DOI: 10.37990/medr.1021148 Med Records 2022;4(2):191-5

MEDICAL RECORDS-International Medical Journal

Araştırma Makalesi / Research Article

Early Diagnosis of Diabetes Mellitus by Machine Learning Methods According to Plasma Glucose Concentration, Serum Insulin Resistance and Diastolic Blood Pressure Indicators

Plazma Glukoz Konsantrasyonu, Serum Insülin Direnci ve Diastolik Kan Basıncı Göstergeleri ile Makine Öğrenme Yöntemleri Kullanılarak Diyabet Hastalığının Erken Tanısı

Mehmet Kivrak

Recep Tayyip Erdogan University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Rize, Turkey

Copyright@Author(s) - Available online at www.dergipark.org.tr/tr/pub/medr
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract

Aim: It is a known fact that diabetes mellitus is increasing frequently and triggering many different diseases. Therefore, early diagnosis of the disease is important. This study was trying to predict the early diagnosis of the disease, according to machine learning methods by measuring plasma glucose concentration, serum insulin resistance, and diastolic blood pressure.

Material and Methods: In the study, the public dataset from a website consists of 768 samples and nine variables. Three different machine learning strategies were used in the early diagnosis of diabetes mellitus (Support Vector Machine, Multilayer Perceptron, and Stochastic Gradient Boosting). 3 repeats and 10 fold cross-validation method was used to optimize the hyperparameters. The model's performance parameters were evaluated based on accuracy, specificity, sensitivity, confusion matrix, positive predictive value (precision), negative predictive value, and AUC (area under the ROC curve).

Results: According to the experimental results (the criteria of accuracy (0.79), sensitivity (0.57), specificity (0.91), positive predictive value (0.79), negative predictive value (0.80), and AUC (0.74)) the Support Vector Machine was more successful than other methods.

Conclusion: Plasma glucose concentration, serum insulin resistance, and diastolic blood pressure markers are important indicators in the early diagnosis of diabetes mellitus. In this study, it was seen that these markers make a significant contribution to the early diagnosis of diabetes mellitus. However, it has been observed that these indicators alone will not be sufficient in the early diagnosis of the disease, especially since age, body mass index and pregnancy contribute significantly.

Keywords: Diabetes mellitus, plasma glucose concentration, serum insulin resistance, diastolic blood pressure, machine learning

Öz

Amaç: Diyabetin sıklıkla arttığı ve bir çok farklı hastalığı tetiklediği bilinen bir gerçektir. Bu nedenle hastalığın erken teşhisi önemlidir. Bu çalışmada plazma glukoz konsantrasyonu, serum insülin direnci ve diyastolik kan basıncı göstergelerinden, makine öğrenmesi yöntemlerine göre hastalığın erken teşhisi öngörülmeye çalışılmıştır.

Materyal Metot: Çalışmada, bir web sitesinden alınan halka açık veri seti 768 örnek ve dokuz değişkenden oluşmaktadır. Diyabetin erken teşhisinde üç farklı makine öğrenme stratejisi kullanıldı (Destek Vektör Makineleri, Çok Katmanlı Algılayıcılar ve Stokastik Gradyan Artırma). Hiper parametre optimizasyonu için 3 tekrarlı 10 kat tekrarlı çapraz doğrulama yöntemi kullanıldı. Modellerin performansı doğruluk, seçicilik, duyarlılık, karışıklık matrisi, pozitif tahmin değeri (kesinlik), negatif tahmin değeri ve AUC (ROC eğrisi altında kalan alan) temel alınarak değerlendirilmiştir.

Bulgular: Deneysel sonuçlara göre (doğruluk (0.79), duyarlılık (0.57), özgüllük (0.91), pozitif tahmin değeri (0.79), negatif tahmin değeri (0.80) ve AUC (0.74) kriterleri), Destek Vektör Makineleri diğer yöntemlere göre daha başarılı çıkmıştır.

