International Journal of Advanced Engineering and Management Research

Vol. 10, No. 02; 2025

ISSN: 2456-3676

Diabetes Prediction Based on Health Indicators Using Machine Learning: Feature Selection and Algorithm Comparison

Ali Küsmüş[1]
[1]Kahramanmaraş Sütçü İmam Univercity, Vocational School of Technical Sciences,
Onikişubat, Kahramanmaraş 46100, Türkiye

doi.org/10.51505/ijaemr.2025.1115 URL: http://dx.doi.org/10.51505/ijaemr.2025.1115

Received: Apr 06, 2025 Accepted: Apr 11, 2025 Online Published: Apr 24, 2025

Abstract

This study aims to predict diabetes status (focused on Type 2 diabetes) using a dataset downloaded from the UCI machine learning repository. Diabetes is a major public health problem worldwide and early diagnosis and determination of risk factors are critical for the management of the disease. In this study, 8 machine learning algorithms (Logistic Regression, Naive Bayes, Decision Trees, Random Forest, AdaBoost, Gradient Boosting, Extra Trees, XGBoost) were compared using health indicators and demographic information obtained from real people. The performance of the models was evaluated using Accuracy, Balanced Accuracy, Precision, Recall, F1-Score and ROC AUC metrics. In addition, the factors that most affect the risk of diabetes were determined using the Random Forest algorithm and the performance of the models was re-evaluated. The results showed that Gradient Boosting and XGBoost algorithms could predict diabetes status with a high performance of 86%. The most important risk factors were found to be high blood pressure (HighBP), body mass index (BMI), general health perception (GenHlth) and age (Age). These findings provide an opportunity for early diagnosis by estimating the probability of people developing diabetes.

Keywords: diabetes prediction, machine learning, risk factors, gradient boosting, xgboost

1. Introduction

Diabetes, especially Type 2 diabetes, is a metabolic disease in which blood sugar levels are chronically elevated because of the pancreas not producing enough insulin or the body's inability to effectively use the insulin produced (insulin resistance) (World Health Organization [WHO], 2023). Affecting millions of people globally, diabetes is a major cause of morbidity and mortality, leading to serious complications such as heart disease, stroke, kidney failure, blindness, and lower extremity amputations (International Diabetes Federation [IDF], 2021). The increasing prevalence of diabetes poses a major economic burden on health systems.

In this context, early diagnosis of diabetes and identification of high-risk individuals are vital for preventing complications and effective management of the disease. The Behavioral Risk Factor

Surveillance System (BRFSS) is the largest, continuously ongoing telephone survey system in the United States (US) that collects data on health-related risk behaviors, chronic health conditions, and healthcare utilization at the state level (Centers for Disease Control and Prevention [CDC], 2024a). BRFSS data are widely used to monitor the prevalence and associated risk factors of various chronic diseases, including diabetes.

Today, machine learning (ML) is used extensively in the medical field due to its ability to make preliminary diagnoses and predictions (Sidey-Gibbons and Sidey-Gibbons, 2019). ML algorithms can handle complex data, extract meaningful data from them, and create predictions based on subsequent future data. ML offers many benefits such as predicting diabetes, determining risk, creating customized risk scores, and combining screening strategies (Kavakiotis et al., 2017).

The main purpose of this study is to provide early diagnosis of diabetes and take precautions using data obtained from real patients. In addition, it is to compare the performances of various machine learning models to obtain more successful results. The most important research questions addressed by the study are:

1. How accurately can machine learning models predict whether an individual has diabetes?
2. What are the risk factors that most strongly determine the risk of diabetes?
3. Which attribute has the highest importance in early diagnosis of diabetes?

After the introduction, this study will detail the relevant literature, the dataset used and the methodology, present and discuss the findings, and end with conclusions and recommendations.
Both traditional statistical methods and machine learning methods have been frequently used in studies conducted to predict diabetes.

