

Predicting Type 2 Diabetes Using Random Forest and XGBoost Algorithms: A Comparative Machine Learning Approach

İlkim Ecem Emre¹

¹Marmara University, Faculty of Business Administration, Department of Management Information Systems, İstanbul, Türkiye

ABSTRACT

Purpose: This study aims to compare the performance of random forest (RF) and XGBoost (XGB) ensemble learning algorithms for predicting Type 2 diabetes.

Methods: The widely known and used PIMA Indians Diabetes dataset was utilized for model development. After data pre-processing, 5-fold cross-validation repeated five times was applied. Model performances were evaluated using accuracy, precision, sensitivity, specificity, F1-score, and AUC metrics.

Results: The RF model achieved 0.817 accuracy and an AUC of 0.874, while the XGB model yielded 0.791 accuracy and an AUC of 0.874. In both models, glucose was identified as the most significant feature for predicting diabetes.

Conclusion: The results showed that RF and XGB models demonstrated comparable discriminative performance under a reproducible analytical framework, with no statistically significant difference in AUC.

Keywords: Diabetes Mellitus, Type 2; Machine Learning; Classification; Ensemble Learning; Random Forest; XGBoost

ÖZET

Amaç: Bu çalışma, Tip 2 diyabetin tahmini amacıyla rastgele orman (RF) ve XGBoost (XGB) topluluk öğrenmesi algoritmalarının performanslarını karşılaştırmayı amaçlamaktadır.

Metot: Model geliştirme sürecinde, yaygın olarak bilinen ve kullanılan PIMA Indians Diyabet veri kümesi kullanılmıştır. Veri ön işleme adımlarının ardından, beş kez tekrarlanan 5 katlı çapraz doğrulama yöntemi uygulanmıştır. Model performansları; doğruluk, kesinlik, duyarlılık, özgüllük, F1 skoru ve AUC ölçütleri kullanılarak değerlendirilmiştir.

Bulgular: RF modeli 0.817 doğruluk ve 0.874 AUC değerine ulaşırken, XGB modeli 0.791 doğruluk ve 0.874 AUC değeri elde etmiştir. Her iki modelde de diyabet tahmininde en önemli öznitelik olarak glikoz belirlenmiştir.

Sonuç: Elde edilen sonuçlar, RF ve XGB modellerinin tekrarlanabilir bir analitik çerçevede karşılaştırılabilir ayırt edici performans sergilediğini ve AUC'de istatistiksel olarak anlamlı bir fark olmadığını göstermiştir.

Anahtar kelimeler: Tip 2 Diyabet; Makine Öğrenmesi; Sınıflandırma; Topluluk Öğrenmesi; Rastgele Orman; XGBoost

İlkim Ecem EMRE
0000-0001-9507-8967

Correspondence: İlkim Ecem Emre
Marmara University, Faculty of Business Administration, Department of Management Information Systems, İstanbul, Türkiye
Phone: +90 216 777 25 76
E-mail: ecem.emre@marmara.edu.tr

Received: 05.08.2025

Accepted: 23.03.2026

Diabetes is one of the most prevalent, serious and leading chronic diseases worldwide. It affects millions of people and imposes a serious burden on patients. Type 2 diabetes is a serious chronic disease that occurs when pancreas fails to produce enough insulin or when the body becomes unable to produce insulin (1,2). Since the insulin is the hormone responsible for regulating blood glucose, when this fails, diabetes may lead to complications such as blindness, kidney failure, heart attacks, stroke and lower limb amputation (3).

Diabetes is a life-threatening chronic disease with a growing global prevalence, necessitating early diagnosis and treatment to prevent severe complications (4). There are different types of diabetes. This study focuses on Type 2 diabetes. Type 2 diabetes is a disorder that disrupts the body's ability to effectively use insulin, which results in elevated blood glucose levels if left untreated (3).

Type 2 diabetes is often considered preventable so that early diagnosis could be the key to prevent the effects of this type (3). Therefore, machine learning (ML) methods can support this process. Early detection and intervention are crucial in managing this disease and preventing its complications. Some studies in the literature have shown that machine learning (ML) techniques can support diabetes prediction and clinical decision making. In this study, ML models are applied to PIMA Indians Diabetes dataset and the performance of each model was compared.

In the literature there are many studies (5) which have focused on the same dataset in order to predict diabetes based on ML models. Some of these studies are reviewed in order to be able to compare and contrast the results obtained from this study. Accuracy (acc), area under the curve (auc) and F1-scores of the reviewed studies were examined and the comparison between studies are examined in the conclusion section. (6) used the same dataset to build prediction model using various models including random forest model (98.7 acc, 98.4 auc, 99.4 F1-score). This study handled the target attribute with three classes as "Normal", "Prediabetes", and "Diabetes", which targets a multi-class classification task. Values of zeros are imputed using mean. Glucose, Blood_Pressure, Insulin, Body_Mass_Index, and Age had zero values in this study. These values were imputed using mean values. Following a min-max scaling method for the normalization of the data. The dataset was divided into three subsets as training (80%), validation (10%), and test (10%). To prevent

the data leakage nested 5-fold cross validation technique was used.