Sonuç: Diyabet hastalığının erken tanısında plazma glukoz konsantrasyonu, serum insülin direnci ve diastolik kan basinci belirteçleri önemli göstergelerdir. Bu çalışmada da bu belirteçlerin diyabetin erken tanısında önemli katkı sağladığı görülmüştür. Ancak tek başlarına bu göstergelerin hastalığın erken tanısında yeterli olmayacağı özellikle yaş, beden kitle indeksi ve gebeliğin de önemli derecede katkı sağladığı görülmüştür.

Anahtar Kelimeler: Diyabet hastalığı, plazma glukoz konsantrasyonu, serum insülin direnci, diyastolik kan basıncı, makine öğrenme

Geliş Tarihi / Received: 09.11.2021 **Kabul Tarihi / Accepted:** 12.03.2022

Sorumlu Yazar /Corresponding Author: Mehmet Kivrak, Recep Tayyip Erdogan University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Rize, Turkey, **E-mail**: mehmet.kivrak@erdogan.edu.tr

INTRODUCTION

Diabetes mellitus is still a very common disease in the world, negatively affecting the daily lives of patients, and continues to be a serious economic burden, especially in countries where obesity is common. It is estimated that diabetes affects 246 million people in the world and about 20-30 million of these patients are affected by symptomatic diabetic polyneuropathy. Considering the increase in obesity rates and the associated increase in type 2 diabetes prevalence, this number will double by 2030 is expected. In young patients with type 1 diabetes, polyneuropathy may occur within a few months of the onset of the disease as a result of poor control of diabetes. Studies show that intense diabetes control reduces the prevalence of clinical neuropathy by 60-69%. Therefore, early diagnosis is very important (1-3).

Machine learning is a system that studies the creation and operation of algorithms that can learn and predict data. Such algorithms work by constructing a model to make and predict decisions based on sample input (1).

The findings of Support Vector Machine (SVM), Multilayer Perceptron (MLP) and, Stochastic Gradient Boosting (SGB) approaches from data mining algorithms for the early diagnosis of diabetes mellitus are presented in this paper.

MATERIAL AND METHOD

Dataset

The dataset used in the study was carried out on the Pima Indians Diabetes Database (PIDD) dataset (4) in the Kaggle database. The data set contains 768 samples and nine variables. These variables are age, pregnancies (PR), plasma glucose (PG) concentration, diastolic blood pressure (BP), tri-fold thick, resting electrocardiography results, serum insulin, body mass index (BMI), Diabetes pedigree (DP) function, and diabetes. A detailed description of the variables is given in Table 1. Ethics committee approval is not required for this study. In this study, the R programming language, and SPSS used.

Table 1. The Detailed Explanation of the Variable

Variables	Abbreviation	VariableType	Role
Age (year)	-	Numerical	Input
Pregnancies	PR	Numerical	Input
Diastolic BP(Blood Pressure (mm/Hg)	BP	Numerical	Input
PG (Plazma Glukoz) Concentration	PG	Numerical	Input
Skin-Fold Thick (mm)	SFT	Numerical	Input
Serum Insilun (mu U / m)	SI	Numerical	Input
Body Mass Index	BMI	Numerical	Input
DP (Diabetes Pedigree) Function	DP	Numerical	Input
Diabet	-	Categorical	Output

Preprocessing of the Data Set

The data set was included in the analysis without splitting. SVM, MLP, and SGB algorithms were used for the classification task. 65 rows with extreme/outlier values were detected in our data set and deleted (figure 1). The optimal hyper-parameters of each model were determined by grid search with 3 repeats and 10-fold repeated k-fold cross-validation. The created models were evaluated with accuracy, specificity, sensitivity, confusion matrix, negative predictive value, positive predictive value, and AUC.

