Orginal Article
Open Access
Iranian Journal of Diabetes and Obesity (IJDO)

# Predicting Diabetes Risk Using Machine Learning: A Comparative Study on the Yazd Health Study (YaHS)

Fateme Sefid[1], Nazanin Norouzi-Ghahjavarestani[2], Malihe Soleymani-Tabasi[2], Jamal Zarepour-Ahmadabadi[2] , Ghasem Azamirad[3], Mohamah yahya Vahidi Mehrjardi [4], Masoud Mirzaei[5], Seyed Mehdi Kalantar[6,7*]

[1]Department of Molecular Medicine,School of Advanced Technologies in Medicine,Shahid Sadoughi University of Medical Sciences Yazd Iran.
[2]Department of Computer Science, Yazd University, Yazd, Iran.
[3]Department of Mechanical Engineering, Yazd University, Yazd, Iran.
[4]Diabetes Research Center, Non-communicable Diseases Research Institute, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.
[5]Yazd Cardiovascular Research Centre, Non-Communicable Diseases Research Centre, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.
[6]Abortion Research Centre, Yazd Reproductive Sciences Institute, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.
[7]Meybod Genetic Research Center, Yazd, Iran.

## Abstract

Diabetes is a chronic disease that can significantly affect health at the global level, highlighting the importance of accurate early risk prediction to support prevention and management efforts. This study aims to evaluate the effectiveness of some efficient machine learning algorithms: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), and Decision Tree (DT) in diabetes risk prediction using dataset acquired from Yazd Health Study (YaHS). Extensive preprocessing steps, including data cleaning, class imbalance handling through Synthetic Minority Oversampling Technique and Edited Nearest Neighbors (SMOTEENN), and feature selection, are applied to enhance the performance of models. Among the evaluated machine learning algorithms, the Random Forest classifier achieved the highest performance with an accuracy of 97%, outperforming other methods in terms of predictive capability. The findings highlight the vital importance of effective data preprocessing and algorithm selection in developing reliable predictive models from healthcare datasets.

**Keywords**: Machine learning, Diabetes, Random forest

QR Code:

**Citation:** Sefid F, Norouzi-Ghahjavarestani N, Soleymani-Tabasi M, Zarepour-Ahmadabadi J, Azamirad G, Vahidi Mehrjardi M Y, et al . Predicting Diabetes Risk Using Machine Learning: A Comparative Study on the Yazd Health Study (YaHS). IJDO 2025; 17 (3) :182-192

**URL:** http://ijdo.ssu.ac.ir/article-1-967-en.html

10.18502/ijdo.v17i3.19267

**Article info:**

**Received:** 29 May 2025
**Accepted:** 30 June 2025
**Published in July 2025**

This is an open access article under the (CC BY 4.0)

**Corresponding Author:**

**Seyed Mehdi Kalantar,** Abortion Research Centre, Yazd Reproductive Sciences Institute, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.

**Tel:** (98) 913 151 8918
**Email:** SMkalantar@yahoo.com
**Orcid ID:** 0000-0002-6994-6449

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]

## Introduction

Diabetes is a persistently chronic and extensive metabolic disorder characterized by elevated levels of blood glucose, which can lead to many serious complications if left untreated. The World Health Organization reports that diabetes is becoming highly prevalent globally, and recently estimated at over 500 million people worldwide according to the latest WHO reports (1). Early diagnosis and intervention are important strategies to prevent or delay diabetes progression and reduce the likelihood of related complications, such as cardiovascular disease, kidney failure, vision defects, and others.

Certain traditional diagnostic methods, which depend on biochemical tests and multiple clinical assessments, can be time-consuming, costly, or inaccessible to some populations. The limitations of traditional diagnostic methods have motivated the implementation of machine learning (ML) models in medical settings. These models analyze and identify meaningful patterns within high-dimensional complex data and predict disease risk effectively.

Machine learning methods have significant potential for predicting diabetes using clinical, demographic, behavioral, and biochemical data. Past studies have demonstrated the potential of many ML algorithms, including Decision Trees, Support Vector Machines, Random Forests, and ensemble models, to be effective in medical classification tasks and to compete with traditional statistical approaches.

