occurs acutely, mostly in children and adolescents, with insulin deficiency due to pancreatic beta cell destruction. Insulin resistance and insulin secretion disorder are prominent in type 2 diabetes. Type 2 diabetes makes up 90-95% of all diabetics. Gestational diabetes defines diabetes that occurs during pregnancy, while other specific types describe high blood sugar that occurs for many reasons that affect the pancreas (Coşansu, 2015).

Such problems in the body can lead to different diseases in other parts of the body. For this reason, early diagnosis in diabetes is vital to avoid many damages. In the study by Güler and Übeyli, multilayer perceptron neural networks trained with four different algorithms were used in the diagnosis of diabetes and they have determined as quick propagation algorithm is the most successful multilayer perceptron training algorithm in the diagnosis of diabetes (Güler & Übeyli, 2006).

Ahmed developed a new estimation method using data mining techniques according to his study. By developing this estimation method, he aimed to divide diabetic patients into two classes as controlled (HbA1C<7%) and uncontrolled (HbA1C>7%). Classification is based on HbA1C measurement. According to results of WEKA Tool experiments, the Logistic algorithm was chosen as the best algorithm with an accuracy rate of 74.8% (Ahmed, 2016).

In another study by Ahmed, a new model was developed to classify diabetic type 2 treatment plans such as insulin, medication, and diet. These treatment plans can help diabetes control blood Glucose level. Since HbA1c is less than 7 in treatment plans, it was considered under control plans. Three categories were chosen for treatment plans, such as insulin, medication, and diet as classification labels. After extensive experiments among data mining algorithms, J48 algorithm was chosen to develop the proposed model based on the accuracy results. The model was implemented using the WEKA application. According to the results obtained, accuracy of the model was found to be 70.8% (Ahmed, 2016).

In the study by Kaya et al., the features of horizontal and vertical Video-Oculography (VOG) signals taken from right and left eyes were used to classify early and late stages of diabetic retinopathy disease. Statistical features obtained from discrete wavelet transform and C4.5 decision tree were applied as inputs to artificial neural networks. It has been concluded that features selected by the C4.5 decision tree algorithm (96.87%) provide better classification performance than features extracted by the discrete wavelet transform (93.75%) (Kaya et al., 2017).

In another study by Bozkurt et al., Pima Indian diabetes dataset was categorized with 8 different classifiers. The data was taken from public website of the University of California Irvine Machine Learning Repository (UCI). Probabilistic neural network (PNN), learning vector quantization (LVQ), feedforward networks (FFN), cascade-forward networks (CFN), distributed time delay networks (DTDN), time delay networks (TDN), the artificial immune system (AIS), and the Gini decision tree algorithms were used as classifiers. When all classifiers in this study were compared, the best accuracy value was 76.00% with DTDN, the best sensitivity value was 63.33%, and the best specificity value was 88.75% with DTDN. The second best accuracy and specificity values after DTDN were obtained with the LVQ network. The second best performance for the sensitivity value was provided by CFN. Since accurate identification of patients is associated with susceptibility, it was concluded that practically, it is more appropriate to use the PNN network that shows the best sensitivity performance (Bozkurt et al., 2014).

The main purpose of this study is to use machine learning techniques that will help experts to diagnose Type 2 diabetes in adults, using only medical laboratory data. The WEKA programme tool was preferred as a machine learning technique in this study.

2. Material and Method

2.1. Type 2 Diabetes Dataset

Information was obtained from the physicians working in Ankara Oncology Hospital about which parameters are required to diagnose type 2 diabetes. In line with these parameters, the necessary ethics committee forms of the hospital were filled and an application was made to obtain related patient data. The patient data used in this study were collected with the approval of Ankara Oncology Hospital Clinical Research Ethics Committee.

Data from 400 patients, 200 of whom were healthy and 200 of whom had type 2 diabetes, were used in this study. Six parameters including age, gender, glucose, glycosylated hemoglobin test (HbA1C), hemoglobin (HGB) and urine were used as input variables of machine learning algorithms. 33% of the total data in the data set is separated as test data. (Training data: 268 and Test data: 132).