(7) used Long Short-Term Memory (0.85 acc, 0.89 AUC), random forest (0.78 acc, 0.81 AUC) and convolutional neural network (0.82 acc, 0.86 AUC) algorithms to build prediction models. Long Short-Term Memory provided the highest accuracy score and AUC values. The dataset in the study was divided as 80% for training, 20% for testing. Standardization method was used for the normalization. For the data imputation, it is not clear which features are included.

(8) used a hybrid approach by combining supervised and unsupervised machine learning techniques. The results showed that the proposed framework achieved random forest (88.5 acc, 0.874 AUC, 0.836 F1-score) extreme gradient boosting (88.5 acc, 0.873 AUC, 0.835 F1-score). Zeros within the numerical attributes were replaced using median values, also outliers were replaced with the medians as well. Mutual information technique was used to identify feature selection. 13 different classifiers were used including random forest and extreme gradient boosting. Insulin, Glucose, Skin_Thickness, Body_Mass_Index, Age, Diabetes_Pedigree_Function was selected as the most influential predictors respectively, whereas number of Pregnancies and Blood_Pressure were eliminated.

(9) developed models with random forest (0.81 acc, 0.785 AUC, 0.720 F1-score) extreme gradient boosting (0.78 acc), logistic regression (0.76 acc). Glucose, Blood_Pressure, Skin_Thickness, Insulin, Body_Mass_Index zero values are imputed using mean values. The dataset in the study was divided as 80% for training, 20% for testing. Along with the Synthetic Minority Oversampling Technique (SMOTE) data balancing technique, standardization method was used for the normalization of the numerical features. Through cross validation techniques hyperparameter optimization was conducted. Glucose, Body_Mass_Index and Age are found as the top predictors of according to random forest algorithm.

(10) developed prediction models using J48 decision tree (0.747 acc, 0.785 AUC, 0.786 F1-score), random forest (0.795 acc, 0.862 AUC, 0.851 F1-score), naïve Bayes (0.786 acc, 0.846 AUC, 0.842 F1-score) algorithms. In order to split data, they have used 70% for training %30 for testing and no scaling method was used. For the feature selection PCA, k-means clustering, and importance ranking were

employed. Zero values in the features Glucose, Blood_Pressure, Skin_Thickness, Insulin, Body_Mass_Index were imputed using median values. They reported that while Glucose, Body_Mass_Index, Age, Insulin and Skin_Thickness is very highly important for the model, Diabetes_Pedigree_Function, Blood_Pressure, and number of Pregnancies rank very low in terms of variable importance.

(11) used random forest (0.79 acc, 0.84 F1-score), decision trees (0.74 acc, 0.80 F1-score), XGBoost (0.74 acc, 0.62 F1-score), support vector machines (0.69 acc, 0.62 F1-score) and k-nearest neighbour (0.70 acc, 0.63 F1-score) algorithms to develop models. Based on the accuracies, random forest outperformed the other models. For the zeros in the attributes mean imputation was conducted. In addition, SMOTE method was used to balance the data.

