Journal of Science Research and Reviews
PRINT ISSN: 1595-9074
E-ISSN: 1595-8329
DOI: https://doi.org/10.70882/josrar.2026.v3i3.164
Homepage: https://josrar.esrgngr.org
Original Research Article

A Comparative Machine Learning Framework for Early Diabetes Risk Prediction

***[1]Abdulrahman Nasiru Sada, [2]Eli Adama Jiya, [3]Yahaya Muhammad Umar and [1]Umar Faruk Mukhtar**

[1]Department of Information Technology, Federal University Dutsin-ma
[2]Department of Computer Science, Federal University Dutsin-ma
[3]Department of General Studies, Katsina State Collage of Health Science and Technology, Katsina
*Corresponding Author's email: asada21@fudutsinma.edu.ng

KEYWORDS
Diabetes prediction,
Machine Learning,
Disease,
Feature Selection.

CITATION
Sada, A. N., Jiya, E. A., Umar, Y. M., & Mukhtar, U. F. (2026). A Comparative Machine Learning Framework for Early Diabetes Risk Prediction. *Journal of Science Research and Reviews*, *3*(3), 207-214.
https://doi.org/10.70882/josrar.2026.v3i3.164

ABSTRACT
Diabetes is one of the leading causes of morbidity and mortality worldwide. To avoid difficult management of the condition, there is need to predict early onset of the condition. This study investigated the application of machine learning techniques and exploratory data analysis to forecast early-onset of diabetes using the PIMA Indian dataset. Preprocessing included handling missing values and standardization, leading to the development and evaluation of 5 models which include Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machine, and Neural Network. Exploratory analysis identified glucose concentration, body mass index, and age as the most influential features. Random Forest achieved the highest accuracy (0.74%) while both Random Forest and Logistic Regression attained the best ROC-AUC score of 0.81%. Feature importance analysis emphasized the predictive significance of glucose and BMI, aligning with clinical knowledge of diabetes risk factors. Despite the promising results the study acknowledged limitations related to the PIMA dataset's demographic scope and the moderate complexity of neural networks, highlighting areas for future enhancement. Ethical considerations, including data privacy and algorithmic bias, were addressed to ensure responsible model development.

INTRODUCTION

The hallmark of diabetes mellitus, a chronic metabolic disease, is persistent hyperglycemia brought on by either decreased insulin action or secretion, or both. With rising prevalence and substantial contributions to morbidity, mortality, and healthcare spending globally, it continues to be a serious public health concern (Firdous et al., 2022) Cardiovascular disease, nephropathy, neuropathy, and retinopathy are among the long-term sequelae of diabetes that significantly lower quality of life and place a substantial burden on healthcare systems (Zafar et al., 2019). As a result, early identification of individuals at high risk of developing diabetes is critical for enabling timely intervention and effective disease management.

Automated patterns can be found in vast, multifaceted healthcare datasets thanks to machine learning algorithms. Random forests, logistic regression, support vector machines, and neural networks are examples of supervised learning algorithms that have been widely used in diabetes research to stratify patient risk and forecast the start of the disease (Khanam & Foo, 2021; Perveen et al., 2020). According to Ganie et al. (2023), ensemble-based models in particular, random forests have proven to be resilient when dealing with medical data that is noisy, unbalanced, and diverse. Notwithstanding these

developments, issues with model interpretability, generalizability, and ethical application continue to be major obstacles to clinical acceptance. While complex "black-box" algorithms may reduce transparency and clinician trust, predictive models developed using population-specific datasets may exhibit limited generalizability across different demographic groups. Therefore, external validation across diverse populations is essential to ensure robust and equitable clinical performance. (Gianfrancesco et al., 2018; Dharmarathne et al., 2024).

To predict diabetes using the Pima Indians Diabetes Dataset, this study uses numerous supervised machine learning methods and exploratory data analysis. This study intends to show the potential of interpretable machine learning models as decision-support tools for early diabetes risk assessment and preventive healthcare planning by methodically comparing model performance and analyzing feature importance in order to identify clinically meaningful predictors of diabetes.