There are many risk factors determined for Diabetes Type 2 disease in the literature. These risk factors include factors such as genetic predisposition, age, obesity, sedentary lifestyle, unhealthy diet, high blood pressure, and high cholesterol (American Diabetes Association, 2023). In many surveys conducted on diabetes, obesity, physical activity, smoking, fruit/vegetable consumption, high blood pressure, and high cholesterol are measured indirectly and directly (CDC, 2024b).
Machine learning algorithms are used as a powerful method in studies conducted on predicting diabetes risk factors. In the studies, algorithms such as Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Trees (DT), Random Forest (RF), Gradient Boosting Machines (GBM) and Artificial Neural Networks (ANN) were used (Kavakiotis et al., 2017; Zou et al., 2018).

In his study in 2021, Bilgin tried to predict diabetes disease by using machine learning algorithms effectively. In the study, he used machine learning algorithms such as K-NN algorithm, Multilayer Artificial Neural Networks, Support Vector Machines, Decision Trees. The highest accuracy rate of 99.81% was achieved by the K-NN algorithm. The study contributed to the development of an early diagnosis kit for diabetes (Bilgin, 2019).

Saiteja et al. In the study conducted by XGBoost algorithm, Logistic Regression, Decision Trees, Random Forest, K-NN, Support Vector Machines and Naive Bayes Machine Learning algorithms were used to predict diabetes disease. Important features in the dataset are used as blood pressure, body mass index and blood sugar rate. Hyperparameter optimization was performed with GridSearchCV and XGBoost algorithm provided the highest accuracy rate (SaiTeja et al., 2025).

The study conducted by Sadaria and Parekh emphasizes the importance of machine learning and data mining techniques for early diagnosis of diabetes. The performances of algorithms such as Logistic Regression, Random Forest and XGBoost were evaluated and XGBoost gave the best results. The study reveals the potential of machine learning in the management of early diagnosis of diabetes (Sadaria & Parekh, 2024).

Perez and Molano compared Logistic Regression, Support Vector Machines, Random Forest, XGBoost and CatBoost machine learning algorithms to predict diabetes based on lifestyle factors. The XGBoost algorithm showed the best success with an accurate rate of 85%. The study emphasizes the importance of lifestyle variables in diabetes prediction(Perez & Avella-Molano, 2025).

Literature research shows that predicting diabetes with Machine Learning methods using data sets collected from real patients is a valid and potentially effective approach. This study aims to contribute to the existing knowledge by systematically comparing various Machine Learning algorithms used in the literature and evaluating the most important risk factors and their performance.

2. Material and Method

In this part of the study, the dataset used, the preprocessing methods applied, the machine learning algorithms used, the feature importance determination and selection methods, and the methods used to evaluate the performance of the models are explained in detail.

2.1 Dataset

The dataset used in this study was obtained from the UCI Machine Learning Repository (University of California Irvine Machine Learning Repository), which is accepted as a reference source in the field of machine learning. UCI datasets are widely used in academic studies and their validity is accepted internationally. The dataset is also available on the Kaggle platform under the title "Diabetes Health Indicators Dataset". The dataset was obtained from the 2015 survey conducted by the US Centers for Disease Control and Prevention (CDC) and has been published publicly. The dataset contains a total of 22 attributes, as shown in Table 1, with the target variable being diabetes status (Diabetes_binary). The target variable is coded as binary (binary) as 0 (no diabetes or prediabetes) and 1 (prediabetes or diabetes).