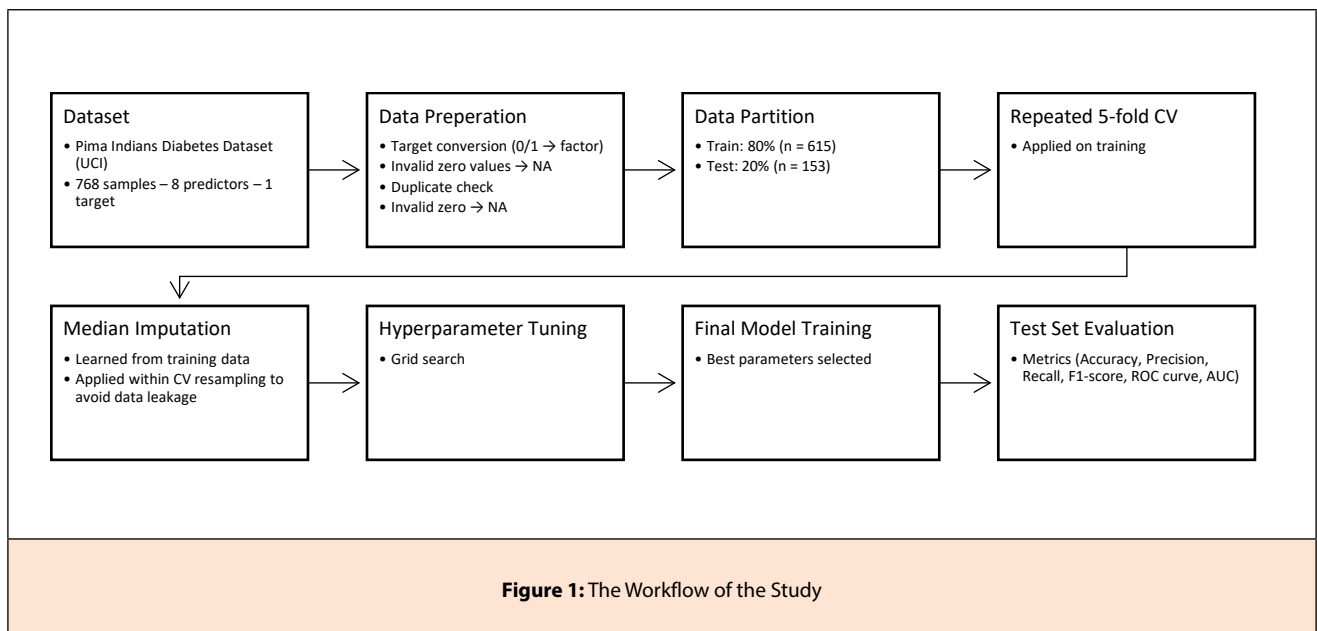
(12) used seven different ML models including decision tree, k-nearest neighbour, random forest, naïve Bayes, adaboost, logistic regression, support vector. The dataset in the study was divided as 85% for training, 15% for testing. Besides 10-fold cross-validation technique was also used to split the dataset. They have conducted models separately for both splitting methods. Standardization method was use for the normalization.

(13) developed model with artificial neural network (0.903 acc, 0.859 F1-score), naive bayes (0.763 acc, 0.616 F1-score), decision tree (0.966 acc, 0.947 F1-score) and deep learning (0.98 acc, 0.968 F1-score) algorithms. Deep

learning outperforms the other models with a major difference and significantly high accuracy.

Overall, previous studies on the PIMA Indians Diabetes dataset demonstrate that machine learning models frequently achieve strong predictive performance. However, there are considerable variations in pre-processing strategies, including different imputation techniques, normalization methods, and data splitting or validation frameworks. In particular, variations in handling zero values, lack of transparent pre-processing pipelines, and differences in validation strategies make direct comparisons between studies difficult. Therefore, this study aims to provide a transparent and reproducible analytical methodology.

In this regard, the purpose of the study is to classify instances as diabetic and non-diabetic based on their various features. Similar to previous studies (14), the underlying algorithm is used to predict the unlabeled data, whether individuals are diabetic or non-diabetic. The workflow of the study is given in Figure 1. Although numerous studies have applied ML models to this dataset, methodological variations exist regarding pre-processing of data (6), validation strategies, and reporting of performance metrics. This study aims to provide a transparent and reproducible workflow addressing these methodological issues while comparing two widely known algorithms.



Material and Methods

Dataset

In the study, widely known Pima Indians Diabetes Dataset (15) is used in order to create models for diabetes classification. This dataset is originally from UCI Machine Learning Repository and used as a common dataset for diabetes prediction. It consists of 768 rows, 9 columns. Instances are women, who are with type 2 diabetes and

without diabetes. There are several medical indicators (independent variables/predictive variables) used for the model development and one target (dependent/target variable) variable indicating each patient's diabetes situation. Independent variables include the number of Pregnancies, Glucose, Blood_Pressure, Skin_Thickness, Insulin, Body_Mass_Index, Diabetes_Pedigree_Function, Age. All predictive variables are numerical variables whereas the target variable is a categorical one. In this dataset, there are 268 diabetic, 500 non-diabetic instances. The variables in the dataset are shown in Table 1.

Table 1. Variables in the Used Dataset

Variable	Explanation	Data Type
Pregnancies	Number of times pregnant	Numerical
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test	Numerical
Blood_Pressure	Diastolic blood pressure (mm Hg)	Numerical
Skin_Thickness	Triceps skin fold thickness (mm)	Numerical
Insulin	2-Hour serum insulin (mu U/ml)	Numerical
Body_Mass_Index	Body mass index (weight in kg/(height in m) ²)	Numerical
Diabetes_Pedigree_Function	A function that scores likelihood of diabetes based on family history	Numerical
Age	Age in years	Numerical
Target Variable	Class variable (0 or 1), indicating whether the patient has Diabetes Mellitus Type 2	Categorical

Table 2. Summary of the Features in the Raw Dataset*.

