Ann Med Res

Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

A clinical decision support system based on machine learning for the prediction of diabetes mellitus

Bahri Evren[a], Zeynep Kucukakcali[b,*]

[a] *Inonu University, Faculty of Medicine, Department of Endocrinology and Metabolism, Malatya, Türkiye*
[b] *Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye*

ARTICLE INFO

Keywords:
Diabetes mellitus
Classification
Machine learning
Stochastic gradient boosting

Received: Oct 05, 2022
Accepted: Oct 13, 2022
Available Online: 22.10.2022

DOI:
10.5455/annalsmedres.2022.10.301

Abstract

Aim: Early diagnosis of diabetes mellitus (DM), one of the most important health problems worldwide, and taking necessary steps are very important. Therefore, it has become very important to develop models for the prediction of the disease. The aim of this study is to create a clinical decision support model with Stochastic Gradient Boosting, a machine learning model for DM prediction.

Materials and Methods: In the study, modeling was done with the Stochastic Gradient Boosting method using an open access data set including the factors associated with DM. Model results were evaluated with accuracy, balanced accuracy, sensitivity, selectivity, positive predictive value, negative predictive value, and F1-score performance metrics. In addition, 5-fold cross-validation method was used in the modeling phase. Finally, variable importance values were obtained by modeling.

Results: Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score from by Stochastic Gradient Boosting modeling were 93.6%, 92.8%, 91.7%, 93.9%, 73.3%, 98.4%, and 81.5%, respectively. According to the variable importance values obtained for the input variables in the data set examined in this study, the most important variables are glucose, age, systolic BP, cholesterol, chol/HDL, BMI, height, waist/hip, HDL, waist, weight, diastolic BP, hip, and gender: male.

Conclusion: In the current study, it was seen that the ML model applied with the results obtained can predict diabetes. Addition, according to the results of the relevant model, the most important risk factors for DM were determined and given in degrees of importance of the risk factors. With these results, necessary precautions can be taken for the disease at early levels.

Introduction

Diabetes mellitus (DM) is a chronic condition that has a significant impact on daily life as well as the quality of life. This disease cannot be totally cured, but its harmful effects in the short and long term can be avoided if it is adequately controlled and measures are taken [1, 2]. Diabetes mellitus (DM) is a chronic and metabolic disease defined by anomalies in protein, carbohydrate, and lipid metabolism caused by absolute or relative insulin insufficiency, as well as clinical and biochemical findings [3]. DM is actually of two main types, Type 1 and Type 2. Type 1 diabetes is an autoimmune illness that results from the loss of pancreatic beta cells. Type 2 diabetes is defined as a combination of insulin resistance and pancreatic beta cell dysfunction in insulin production [4]. Patients with type 1 diabetes are often younger, with the majority being under the age of 30. Increased thirst and frequent urination are typical clinical symptoms, as are elevated blood glucose levels. This kind of diabetes cannot be properly treated with oral drugs alone, and patients must receive insulin therapy [5]. Type 2 diabetes is more common in persons in their forties and fifties, and it is frequently related to obesity, hypertension, dyslipidemia, arteriosclerosis, and other disorders [6].

DM is one of the fastest-growing diseases worldwide and is expected to affect 693 million adults by 2045. Vascular complications of both the macrovascular (cardiovascular disease (CVD)) and microvascular (diabetic kidney disease (DKD), diabetic retinopathy, and neuropathy) systems are the leading cause of morbidity and mortality in diabetics, imposing a massive financial burden due to disparities in healthcare expenditure and treatment access between developed and developing countries [7]. DM, which is a

[*] Corresponding author:
Email address: zeynep.tunc@inonu.edu.tr (Zeynep Kucukakcali)

concerning scenario given the impacts indicated and the predicted number of patients, needs to be managed. As a result, how to rapidly and accurately identify and assess diabetes is a topic worth researching.

Machine learning (ML) is a subfield of artificial intelligence (AI) that aims to make predictions about new data when exposed to new data by performing data-driven learning. AI/ML methods are one of the most commonly utilized technologies in illness detection and clinical decision support systems in recent years, with a wide range of applications. ML, which has a wide application area in health, constitutes the basic infrastructure of applications in determining genetic diseases, early diagnosis of cancer diseases, and the identification of patterns in medical imaging. In the last decade, with the availability of large datasets and greater computing power, ML methods have achieved high performance in various situations [8, 9]. ML techniques have been used to achieve clinical goals by utilizing their unique characteristics [10]. In everyday medical practice, AI in medicine has been linked to the creation of programs to help doctors with their daily responsibilities, such as making diagnoses, making therapeutic decisions, and anticipating emergency situations or the deterioration of patients [11, 12].