Table 1 Dataset Attribute Information

Variable Name	**Description**
ID	Patient ID
Diabetes_binary	0 = no diabetes 1 = prediabetes or diabetes
HighBP	0 = no high BP 1 = high BP
HighChol	0 = no high cholesterol 1 = high cholesterol
CholCheck	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
BMI	Body Mass Index
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
Stroke	(Ever told) you had a stroke. 0 = no 1 = yes
HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
PhysActivity	physical activity in past 30 days - not including job 0 = no 1 = yes
Fruits	Consume Fruit 1 or more times per day 0 = no 1 = yes
Veggies	Consume Vegetables 1 or more times per day 0 = no 1 = yes
HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes
AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
GenHlth	Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
MentHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk	Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
Sex	0 = female 1 = male

Age	13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
Education	Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
Income	Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

The dataset was created with information from a total of 253,680 patients. Missing data analysis was also performed on the dataset, it was determined that there was no missing data. When the data distribution of the target variable was examined, it was determined that there were 86% non-diabetic patients and 14% diabetic patients. This distribution shows us that there is a significant class imbalance.

2.2 Machine Learning Models

In this study, 8 different Machine Learning algorithms were used, and their success values were examined. Machine Learning Algorithms Used:
Logistic Regression (LR): It is a high probability estimation model used for binary classification problems. In the algorithm that provides high interpretability, it creates a linear decision boundary through the Sigmoid function(Hosmer, Lemeshow, & Sturdivant, 2000).

Naive Bayes Classifiers: This method, based on Bayes theorem, works with the assumption that the independent variables are unconditionally independent(Murphy, 2012).

Decision Tree (DT): Decision trees classify by dividing the feature space with sequential branching rules. It requires pruning techniques due to the tendency to over-learning (Breiman, 2001).

Random Forest (RF): It is based on combining many decision trees with the bagging method (Breiman, 2001).
AdaBoost: Iteratively optimizes the weighted combination of weak learner (Friedman, 2001)s.
Gradient Boosting (GB): Builds sequential trees by minimizing error gradients (Friedman, 2001).
Extra Trees (ET): Randomizes branching thresholds, unlike RF (Geurts et al., 2006).
XGBoost: It is a GB implementation optimized with regularization and parallel processing (Chen & Guestrin, 2016).

Vol. 10, No. 02; 2025

ISSN: 2456-3676

2.3 Evaluation Metrics

The following metrics were used to evaluate and compare the performance of the models developed within the scope of the study.

Accuracy: The ratio of correctly classified samples to the total samples (Murphy, 2012).

Accuracy = (TP + TN) / (TP + TN + FP + FN) (1)

Precision: How many of the samples predicted as positive were truly positive. How many of the predicted as diabetes were truly diabetic(Manning, 2008).

Precision = TP / (TP + FP) (2)

Recall: How many of the truly positive samples were correctly predicted as positive. How many of the truly diabetic samples were correctly detected.

Recall = TP / (TP + FN) (3)

F1-Score: The harmonic average of Precision and Recall. It provides a balance between these two metrics(Powers, 2020).

F1-Score = 2 * (Precision * Recall) / (Precision + Recall) (4)

ROC AUC (Area Under the Receiver Operating Characteristic Curve): It is a measure of how well the model can distinguish positive and negative classes at different threshold values. The closer it is to 1, the better. It is more resistant to imbalance(Fawcett, 2006).

Balanced Accuracy: It is the average of the Recall values for each class. It is a fairer performance measure than Accuracy for imbalanced data sets(Krawczyk, 2016).

(Recall_Class_0 + Recall_Class_1) / 2

3. Results

In this section, the results of the analysis carried out in the study are presented and the findings are discussed in the light of the literature.

3.1 Descriptive Statistics

When the distribution of the target variable "Diabetes_binary" in the dataset is examined, it is determined that approximately 13.9% of the participants are prediabetes or diabetic (Value 1), and 86.1% are not diabetic (Value 0). This situation confirms the significant class imbalance in the dataset by informing us about it. It is also seen that it increases the importance of metrics such as Balanced Accuracy in model evaluation. When the main risk factors were examined, "HighBP", "BMI", "GenHlth" and "Age" were identified as the most important attributes, as seen in Figure 1.