Variable	Min	Max	Median	Mean
Pregnancies	0.00	17.00	3.00	3.845
Glucose	0.00	199.00	117.00	120.90
Blood_Pressure	0.00	122.00	72.00	69.11
Skin_Thickness	0.00	99.00	23.00	20.54
Insulin	0.00	846.00	30.50	79.80
Body_Mass_Index	0.00	67.10	32.00	31.99
Diabetes_Pedigree_Function	0.0780	24.200	0.3725	0.4719
Age	21.00	81.00	29.00	33.24

* These statistics refer to the raw dataset. The zero values were treated as NA during pre-processing.

Data Preparation

Data types of all predictive variables were numerical, whereas target variable is a categorical variable. The target variable is converted to factor and labelled as "Diabetic" for diabetes indication, "Non-diabetic" for non-diabetes indication instead of 1 and 0 respectively. No duplicated values were found.

During pre-processing missing values were identified. Several variables (Glucose, Blood_Pressure,

Skin_Thickness, Insulin, Body_Mass_Index) contained zero values which are physiological implausible for living patients. Therefore, these values were treated as missing values and recorded as NA. Median imputation was applied within the cross-validation pipeline. The imputation was conducted using the training dataset and applied within the resampling process to avoid data leakage. For the final model, median values were estimated from the training dataset and automatically applied to the test dataset through the same pre-processing pipeline during prediction phase. According to clinical reference

ranges, physiological measurements such as blood glucose (16) or blood pressure (17) cannot take a value of zero, as such values are considered incompatible with life, physiologically implausible and infeasible (10,18,19). Therefore, zero values within Glucose, Blood_Pressure, Skin_Thickness, Insulin, Body_Mass_Index columns were treated as missing values.

To ensure reproducibility, a fixed random seed (set.seed(8)) was used during data partitioning, hyperparameter tuning, and model training.

Model Performance Evaluation Method

The dataset is partitioned into a training dataset and a test dataset. Based on hold-out method (20), 80% of the dataset was used for training, 20% was used to evaluate model performance using stratified sampling. So, 615 samples were used for training, 153 samples for testing while preserving the proportion of diabetes and non-diabetes cases across partitions.

Hyperparameter Tuning

The 5-fold cross validation (CV) method (21) was conducted for 5 times repeatedly on the training dataset. Hyperparameter tuning was conducted using grid search within CV. The hyperparameter search is intentionally limited to ensure computational efficiency and to avoid excessive model complexity. For the random forest model, the mtry parameter was tuned with the values of 2, 3, 4, and 5. For XGBoost model, nrounds (100, 200), max_depth (3, 6), and eta (0.05, 0.1) were explored. Based on CV results, the best performing parameters were selected as mtry = 2 for random forest, nrounds = 100, max_depth = 3, and eta = 0.05 for XGBoost based on the mean ROC-AUC which were obtained across all folds of repeated CV procedure (Table 3). These optimal parameters were used to train the final models on the training dataset, and the predictive performance of the models was evaluated on the test dataset (Table 4).

Table 3. Model Performances with The Training Dataset

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score	NPV	AUC
RF	0.745	0.657	0.568	0.841	0.609	0.784	0.808
XGB	0.745	0.660	0.563	0.844	0.607	0.782	0.815

Table 4. Model Performances with The Test Dataset

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score	NPV	AUC
RF	0.817	0.766	0.679	0.89	0.720	0.840	0.874
XGB	0.791	0.744	0.604	0.89	0.677	0.809	0.874

Machine Learning Models

Two ensemble learning algorithms, random forest (RF) and extreme gradient boosting / XGBoost (XGB) from ensemble models were chosen for the model development. Due to their improved prediction performance and robustness, ensemble models are chosen for the analyses (22). All models developed using R programming language (version 4.5.2) and RStudio (version 2026.01.1). Used packages were readxl (23) (version 1.4.5), caret (24) (version 7.0-1), randomForest (25) (version 4.7-1.2), xgboost (26) (version 1.7.11.1), pROC (27) (version 1.19.0.1), recipes (28) (version 1.3.1).

After hyperparameter tuning, the optimal parameter values were used to train the final models and evaluate their predictive performance on the independent test dataset.

