Diabetes Prediction Using Machine Learning Techniques
Makine Öğrenmesi Tekniklerini Kullanarak Diyabet Tahmini

Şeyma Kızıltaş Koç [1], Mustafa Yeniad[1]
[1]Department of Computer Engineering, Ankara Yildirim Beyazit University, Ankara, Turkey
ORCIDs: 0000-0001-5224-3598, 0000-0002-9422-4974
E-mails: kiziltasseyma@gmail.com, myeniad@ybu.edu.tr

***Abstract*—Technologies which are used in the healthcare industry are changing rapidly because the technology is evolving to improve people's lifestyles constantly. For instance, different technological devices are used for the diagnosis and treatment of diseases. It has been revealed that diagnosis of disease can be made by computer systems with developing technology.Machine learning algorithms are frequently used tools because of their high performance in the field of health as well as many field. The aim of this study is to investigate different machine learning classification algorithms that can be used in the diagnosis of diabetes and to make comparative analyzes according to the metrics in the literature. In the study, seven classification algorithms were used in the literature. These algorithms are Logistic Regression, K-Nearest Neighbor, Multilayer Perceptron, Random Forest, Decision Trees, Support Vector Machine and Naive Bayes. Firstly, classification performance of algorithms are compared. These comparisons are based on accuracy, sensitivity, precision, and F1-score. The results obtained showed that support vector machine algorithm had the highest accuracy with 78.65%.**

***Keywords*—Machine learning, diabetes prediction, classification algorithms, accuracy**

***Özetçe*—Teknoloji, insanların yaşam biçimini iyileştirmek için sürekli gelişmekte olduğundan, bunun bir sonucu olarak sağlık sektöründe kullanılan teknolojiler de hızla değişmektedir. Örneğin hastalıkların teşhis ve tedavisinde farklı teknolojik cihazlar kullanılmaktadır. Gelişen teknoloji ile birlikte bilgisayar sistemleri ile hastalıkların teşhisinin yapılabileceği ortaya çıkmıştır. Makine öğrenmesi algoritmaları birçok alanda olduğu gibi sağlık alanında da yüksek performans göstermesi nedeniyle sıklıkla başvurulan araçlardır. Bu çalışmanın amacı, diyabet teşhisinde kullanılabilecek farklı makine öğrenmesi sınıflandırma algoritmalarının araştırılması ve literatürdeki metriklere göre karşılaştırmalı analizlerini yapmaktır. Çalışmada, literatürde sıkça kullanılan yedi sınıflandırma algoritması kullanılmıştır. Bu algoritmalar, Lojistik Regresyon, K-En Yakın Komşu, Çok Katmanlı Algılayıcılar, Rastgele Orman, Karar Ağaçları, Destek Vektör Makinesi ve Naive Bayes sınıflandırma algoritmalarıdır. İlk olarak algoritmaların sınıflandırma başarıları karşılaştırılmıştır. Bu karşılaştırmalar, doğruluk, duyarlılık, kesinlik ve F1 skoru oranları üzerinden yapılmıştır. Elde edilen sonuçlar, destek vektör makinesi algoritmasının %78.65 ile en yüksek doğruluk oranına sahip olduğunu göstermiştir.**

***Anahtar Kelimeler*—Makine öğrenmesi, diyabet tahmini, sınıflandırma algoritmaları, doğruluk**

I. Introduction

Diabetes mellitus is generally well-known as diabetes. Diabetes mellitus occurs when the pancreas cannot produce enough insulin or the one it produces cannot be used effectively. Insulin provides that sugar is stored as glycogen in the cell. Diabetics cannot use the glucose that passes from food to the blood, so blood sugar level rises. This event is called hyperglycemia. This situation causes damage to many tissues and organs in the long term such as eyes, kidneys, nerves, heart and blood vessels. Diabetes can be classified into three different types such as diabetes 1, diabetes 2, and gestation diabetes. Diabetes is a major health issue that has reached alarming levels in the world. Recent researches by World Health Organization (WHO) showed a great increase in number of diabetic patients and the deaths that are attributed to diabetes each year. In 2014, 8.5% of adults that aged 18 years and older had diabetes. In addition, according to the International Diabetic Federation (IDF) Diabetes Atlas 9th Edition 2019, 578 million people will have diabetes in 2030. Diabetes is a chronical disease. Therefore, it is significant that diabetes might be uncover at an early stage [1].

Machine learning is a branch of artificial intelligence (AI) and computer science which provides systems with the ability to learn and improve from its own experience. Machine learning methods are used in the health sector because the number of data is very large and the analysis takes time. Researches have shown that machine learning techniques can be used for diabetes prediction such as [2].

In addition, there are many studies for diabetes in the field of machine learning. Because diabetes is a chronical disease and should be diagnosed at an early stage. Various machine learning-based methods have been proposed on diabetes disease recognition. Islam and Jahan applied various machine learning methods which can be used in diabetes prediction [3]. Naive Bayes (NB), Logistic Regression (LR), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Decision

Tree (DT), Random Forest (RF), AdaBoost, One Rule, K-Nearest Neighbor (K-NN) methods were used by the authors. Pima Indian Diabetes Dataset (PIDD) was used. The highest accuracy was obtained by the LR which was 78.01%. Deepti Sisodia and Dilip Sisodia applied NB, DT and SVM algorithms to diagnose diabetes disease [4]. PIDD was used and the performances of the classifications were tested by the WEKA tool. The highest accuracy was obtained by the NB which was 76.30%. Saru and Subashree analyzed machine learning techniques using PIDD to predict diabetes [5]. LR with SVM, DT, K-NN (k=1) and K-NN (k=3) are the classifiers. The highest accuracy was obtained by the DT which was 94.4%. Kumari and Chitra applied SVM to diagnosis of diabetes [6]. The accuracy was recorded as 78% for PIDD.