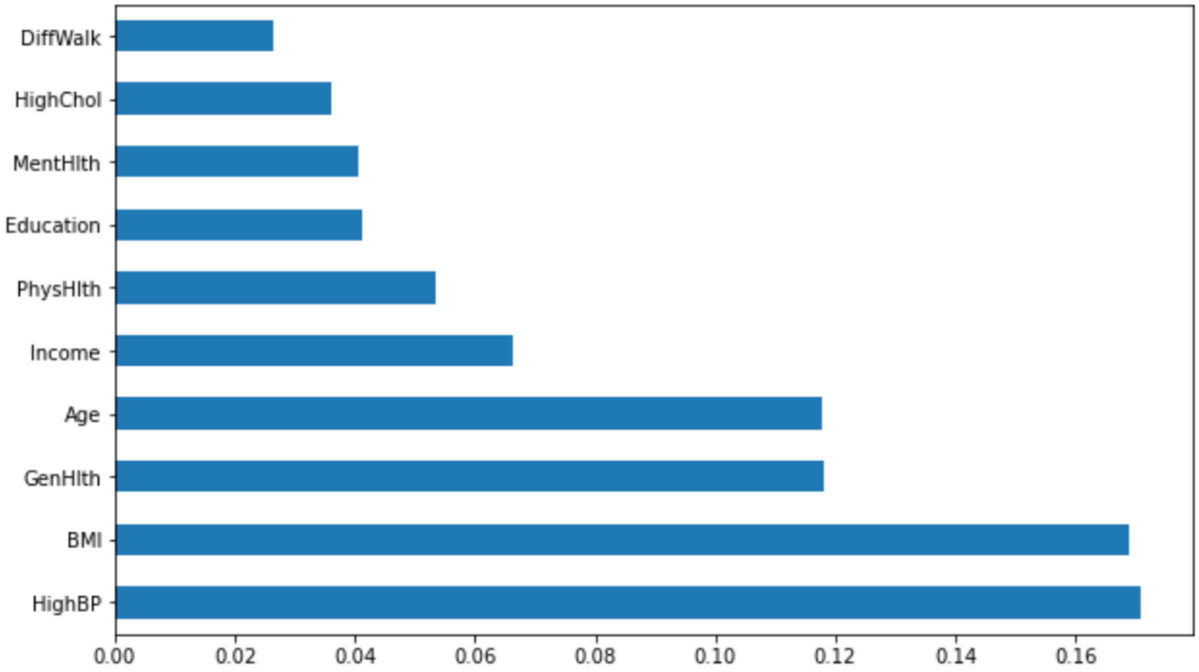

Figure 1 Feature Importances Comparison

The fact that factors such as high blood pressure (HighBP), body mass index (BMI), general health perception (GenHlth), and age (Age) are at the top of the list are consistent with the findings in the literature (CDC, 2024b). The importance of factors such as physical health status (PhysHlth) and walking difficulty (DiffWalk) may reflect the effect of diabetes on physical functions or the contribution of these conditions to the development of diabetes. The fact that socioeconomic factors such as income and education are also on the list indicates the effect of health inequalities and lifestyle factors on the risk of diabetes (Braveman & Gottlieb, 2014). It has been observed that behavioral factors such as smoking, fruit (Fruits) and vegetable (Veggies) consumption are also important but not as dominant as the physiological and demographic factors at the top.

3.2 Model Performance

Model Performance (All Features): 8 Machine Learning models were trained using all 21 features and their performance was evaluated on the test set. 80% of the data set was set as training and 20% as testing. The results are shown in Table 2.