In this paper, we develop a machine learning-based framework for predicting diabetes risk using data from the Yazd Health Study (YaHS), a large-scale population-based cohort conducted in Iran. This study aims to explore the potential of various supervised learning algorithms applied to a diverse and representative dataset, with the goal of identifying the most accurate and robust predictive model. To ensure optimal performance, we implement systematic data preprocessing, apply appropriate feature selection strategies, and conduct comprehensive model evaluation. The resulting model may contribute to early diagnosis efforts and provide healthcare professionals with a practical tool for real-time diabetes risk assessment.

The remainder of this paper is structured as follows: Section 2 reviews related research in diabetes prediction using machine learning. Section 3 describes the proposed method, including the dataset, preprocessing steps, and the machine learning models. Section 4 presents the experimental results and compares the performance of the evaluated models. Finally, Section 5 concludes the study and outlines directions for future research.

### Literature reviews

Machine learning models have been widely used to address medical problems in recent years. These methods can quickly and accurately identify the risk of developing a disease. This ability of machine learning models has led to an increase in their use by healthcare professionals in the medical field. The risk of developing diabetes is one area where these models have been employed to predict its probability, which will be discussed in the following sections

Alkalifah (2) examined various types of machine learning regression models to predict changes in blood glucose using Continuous Glucose Monitoring (CGM) data. The study applied several models, including Binary Decision Tree (BDT), Support Vector Machine (SVM), Artificial Neural Network (ANN), Gaussian Process Regression (GPR), and Boosting Tree Ensembles on a dataset comprising 14,733 observations of real and synthetic physiological data: heart rate, body temperature and blood pressure. In this study, the BDT achieved the highest accuracy of 92.58%, followed by Boosting Tree Ensembles at 92.04%, and GPR at 88.59%. Although the models were regression-based,

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]



(CC BY 4.0)

the authors used predicted glucose levels to classify glycemic conditions (hypoglycemia, hyperglycemia, and normoglycemia), concluding that tree-based models were most effective in this context.

In another study (3), the authors developed and evaluated a machine learning model to predict short-term iatrogenic hypoglycemia in hospitalized patients using a comprehensive dataset of more than 35,000 inpatient admissions. The models evaluated in the study included Random Forest, Logistic Regression, Naive Bayes, and Gradient Boosting classifiers. Among them, Random Forest was determined to be the best-performing classifier, achieving a remarkable accuracy of 95%. The model demonstrated that it was well-suited for complex hospital data.

The J48 decision tree model utilized by Chen (4) was based on the C4.5 algorithm. It was implemented in a high-risk adult cohort for diabetes prediction in China. Eight clinical characteristics, such as body mass index and glucose levels, have been used to develop a model that is both interpretable and computationally efficient. The authors achieved an accuracy of 90.04% in predicting a diabetes diagnosis, concluding that decision tree methods, applied to organized clinical data, are effective. While J48 could provide many features and decision pathways, it relies on only one classifier and is limited by the simplicity of the dataset.

As mentioned earlier, various machine learning models have been developed in the field of medicine, especially for diagnosing diabetes and related diseases; each of them has advantages and disadvantages. In this study, we will examine machine learning models that can assess the risk of developing diabetes more accurately.

In a recent study, Kumar et al. (2023) applied several machine learning models-including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF)-to predict diabetes using clinical datasets. Their analysis focused on evaluating these algorithms based on accuracy and reliability. The results showed that Random Forest outperformed other models, demonstrating strong predictive capability and robustness for diabetes classification tasks. The study reinforces the effectiveness of ensemble methods in handling structured medical data for early disease detection (5).

## Material and methods

In this section, the dataset, data preprocessing, and machine learning algorithms are presented in detail. An overview of the steps involved in the proposed method is also illustrated in Figure 1.

### Dataset

The dataset used in this paper is the Yazd Health Study (YaHS), which is a population-based cohort of over 9,000 individuals aged 20 to 70 years from the Greater Yazd Area of Iran. This dataset contains more than 300 features covering a wide range of domains such as demographics, health history, health behaviors, food and lifestyle habits, indicators of mental health, health-related physical activity, and anthropometric data, among others.

It also includes laboratory test results and biological samples (stored in a biobank) that can be utilized for genetic and biochemical analysis. Consequently, this dataset represents a comprehensive collection of data gathered through diverse methodologies (i.e., structured interviews, clinical assessments, and follow-ups) and offers substantial potential for elucidating non-communicable diseases and associated risk factors from a Middle Eastern perspective.

Following initial preprocessing steps, including missing value management and outlier removal, a focused subset was selected to better capture diabetes-related factors. This modified subset consisted of 105 features and 3911 samples, which were initially classified into diabetic and non-diabetic groups (6).