In this study, machine learning classification algorithms which are NB, LR, MLP, SVM, K-NN, DT and RF were implemented. The dataset is Pima Indians Diabetes Data.

II. Materials and Methods

A. Dataset

The dataset used in the study is Pima Indian Diabetes. The dataset can be found on the Kaggle website [7]. The original owner of this dataset is the National Institute of Diabetes and Digestive and Kidney Diseases. PIDD has been gathered among the Pima Indian female population aged at least 21 years near Phoenix, Arizona. This dataset contains 768 samples with 9 attributes. This dataset has 8 specific variables. The description of the data is given below in Table I [3].

Attribute	Description	Type
Pregnancies	Number of times pregnant	Numeric
Plasma-Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
BloodPressure	Diastolic blood pressure (mm Hg)	Numeric
SkinThickness	Triceps skin fold thickness (mm)	Numeric
Insulin	2-Hour serum insulin (mu U/ml)	Numeric
BMI	Body mass index ((weight in kg)/(height in m))	Numeric
DiabetesPedigreeFunction	Diabetes pedigree function	Numeric
Age	Age of the patient (years)	Numeric
Outcome	Class variable (0 or 1)	Nominal

Table I: Attribute's name and their types

There are 500 instances of class 0 and 268 class 1. Waikato Environment for Knowledge Analysis (WEKA) tool has been used to categorize the data in this paper. WEKA is developed at University of Waikato. WEKA version 3.8 was used in this study.

B. Data Preprocessing

Data may not always be complete and there may be abnormal values, impossible data combinations, missing values, duplicate data in the data set. Data preprocessing is required task to clean data and increase the accuracy and effectiveness of a machine learning model.

When the Pima Indian Diabetes dataset was analyzed, it was found that many attributes had impossible values with 0's. The numbers of missing values for each attribute are as follows:

- Pregnancies: 110
- Glucose: 5
- BloodPressure: 35
- SkinThickness: 227
- Insulin: 374
- BMI: 11

To eliminate this missing values in this study, zero values for the pregnant attribute were left assuming they were real values and 234 samples that have at least two impossible value for attribute of the glucose, blood pressure, skin thickness, insulin and bmi were removed. After the pre-processing 534 instances are remain out of 768.

C. Performance Evaluation

While classifying the data in this study, 10-fold cross validation was applied as a test option. Cross-validation is preferred for overfitting problem and small datasets. In 10-fold cross validation, the data file is divided into ten and nine parts are used for training and one part for testing, this process is repeated ten times. In this study, WEKA tool was used and results of algorithms were compared. To compare the results, the number of values which are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were used in the confusion matrix and applying the following equations with these numbers, accuracy, sensitivity, precision and F1-score ratios were calculated.

- Accuracy: Shows ratio of correctly classified samples to the total number of tested samples (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Sensitivity (Recall): Shows ratio of positive classification of instances i.e. TP to the sum of TP and FN (2).

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

- Precision: Shows ratio of positive sample that were correctly classified to the total number of positive predicted samples (3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- F1-score: It is a way of combining the precision and recall of the model (4).

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

III. RESULT

The results of the different classification methods tested with the WEKA tool are shown in Table II.

SVM had the highest accuracy (78.65%) whereas K-NN obtained lowest accuracy (71.16%). Total accuracy is above 71% in all cases. The second highest accuracy (77.71%) obtained from LR. Sensitivity and precision are quite gladsome. Also, RF acquired third highest accuracy (76.77%).

Algorithm	Accuracy (%)	Sensitivity (%)	Precision (%)	F1-score (%)
Naïve Bayes	75.65	82.35	81.44	81.89
Support Vector Machine	78.65	89.91	80.45	84.91
Decision Tree	74.71	78.99	82.45	80.68
Logistic Regression	77.71	89.07	79.89	84.23
Random Forest	76.77	85.99	80.57	83.19
K-Nearest Neighbor	71.16	80.39	77.35	78.84
Multilayer Perceptron	74.71	82.35	80.32	81.32

Table II: Result of algorithms

IV. CONCLUSION

Millions of people around the world suffer from diabetes. However most of these people don't even know if they have the disease. Early diagnosis of diabetes can abate long-term complications and cost. Therefore, multiple machine learning algorithms applied and analyzed for PIDD. The results show that the best performance was produced by an SVM algorithm. Generally, all techniques produced an accuracy score of around 70 %. Further analysis of attributes and different combination of feature selection is necessary to achieve higher accuracy. Also, much more datasets can be generated, real datasets can be taken or deep neutral networks can be applied to consider the real impact of the performance of the algorithms.

AUTHOR CONTRIBUTIONS

This paper is a part of *Ş.K.*'s MSc thesis. *M.Y.* is the advisor of the thesis. All authors equally contributed on writing the paper.

REFERENCES

[1] Selek MB, Ciftciogullari UA, Yuce YK, Isler Y. Developing an educational mobile game to provide diabetes-awareness among children. Journal of Intelligent Systems with Applications 2021; 4(1): 20-23.

[2] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics 2018; 9: 515.

[3] Islam MA, Jahan N. Prediction of onset diabetes using machine learning techniques. International Journal of Computer Applications 2017; 180(5): 7-11.

[4] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Computer Science 2018; 132: 1578-1585.

[5] Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. International Journal of Emerging Technology and Innovative Engineering 2019; 5(4): 3368308.

[6] Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications 2013; 3(2): 1797-1801.

[7] Kaggle Datasets. Pima indians diabetes database. 2016. Retrieved from https://www.kaggle.com/uciml/pima-indians-diabetes-database at October 10, 2020.