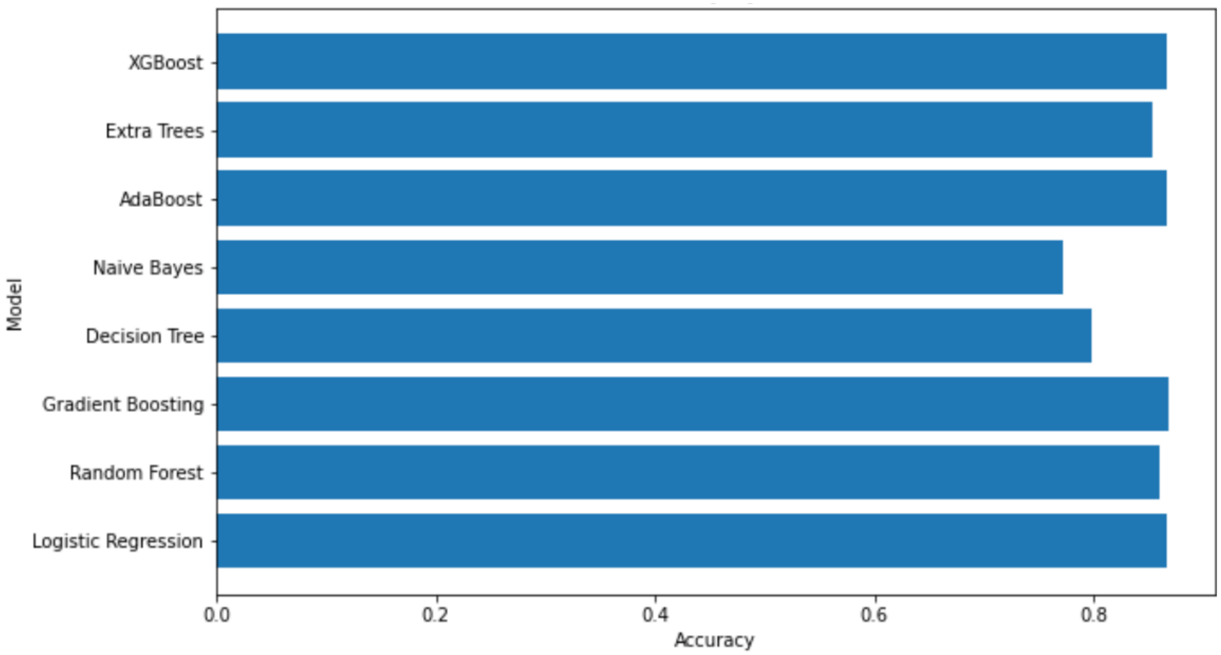

Figure 2 Dataset Attribute Importance Levels

Table 2 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0,86	0,72	0,58	0,6
XGBoost	0,86	0,72	0,58	0,6
AdaBoost	0,85	0,71	0,58	0,61
Logistic Regression	0,85	0,71	0,57	0,59
Random Forest	0,84	0,68	0,57	0,59
Extra Trees	0,83	0,65	0,57	0,59
Decision Tree	0,79	0,59	0,6	0,6
Naive Bayes	0,77	0,62	0,69	0,63

