STM JOURNALS

Research & Reviews: A Journal of
Bioinformatics

ISSN: 2393-8722
Volume 11, Issue 3, 2024
September—December
DOI (Journal): 10.37591/RRJoBI

https://journals.stmjournals.com/rrjobi

Review RRJoBI

Diabetes Prediction Using ML Techniques

Yash Patil[1,*], Shubham Gosavi[2], Nikhita Mangaonkar[3]

Abstract

Diabetes mellitus, often called diabetes, refers to a group of common endocrine disorders that result in consistently high blood sugar levels. It is mainly categorized into two types: Type 1 Diabetes: This form arises when the body is unable to produce insulin. Type 2 Diabetes: This type is characterized by insulin resistance or insufficient insulin production. Early detection and timely intervention are essential to reduce the risk of complications and manage the condition effectively. The study delves into the efficiency of diverse Machine learning algorithms utilized to predict the onset of diabetes by analyzing demographic, clinical, and lifestyle factors. The study utilises a comprehensive dataset comprising health records of individuals, encompassing parameters, such as age, body mass index (BMI), glucose levels, insulin sensitivity, family history, and lifestyle habits. Feature selection plays a crucial role in diabetes research by identifying the key variables that contribute significantly to the development and management of the condition. These techniques help researchers and healthcare professionals focus on the most relevant factors. Various machine learning algorithms are used to forecast outcomes, such as the onset of diabetes and other health conditions. Logistic regression, known for its simplicity and interpretability, is compared against more complex models like random forests, decision trees, and support vector machines. Evaluation of these models relies on performance metrics, such as accuracy, sensitivity, specificity, and AUC-ROC, which help determine their appropriateness for various prediction tasks.

Keywords: Heart disease, classification, prediction, diabetes mellitus, insulin, glucose

INTRODUCTION

Diabetes is a pressing worldwide health issue, with a 3 percent rise in mortality rates from 2000 to 2019 and approximately 2 million deaths in 2019 attributed to diabetes and related kidney diseases. Early detection is crucial for managing diabetes and preventing its severe complications like heart disease and stroke. Machine learning models offer promise in predicting diabetes onset by analysing diverse health data, integrating demographics, clinical details, and lifestyle factors.

***Author for Correspondence**
Yash Patil
E-mail: yash.patil22@spit.ac.in

[1]Student, Department of Computer Applications, Sardar Patel Institute of Technology, Mumbai, Maharashtra, India
[2]Student, Department of Computer Applications, Sardar Patel Institute of Technology, Mumbai, Maharashtra, India
[3]Assistant Professor, Department of Computer Applications, Sardar Patel Institute of Technology, Mumbai, Maharashtra, India

Received Date: September 07, 2023
Accepted Date: October 08, 2024
Published Date: October 26, 2024

Citation: Yash Patil, Shubham Gosavi, Nikhita Mangaonkar. Diabetes Prediction Using ML Techniques. Research & Reviews: A Journal of Bioinformatics. 2024; 11(3): 1–9p.

This research investigates diverse machine learning algorithms' effectiveness in predicting diabetes onset using an extensive dataset encompassing variables like age, BMI, glucose levels, family medical history, and more. Advanced techniques will pinpoint key factors influencing diabetes development.

METHODOLOGY

Data Collection

The dataset used in the analysis was retrieved from Kaggle, a popular platform for datasets and machine learning competitions. It comprises 768 instances, each with 9 attributes with the details provided in (Table 1).

Table 1. Description of dataset attributes retrieved from Kaggle.

Data Element	Type	Description
Pregnancies	Num	Number of pregnancies.
Glucose	Bi	Plasma glucose concentration at 2 hrs.
Blood Pressure	Num	Diastolic blood pressure (mm Hg).
Skin Thickness	Num	Triceps skin fold thickness (mm).
2-Hour seruminsulin (mu U/ml)	Num	Cholesterol level.
BMI	Num	Body mass index (BMI).
Diabetes Pedigree Function	Nom	Diabetes pedigree function.
Age of patient	Num	Age (years).
Outcome	Bi	Class variable (0/1).