Model Performance Evaluation Metrics

For the evaluation of the model performances, metrics obtained from confusion matrix were used. All metrics were obtained on the independent test dataset. "Diabetics" class was chosen as positive class. All metric calculations are made accordingly. Accuracy, precision, recall, F1-score,

AUC metrics are given in the findings section. Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values are also presented. Discriminative ability of a binary event is quantified by AUC in ROC curve (29). AUC serve as a metric for binary diagnostic classification in order to discriminate between normal and abnormal conditions (30). In the study, classification between diabetic and non-diabetic instances is tried to be represented by ML models, therefore AUC is evaluated as one of the performance metrics. In addition, to statistically compare the discriminative performances of the models, DeLong's test (31) for two ROC curves was conducted to evaluate whether the difference between the AUC values of the RF and XGB models was statistically significant. F1-score is also evaluated as one of the metrics, which is preferable in datasets with imbalanced target attribute distribution (32). Based on the final models, variable importance ranking for each model is examined. Feature importance for RF was computed using the mean decrease Gini index metric in the randomForest library. For the XGB model, feature importance was calculated based on the Gain metric. The importance values were obtained using varImp() function of the caret library (24). Mean Decrease Gini, measures impurity reduction, while Gain reflects the improvement in model performance achieved by each split.

Results

Firstly, the descriptive statistics of the dataset are given in Table 2. In the study, ML models were compared for predicting Diabetes Mellitus, Type 2 using Pima Indians Diabetes dataset. The models were evaluated through 5-fold CV which is repeated for 5 times, which was employed both for hyperparameter tuning and internal performance evaluation. Two ensemble learning algorithms (RF, XGB) were selected for model development. During the training process, different hyperparameters were evaluated within the cross-validation framework, and the best performing parameters were selected on the ROC-AUC metric. The performance metrics obtained from the training dataset, are reported in Table 3 represent the mean CV results from each fold obtained with the optimal hyperparameter values. The RF model achieved an accuracy of 0.745, precision of 0.657, sensitivity of 0.568, specificity of 0.841, F1-score of 0.609, and AUC of 0.808. Similarly, the XGB model achieved an accuracy of

0.745, precision of 0.660, sensitivity of 0.563, specificity of 0.844, F1-score of 0.607, and AUC of 0.815. The obtained results on the training dataset, indicate that both models demonstrate comparable discriminative performance, with nearly identical metric values.

Six metrics were evaluated and compared with each other in order to evaluate the final model performance on the test dataset (Table 4). RF model achieved an accuracy of 0.817, precision of 0.766, sensitivity of 0.679, specificity of 0.89 and F1-score of 0.72. The area under the ROC curve (AUC=0.874) also indicated strong discriminative ability between diabetic and non-diabetic instances in the dataset. XGB also performed well with reasonable results. XGB model achieved an accuracy of 0.791, precision of 0.744, sensitivity of 0.604, specificity of 0.89 and F1-score of 0.677. The area under the ROC curve (AUC=0.874) of the XGB model is achieved the same AUC as the RF model. Among these two models, RF model achieved slightly higher values in terms in some classification metrics. To statistically compare the ROC performances of the models, DeLong's test was applied. The test indicated that the differences between the AUC values of both models (AUC=0.87) was not statistically significant ($Z = 0.04$, $p\text{-value} = 0.96$), suggesting that both models demonstrated comparable discriminative performance. In addition, the model performances on training dataset were consistent with the performances obtained on the independent test dataset, indicating that the models generalize well and do not indicate substantial overfitting (33,34).

Figure 2 illustrates the ROC curve comparison of RF and XGB models for prediction of Type 2 Diabetes. The ROC curve demonstrates each model's ability to distinguish between positive and negative classes. As seen in the figure, both models perform well above the diagonal line that represents random classification. However, RF (represented with blue curve) achieved an AUC value of 0.874. In addition, the XGB model (represented with red curve) also has achieved the same AUC value of 0.874. Since AUC value, between 0.8 - 0.9 considered as good (30), the discrimination between two groups can be regarded as satisfactory and clinically considerable.