2.2. Machine Learning Algorithms

Decision trees are widely used because they have a tree-shaped decision structure, which is learned by induction from a data set of known classes, is easy to interpret, low in cost, and can be easily integrated with database systems (Uzun et al., 2019). J48, Random Forest, Random Tree and IBk decision tree algorithms were used in this study.

J48 algorithm aims to optimize decision tree by utilizing the entropy value of variable and Shannon's Information Theory.

Entropy is a measure of the uncertainty of a random variable. Information gain is a measure of how much the uncertainty in the target variable changes when data is partitioned using a predictor variable (Taşcı & Şamlı, 2020).

In Random Forest algorithm, a forest is determined by creating more than one decision tree. The results of each tree are eliminated by voting or the average is taken to reach the solution of problem. The Random Forest algorithm emerged from the blending of the bagging and Random Subspace methods. As in the bagging method, the data set is divided into subsets (Sarica et al., 2017).

Random Tree algorithm is a type of supervised classifiers. It has an ensemble learning algorithm that produces many individual learners. It uses a bagging idea to generate a random dataset for creating a decision tree (Kalmegh, 2015).

K-Nearest Neighbour (KNN) algorithm is a classification process made by considering the closeness of observations that do not have label information. In the implementation phase of this algorithm, the system is first trained using the training data. The training set contains data with classification information. After the training set is given, a k value is determined by the user. Methods such as Euclidean distance, Jaccard Distance, Simple Matching Distance, Manhattan Distance are used to measure the closeness of data to each other. KNN algorithm is referred as IBk algorithm under the Lazy methodology in Weka tool (Timuçin & Düzdar Argun, 2021).

2.3. Classification Performance Metrics

Seven performance metrics were used for the classification performances of machine learning algorithms. These are:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$F-Measure = 2 \times \left(\frac{Precision \times Recall}{Precision+Recall}\right) \quad (5)$$

$$MCC = \frac{((TP \times TN)-(FP \times FN))}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (6)$$

In Equations (1) – (6); “True Positive (TP)” is the correctly predicted positive class value, “False Negative (FN)” is the incorrectly predicted negative class value, “False Positive (FP)” is the incorrectly predicted positive class value, and “True Negative (TN)” represents the correctly predicted negative class value, respectively in confusion matrix (Kaya et al., 2018). Graphical representation of the relationship between TP and FP is Receiver Operating Characteristics Curve (ROC). As the area under the ROC curve (ROC AUC) value approaches 1, the diagnostic value increases.

3. Results and Discussion

The reference values of six input parameters, which vary according to age and gender, are given in Table 1.

Table 1. Input data and reference ranges

Variables	Age	Gender	Reference Range
Age	-	-	-
Gender	-	-	-
HbA1C	Adult	Female / Male	3.5-5.6 mmol/mol
Glucose	Adult	Female / Male	100-126 mg/dL
HGB	Adult	Female Male	12.5-15.5 gr/dL 13.5-17.5 gr/dL
Urine	Adult	Female / Male	Positive / Negative

The distribution of patient data subjected to testing and training processes in machine learning algorithms used are shown in Table 2.

Table 2. Distribution of training and test datasets

	Train	Test
Healthy	66	66
Type 2 Diabetes Patient	134	134

In this study, Type 2 diabetes data set was classified using J48, Random Forest, Random Tree and IBk algorithms in the WEKA tool. In the testing phase of all algorithms, the cross-validation method was applied as "10-fold".

The performance metrics of accuracy, specificity, ROC AUC, Matthews Correlation Coefficient (MCC), Recall, Precision and F-Measure obtained from four different algorithms used are presented in detail in Table 3.

Table 3. Performance metrics of all algorithms

	J48	Random Forest	Random Tree	IBK
Accuracy	0,9545	0,9697	0,9091	0,8788
Recall	0,9696	0,9847	0,9091	0,7727
Specificity	0,9394	0,9394	0,9091	0,9848
ROC AUC	0,9366	0,9840	0,9091	0,9099
MCC	0,9095	0,9398	0,8182	0,7754
Precision	0,9412	0,9412	0,9091	0,9808
F-Measure	0,9552	0,9624	0,9091	0,8644

According to Table 3 given above, it is seen that the average accuracy values of all algorithms are above 87%. Random Forest algorithm has the highest accuracy result with 96.97%. The closest follower of this algorithm is the J48 algorithm, and it gives the second best result with 95.45%. Random Tree algorithm comes in the third place with 90.91% accuracy values. The IBk algorithm, on the other hand, ranks fourth with an accuracy rate of 87.88%.