Data-Preprocessing

The data preprocessing process involves dealing with missing data, standardisation or normalizing modifying numerical characteristics, encoding the categorical variables, and managing outliers to enhance the quality and usability of raw data for accurate and meaningful analysis.

Data Visualization and Data Cleaning

Initially, conducted an assessment for missing values and discovered none.

Then we plotted some histograms based on the dataset features to check how the data is distributed. correlation matrix (Figure 1).

Assessing for Imbalances

Discrepancies in the output may skew prediction accuracy. Thus, we validated the balance of the output "target" as depicted in Figure 2. Upon examination, the data exhibited balance with a 9:11 ratio between the two categories, obviating the necessity for data resampling.

APPLIED ALGORITHMS

Random Forest Regression

The Random Forest (RF) technique excels in supervised machine learning, proficiently handling classification and regression tasks by leveraging ensemble learning – a powerful technique that combines multiple classifiers to significantly boost prediction accuracy, especially in complex scenarios.

1. *Logistic regression:* Logistic regression, a supervised machine learning methodology, tackles binary classification tasks by estimating the probability of an outcome, event, or observation. The model furnishes a binary outcome with only two potential results: affirmative/negative, 0/1, or true/false (Figure 3).
2. *Support vector machines (SVM):* SVM, an adaptable supervised learning approach, serves in classification, regression, and outlier detection. Its principal aim lies in establishing a decisive boundary between disparate classes to ensure precise label predictions utilizing one or more feature vectors, as portrayed in (Figure 4). This decision boundary, often denoted as the hyperplane, strategically optimizes the distance from adjacent data points. These support vectors play a pivotal role in determining the optimal hyperplane, thus bolstering SVM's predictive resilience.
3. *Decision tree algorithm:* Employing a tree-like model, the decision tree algorithm in machine learning navigates decisions and their potential ramifications to make predictions. It iteratively segregates data into subsets based on the most influential feature at each tree node. Figure 5 elucidates the construction of a decision tree for a weather dataset.
4. *XGBoost:* XGBoost, a gradient boosting algorithm, endeavors to accurately forecast a target variable by amalgamating a set of estimates from simpler and weaker models (Figures 6 and 7).
5. *Evaluation metrics:* Evaluation metrics gauge the quality and effectiveness of machine learning models. This paper selects the optimal model based on distinct evaluation metrics (Table 2).
6. *Confusion matrix:* An N * N matrix, where N represents the number of classes under prediction. In a binary prediction scenario, such as in this study, the confusion matrix dimension is 2 * 2 (Figure 8).

Research & Reviews: A Journal of Bioinformatics
Volume 10, Issue 3
ISSN: 2393-8722