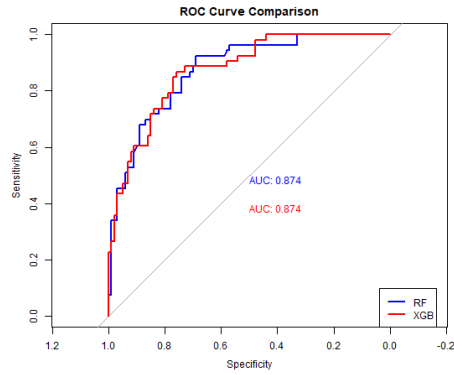


Figure 2: ROC Curve Comparison of the Models

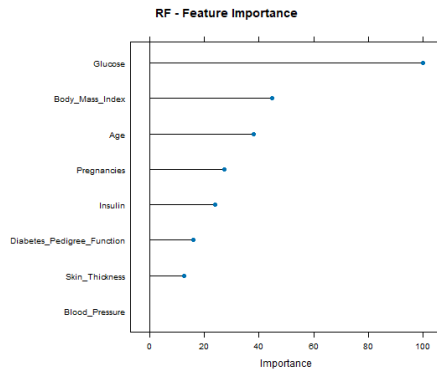


Figure 3: Feature Importance Plot of RF Model

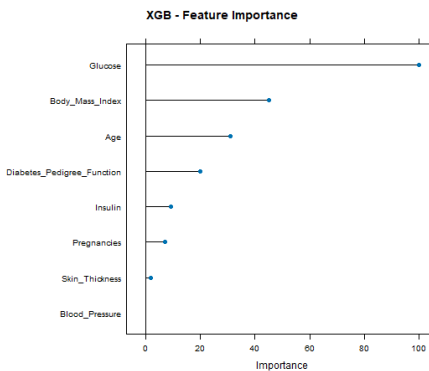


Figure 4. Feature Importance Plot of XGB Model

RF model achieved slightly higher values in some classification metrics (accuracy, sensitivity), whereas both models demonstrated comparable discriminative performance in terms of AUC, as also supported by

DeLong's test. To summarize, it can be said that, the ROC analysis confirms that the RF and XGB models provide a satisfactory overall classification performance.

Table 5. Variable Importance of the Models

RF - Feature Importance	Overall	XGB - Feature Importance	Overall
Glucose	100	Glucose	100.00
Body_Mass_Index	44.80	Body_Mass_Index	44.98
Age	38.09	Age	31.07
Pregnancies	27.32	Diabetes_Pedigree_Function	19.88
Insulin	24.01	Insulin	9.43
Diabetes_Pedigree_Function	16.12	Pregnancies	7.01
Skin_Thickness	12.64	Skin_Thickness	1.82
Blood_Pressure	0.00	Blood_Pressure	0.00

Table 5 shows the variable importance of each model in an order. Both models are consistent at highlighting "Glucose" variable as the most influential predictor in predicting Type 2 diabetes risk. Visualized version of the feature importance is presented in figures as well. Figure 3 and Figure 4 present the feature importance rankings obtained from the RF and XGB models, respectively.

In RF model (Figure 3), the most influential predictor is Glucose (100). Body_Mass_Index (44.80) and Age (38.09) are the second and third most important predictive features, followed by Pregnancies (27.32), Insulin (24.01), Diabetes_Pedigree_Function (16.12), and Skin_Thickness (12.64) show moderate contributions to the model. In addition, Blood_Pressure (0) shows negligible predictive value in this model.

In XGB model (Figure 4), a similar pattern of feature importance is observed. Glucose remains the influential predictor with an importance score of 100, which is followed by Body_Mass_Index (44.98), Age (31.07), and Diabetes_Pedigree_Function (19.88). Interestingly, Insulin (9.43) and Pregnancies (7.01) exhibit lower importance roles in XGB compared to RF model. Skin_Thickness (1.82) shows minimal contribution. In addition, it is holding less importance than it is in RF model. However, Blood_Pressure remains the least influential (0) predictive variable in this model as it is in RF model.

These results reinforce the relevance of metabolic and hereditary factors in diabetes prediction and provide insight into which clinical variables may be prioritized in decision support systems. Furthermore, the consistency in variable rankings between the two models increases the reliability of the findings.

Discussion

In this study, two different ensemble learning models with a very well-known dataset were developed. Although this dataset is widely used, it provides a good example for classification models and therefore preferred by the researchers. The findings of the study indicate that ML models can effectively predict diabetes using this dataset. RF showed slightly better performance in some metrics compared to XGB in prediction of diabetics and non-diabetic instances in the dataset. Both models demonstrated comparable discriminative performance in terms of AUC. Feature importance analyses and ROC visualizations provide insights into model behavior and support both models' interpretability. In this regard, the consistency in variable rankings between these two models increases the reliability of the findings. In addition, the model performances on training dataset were consistent with the performances obtained on the independent test dataset, indicating that the models generalize well and do not indicate substantial overfitting (33,34).

When compared with recent studies published after 2020, RF model developed in this study (0.817 acc, 0.874 AUC) outperformed similar models reported by (7,9–12), both in terms of classification accuracy and discrimination power represented with AUC. (6) outperformed all of the studies reviewed in the scope of this study with a significantly higher accuracy. The obtained results from RF model shows a strong performance for predicting diabetes. (13) also showed significantly higher accuracy, however, they have not used RF instead this result was obtained using deep learning. In health-related predictive modeling, it is important to distinguish between healthy samples and unhealthy samples. Given the critical nature of medical applications, where minimizing false negatives

is often vital, RF appears to provide slightly more stable performance in this study.