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]



(CC BY 4.0)

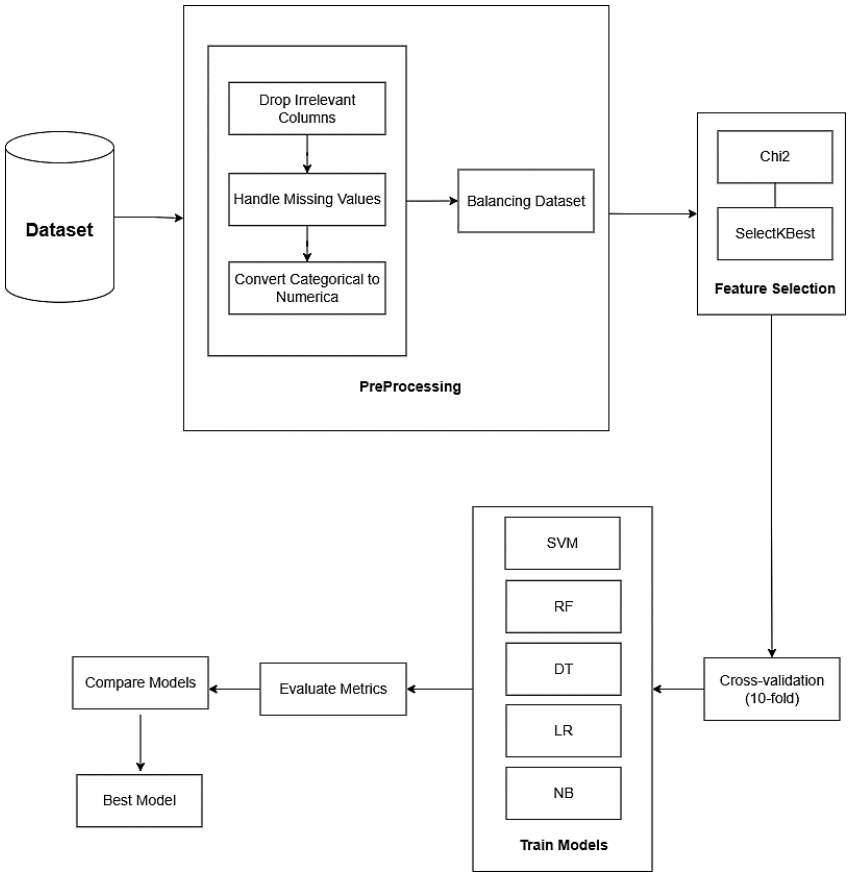


**Figure 1. An overview of the proposed method**

The final dataset used to train the model consisted of 595 diabetic and 3316 non-diabetic individuals. The purpose of this selection was to ensure higher quality and relevance of the data for predicting diabetes risk.

As the YaHS dataset includes a wide range of features, various studies have been conducted on it. Some of these studies have focused on heart disease (7), others on thyroid disorders (8), COVID-19 (9), and other health conditions. In all these studies, irrelevant features are discarded, and only relevant ones are retained.

## Data-preprocessing

One of the important steps in working with machine learning models is data preprocessing. Properly preprocessed data can significantly improve the performance of the models.

The first step of data preprocessing involves removing unrelated features from the dataset. In this study, features that were not relevant to diabetes prediction were eliminated. After that, missing values in the remaining features were addressed. Since the YaHS dataset includes null values, the mean imputation technique was applied. This method computes the mean of each column and replaces the null values with the corresponding mean (10). Since machine learning models require numerical input, categorical features in YaHS dataset were converted into numerical format.

Another important preprocessing step before applying the machine learning models is addressing class imbalance. In this study, the SMOTEENN method is applied for this purpose. This approach combines the Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN) to enhance both the representation of the minority class and the quality of the

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]



(CC BY 4.0)

synthetic samples. SMOTE generates synthetic examples by interpolating between existing minority class instances, thereby balancing class distribution. After generating synthetic data, ENN is applied to eliminate noise by removing samples that differ substantially from their neighbors, especially when the majority of neighbors do not belong to the minority class. This combined approach improves class balance while preserving meaningful data points, reducing the likelihood of misclassification and enhancing the robustness of the model input.

Although using such techniques may introduce synthetic data, it was deemed necessary in this study due to the highly imbalanced nature of the dataset (11).