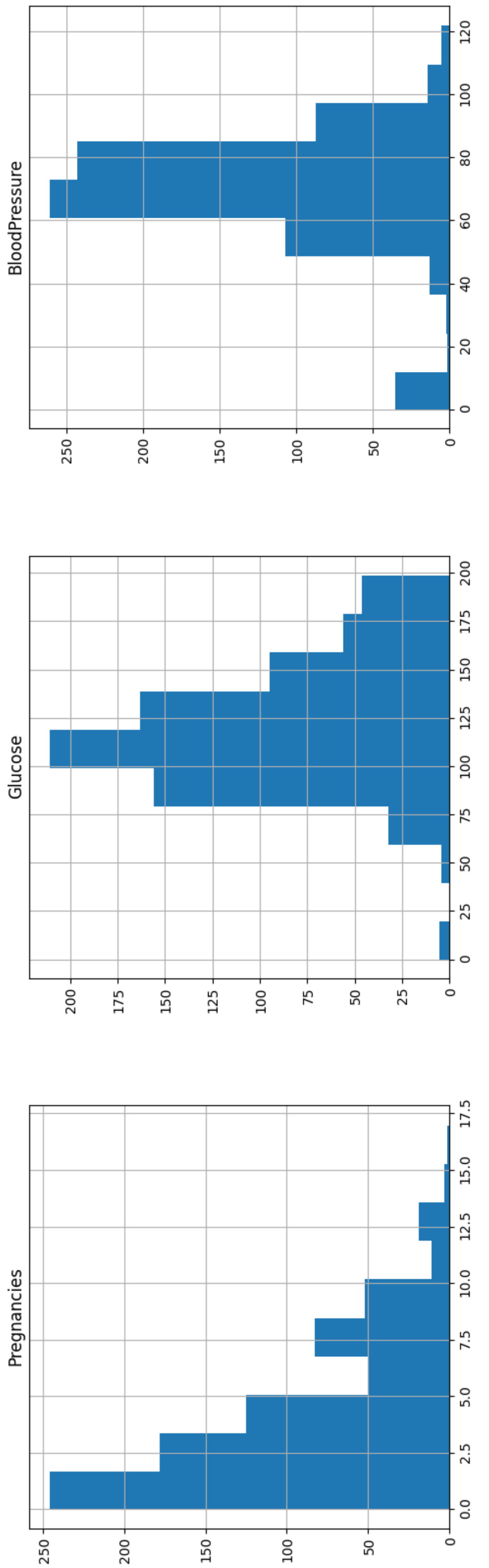


Figure 1. Pregnancies, glucose, blood pressure levels.

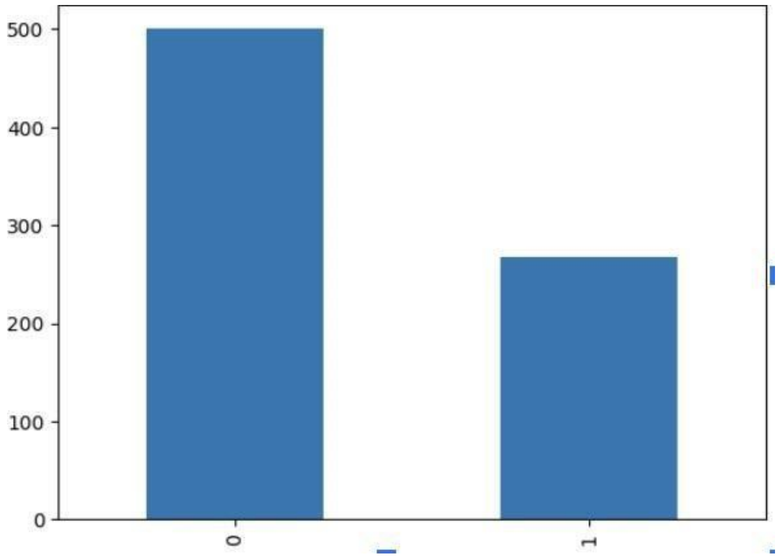
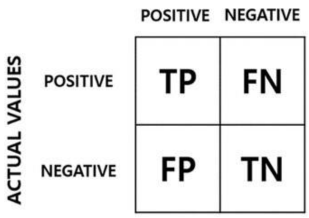


Figure 2. Output target balance with a 9:11 ratio between the two categories, indicating no need for data resampling.



$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Table 2. Performance metrics for various machine learning algorithms.

Algorithm Names	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Random forest regression	0.80	0.85	0.82	0.76
Decision tree	0.76	0.80	0.78	0.71
XgBoost	0.79	0.79	0.79	0.73
SVM	0.76	0.90	0.82	0.75
Logistic regression	0.81	0.88	0.84	0.79

LITERATURE REVIEW

Literature regarding the prediction of diabetes utilizing machine learning algorithms encompasses a diverse array of research endeavours, all directed towards one common goal: improving early detection and intervention strategies for diabetes. Central to this pursuit are techniques for feature selection and dimensionality reduction, as emphasized in prior studies [1]. These studies underscore the importance of meticulously identifying relevant variables from vast health records. By doing so, researchers aim to refine prediction models, ultimately enhancing the accuracy and efficiency of diabetes prediction [2].