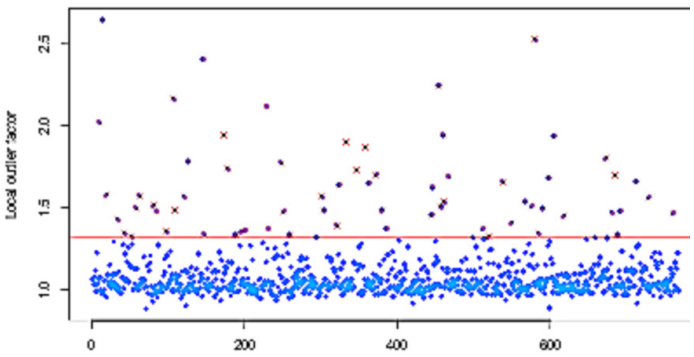


Figure 1. Extreme/Outlier Values Analysis

Support Vector Machine (SVM)

Support Vector Machine is a set of supervised learning algorithms that detect patterns. Makes model estimation by running the support vector machine at each stage of the smallest optimization problem involving two lagrangian multipliers (5). SVM generates linear and nonlinear estimates of the target variable in classification and estimation problems with different kernel functions to determine the best planes. Determining the optimal kernel function is an important criterion for the accuracy performance of the model (6,7). Kernel functions such as radial, linear, Laplace are used in the SVM algorithm. Thanks to its optimized technique, SVM offers optimal solutions in large and complex data sets (8,9). The hyperparameters of the SVM classifier are C, sigma, and interaction depth. The hyperparameters of models are presented in Table 2.

Stochastic Gradient Boosting (SGB)

Boosting is an ensemble-based data mining meta-algorithm that improves the performance of prediction and classification of any learning approach (10). Stochastic gradient boosting (SGB) is a data processing approach introduced by (11). SGB is a crucial technique accustomed to creating forecasts and classification tasks and adjusting forecast performance through the appliance of preprocessing procedures. SGB was implemented in R by the Generalized Boosted Regression Models (GMB) Package (12). The hyperparameters of the SGB classifier are n.trees, shrinkage, and n.minobsinnode. The hyperparameters of the model are illustrated in table 2.

Multilayer Perceptron (MLP)

The most commonly used artificial neural network model

is the MLP network, which has also been comprehensively analyzed and lots of learning algorithms are developed from it (13). MLP is a feed-forward, fully neural network model that maps input data set to a convenient output set by adjusting the weight between internal data nodes. The hyperparameters of the MLP classifier are hiding the layer size, activation, alpha, and learning rate. The hyperparameters of the model are illustrated in Table 2.

Table 2. The Hyper Parameters of Models

Model	Hyper Parameters	Range	Number of Combination
	C	(2-5–215)	
SVM	Sigma	(2-15–23)	300
	Interaction Depth	(1–100)	
	n.treesa	(50–1500)	
SGB	Shrinkage	0.1	3000
	n.minobsinnodeb	20	
	Hidden layer size	(50–100)	
MLP	Alpha	(0.0001, 0.05)	100
	Learning rate	(constant-adaptive)	

[a] A total number of trees.
[b] A minimum number of observations in the trees terminal nodes.

RESULTS

Statistical Analysis

Quantitative data were summarized as the arithmetic means with standard deviation, qualitative data as numbers by percentage, and median with min and max values. After suitability of the data to multiple normal distributions, the difference between the groups in normally distributed groups was examined by t-test in independent samples and the Mann-Whitney U test for variables that didn't normally distribute. For statistical analysis, IBM SPSS version 22 (14) and R Studio version 1.1.463 (15) were used. In the diabetes data set, the health status of individuals is shown as '0' or '1'.'0' indicates that the individual does not have diabetes, and '1' indicates that the individual has diabetes. In the data set, 500 individuals are diabetic and 268 individuals are not diabetic. The distributions of the features in the data set are presented in Figure 2 and the correlation matrix in Figure 3. In Figure 3, when the relationship of the features with the class label (cl) is examined, it is observed that diabetes mellitus is associated with the highest (0.47) PG concentration, followed by BMI, age, and PR.

The performance metrics of each model are shown table 3. The accuracy values were 0.79 for SVM, 0.78 for SGB and 0.65 for MLP. The sensitivity values were 0.57 for SVM, 0.55 for SGB and 0.00 for MLP. The specificity values were 0.91 for SVM, 0.91 for SGB and 1.00 for MLP. The positive predictive values were 0.79 for SVM, 0.76 for SGB and non-computed for MLP. The negative predictive values were 0.80 for SVM, 0.80 for SGB and 0.65 for MLP. The AUC values were 0.74 for SVM, 0.73 for SGB and 0.50 for MLP.