## Feature selection

Given that the aim of this research is to predict the risk of diabetes in individuals, and since the YaHS is a high-dimensional dataset, a feature selection procedure was performed to highlight the most relevant attributes influencing the classification task. This process underwent several iterations and incorporated expert knowledge, statistical feature selection methods, and correlation analysis.

To begin the feature selection process, missing values were imputed using mean imputation, and irrelevant variables not associated with diabetes prediction were removed. Following this, the SelectKBest method with a Chi-squared scoring function was applied to evaluate the statistical dependency between each input feature and the output class (i.e., diabetes status). The Chi-squared test was chosen because it is well-suited for categorical classification tasks and provides a measure of how strongly each independent feature is associated with the target variable. This enables ranking features based on their relevance. Features with a Chi-squared score greater than 50 were retained for further analysis, Features with a chi-square score below this value had little or no impact on model performance and were often weakly associated with diabetes risk. Therefore, only features scoring above 50 were retained for further analysis. Figure 2 shows some of these features.

To further optimize the feature selection and avoid redundancy, a correlation matrix was computed among the remaining features. A correlation threshold was applied to eliminate highly correlated features, ensuring that only one representative feature was retained from each correlated group. This step helps to ensure that the selected features are both informative and relatively independent, thereby reducing the risk of multicollinearity and enhancing model's generalizability (Figure 2).

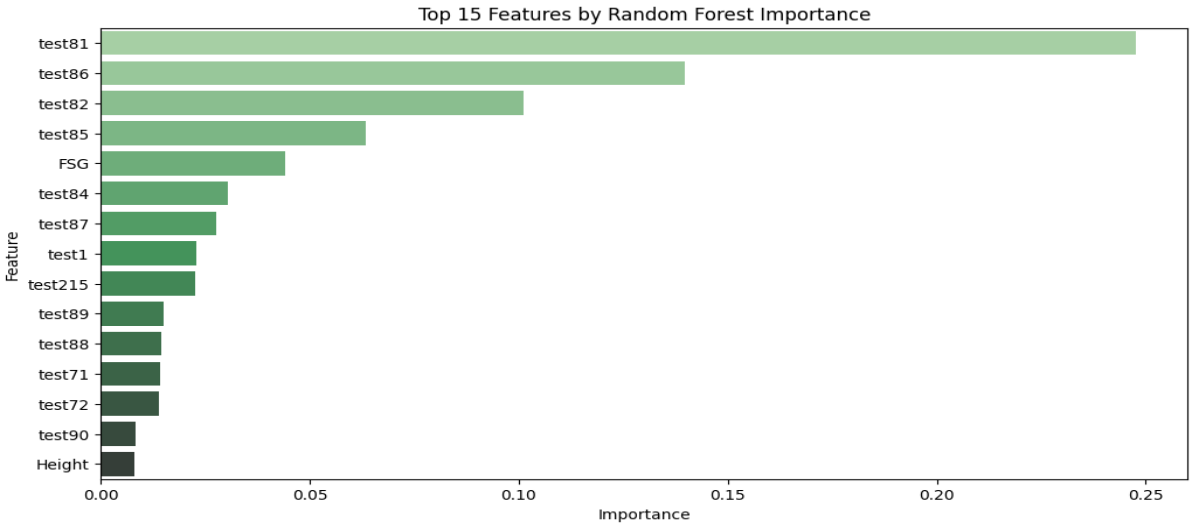


**Figure 2. Top 15 features by random forest importance**



(CC BY 4.0)

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]

The most relevant features (30 in total) were primarily anthropometric and biochemical factors, such as Fasting Serum Glucose (FSG), Weight, Height and Triglycerides (TG). However, several clinically derived variables from structured assessments also showed a strong association with diabetes outcomes.

## Model evaluation via k-fold cross-validation

K-fold cross-validation is a widely used technique for splitting data into training and testing sets. In this approach, the dataset is partitioned into k equal-sized subsets (folds). During each iteration, k- 1 folds are used for training, and the remaining fold is used for testing the machine learning algorithm. This process is repeated k times, with each fold used exactly once as the test set. The final model performance is computed by averaging the evaluation metrics obtained from each fold. In this study, 10-fold cross-validation was employed to assess the performance of the machine learning models (12).

## Machine learning models

In this research, five supervised machine learning algorithms were utilized to classify individuals as being at risk for diabetes or not, based on multiple clinical and physiological features. The primary objective for using various supervised algorithms was to compare their performance and identify the most accurate and generalizable model for diabetes prediction.