XGB model developed in this study (0.791 acc, 0.874 AUC) also demonstrates a competitive performance when compared to recent studies. While (8) outperformed these findings, (11) reported a lower accuracy and F1-score and (9) slightly lower performance with accuracy value. These findings highlight that the proposed XGB model performs robustly in both classification performance and discriminatory capacity, while demonstrates comparable performance in terms of AUC which is a critical indicator in imbalanced medical datasets. Although the RF model showed slightly higher performance values than XGB in some metrics, the statistical comparison indicated that the difference between the AUC values was not statistically significant. Therefore, both algorithms can be considered to have comparable predictive performance for Type 2 diabetes prediction in this dataset. The results in this study suggest that both RF and XGB models are indicative of good discriminative performance for Type 2 Diabetes prediction within the dataset.

In terms of pre-processing methods or modeling framework there are similarities and differences with the other studies in the literature. Used hyperparameter framework in (5,9) is similar to this study. In different studies researchers followed various steps in order to pre-process the dataset. In some studies normalization (5,9,12) and data balancing (9,11) techniques were applied in contrast with this study. In terms of data imputation, while some researchers preferred mean imputation (5,9,11), some (8,10) preferred median imputation similarly with this study. For data splitting, (7) used only hold-out method in contrast while (12) used only cross-validation technique.

The obtained results suggest that both RF and XGB models prioritize Glucose as the dominant factor in Type 2 diabetes prediction, which aligns with clinical expectations given the direct association between glucose levels and diabetes diagnosis (35). This finding is also supported by other reviewed studies (6,9,10) in the literature. In addition, Type 2 diabetes impairs the body's ability to manage glucose levels, however, as type 2 diabetes is largely preventable, early detection could be the most effective approach to mitigating the progression and complications of the disease (3). The finding that blood pressure has the lowest importance is also noteworthy which is consistent with other studies (8,10).

Conclusion

In light of the results, ML models could be helpful for clinicians to detect the risky groups or prevent diabetes before it occurs or in its early phases. Despite the good predictive performance of the models, several limitations should be acknowledged. Firstly, since this dataset is about a very specific group, different datasets with more diversity could be used in model development. This may help to enhance the potential generalizability of the findings to more diverse populations. Secondly, hyperparameter search was also limited. A more comprehensive tuning strategy could further improve model performance. Additionally, alternative imputation methods and feature engineering techniques, may further improve model performance. Moreover, the study relied on a single publicly available dataset, and external validation using independent clinical datasets was not conducted. Various datasets like "Early-Stage Diabetes Risk Prediction Dataset (<https://doi.org/10.24432/C5VG8H>)" or "Diabetes Health Indicators Dataset (<https://doi.org/10.24432/C53919>)" which have more instances and features could be also used for diabetes prediction and model development. Finally, although the evaluated variables represent common clinical indicators, additional factors such as lifestyle, genetic markers, and longitudinal health records may further improve predictive performance.

Future work may involve advanced models and real-world clinical data. The findings suggest that, following appropriate missing handling and a reproducible modeling framework, RF and XGB provide comparable discriminative performance for Type 2 diabetes prediction. While the results demonstrate the potential of these models, further validation using diverse and real-world clinical datasets is required before practical implementation in decision-support systems.

Declarations

Funding

Not applicable

Conflicts of interest/Competing interests

Not applicable

Ethics approval

Ethical approval was not required for this study because the study used publicly available secondary data.

Availability of data and material (data transparency)

The dataset can be accessed publicly through the link: <https://data.mendeley.com/datasets/7zcc8v6hvp/1>

Authors' contributions

The author (IEE) fulfilled all authorship criteria, including conception, design, data acquisition, analysis, and interpretation of the study; drafting and critical revision of the manuscript; final approval of the version to be published; and full accountability for the integrity and accuracy of the work.