Table 4 and figure 4 presents the relative importance values of the best classifier model (SVM) which was chosen by the majority of measurements metrics. The furthest interrelated variables with diabetes mellitus were sorted from highest to smaller by the significance values. Figure 5 presents the model's comparison of ROC curves. Figure 6 illustrates the confusion matrix for the best model (SVM).

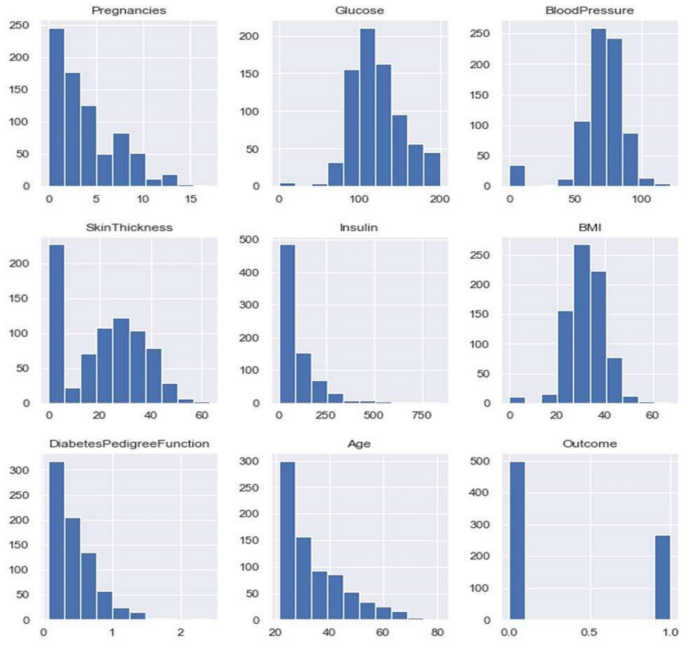


Figure 2. Distributions of Variables

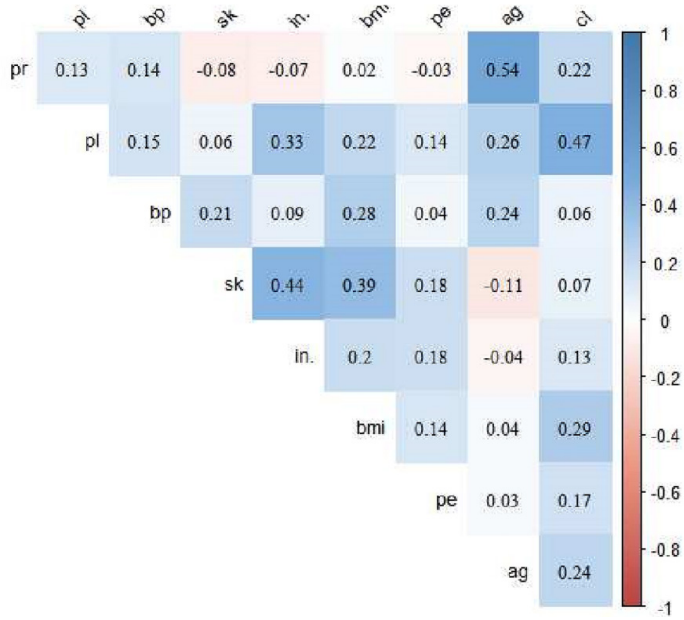


Figure 3. Correlation Matrix

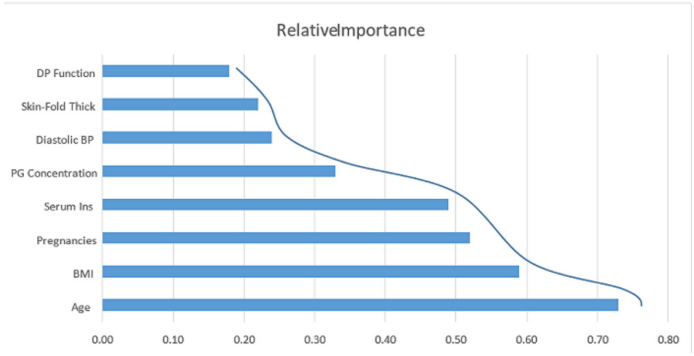


Figure 4. The Variables Importance Values of the Best Classifier

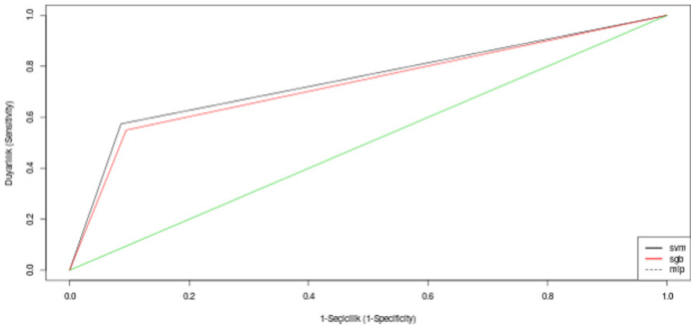


Figure 5. Comparison of ROC Curves

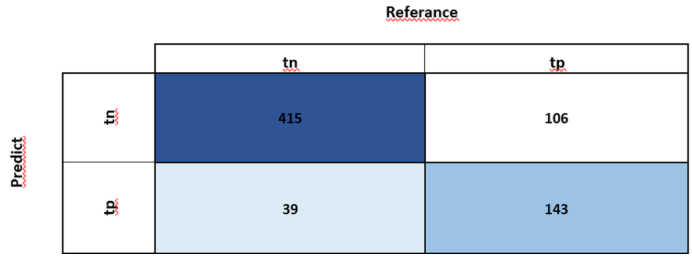


Figure 6. Confusion Matrix for Best Model

Table 3. Performance Metrics of Models

Performance Metrics	Models		
	SVM	SGB	MLP
Accuracy	0.79	0.78	0.65
Sensitivity	0.57	0.55	NaN
Specificity	0.91	0.91	1.00
Positive predictive value (Precision)	0.79	0.76	NaN
Negative predictive value	0.80	0.80	0.65
AUC	0.74	0.73	0.50

Table 4. The Variables Importance values of the Best Classifier

Variable	Relative Importance
Age	0.73
BMI	0.59
Pregnancies (PR)	0.52
Serum Insilun (SI)	0.49
PG (Plazma Glukoz) Concentration	0.33
Diastolic BP (Blood Pressure)	0.24
Skin-Fold Thick (SFT)	0.22
DP (Diabetes Pedigree) Function	0.18

DISCUSSION

Early detection of diabetes mellitus is an important medical problem. Machine learning methods have made a place for themselves in early diagnosis and planning in the field of health. Especially in chronic diseases with high costs, machine learning methods become very useful. In this study, the performances of different machine learning classification methods in predicting diabetes were compared.

According to the experimental results, the SVM was more successful than other methods according to the criteria of accuracy (0.79), specificity (0.91), sensitivity (0.57), positive predictive value (0.79), negative predictive value (0.80), and AUC (0.74).

Plasma glucose concentration, serum insulin resistance, and diastolic blood pressure markers are important indicators in the early diagnosis of diabetes mellitus. In this study, it was seen that these markers make a significant contribution to the early diagnosis of diabetes mellitus. However, it has been observed that these indicators alone will not be sufficient in the early diagnosis of the disease, especially age, BMI and pregnancy contribute significantly.

In similar studies in the literature, Bahat et al., with five different machine learning algorithms (decision tree, support vector machine, random forest, logistic regression, and k nearest neighbor) only examined the classification performances in the early diagnosis of diabetes in terms of accuracy metric. In terms of accuracy values, the decision tree was 0.79, the support vector machine was 0.72, the random forest was 0.75, logistic regression was 0.76, and the k nearest neighbor was 0.80 (16).

In his work on the impact of machine learning and feature selection on type 2 diabetes risk prediction, Riihimaa looked at the area under the ROC curve as a model performance measure. In the study, AUC values according to logistic regression and machine learning methods were found to be 0.64 and 0.85, respectively (17).