The selected models belong to different algorithmic families, including linear models, probabilistic classifiers (mostly generative and variational methods), tree-based algorithms, and instance-based learners. Because the model families are diverse, this enables performance comparisons under different learning assumptions and decision boundaries. Each algorithm is briefly described below. More complex models such as XGBoost, AdaBoost, and deep neural networks were not included in this study due to considerations related to interpretability, computational constraints, and the primary focus on evaluating the performance of more classical and widely used machine learning algorithms in clinical prediction tasks.

Logistic Regression (LR): Logistic regression is a popular linear model for binary classification that represents the probability of a class using a logistic function. It assumes a linear relationship between the input features and the log-odds of the target class. Logistic regression is fast, interpretable, and serves as a strong baseline for classification tasks (13).

Decision Tree Classifier (DT): Decision trees are non-parametric models that recursively partition the feature space into cells and maximize information gain by creating thresholds based on the feature space. Decision trees can capture non-linear relationships and can be visualized to show the logical decision rules (14).

Random Forest Classifier (RF): Random forests are an ensemble (i.e., bagging) of decision trees created during the training stage and will output the class that is the mode of 'votes' from the set of predictions for the individual trees. Random forests reduce the effects of overfitting and improve accuracy through averaging, and they also work very well for structured data, like tabular data (15).

Support Vector Machine (SVM): SVMs are powerful classifiers and find the hyperplane that best separates the two classes while maximizing the margin. A linear kernel was selected for the SVM model to minimize computation time while taking advantage of the linear kernel's robustness in high dimensional feature spaces. SVMs generally have good generalization performance (16).

Gaussian Naive Bayes (GNB): A probabilistic classification model based on Bayes' theorem under the assumption that the features of a sample are independent of each other and that feature values are normally distributed. Given the assumptions of independence and normality, Naive Bayes may not be able to fit complex boundaries on the data; however, Naive Bayes often produces results that are not much worse than more

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]



(CC BY 4.0)

advanced models, particularly with small datasets (17).

The algorithms were fitted on the preprocessed training dataset, then the held-out/unseen test dataset was evaluated on the classifier models. Implementing multiple learning algorithms allowed for a comprehensive performance comparison, which ultimately revealed that the Random Forest model was the most accurate and well-balanced classifier for the specifics of the dataset.

## Ethical considerations

This study uses a dataset that contains sensitive health information, which was handled in accordance with ethical guidelines to protect participants' privacy. Applying machine learning in healthcare raises ethical concerns such as bias, fairness, and the responsible use of medical data. Care was taken to minimize bias, and the model is intended to support, not replace, clinical judgment.

## Results

### Experimental results

This section presents the performance evaluation of five supervised machine learning algorithms-Random Forest, Decision Tree, Support Vector Machine, Gaussian Naïve Bayes, and Logistic Regression—for predicting diabetes risk using features extracted from the Yazd Health Study (YaHS) dataset.

To assess the classification effectiveness of each model, four key performance metrics were used: accuracy, precision, recall, and F1-score. These metrics offer a comprehensive assessment of the model's performance, particularly under conditions of class imbalance. To enhance the robustness of the evaluation, standard deviations (SD) are reported alongside the performance metrics, thereby providing insights into the consistency and stability of the model across multiple runs. Furthermore, from a clinical standpoint, minimizing false negatives is of paramount importance in the context of diabetes detection, which underscores the relevance and appropriateness of the selected evaluation metrics.

The implementation of multiple learning algorithms enabled a thorough comparative analysis, which revealed that the Random Forest classifier consistently outperformed other models across all evaluation metrics. As shown in Table 1, the Random Forest achieved an overall accuracy of 97.45%, precision of 95.12%, recall of 97.38%, and an F1-score of 96.11%. These results highlight the Random Forest model's strong generalizability and robustness in distinguishing between diabetic and non-diabetic individuals. Given its balanced performance, this model is particularly well-suited for clinical applications where it is crucial to reduce both false positives and false negatives.

Other models such as the Decision Tree classifier also demonstrated solid performance, with an accuracy of 93.96%. However, its slightly lower precision and F1-score, compared to Random Forest, can be attributed to the model’s tendency toward overfitting, especially when hyperparameters are not regularized.

In contrast, the Support Vector Machine model showed fewer promising results, particularly due to a higher number of false negatives. This makes it a less ideal candidate for clinical settings where missing a diabetic case could have serious consequences.