Related Work

Machine learning is a data-driven analytical approach that integrates multiple risk factors into predictive algorithms(Xu et al., 2020). The existing literature has demonstrated the potential of ML techniques to enhance the accuracy and efficiency of diabetes screening, enabling earlier diagnosis and treatment. (Chari et al., 2019). Several studies have focused on developing predictive models for diabetes using clinical data and ML algorithms (Perveen et al., 2020). Various ML algorithms, such as artificial neural networks, decision trees, and support vector machines, have been evaluated for their ability to predict diabetes risk. This study has shown that ML models can achieve high accuracy in identifying individuals at risk of developing diabetes, outperforming traditional statistical methods. Researchers have also explored the use of ensemble methods, including random forests and gradient boosting, to improve the performance of diabetes prediction models (Ogunpola et al., 2024). The application of ML facilitates the discovery of previously unidentified heterogeneity in diabetes, which aids in the sensitive recognition of complex data patterns (Cho et al., 2019).Previous studies have demonstrated that ML models can attain high accuracy in predicting the risk of various diseases, including heart failure (Awan et al., 2019). Previous research has demonstrated that integrating clinical and demographic variables into machine learning models can significantly improve the accuracy and reliability of diabetes risk prediction. Nonetheless, despite growing research interest in this field, significant gaps persist in the existing body of literature. Firstly, although model interpretability is essential for clinical trust and practical adoption, it is often neglected in favor of predictive accuracy. As a result, many complex models operate as “black boxes,” thereby limiting their usability in clinical decision-support systems.

MATERIALS AND METHODS

Data Collection

This study leverages the Pima Indian Diabetes Dataset, a widely recognized benchmark in the machine learning community, sourced from Kaggle. The dataset comprises health records of female Pima Indian individuals, residing near Phoenix, Arizona, all of whom are at least 21 years old, providing a foundational dataset for investigating diabetes prediction The dataset encompasses 768 instances and 8 input features, with a single target variable denoting the presence or absence of diabetes. The Pima Indian population has a notably high prevalence of type 2 diabetes, rendering this dataset particularly valuable for studying diabetes risk factors and developing predictive models (Yeh & Kong, 2013). Ethical approval was prioritized throughout the research process, ensuring adherence to the principles of beneficence, non-maleficence, autonomy, and justice by following the guidelines for data privacy and security, and compliance with relevant regulations to protect the confidentiality and rights of the individuals represented in the dataset(Avoi & Liaw, 2021).

Data Preprocessing

Data preprocessing constituted a critical step in preparing the dataset for analysis, aiming to enhance data quality, address missing values, and transform features into a suitable format for machine learning algorithms. Missing values were handled using mean imputation to preserve dataset size. All features where standardized using Z-score normalization via the Standard Scaler from Scikit-learn. Feature scaling was applied to normalize the range of values across different features, preventing features with larger values from dominating the analysis and ensuring that all features contribute equally to the models. This step was essential for models sensitive to feature scales, such as KNN and SVM.

Exploratory Data Analysis (EDA)

EDA was performed to gain comprehensive insights into the distribution, patterns, and interrelationships among the variables. Univariate statistical techniques, such as histograms and box plots, were leveraged to visualize the individual feature distributions and identify any skewness, as well as to analyze the outcome variable's class distribution for potential imbalance. Bivariate analysis, including scatter plots and pair plots, facilitated the visualization and examination of correlations between the features. Additionally, correlation matrices and heat maps were utilized to provide a more holistic understanding of the feature relationships.

Feature Selection

Feature selection is an important step to reduce the dimensionality of the data, improve model performance, and enhance interpretability ((AlSagri & Ykhlef, 2020). It played a crucial role in identifying the most relevant and informative features for predicting diabetes, aiming to reduce dimensionality, improve model interpretability, and enhance generalization performance. Feature importance was assessed using techniques such as Random Forest which quantify the contribution of each feature to the predictive power of the model. Univariate feature selection methods, such as SelectKBest, were employed to select the top K features based on statistical tests, such as chi-squared test, to assess the relationship between each feature and the target variable

Model Development

Model development involved the evaluation of multiple supervised machine learning algorithms for diabetes prediction, with model selection based on predictive performance, interpretability, and generalization capability. The algorithms were chosen to represent a range of modeling approaches commonly applied in healthcare predictive analytics. Logistic Regression was selected as a baseline model due to its simplicity, interpretability, and widespread use in clinical risk prediction. Support Vector Machines (SVMs) were included because of their ability to model complex non-linear relationships and perform effectively in high-dimensional feature spaces. Random Forests were selected for their robustness to over fitting, ability to capture non-linear interactions among variables, and provision of feature importance measures that enhance interpretability. By comparing these algorithms, the study aimed to identify a model that balances predictive accuracy with clinical interpretability and practical applicability. All models were implemented using the Scikit-learn library in Python.