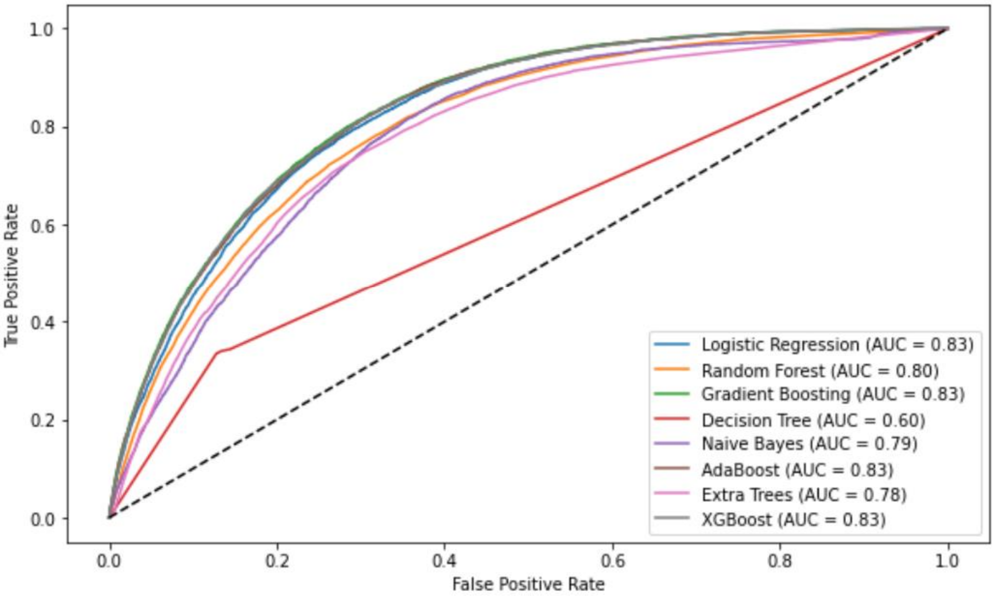

Figure 3 ROC Curves for All Models

When the results are examined, it is seen that Gradient Boosting, XGBoost, AdaBoost and Logistic Regression models exhibit the highest performance especially in terms of Balanced Accuracy and ROC AUC metrics. For example, Gradient Boosting and XGBoost achieved Balanced Accuracy and 83% ROC AUC scores. While the performance of simpler models such as Extra Trees was 78%, algorithms such as Gradient Boosting, XGBoost, AdaBoost and Logistic Regression generally gave superior results. This finding may indicate the existence of complex and non-linear relationships between the factors affecting the risk of diabetes and may indicate that ensemble models are more successful in capturing these relationships (Maniruzzaman et al., 2018). Low Recall scores (especially when Precision is high) may indicate that some models have difficulty in detecting patients with diabetes (high False Negative rate), which may reflect class imbalance and is a point that should be considered in clinical practice. Using Balanced Accuracy allows us to see this situation more clearly.

3.3 Limitations of the Study

- Data Structure: The dataset was created based on the statements of the patients because of interviews. The data was not verified with medical records.
- Feature Engineering: Existing features were used directly in the study. Combining features, creating interaction terms, or different coding methods (e.g. using continuous instead of categorical for age and BMI or vice versa) may affect performance.
- Model Optimization: The models were run with default parameters. For comprehensive model optimization, analyses were performed according to the models and their performance could be further improved by using specific parameters.
- External Validity: The dataset was conducted in a specific region and a specific period. It was not conducted in a different region and in the current period.

- Prediabetes Separation: Combining prediabetes and diabetes in the target variable in the dataset (if so) may produce different results compared to predicting only diabetes or only prediabetes.

Future Studies: Future studies can check the validity of the findings by creating newer data sets. Different feature engineering techniques or artificial intelligence-supported machine learning models can be tried.

3.4 Discussion

The findings obtained because of this study show that the data in the dataset obtained from the UCI Machine Learning repository is an effective source in predicting diabetes status using machine learning algorithms. It is shown that the Machine Learning algorithms applied to the dataset produce very close results to each other and consistent results are obtained in model performance values.

The analysis of the importance of the features confirmed that factors such as general health status, high blood pressure, BMI and age are the most important determinants for diabetes, in line with the literature. This finding confirms the importance of public health interventions and clinical screening programs focusing on these factors.

In addition, in the light of the results obtained in the study, simpler and more efficient models can be developed by reducing the number of features. For example, an assessment tool can be created by performing a diabetes risk analysis based on only a few critical questions.

4. Conclusion

This study successfully demonstrated the effectiveness of various machine learning models to predict diabetes status using data obtained from the UCI Machine Learning repository. The findings reveal the following key conclusions:

1. Health and demographic indicators in the dataset can be used to predict individuals' diabetes status with high accuracy through machine learning models.
2. Factors that largely determine diabetes risk include general health status perception, high blood pressure, body mass index, age, physical health problems, and walking difficulty. These findings are largely consistent with the existing medical literature and confirm the critical role of these factors in diabetes screening and prevention.
3. A prediction performance close to that achieved using all the features can be achieved using a subset of the most important risk factors (e.g., the top 10 factors). This offers the potential to reduce model complexity and develop more practical, focused risk assessment tools.