References

1. Mayo Clinic Staff. Diabetes - Symptoms and causes - Mayo Clinic. Mayo Clinic [Internet]. 2020 [cited 2025 Apr 25]; Available from: <https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444>
2. American Diabetes Association (ADA). Understanding Type 2 Diabetes [Internet]. 2025 [cited 2025 Apr 25]. Available from: <https://diabetes.org/about-diabetes/type-2>
3. World Health Organization (WHO). Diabetes [Internet]. 2024 [cited 2025 Apr 25]. Available from: <https://www.who.int/news-room/fact-sheets/detail/diabetes>
4. Abnoosian K, Farnoosh R, Behzadi MH. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC Bioinformatics*. 2023 Dec 1;24(1):1–24.
5. Sharma T, Shah M. A comprehensive review of machine learning techniques on diabetes detection. *Visual Computing for Industry, Biomedicine, and Art*. 2021 Dec 1;4(1):30.
6. Zaferani N, Afrash MR, Moulaei K. Predicting and classifying type 2 diabetes using a transparent ensemble model combining random forest, k-nearest neighbor, and neural networks. *Scientific Reports*. 2026 Dec 19;16(1):1892–.
7. Jayakumar A, Saji AK, Tom P, Thomas J. A Detailed Study on Diabetes Detection using The PIMA Indian Diabetes Database. *International Research Journal of Modernization in Engineering*. 2025;7(3):10353–8.
8. Abu-Shareha AA, Mosleh Abualhaj, Abdelrahman H. Hussein, Amal Amer, Anusha Achuthan, Alfian Abdul Halin. Diabetes Prediction Using Hybrid Supervised and Unsupervised Techniques Based on PIMA Dataset. *Journal of Artificial Intelligence and Technology*. 2025 Nov 23;6:79–87.
9. Nassiwa F, Zeng J. Evaluating Traditional Machine Learning Models for Predicting Diabetes Onset Using the Pima Indians Dataset. *Annals of Medical and Health Sciences Research*. 2024;14(7):1010–5.
10. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*. 2023 Aug 1;35(22):16157–73.
11. Andrabi SAB, Singh I. A Comparative Study of Machine Learning Techniques for Diabetes Prediction. In: 4th International Conference on Inventive Research in Computing Applications, ICIRCA 2022 - Proceedings. Institute of Electrical and Electronics Engineers Inc.; 2022. p. 741–5.
12. Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. *ICT Express*. 2021 Dec 1;7(4):432–9.
13. Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes and Metabolic Disorders*. 2020 Jun 1;19(1):391–403.
14. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*. 2019 Dec 21;19(1):1–16.
15. Joseph LP, Joseph EA, Prasad R. Diabetes Datasets [Internet]. Mendeley Data; 2022 [cited 2026 Mar 7]. Available from: <https://data.mendeley.com/datasets/7zcc8v6hvp/1>
16. American Heart Association. Life's Essential 8™ - How to Manage Blood Sugar Fact Sheet [Internet]. 2025 [cited 2026 Mar 22]. Available from: <https://www.heart.org/en/healthy-living/healthy-lifestyle/lifes-essential-8/how-to-manage-blood-sugar-fact-sheet>
17. American Heart Association. Understanding Blood Pressure Readings [Internet]. Aha. 2017 [cited 2026 Mar 22]. p. 01. Available from: <https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings>
18. Thakur D, Gera T, Bhardwaj V, AlZubi AA, Ali F, Singh J. An enhanced diabetes prediction amidst COVID-19 using ensemble models. *Frontiers in Public Health*. 2023;11:1331517.
19. Houssein EH, Ibrahim IA, Mostafa A, Albarrak AM, Younan M. SMENN-hybrid: an efficient technique combining the synthetic minority oversampling technique with ensemble learning for diabetes prediction. *Scientific Reports*. 2025 Dec 3;15(1).
20. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: *International Joint Conference on Artificial Intelligence*. 1995. p. 1137–45.
21. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society Series B (Methodological)*. 1974 Jun 4;36(2):111–47.
22. Miénye ID, Sun Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. Vol. 10, *IEEE Access*. Institute of Electrical and Electronics Engineers Inc.; 2022. p. 99129–49.
23. Wickham H, Bryan J. readxl: Read Excel Files [Internet]. 2025. Available from: <https://cran.r-project.org/package=readxl>
24. Kuhn, Max. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* [Internet]. 2008;28(5):1–26. Available from: <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>
25. Liaw A, Wiener M. Classification and Regression by randomForest. *R News* [Internet]. 2002;2(3):18–22. Available from: <http://cran.r-project.org/doc/Rnews/>
26. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. xgboost: Extreme Gradient Boosting [Internet]. 2025. Available from: <https://cran.r-project.org/package=xgboost>
27. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
28. Kuhn M, Wickham H, Hvitfeldt E. recipes: Preprocessing and Feature Engineering Steps for Modeling [Internet]. 2025. Available from: <https://cran.r-project.org/package=recipes>
29. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*. 2022 Dec 1;4(12):e853–5.

30. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology* [Internet]. 2022 Feb 1 [cited 2025 Apr 24];75(1):25. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8831439/>
31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*. 1988 Sep;44(3):837.
32. Christen P, Hand DJ, Kirielle N. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Computing Surveys*. 2023 Mar 31;56(3).
33. Bro R. How to overfit. *Chemometrics and Intelligent Laboratory Systems*. 2025 Sep 15;264:105461.
34. López OAM, López AM, Crossa DJ. Overfitting, Model Tuning, and Evaluation of Prediction Performance. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. 2022 Jan 14;109–39.
35. American Diabetes Association Professional Practice Committee. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2024. *Diabetes Care* [Internet]. 2023;47(Supplement_1):S20–42. Available from: <https://doi.org/10.2337/dc24-S002>