Moreover, a plethora of investigations, such as those conducted by [3–9] delve into the application of various machine learning algorithms. Through comparative analyses, these studies evaluate the performance of models including logistic regression, decision trees, support vector machines, and deep learning techniques [10–13]. Their assessments extend to key metrics, such as accuracy, sensitivity, and specificity, providing valuable insights into the efficacy of different algorithms in predicting the onset of diabetes [14, 15].

Further contributions come from [16], who offer insights into the effectiveness of ensembling techniques. By combining multiple machine learning classifiers, they demonstrate how prediction accuracy and robustness can be significantly enhanced. Additionally, [12] delve into the impact of lifestyle indicators on diabetes prediction. Their research underscores the importance of incorporating personal habits and behaviours into predictive models, thereby providing a more comprehensive understanding of diabetes risk factors.

Collectively, this body of literature underscores the interdisciplinary efforts aimed at leveraging machine learning techniques to enhance diabetes prediction. By refining feature selection, algorithm selection, and the integration of diverse data sources, researchers strive to develop more accurate and actionable prediction models to mitigate the adverse effects of diabetes.

RESULT AND ANALYSIS

The construction of the diabetes prediction model involved five selected machine learning techniques, and the outcomes were attained through three distinct stages to arrive at the optimal final model. Initially, predictions were made using all dataset features. Subsequently, predictions were performed using only the selected features in the second stage. To further enhance performance, the last stage involved prediction after implementing hyperparameter tuning to develop the best-performing model. The results of each model, based on the confusion matrix and accuracy, are delineated in below Figures [3–7].

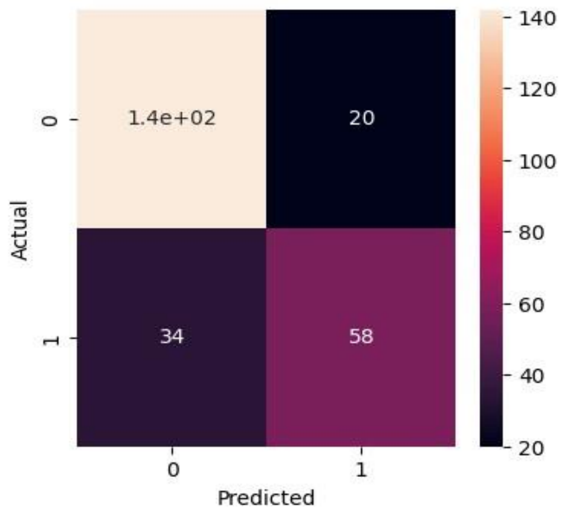


Figure 3. Logistic regression.

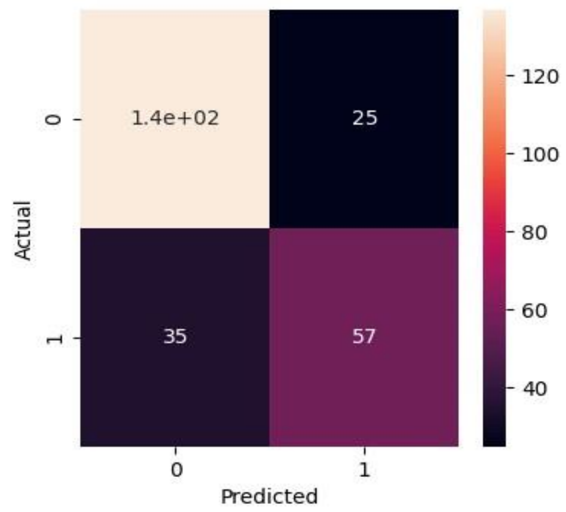


Figure 4. Random forest.