Model Training and Evaluation

The machine learning models Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machine, and a Neural Network classifier were trained on the dataset. The preprocessed dataset was partitioned into training and testing subsets, allocating 80% for model training and the remaining 20% for model evaluation. To ensure uniform feature contributions, particularly crucial for distance-based models like KNN and SVM, feature scaling using standardization was applied. Additionally, grid search with cross-validation was employed to optimize the models' hyperparameters. For instance, in the case of the KNN model, the optimal number of neighbors (k) was determined by evaluating different values between 1 and 14, while for the Random Forest model, parameters such as the number of estimators and maximum tree depth were fine-tuned. Each model was trained exclusively on the training set, preventing any data leakage into the test set and enabling the accurate assessment of model generalization.

The performance of the models was evaluated using several classification metrics, including accuracy, precision, recall, F1-score, and the Area under the Receiver Operating Characteristic Curve (AUC-ROC), to provide a comprehensive assessment of their ability to distinguish between diabetic and non-diabetic individuals. These metrics were computed based on the confusion matrix comprising True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy measures the proportion of correctly classified instances and is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision measures the proportion of positive predictions that are correct:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall (Sensitivity) measures the proportion of actual positive cases correctly identified by the model:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1-score provides a harmonic mean of precision and recall:

$$F_1 = 2 \times \frac{PrecisionRecall}{Precision+Recall} \quad (4)$$

The AUC-ROC evaluates the model's ability to discriminate between classes across different classification thresholds. The ROC curve is generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR), where:

$$TPR = \frac{TP}{TP+FN} \quad (5)$$

$$FPR = \frac{FP}{FP+TN} \quad (6)$$

A higher AUC value indicates better discriminative performance.

To evaluate the generalization capability of the models, k-fold cross-validation was employed. In this approach, the dataset was partitioned into (k) equal subsets. During each iteration, one subset was used for validation while the remaining (k-1) subsets were used for training. The final performance was obtained by averaging the results across all folds. Furthermore, all performance metrics were computed on a held-out testing dataset to provide an unbiased assessment of each model's predictive performance and generalization ability

RESULTS AND DISCUSSION

Descriptive Statistics and Data Exploration

Preliminary exploratory data analysis was performed to gain comprehensive insights into the distribution, patterns, and interrelationships among the variables in the dataset,as shown in figure1 The dataset comprises 768 observations and 9 variables, including measures of glucose concentration, body mass index, insulin levels,

age, as well as a binary outcome variable denoting the presence or absence of diabetes. The central tendency analysis revealed that the mean glucose level was approximately 121 mg/dL, while the average body mass index was around 32. Furthermore, insulin levels exhibited substantial variation, suggesting the potential presence of outliers.Several variables, such as insulin and skin thickness, contained zero values, suggesting the presence of missing data that may have been addressed through mean imputation technique.

Visualizations

Exploratory data visualizations were employed to gain a comprehensive understanding of the data. Histograms and box plots revealed the presence of skewness and outliers particularly in the insulin and skin thickness variables from figure 2 and Figure 3. Additionally, a correlation heatmap was utilized to identify strong correlations between glucose levels and the outcome variable, as well as moderate correlations with body mass index and age.