Gaussian Naïve Bayes and Logistic Regression yielded comparable results, both achieving F1-scores above 83%. These models, while simpler and faster to train, may be more suitable for scenarios where interpretability and computational efficiency are prioritized over predictive performance (Table 1).

As illustrated in Figure 3, the confusion matrices for all models reinforce the quantitative results.

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]



(CC BY 4.0)

**Table 1. Comparison the Machine Learning models' performance**

| Model | Accuracy (%)±SD | Precision (%)±SD | Recall (%)±SD | F1-Score (%)±SD |
|---|---|---|---|---|
| Random forest | 97.45 (±0.004) | 95.12 (±0.01) | 97.38 (±0.01) | 96.11 (±0.007) |
| Decision tree | 93.96 (±0.006) | 89.48 (±0.015) | 94.08 (±0.011) | 91.95 (±0.012) |
| Support vector machine | 76.36 (±0.009) | 95.65 (±0.009) | 34.90 (±0.024) | 51.15 (±0.018) |
| Naive Bayes | 88.62 (±0.008) | 84.36 (±0.029) | 83.25 (±0.015) | 83.69 (±0.013) |
| Logistic regression | 89.87 (±0.013) | 87.17 (±0.017) | 83.84 (±0.018) | 85.36 (±0.012) |