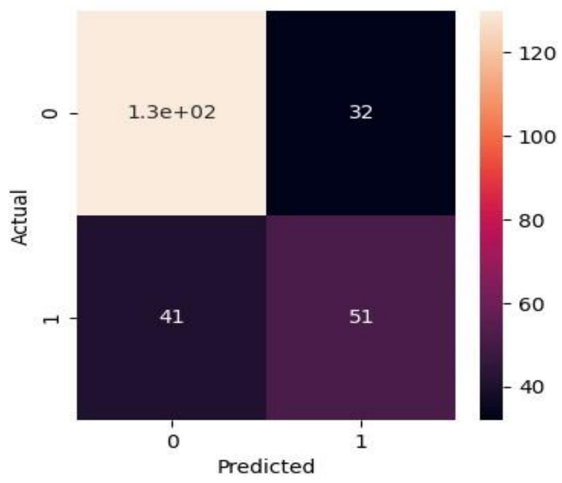


Figure 5. Decision tree.

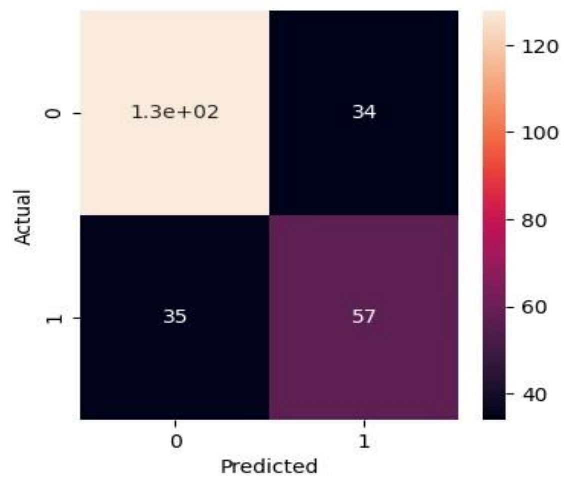


Figure 6. XGBoost.

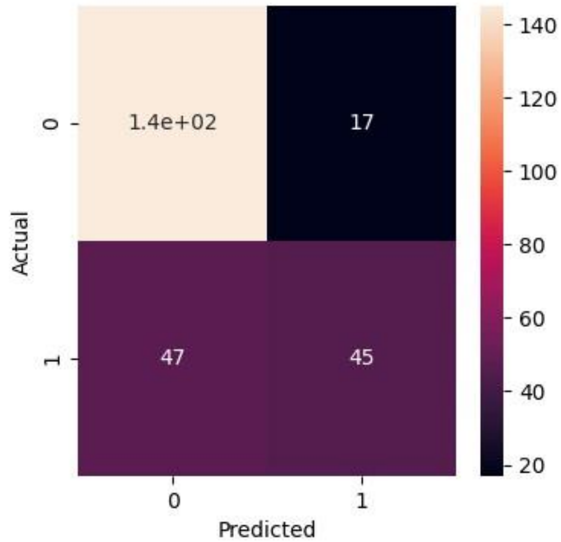


Figure 7. Support vector matrix.

Based on the analysis above, it was found that both logistic regression and random forest regression models performed well on the dataset, achieving accuracies of 79% and 76%, respectively. In the second stage, the focus shifts to identifying crucial features from these models and eliminating less significant ones. Following the application of feature selection techniques on the logistic regression model, it was determined that Diabetes Pedigree Function, Pregnancies, BMI, Glucose, and Age are the pivotal features, as depicted in Figure 9.