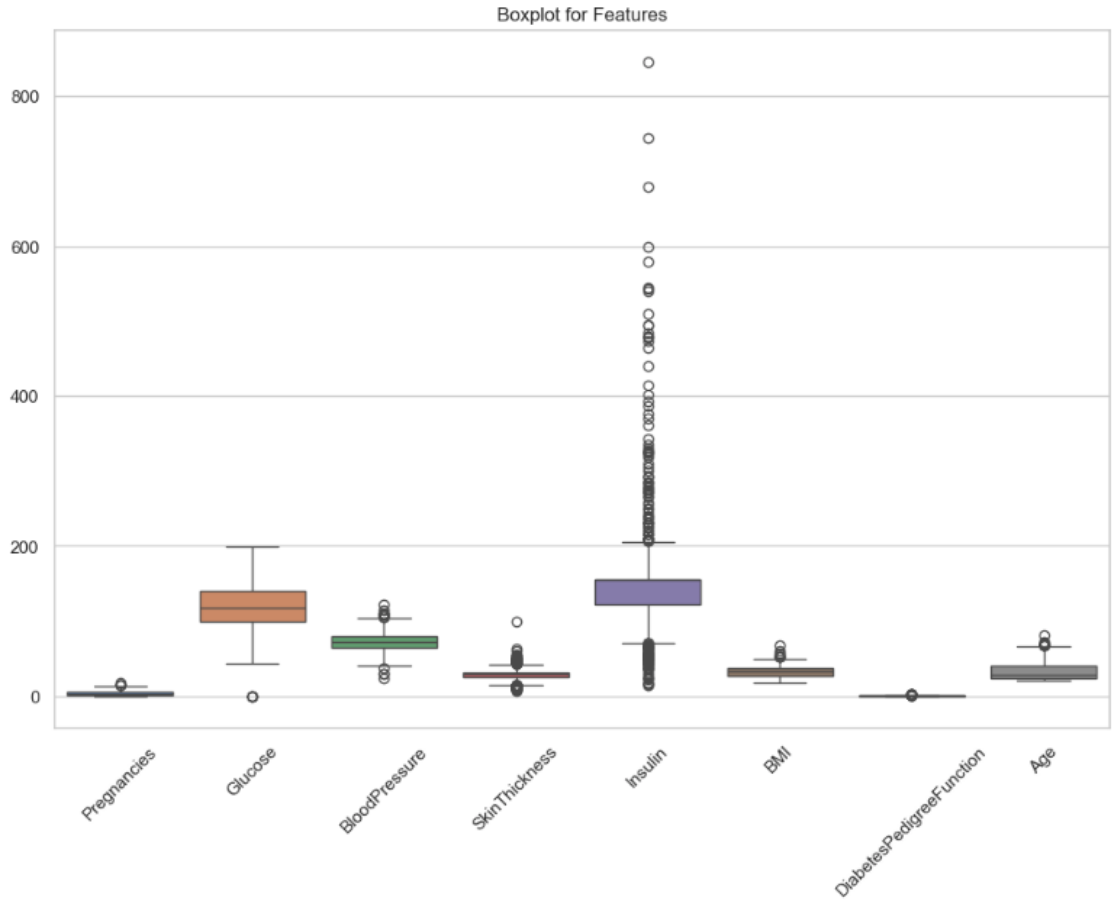

Figure 1: Boxplot for features

The boxplots reveals substantial variability in the Dataset, with numerous outliers and extreme values in features like Insulin, SkinThickness, and BloodPressure. The Insulin feature, in particular, exhibits significant right-skewness and extreme values exceeding 600 units, which may impact model stability if not properly addressed. In contrast, Glucose and BMI exhibited relatively stable distributions, although occasional high outliers were observed. Additionally, the presence of zero values in Glucose, BloodPressure, and SkinThickness indicated potential missing data, which were subsequently addressed through data imputation. Overall, this analysis highlighted the importance of thorough data cleaning and normalization in enhancing the robustness and reliability of the predictive models developed using the dataset.