The aim of this study is to identify the factors that may be related to DM that can be transformed into useful information for the advancement of clinical practice and health care through statistical methods and machine learning capable of self-improvement.

Materials and Methods

Dataset and variables

The present research was performed as a retrospective case-control study. The data set used in the study includes demographic and laboratory variables of 390 patients. In the data set, those with a hemoglobin A1c value above 6.5 were considered diabetic, and those below it were considered non-diabetes. The data set included in the study was obtained from "https://data.world/informatics-edu/diabetes-prediction".

Explanatory information about the variables in the data set is given in Table 1.

Stochastic gradient boosting algorithm

Stochastic gradient boosting is a method developed by Friedman by incorporating randomness into the gradient boosting method. In stochastic gradient boosting, a random subsample is selected with permutation sampling strategy at each refresh. This selected subsample is used to calculate the model update instead of all learners and to reduce the correlation between trees [13]. As with other ensemble learning methods, large trees are not created in this method, but instead, each tree (usually 100–200 trees) developed during the process is summarized and each observation is grouped according to the most common classification among trees. These differences cause the stochastic gradient boosting method to differ from other augmentation methods and reduce its sensitivity to outliers and unbalanced data sets [13, 14].

Biostatistical analysis

The median (minimum-maximum) and count (percentage) were used to summarize variables. Shapiro-Wilk test of normality was used to determine normal distribution. Whether there is a statistically significant difference between the output variable and input variables was evaluated by using where appropriate Mann-Whitney U test, and Pearson Chi-square test. It was determined that a value was statistically significant if it had a p-value of less than 0.05 ($p<0.05$). IBM SPSS Statistics 26.0 was utilized in all analyses.

Machine learning modeling and performance evaluation

In the current study, Stochastic Gradient Boosting was used in the modeling stage for the dataset in question. The data set was divided as 80:20 as a training and test dataset. The n-fold cross-validation approach was used for the analyses. The data is separated into n parts in the n-fold cross-validation procedure, and the model is applied to n parts. One of the n parts is used for testing, while the remaining n-1 parts are used to train the model. In this study, 5-fold cross-validation was employed for the modeling process. Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score were used as performance evaluation criteria. In addition, variable importances were calculated, which gives information about how much the input variables contribute to the output variable. Modeling was done using R studio 4.2.1.

Results

The dataset used in the current study included demographic and clinical information of 60 diabetic and 330 nondiabetic patients. The mean age of the patients was 46.77 ±16.44 years. Of the patients in the study, 228 (58.5%) were female and 162 (41.5%) were male. The results of the statistical tests performed with the explanatory

Table 1. Explanatory information about the variables.

Column attribute	Description
Cholesterol	Total cholesterol
Glucose	Fasting blood sugar
HDL	HDL or good cholesterol
Chol/HDL	Ratio of total cholesterol to good cholesterol Desirable result is < 5
Age	All adult African Americans
Gender	162 males, 228 females
Height	In inches
Weight	In pounds (lbs)
BMI	703 x weight (lbs)/ [height(inches]2
Systolic	BP The upper number of blood pressure
Diastolic	BP The lower number of blood pressure
Waist	Measured in inches
Hip	Measured in inches
Waist/hip	Ratio is possibly a stronger risk factor for heart disease than BMI
Diabetes	Yes (60), No (330)

Table 2. Descriptive statistics for categoric input variable.

Variable		Diabetes		p*
		Diabetes	No diabetes	
Gender	Female	34 (56.67)	194 (58.79)	0.759
	Male	26 (43.33)	136 (41.21)	

*:Pearson Chi-Square Test.

Table 3. Descriptive statistics for numeric input variables.

Variables	Diabetes	No diabetes	p*
	Median (Min-Max)	Median (Min-Max)	
Cholesterol	219(115-443)	199(78-347)	<0.001
Glucose	186(60-385)	86.5(48-371)	<0.001
Hdl Chol	42(23-114)	47(12-120)	0.005
Chol/HDL Ratio	5.2(2-19.3)	4.1(1.5-10.6)	<0.001
Age	59.5(26-91)	42(19-92)	<0.001
Height	67(59-75)	66(52-76)	0.617
Weight	189(123-320)	170(99-325)	0.001
Bmı	30.2(21.5-51.4)	27.5(15.2-55.8)	0.002
Systolic BP	145(100-200)	132(90-250)	<0.001
Diastolic BP	87(50-118)	82(48-124)	0.257
Waist	40.5(30-56)	37(26-53)	<0.001
Hip	44.5(37-62)	42(30-64)	0.003
Waist/Hip Ratio	0.905(0.75-1.14)	0.87(0.68-1.14)	0.001

*:Mann Whitney U Test.

Table 4. Performance metrics of the Stochastic Gradient Boosting model.