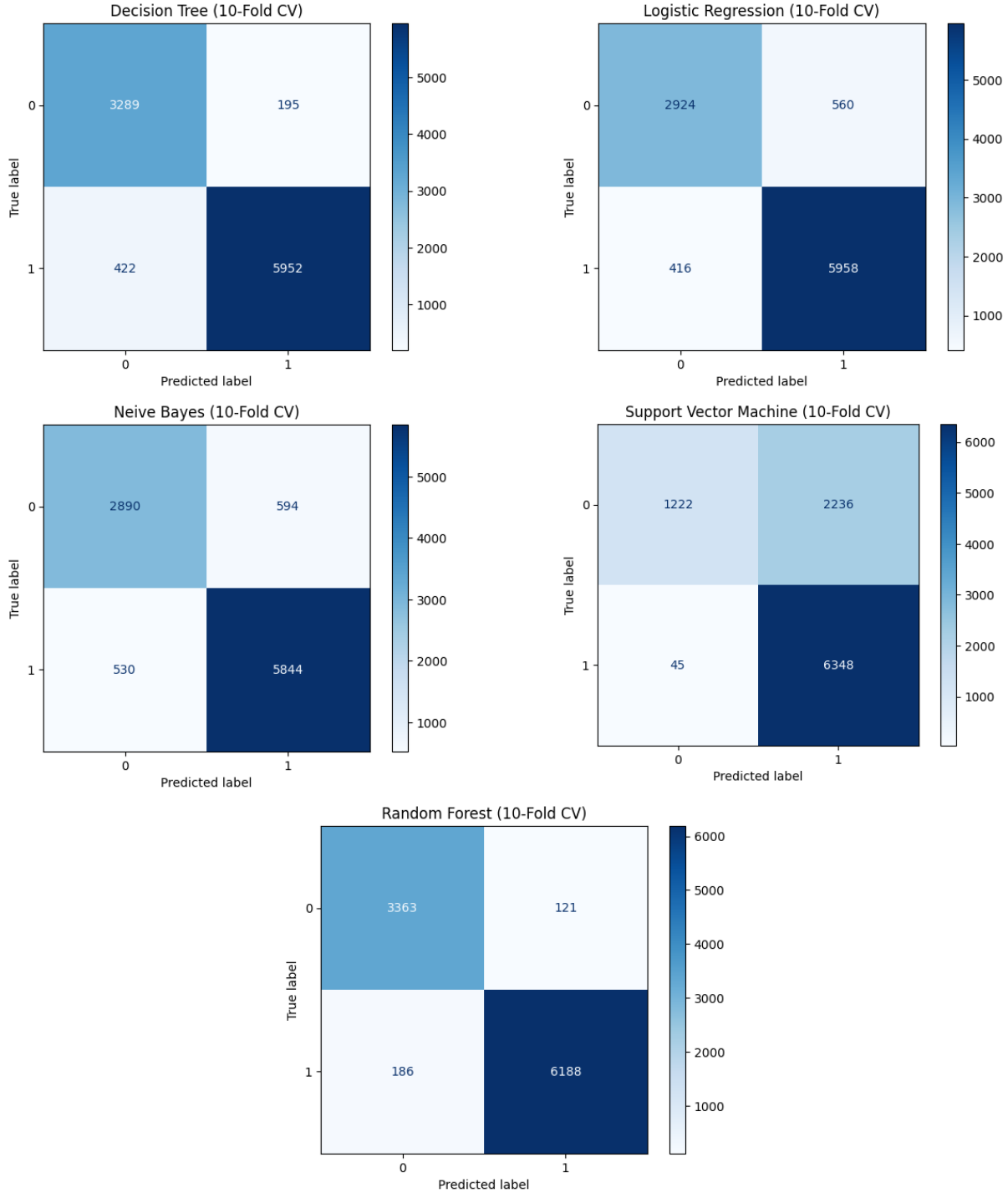


**Figure 3. Confusion matrix of the machine learning models**

The Random Forest classifier achieved a balanced distribution of true positives and true



(CC BY 4.0)

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]

negatives, while SVM showed imbalance due to the relatively high number of false negatives.

Figure 4, presents a comparison of the machine learning models for diabetes risk prediction using the Receiver Operating Characteristic (ROC) curves and their corresponding Area Under the Curve (AUC) values. The ROC curve is a diagnostic tool that visualizes the performance of binary classification models by illustrating the trade-off between the True Positive Rate and the False Positive Rate across varying decision thresholds. The AUC provides a scalar summary of the ROC curve, with values ranging from 0 to 1, where higher AUC values reflect greater model sensitivity and lower false positive rates, indicating a more effective classifier. Among the evaluated models, the Random Forest algorithm attained the highest AUC score, suggesting superior performance in predicting diabetes risk relative to the other machine learning approaches examined.

## Discussion

This research presented and evaluated a machine learning–based framework using data from the Yazd Health Study (YaHS) to predict individuals' risk of developing diabetes. A comprehensive data preprocessing pipeline was applied, including the elimination of irrelevant features, imputation of missing values, statistical feature selection using the Chi-squared test, and dataset balancing through the SMOTENN technique. These steps ensured that the models were trained on a clean, informative, and well-distributed dataset.

Five supervised machine learning algorithms-Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gaussian Naive Bayes-were implemented and their results were compared. Among them, the Random Forest classifier demonstrated superior performance across all evaluation metrics, achieving an overall accuracy of 97.45%, along with the highest precision, recall, and F1-score, indicating its robustness and generalizability for diabetes prediction on YaHS dataset.

The results highlight the effectiveness of ensemble learning methods, particularly Random Forests, when applied to complex and high-dimensional medical datasets.

The proposed framework demonstrates strong potential for early risk prediction and preventive intervention by identifying key

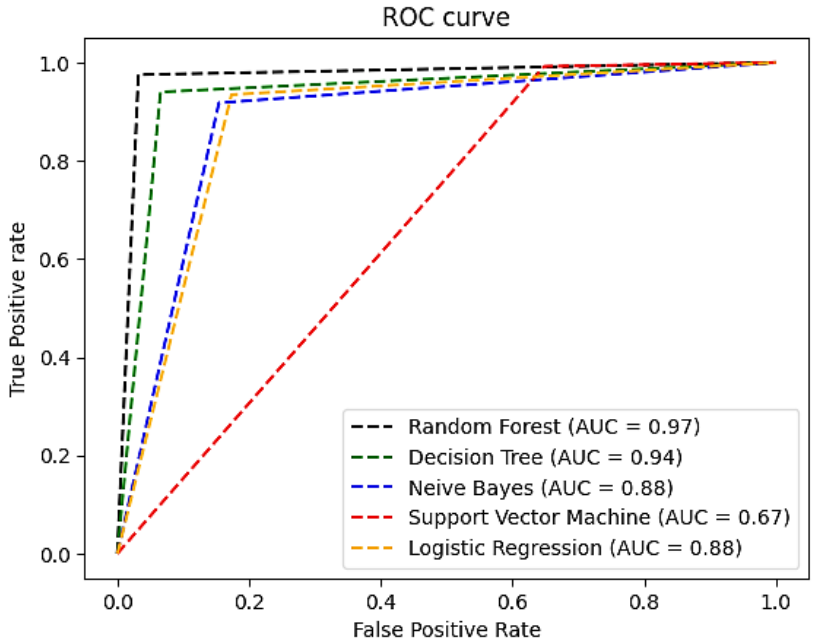


**Figure 4. The comparison of machine learning models using ROC curves**



(CC BY 4.0)

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]

clinical, anthropometric, and biochemical indicators associated with diabetes onset.