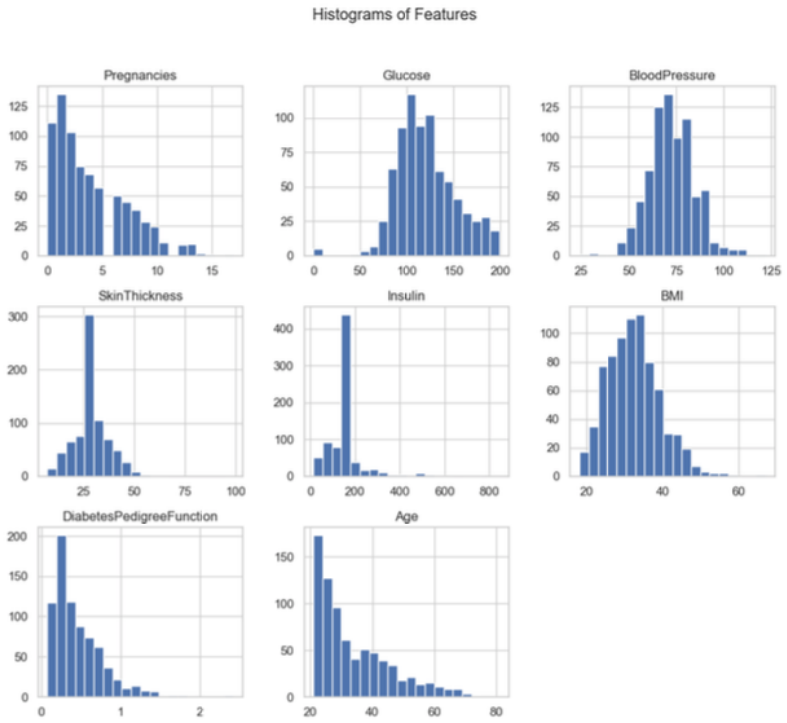

Figure 2: Showing Histograms of Features

From figure 2The feature distributions in the Dataset exhibit substantial variability. Most variables display right-skewed patterns, indicating data points concentrated at lower values with extended tails. Glucose and BloodPressure show more symmetric, bell-shaped distributions. BMI has a slightly right-skewed normal distribution, reflecting the prevalence of overweight individuals in diabetic populations. The SkinThickness and Insulin histograms present pronounced peaks at low values, suggesting potential data quality issues. These diverse feature distributions emphasize the need for robust data preprocessing to ensure optimal predictive model performance.

The figure 3 present the correlations among rhe various variables with the prediction class

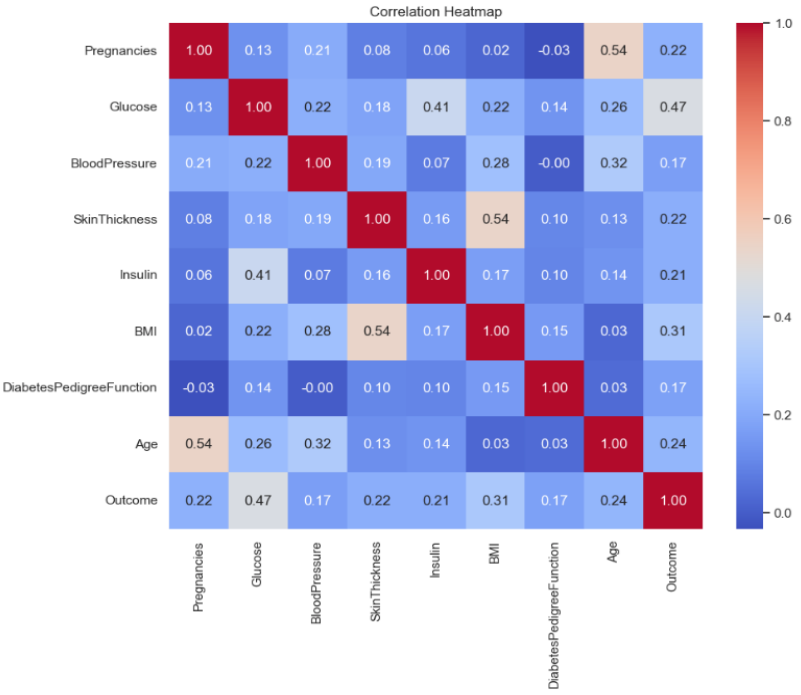

Figure 3: Correlation Heatmap

The correlation heatmap in figure 4 reveals the strength and directionality of linear relationships among the features in the Pima Indians Diabetes Dataset. As evident, Glucose exhibits the most robust positive correlation with the target variable Outcome(r = 0.47), indicating that elevated glucose levels are moderately associated with a higher probability of diabetes. Additionally, BMI and Age also display positive but weaker correlations with Outcome(r = 0.31 and r = 0.24, respectively), aligning with established meAdical evidence regarding risk factors for diabetes. Interestingly, Pregnancies is modestly correlated with both Age(r = 0.54) and Outcome(r = 0.22), suggesting that older women tend to have experienced more pregnancies, and that pregnancy history may marginally contribute to diabetes risk. Other features, such as BloodPressure**,** SkinThickness, and Insulin, demonstrate relatively low correlations with Outcome (all below 0.25), implying a less direct impact on diabetes prediction within this dataset. Notably, a moderate correlation(r = 0.54) between BMI and SkinThickness was observed, which is expected given their shared relationship to body fat measurements. Overall, the heatmap highlights that not all features contribute equally to predicting diabetes, underscoring the importance of focusing on Glucose, BMI, and Age during model development(Deberneh & Kim, 2021).