Performance Metrics	Testing Stage Value (%)
Accuracy	93.6
Balanced accuracy	92.8
Sensitivity	91.7
Specificity	93.9
Positive predictive value	73.3
Negative predictive value	98.4
F1-score	81.5

variables for the diabetes input variable are given in Table 2 and Table 3.

According to the statistical analyzes performed, the analyzes performed with the other variables except the age and diastolic BP variables were found to be statistically significant. However, statistically significant results were not obtained for these age and diastolic BP variables.

The findings of the performance metrics from the Stochastic Gradient Boosting model are given in Table 4.

Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score obtained from the Stochastic Gradient Boosting model as a result of the modeling were 93.6%, 92.8%, 91.7%, 93.9%, 73.3%, 98.4%, and 81.5%, respectively.

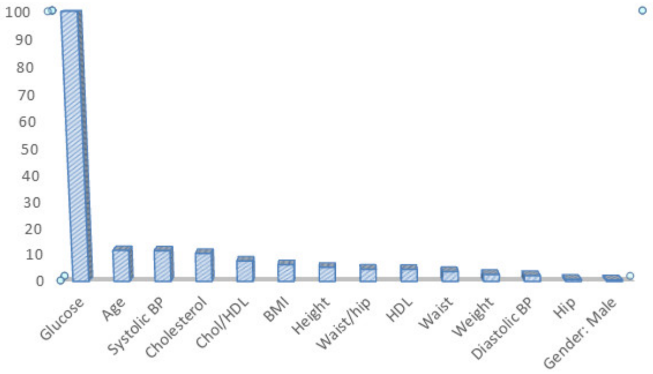

Figure 1. Graph of values for performance metrics obtained from Stochastic Gradient Boosting model.

The graph of the values related to the performance metrics obtained as a result of the modeling is given in Figure 1.

Discussion

According to recent estimates, the human population worldwide is battling an epidemic of diabetes and the disease is on the rise. Despite great strides in the understanding and management of diabetes, the disease and its complications continue unabated and its harms increase. In parallel, recent advances in understanding the pathophysiology of the disease process have paved the way for new treatment strategies to combat diabetes [15]. However, there is a great need for early prevention strategies to reduce disease-related mortality and morbidity. Recent studies show that DM can be prevented and early screening and diagnosis are therefore central to effective prevention strategies. Therefore, given the huge health burden, more attention should be paid to the early detection of DM [16, 17]. For this purpose, a number of clinical prediction methods have been developed in recent years to identify individuals with unknown diabetes or at high risk of developing diabetes. However, these models may not be readily applicable to patients reporting to a hospital for different types of services. However, it may be important to develop a risk estimation system that can be used and expanded [18, 19].

Machine learning and data mining models are increasingly used and preferred in different disciplines. The main purpose of these models is to determine the effective variables and the relationship between them and these models can also be used for prediction. In fact, machine learning models can be defined as the process of designing a model learned through experience and the process of improving model performance. These models are a sub-field of artificial intelligence and can be used as an active research and application area in different sciences. In addition, machine learning techniques are widely used and applied for the diagnosis of diseases in medical science [20-22]. In addition , machine learning techniques have been widely used for the diagnosis of DM in many studies conducted today [23, 24]. The aim of this study is to create a risk estimation model with the Stochastic Gradient Boosting model, which is one of the machine learning methods for DM, whose effects and management and treatment strategies will be easier if diagnosed early.

Among the performance criteria obtained from the Stochastic Gradient Boosting result, accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were obtained as 93.6%, 92.8%, 91.7%, 93.9%, 73.3%, and 98.4%, and 81.5%, respectively. Successful results were obtained in the diagnosis of AAp with the modeling performed, and according to the variable importance obtained as a result of the model, the variables most associated with the diagnosis were glucose, age, systolic BP, cholesterol, chol/HDL, BMI, height, waist/hip, HDL, waist, weight, diastolic BP, hip, and gender: male respectively.

In a study using the same data set, Logistic Regression, KNN, AdaBoost, and Multilayer Perceptron from machine learning models were used, and then stacking from ensemble models was used to combine these models. The highest accuracy value obtained from the study was obtained as 93%. In addition, the variable importance was not included in this study [25]. However, in the current study, the variable importance was included and the variables associated with the disease were ranked according to their importance. In a study using the same data set, many machine learning models were used and experiments were conducted under different conditions. When the results obtained were examined, the highest accuracy value was obtained as 91% [26]. With the present study, a prediction model was created with an accuracy of 93.6% and risk factors associated with diabetes were determined in order of importance.