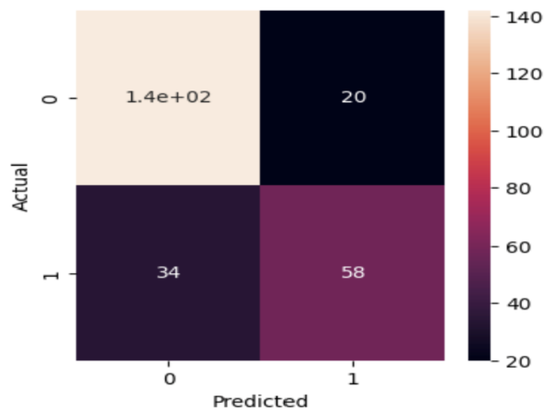


Figure 8. Confusion matrix.

Following the removal of all other features, a new model was constructed. However, this model exhibited underperformance compared to the initial model, achieving only 75% accuracy. Subsequently, feature selection techniques were applied to the random forest regressor models, revealing that the important features are trestbps, chol, thalach, cp, exang, oldpeak, slope, thal, and ca, as illustrated in Figure 9.

After employing GridSearchCV, the logistic regression model revealed the best parameters as depicted in Figure 10. Utilizing these parameters, the model achieved an accuracy of 85%. Similarly, for the Random Forest Regressor, the following parameters were explored:

Optimal hyperparameters

- “n_estimators”: (100, 200, 300),

- “max_depth”: (None, 10, 20, 30),
- “min_samples_leaf”: (2, 5, 10),
- “min_samples_split”: (1, 2, 4),
- “max_features”: (“auto”, “sqrt”, “log2”).

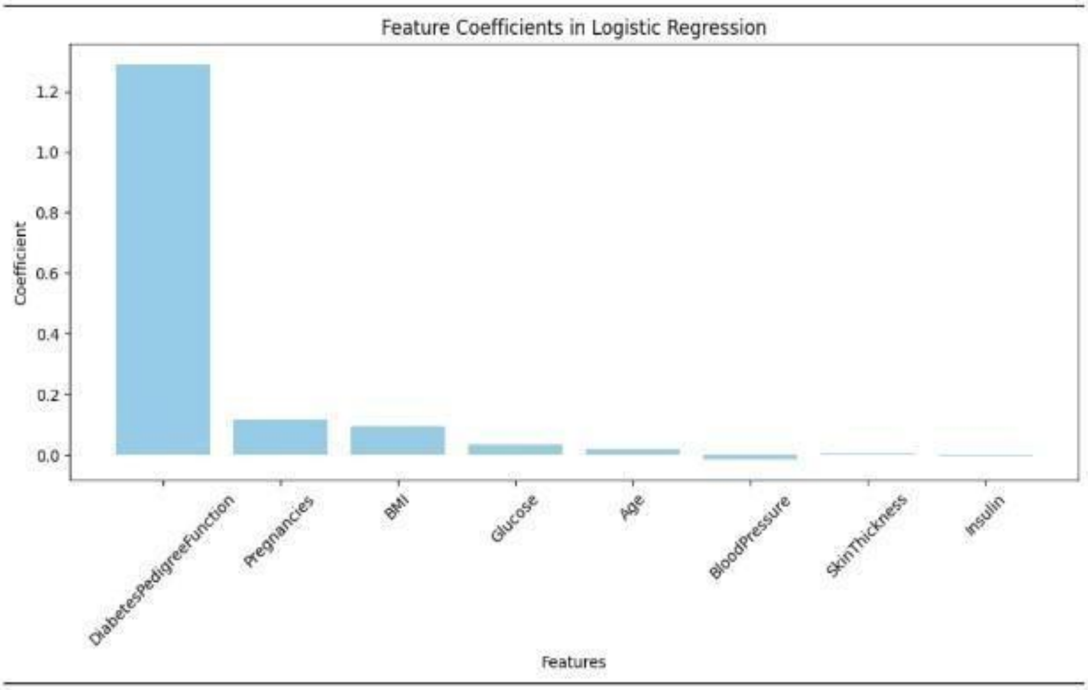


Figure 9. Feature coefficients in logistic regression.