While all algorithms are evaluated according to the recall criterion, it is observed that the algorithm giving the best result is the Random Forest algorithm with a value of 98.47%. J48 is in the second place with 96.96%, and Random Tree algorithm is in the third place with a value of 90.91%.

When the results are observed in terms of specificity, IBk is in the first place with the value of 98.48%, J48 and Random Forest

are in the second place with the value of 93.94%, and Random Tree algorithm is in the third place with the value of 90.91%.

Looking at the ROC area values, it is seen that the best algorithm is Random Forest with 98.4% AUC value. J48 is in the second place with 93.66% AUC value and IBk algorithm is in third place with 90.99% AUC value.

4. Conclusion

In order to increase the performance of study, different data analysis techniques and other expert systems can be studied. Accuracy of the study can be increased by using more data and input parameters. Except in cases where there are no severe diabetes symptoms, the diagnosis can be confirmed the next day with the same or a different method. Success evaluations can be made by using different machine learning tools and applying different algorithms. In near future, where machine learning is expected to be used more actively in the field of health, new algorithms can be developed and used for needs.

5. Acknowledge

We would also like to thank Dilara Bilim, Güllü Çıtak and Meryem Ağca for valuable contribution to the study.

References

Ahmed, T. M. (2016). Using data mining to develop model for classifying diabetic patient control level based on historical medical records. *Journal of Theoretical and Applied Information Technology, 87*(2), 316-323.

Ahmed, T. M. (2016). Developing a predicted model for diabetes type 2 treatment plans by using data mining. *Journal of Theoretical and Applied Information Technology, 90*(2), 181-187.

Bozkurt, M. R., Yurtay, N., Yılmaz, Z., & Sertkaya, C. (2014). Comparison of different methods for determining diabetes. *Turkish Journal of Electrical Engineering & Computer Sciences,* 22, 1044-1055.

Coşansu, G. (2015). Diyabet: küresel bir salgın hastalık. *Okmeydanı Tıp Dergisi, 31*(Ek Sayı), 1-6.

Güler, İ., & Übeyli, E. (2006). Çok katmanlı perseptron sinir ağları ile diyabet hastalığının teşhisi. *Journal of the Faculty of Engineering and Architecture of Gazi University, 21*(2), 319-326.

Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. *International Journal of Innovative Science, Engineering & Technology, 2*(2), 438-446.

Kaya, C., Erkaymaz, O., Ayar, O., & Özer, M. (2017). Classification of diabetic retinopathy disease from Video-Oculography (VOG) signals with feature selection based on C4.5 decision tree. *Proceedings of 2017 Medical Technologies National Congress (TIPTEKNO*), 1-4. https://ieeexplore.ieee.org/document/8238093.

Kaya, C., Erkaymaz, O., Ayar, O., & Özer, M. (2018). Impact of hybrid neural network on the early diagnosis of diabetic retinopathy disease from video-oculography signals. *Chaos, Solitons & Fractals, 114*, 164-174.

Özlüer Başer, B., Yangın, M., & Sarıdaş, E. (2021). Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması. *Süleyman Demirel University Journal of Natural and Applied Sciences, 25*(1), 112-120.

Sarica, A., Ceresa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: a systematic review. *Frontiers in Aging Neuroscience, 9*(329), 1-12.

Taşcı, M. E., & Şamlı, R. (2020). Veri madenciliği ile kalp hastalığı teşhisi. *European Journal of Science and Technology, (Special Issue)*, 88-95.

Timuçin, T., & Düzdar Argun, İ. (2021). Initial seed value effectiveness on performances of data mining algorithms. *Düzce University Journal of Science and Technology, 9*, 555-567.

Uzun, R., İşler, Y., & Toksan, M. (2019). WEKA yazılım paketinin siğil tedavi yöntemlerinin başarısının tahmininde kullanımı. *Düzce University Journal of Science and Technology, 7*, 699-708.