DM, which is one of the most important health problems all over the world, should be detected at an early stage and the necessary steps for disease management should be taken earlier. For this purpose, there is a need to develop prediction models that can predict diabetes using retrospective medical records. As a result, the successful applications of machine learning models in medicine paved the way for these studies. In this study, it was seen that the ML model applied with the results obtained can predict DM. Addition, according to the results of the relevant model; the most important risk factors for DM were determined and given in degrees of importance of the risk factors. With this association, early signs of the disease can be detected and necessary precautions can be taken at early levels.

Ethics approval

The dataset used in the study is open access and does not require ethics committee approval.

References

1. Cramer JA. A systematic review of adherence with medications for diabetes. Diabetes care. 2004;27(5):1218-24.
2. Shobhana R, Begum R, Snehalatha C, Vijay V, Ramachandran A. Patients' adherence to diabetes treatment. The Journal of the Association of Physicians of India. 1999;47(12):1173-5.
3. Başkal N. Diabetes Mellitus Tanım, Klasifikasyon, Tanı, Klinik, Laboratuar ve Patogenez. Erdoğan G Klinik Endokrinoloji Anıtıp AŞ yayınları, Ankara. 2003:207-33.
4. Cameron A. The metabolic syndrome: validity and utility of clinical definitions for cardiovascular disease and diabetes risk prediction. Maturitas. 2010;65(2):117-21.
5. Iancu I, Mota M, Iancu E, editors. Method for the analysing of blood glucose dynamics in diabetes mellitus patients. 2008 IEEE international conference on automation, quality and testing, robotics; 2008: IEEE.
6. Robertson G, Lehmann ED, Sandham W, Hamilton D. Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. Journal of Electrical and Computer Engineering. 2011;2011.
7. Cole JB, Florez JC. Genetics of diabetes mellitus and diabetes complications. Nature reviews nephrology. 2020;16(7):377-90.
8. Polikar R. Ensemble learning. Ensemble machine learning: Springer; 2012. p. 1-34.
9. Akman M, Genç Y, Ankarali H. [Random Forests Methods and an Application in Health Science]. Turkiye Klinikleri J Biostat. 2011;3(1):36-48.
10. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017;15:104-16.
11. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine learning methods to predict diabetes complications. Journal of diabetes science and technology. 2018;12(2):295-302.
12. Rigla M, García-Sáez G, Pons B, Hernando ME. Artificial intelligence methodologies and their application to diabetes. Journal of diabetes science and technology. 2018;12(2):303-10.
13. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002;38(4):367-78.
14. Lawrence R, Bunn A, Powell S, Zambon M. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote sensing of environment. 2004;90(3):331-6.
15. Tiwari AK, Rao JM. Diabetes mellitus and multiple therapeutic approaches of phytochemicals: Present status and future prospects. Current science. 2002:30-8.
16. Xiong X-l, Zhang R-x, Bi Y, Zhou W-h, Yu Y, Zhu D-l. Machine learning models in type 2 diabetes risk prediction: results from a cross-sectional retrospective study in Chinese adults. Current medical science. 2019;39(4):582-8.
17. Knowler WC, Fowler SE, Hamman RF, Christophi CA, Hoffman HJ, Brenneman AT, et al. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. Lancet (London, England). 2009;374(9702):1677-86.
18. Buijsse B, Simmons RK, Griffin SJ, Schulze MB. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. Epidemiologic reviews. 2011;33(1):46-62.
19. Thoopputra T, Newby D, Schneider J, Li SC. Survey of diabetes risk assessment tools: concepts, structure and performance. Diabetes/metabolism research and reviews. 2012;28(6):485-98.
20. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J-F, et al. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems. 2012;36(4):2431-48.
21. Mansour R, Eghbal K, Amirhossein H. Comparison of artificial neural network, logistic regression and discriminant analysis efficiency in determining risk factors of type 2 diabetes. 2013.
22. Wang C, Li L, Wang L, Ping Z, Flory MT, Wang G, et al. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach. Diabetes research and clinical practice. 2013;100(1):111-8.
23. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee Y-h, et al. Screening for prediabetes using machine learning models. Computational and mathematical methods in medicine. 2014;2014.
24. Barber SR, Davies MJ, Khunti K, Gray LJ. Risk assessment tools for detecting those with pre-diabetes: a systematic review. Diabetes research and clinical practice. 2014;105(1):1-13.
25. Liza FR, Samsuzzaman M, Azim R, Mahmud MZ, Bepery C, Masud MA, et al., editors. An Ensemble Approach of Supervised Learning Algorithms and Artificial Neural Network for Early Prediction of Diabetes. 2021 3rd International Conference on Sustainable Technologies for Industry 40 (STI); 2021: IEEE.
26. Balasubramanian S, Kashyap R, CVN ST, Anuradha M, editors. Hybrid prediction model for type-2 diabetes with class imbalance. 2020 IEEE international conference on machine learning and applied network technologies (ICMLANT); 2020: IEEE.