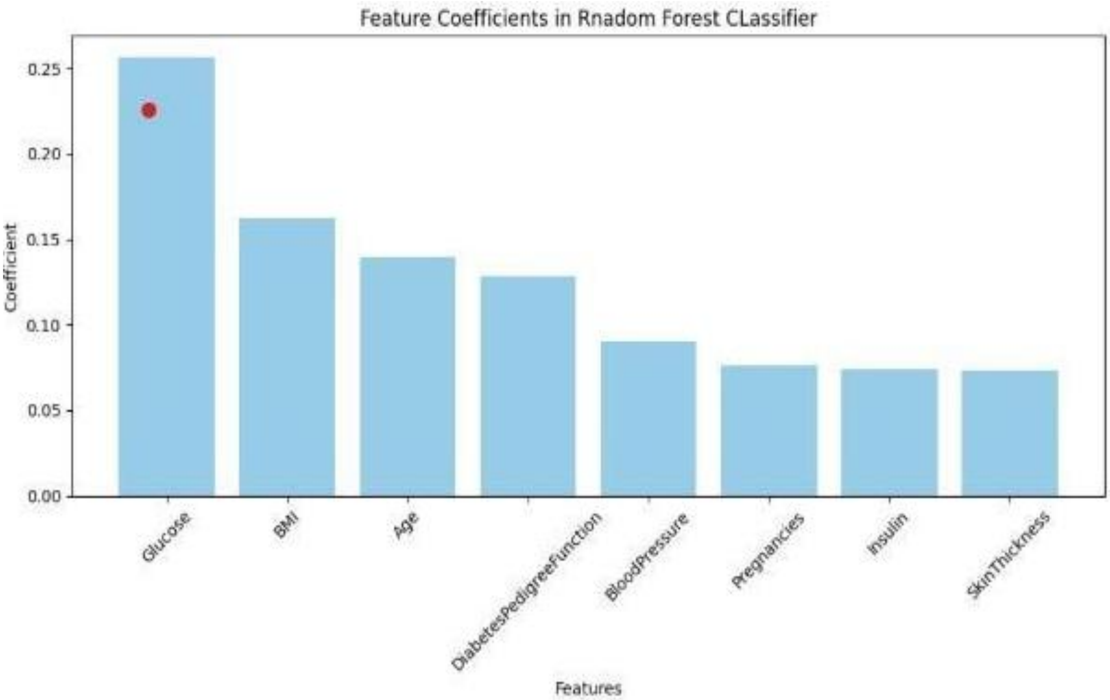


Figure 10. Parameters in logistic regression model.

Following another round of GridSearchCV, the optimal hyperparameters were identified as “max depth”: None, “max features”: auto, “min samples leaf”: 2, “min samples split”: 10, “n_estimators”: 300. Implementing these parameters, the model attained an accuracy of 80%.

CONCLUSIONS

In summary, the integration of machine learning into diabetes disease prediction represents a significant advancement in healthcare. By leveraging advanced algorithms with extensive datasets, these tools demonstrate promise in early detection and risk assessment. The ability to discern subtle patterns in diverse patient data enables more accurate predictions, facilitating timely interventions and improved outcomes. However, continuous refinement and validation with real-world clinical data remain essential for reliability and applicability across diverse populations. As technology evolves, collaboration among healthcare professionals, data scientists, and policymakers remains crucial for driving innovation in healthcare delivery and enhancing patient outcomes.