As a result, this study emphasizes that large-scale health datasets, when combined with machine learning algorithms, can make significant contributions to the understanding and management of chronic diseases such as diabetes. The models obtained and the identified risk factors can help develop public health policies, design targeted screening programs, and raise awareness of individuals about their own risks. Early diagnosis and intervention at the initial stage of the disease can improve the quality of life of individuals. Considering the limitations of the study, it is recommended that the findings be supported by further research and validation studies. It is also expected to contribute greatly to subsequent studies.

References

American Diabetes Association (ADA). (2023). Standards of Medical Care in Diabetes—2023. Diabetes Care, 46(Supplement_1).

Bilgin, M. (2019). Türkçe metinlerin siniflandirma başarisini artirmak için yeni bir yöntem önerisi. *Uludağ Üniversitesi Mühendislik Fakültesi Dergisi*, *24*(1), 125-136.

Breiman, L. (2001). [No title found]. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Centers for Disease Control and Prevention (CDC). (2024a). Behavioral Risk Factor Surveillance System (BRFSS). Retrieved from https://www.cdc.gov/brfss/data_documentation/index.htm

Centers for Disease Control and Prevention (CDC). (2024b). National Diabetes Statistics Report. Retrieved from https://www.cdc.gov/brfss/data_documentation/index.htm

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. San Francisco California USA: ACM. https://doi.org/10.1145/2939672.2939785

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189-1232.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). *Applied logistic regression. Hoboken*. NJ: John Wiley & Sons.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0

Maniruzzaman, Md., Rahman, Md. J., Al-MehediHasan, Md., Suri, H. S., Abedin, Md. M., El-Baz, A., & Suri, J. S. (2018). Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. *Journal of Medical Systems*, *42*(5), 92. https://doi.org/10.1007/s10916-018-0940-7

Manning, C. D. (2008). *Introduction to information retrieval*. Cambridge university press. Geliş tarihi gönderen http://diglib.globalcollege.edu.et:8080/xmlui/bitstream/handle/123456789/1096/Manning_introduction_to_information_retrieval.pdf?sequence=1&isAllowed=y

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press. Geliş tarihi gönderen https://books.google.com/books?hl=tr&lr=&id=RC43AgAAQBAJ&oi=fnd&pg=PR7&dq=Murphy,+K.+P.+(2012).+Machine+Learning:+A+Probabilistic+Perspective.+MIT+Press.&ots=unjxaFLqZc&sig=tbzANVBkhXH5wmcZ6yd3ht7ru2Y

Perez, E. R., & Avella-Molano, B. (2025). Learning from the machine: İs diabetes in adults predicted by lifestyle variables? A retrospective predictive modelling study of NHANES 2007–2018. *BMJ open*, *15*(3), e096595.

Powers, D. M. W. (2020, Ekim 11). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*. arXiv. https://doi.org/10.48550/arXiv.2010.16061

Sadaria, P., & Parekh, R. (2024). An Analysis of Machine Learning Approaches for Diabetic Prediction. Içinde V. Goar, A. Sharma, J. Shin, & M. F. Mridha (Ed.), *Deep Learning and Visual Artificial Intelligence* (ss. 49-57). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-4533-3_5

SaiTeja, L., Regulwar, G. B., Sai Anish Reddy, G., Satwick, T., Kulkarni, V., Singh, S., & Shalini, K. (2025). Diabetes Prediction by Using Various Machine-Learning Algorithms. Içinde V. Bhateja, P. Patel, & M. Simic (Ed.), *Intelligent Data Engineering and Analytics* (ss. 233-243). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-96-0139-4_19

World Health Organization (WHO). (2025). Diabetes. Retrieved from https://www.who.int/health-topics/diabetes#tab=tab_1

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, *9*, 515.