## Conclusions

Among the evaluated machine learning algorithms, the Random Forest classifier achieved the highest performance with an accuracy of 97%, outperforming other methods in terms of predictive capability.

## Acknowledgments

We thank Shahid Sadughi University of Medical Science.

## Funding

This publication resulted (in part) from Technological plan supported by the Shahid sadoughi University of Medical Science.

## Conflict of Interest

All authors declare no Potential Conflicts of Interest.

## Authors' contributions

F.S, SM.K, Gh.A, MY.V and M.M laid out the main idea and participated in the design of the study and conducted coordination. N.NGh, M.ST, J.ZA participated in the data collection, and analysis, and drafted the manuscript. All authors read and approved the final manuscript. They agreed to be fully accountable for the integrity and accuracy of the study.

## References

1. World Heath Organization. 14 November 2024; Available from: https://www.who.int/en/news-room/fact-sheets/detail/diabetes.
2. Alkalifah B, Shaheen MT, Alotibi J, Alsubait T, Alhakami H. Evaluation of machine learning-based regression techniques for prediction of diabetes levels fluctuations. Heliyon. 2025;11(1):e41199.
3. Mathioudakis NN, Abusamaan MS, Shakarchi AF, Sokolinsky S, Fayzullin S, McGready J, et al. Development and validation of a machine learning model to predict near-term risk of iatrogenic hypoglycemia in hospitalized patients. JAMA Network Open. 2021;4(1):e2030913.
4. Pei D, Yang T, Zhang C. Estimation of diabetes in a high-risk adult Chinese population using J48 decision tree model. Diabetes, Metabolic Syndrome and Obesity. 2020:4621-30.
5. Kumar N, Singh P, Kumari S, Singh BK. Predicting Diabetes Using Machine Learning. In2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) 2023 (pp. 1737-1742).
6. Mirzaei M, Salehi-Abargouei A, Mirzaei M, Mohsenpour MA. Cohort Profile: The Yazd Health Study (YaHS): a population-based study of adults aged 20–70 years (study design and baseline population data). International journal of epidemiology. 2018;47(3):697-8h.
7. Ahmadabadi JZ, Mehrjardi FZ, Ghanbary M, Mirzaei M. Identification of Effective Factors and Prediction of Ischemic Heart Disease Using Machine Learning Methods and Data from the Yazd Health Study (YaHS). Journal of Shahid Sadoughi University of Medical Sciences. 2024; 32(7): 8067-79.(in Persian)
8. Khosravi M, Azizi R, Fallahzadeh H, Mirzaei M. Prevalence, Incidence, and Risk Factors of Hypothyroidism in Adult Residents of Yazd Greater Area, 2015–2021: Results of Yazd Health Study. Iranian Journal of Medical Sciences. 2024;49(10):623.
9. Darand M, Golpour-Hamedani S, Karimi E, Hassanizadeh S, Mirzaei M, Arabi V, et al. The association between adherence to unhealthy plant-based diet and risk of COVID-19: a cross-sectional study. BMC Infectious Diseases. 2024;24(1):1-8.
10. Han J, Kamber M, Pei J. Data mining: Concepts and. Techniques, Waltham: Morgan Kaufmann Publishers. 2012.
11. Muntasir Nishat M, Faisal F, Jahan Ratul I, Al-Monsur A, Ar-Rafi AM, Nasrullah SM, et al. A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers With SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. Scientific Programming. 2022;2022(1):3649406.
12. Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access. 2020;8:76516-31.
13. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]



(CC BY 4.0)

logistic regression: an overview. Journal of thoracic disease. 2019;11(Suppl 4):S574.
14. Ying LU. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 2015;27(2):130.
15. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. BMJ Nutrition, Prevention & Health. 2021;4(1):140.
16. Khokhar PB, Gravino C, Palomba F. Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artificial intelligence in medicine. 2025:103132.
17. Edeh MO, Khalaf OI, Tavera CA, Tayeb S, Ghouali S, Abdulsahib GM, et al. A classification algorithm-based hybrid diabetes prediction model. Frontiers in Public Health. 2022;10:829519.

[ Downloaded from ijdo.ssu.ac.ir on 2026-02-09 ]

[ DOI: 10.18502/ijdo.v17i3.19267 ]



(CC BY 4.0)