REFERENCES

1. Sivaranjani S, Ananya S, Aravinth J, Karthika R. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In:2021 7th international conference on advanced computing and communication systems (ICACCS); 2021 Mar 19; Vol. 1. IEEE; 2021. p. 141–146. doi: 10.1109/ICACCS51430.2021.9441935.
2. Parimala G, Kayalvizhi R, Nithiya S. Diabetes Prediction using Machine Learning. In: 2023 International Conference on Computer Communication and Informatics (ICCCI); 2023 Jan 23. IEEE; 2023. p. 1–10. doi: 10.1109/ICCCI56745.2023.10128216.
3. Almahdawi A, Naama ZS, Al-Taie A. Diabetes prediction using machine learning. In: 2022 3rd Information Technology to Enhance E-learning and Other Applications (IT-ELA); 2022 Dec 27. IEEE; 2022. p. 186–190. doi: 10.1109/IT-ELA57378.2022.10107919.
4. Sarwar MA, Kamal N, Hamid W, Shah MA. Prediction of diabetes using machine learning algorithms in healthcare. In: 2018 24th International Conference on Automation and Computing (ICAC); 2018 Sep 6. IEEE; 2018. p. 1–6. doi: 10.23919/IConAC.2018.8748992.
5. Lyngdoh AC, Choudhury NA, Moulik S. Diabetes disease prediction using machine learning algorithms. In: 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES); 2021 Mar 1. IEEE; 2021. p. 517–521. doi: 10.1109/IECBES48179.2021.9398759.
6. Dubey Y, Wankhede P, Borkar T, Borkar A, Mitra K. Diabetes prediction and classification using machine learning algorithms. In: 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON); 2021 Dec 4. IEEE; 2021. p. 60–63. doi: 10.1109/BECITHCON54710.2021.9893653.
7. Rady M, Moussa K, Mostafa M, Elbasry A, Ezzat Z, Medhat W. Diabetes prediction using machine learning: A comparative study. In: 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES); 2021 Oct 23. IEEE; 2021. p. 279–282. doi: 10.1109/NILES53778.2021.9600091.
8. Woldemichael FG, Menaria S. Prediction of diabetes using data mining techniques. In: 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI); 2018 May 11. IEEE; 2018. p. 414–418. doi: 10.1109/ICOEI.2018.8553959.
9. Sonar P, Jaya Malini K. Diabetes prediction using different machine learning approaches. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC); 2019 Mar 27. IEEE; 2019. p. 367–371. doi: 10.1109/ICCMC.2019.8819841.
10. Jana S, Bharanidharan N, Shanmukha Nagasai P, Kumar S, Nageshwar VM. Diabetes prediction using machine learning algorithms. In: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS); 2022 Mar 25; Vol. 1. IEEE; 2022. p. 46–51. doi: 10.1109/ICACCS54159.2022.9785073.
11. Yahyaoui A, Jamil A, Rasheed J, Yesiltepe M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK); 2019 Nov 6. IEEE; 2019. p. 1–4. doi: 10.1109/UBMYK48245.2019.8965556.

12. Anand A, Shakti D. Prediction of diabetes based on personal lifestyle indicators. In: 2015 1st International Conference on Next Generation Computing Technologies (NGCT); 2015 Sep 4. IEEE; 2015. p. 673–676. doi: 10.1109/NGCT.2015.7375206.
13. Hasan SM, Rabbi MF, Champa AI, Zaman MA. An effective diabetes prediction system using machine learning techniques. In: 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT); 2020 Nov 28. IEEE; 2020. p. 23–28. doi: 10.1109/ICAICT51780.2020.9333497.
14. Mir A, Dhage SN. Diabetes disease prediction using machine learning on big data of healthcare. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA); 2018 Aug 16. IEEE; 2018. p. 1–6. doi: 10.1109/ICCUBEA.2018.8697439.
15. Dutta D, Paul D, Ghosh P. Analysing feature importances for diabetes prediction using machine learning. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON); 2018 Nov 1. IEEE; 2018. p. 924–928. doi: 10.1109/IEMCON.2018.8614871.
16. Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access. 2020 Apr 23;8:76516–31. doi: 10.1109/ACCESS.2020.2989857.