Islam et al., in their research on prediction of onset diabetes using machine learning techniques, using different machine learning methods (naiveBayess, logistic regression, multilayer perceptron, support vector machines, lazy and meta classifiers, rules, and trees). They compared classification performances in terms of accuracy, sensitivity, selectivity, positive predictive value, negative predictive value, and AUC metrics. The logistic regression model provided the highest performance with an accuracy value of 0.78 (18-19).

CONCLUSION

The use of machine learning methods in the early diagnosis of diabetes is increasing day by day. With the development of independent classifiers or ensemble learning algorithms, the number of current algorithms used in medicine is increasing. Studies to be carried out based on more than one performance measure, without being dependent on a single performance criterion will enable more meaningful comparisons to be made. In this study, an evaluation was made according to more than one performance criteria in machine learning methods and the best model was determined. In general, the model gave more successful results in separating the healthy than in separating the patients. The model is at a reasonably acceptable level according to the general criteria, but it is thought that there will be improvements in the model performance criteria to be used by increasing the number of variables. In addition, classification success can be increased by using methods that can provide more success with ensemble learning and hybrid methods.

Financial disclosures: *The authors declared that this study hasn't received no financial support.*

Conflict of Interest: *The authors declare that they have no competing interest.*

Ethical approval: Ethics committee approval is not required for this study. In this study, the R programming language, and SPSS used.

REFERENCES

1. Said G. Diabetic neuropathy-A Review. Nat Clin Prac Neurol. 2007;3:331-40.
2. Albers JW. Diabetic Neuropathy: Mechanisms, emerging treatment,s and subtypes. Curr Neurol Neurosci Rep. 2014;14:473.
3. Charnogursky G. Neurological complications of diabetes. Curr Neurol Neurosci Rep. 2014;14:457.
4. Prima Indians Diabetes Database (PIDD), ttps://www.kaggle.com/saurabh00007/diabetescsv access date 11.05.2021
5. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers. 1999;10:61-74.6.
6. Birjandi SM, Khasteh SH. A survey on data mining techniques used in medicine. J Diabetes Metab Disord. 2021;20:2055-71.
7. Nitze I, Schulthess U, Asche H. Comparison of machine learning algorithms random forest, artificial neural network and support vector machine the o maximum likelihood for supervised crop type classification. Proc of the 4th GEOBIA 2012, p. 35.
8. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273-97.
9. Ayhan S, Erdogmus S. Kernel function selection for solving classification problems with support vector machines. Eskisehir Osmangazi University. Journal of economics and administrative sciences. 2014, 9: 175-201.
10. Arslan A, Sen B. Detection of non-coding RNA's with optimized support vector machines. 23nd Signal Processing and Communications Applications Conference (SIU) IEEE. 2015:1668-71.
11. Schapire RE. The boosting approach to machine learning: an overview, nonlinear estimation and classification. Springer. 2003, p.149-71.
12. Friedman JH. Stochastic gradient boosting Comput. Stat Data Anal. 2002;38:367-78.
13. Ridgeway G. Generalized Boosted Regression Models: A guide to the gbm package. Update, 1(1), 2007.
14. Rosenblatt, F. Two theorems of statistical separability in the perceptron. United States Department of Commerce. 1958.
15. Yasar S, Arslan A, Colak C. et al. A Developed Interactive Web Application for Statistical Analysis: Statistical Analysis Software. Middle Black Sea J Health Science. 2020;2:227-39.
16. Campbell, M. RStudio Projects. In Learn RStudio IDE. 2019, p. 39-48.
17. Sarwar MA, Kamal N, Hamid W et al. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. ICAC - IEEE. 2018: p. 1-6.
18. Riihimaa, P. Impact of machine learning and feature selection on type 2 diabetes risk prediction. J Med Artif Intell. 2020;3: 20-4.
19. Islam MA, Jahan N . Prediction of onset diabetes using diabetes using machine learning techniques. Int J Computer Applications. 2017;180:7-11.