Model Perfomance

Table 1 presents the performance of the evaluated machine learning models for diabetes prediction. Among the models, the Random Forest classifier achieved the highest accuracy (74.7%), indicating its superior ability to correctly classify diabetic and non-diabetic individuals. This performance may be attributed to its ensemble-based architecture, which combines multiple decision trees to capture complex, non-linear relationships among predictor variables while reducing the risk of over fitting. The model also achieved a competitive ROC-AUC score of 0.803, demonstrating strong discriminative capability and good generalization performance.

Although Logistic Regression achieved a slightly lower accuracy (72.7%), it recorded the highest ROC-AUC score (0.810). This suggests that the model was more effective at distinguishing between diabetic and non-diabetic cases across different classification thresholds. The strong performance of Logistic Regression indicates that the relationships between the predictor variables and diabetes status may be reasonably represented by a linear decision boundary. Furthermore, its interpretability makes it particularly attractive for clinical decision-support applications where transparency is essential.

The K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) models demonstrated competitive performance, achieving accuracies of 72.7% and 70.8%, respectively. The relatively strong results obtained by these models suggest that local neighborhood structures and non-linear feature interactions contribute to diabetes prediction. However, their performance remained slightly below that of Random Forest and Logistic Regression, indicating that the dataset characteristics may be better captured by ensemble and linear modeling approaches.

In contrast, the Neural Network classifier recorded the lowest accuracy (67.5%). This outcome may be attributed to the relatively small dataset size and the limited complexity of the neural network architecture. Neural networks typically require larger datasets and extensive hyperparameter optimization to fully exploit their learning capabilities. Consequently, the simpler machine learning models were better suited to the available data.

Overall, the findings indicate that Random Forest and Logistic Regression provided the best balance between predictive performance and practical applicability. While Random Forest achieved the highest classification accuracy, Logistic Regression offered superior discriminative ability and greater interpretability. These results are consistent with previous studies that have reported strong performance of ensemble learning methods and traditional statistical models in diabetes prediction tasks. The confusion matrices and ROC curves further supported these findings by illustrating the trade-offs between sensitivity and specificity across the evaluated models and confirming the robustness of the selected classifiers.

Table 1: Showing Models Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logic Regression	0.727	0.750	0.727	0.733	0.810
KNN	0.727	0.741	0.747	0.742	0.798
Random Forest	0.747	0.720	0.727	0.722	0.803
SVM	0.708	0.733	0.740	0.733	0.795
Neural Network	0.675	0.721	0.727	0.723	0.798

ROC Curve Analysis

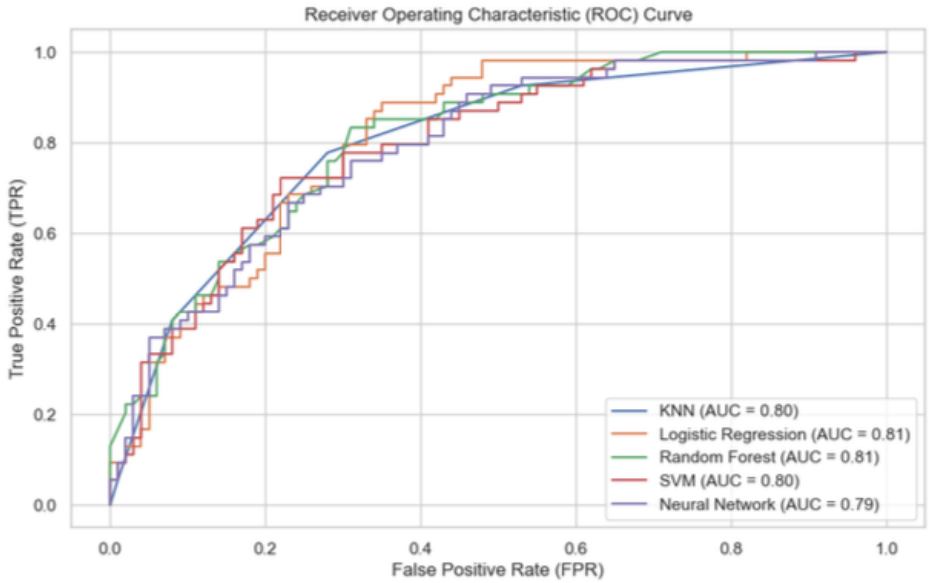

Figure 4: ROC curve analysis

The ROC curve analysis revealed that all the models exhibited predictive performance well above random chance levels. The Logistic Regression model demonstrated the most pronounced ROC curve, with the highest Area Under Curve statistic, indicating superior ability to differentiate between diabetic and non-diabetic cases(Ahmed et al., 2023). The clear visual separation between the model ROC curves further corroborated the superior classification capabilities of the Logistic Regression and Random Forest algorithms(Safaripour & Lim, 2022).

CONCLUSION

This study evaluated the predictive performance of Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and Neural Network models for diabetes prediction using the Pima Indians Diabetes Dataset. Among the evaluated models, Random Forest achieved the highest classification accuracy (74.7%), indicating its effectiveness in capturing complex relationships among clinical variables. This finding is consistent with previous studies that have reported superior performance of ensemble learning techniques for diabetes prediction due to their ability to model non-linear interactions and reduce overfitting (Iparraguirre-Villanueva et al., 2023).

Logistic Regression achieved a slightly lower accuracy (72.7%) but recorded the highest ROC-AUC score (0.810), demonstrating strong discriminative ability across different classification thresholds. This result agrees with previous research that has identified Logistic Regression as a reliable and interpretable baseline model for medical diagnosis and risk prediction tasks. The competitive performance of Logistic Regression suggests that a substantial proportion of the relationship between the predictor variables and diabetes status can be effectively captured using a linear decision boundary.

The KNN and SVM models also demonstrated competitive performance, with accuracies of 72.7% and 70.8%, respectively. Although these models performed reasonably well, their results were slightly inferior to those of Random Forest and Logistic Regression. Similar findings have been reported in the literature, where KNN and SVM often provide satisfactory classification performance but may be more sensitive to parameter selection and data distribution characteristics.

In contrast, the Neural Network model achieved the lowest accuracy (67.5%). This finding differs from some studies that have reported superior neural network performance in diabetes prediction. However, neural networks generally require larger datasets and extensive hyperparameter tuning to achieve optimal performance. The relatively modest size of the Pima Indians Diabetes Dataset and the selected network architecture may therefore have limited the model's predictive capability in this study.

Overall, the results indicate that Random Forest and Logistic Regression provided the most effective balance between predictive performance and practical applicability. The findings support existing evidence that ensemble-based methods and interpretable statistical models remain strong candidates for diabetes risk prediction, particularly when working with moderately sized clinical datasets. These models could assist healthcare practitioners in identifying high-risk individuals and enabling earlier interventions to reduce the burden of diabetes-related complications.

REFERENCES

Chari, K. K., Babu, M. C., & Kodati, S. (2019). Classification of diabetes using Random Forest with feature selection. *International Journal of Innovative Technology and Exploring Engineering*, 9(1), 1295–1300. https://doi.org/10.35940/ijitee.l3595.119119

Dharmarathne, G., Jayasinghe, T. N., Bogahawaththa, M., Meddage, D. P. P., & Rathnayake, U. (2024). A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Healthcare Analytics*, 5, 100301. https://doi.org/10.1016/j.health.2024.100301

Firdous, S., Wagai, G. A., & Sharma, K. (2022). A survey on diabetes risk prediction using machine learning approaches. *Journal of Family Medicine and Primary Care*, 11(11), 6929–6934. https://doi.org/10.4103/jfmpc.jfmpc_502_22

Ganie, S. M., Pramanik, P. K. D., Malik, M. B., Mallik, S., & Qin, H. (2023). An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics*, 14, 1252159. https://doi.org/10.3389/fgene.2023.1252159

Gianfrancesco, M., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11), 1544–1554. https://doi.org/10.1001/jamainternmed.2018.3763

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432–439. https://doi.org/10.1016/j.icte.2021.02.004

Kaur, H., & Kaur, G. (2019). Voting-based classification for diabetes prediction. *International Journal of Recent Technology and Engineering*, 8, 913–918. https://doi.org/10.35940/ijrte.b1172.0782s619

Perveen, S., Shahbaz, M., Ansari, M. S., Keshavjee, K., & Guergachi, A. (2020). A hybrid approach for modeling type 2 diabetes mellitus progression. *Frontiers in Genetics*, 10, 1076. https://doi.org/10.3389/fgene.2019.01076

Zafar, F., Raza, S., Khalid, M. U., & Tahir, M. (2019). Predictive analytics in healthcare for diabetes prediction. *Proceedings of the International Conference on Data Science*, 253–259. https://doi.org/10.1145/3326172.3326213