Prediction of Clinical Risk Factors of Diabetes Using Multiple Machine Learning Techniques

University of Barishal

A Project Report under the Training Program of "Data Science with Python"

Prepared By:

Name: Md. Abu Talha Tanvir

Course Name: Data Science with Python

Batch: DS-04

Roll: 09-004-23

Dept. of Biochemistry & Biotechnology

Supervised By:

Md. Mahbub E Noor

Lecturer

Computer Science & Engineering

University Of Barishal

EDGE
ENHANCING DIGITAL GOVERNMENT AND ECONOMY

EDGE: BU-CSE Digital Skills Training Program

Abstract

Diabetes is a rapidly growing global health concern, affecting individuals across all age groups and significantly reducing life expectancy. It is often accompanied by severe complications, including cardiovascular disease, kidney failure, stroke, and organ damage. Early diagnosis of diabetes is critical to mitigating its progression and reducing associated risks. This research analyzes diabetes risk factors using a dataset comprising clinical attributes such as glucose levels, BMI, blood pressure, and age. Logistic Regression, Random Forest and Neural Network were employed to predict diabetes risk. To address data challenges, preprocessing steps included replacing invalid values, scaling features, and resolving class imbalance. The impact of outliers was also assessed and managed to improve model reliability. Results from the analysis identified glucose levels, BMI, and age as the most significant predictors of diabetes. Among the machine learning models tested, Random Forest achieved the highest accuracy of 75%, surpassing Logistic Regression (73%) and Neural Network showed 70% of accuracy. These findings highlight the potential of machine learning in early diabetes diagnosis, aiding clinicians in timely intervention and improved patient outcomes.

Index Terms—Diabetes Prediction, Logistic Regression, Risk Factor Analysis, Random Forest, Neural Network, Class Imbalance, Outlier Handling.

1 Introduction

1.1 Background of the Study

Diabetes, often described as a “silent killer,” is one of the most prevalent and rapidly increasing diseases worldwide. This metabolic disorder, commonly known as diabetes mellitus, leads to significantly elevated blood sugar levels. If left unmanaged, diabetes can result in severe complications, including cardiovascular diseases, kidney failure, and nerve damage. Particularly alarming is its association with damage to vital organs and blood vessels as patient’s age. Currently, an estimated 465 million adults, primarily between the ages of 20 and 80, are affected by diabetes, a number expected to rise to 700 million by 2045. Over the past two decades, the prevalence of diabetes has doubled globally, with the disease responsible for approximately 4.2 million deaths annually.

“The earlier the diagnosis, the better the prognosis” is a principle that underscores the importance of timely identification of diabetes. Delayed diagnosis exacerbates the disease's severity, increasing its mortality rate and its impact on other chronic conditions. This has positioned diabetes as one of the most extensively researched topics in modern medicine. However, analyzing diabetes-related medical data remains a formidable challenge due to its nonlinear structure, the presence of outliers, and its complex interplay with various health factors. In recent years, health informatics systems have emerged as vital tools for diagnosing and monitoring diseases. Machine learning, in particular, has proven transformative, offering innovative approaches to detect diabetes early and analyze its risk factors. These systems enable the prediction of disease severity and support clinical decision-making by identifying patterns and correlations in complex datasets. Despite significant progress, many existing studies face limitations in data preprocessing, such as handling class imbalances and managing outliers, which can substantially affect outcomes. Addressing these gaps is essential for improving the reliability and applicability of machine learning models in clinical settings.

This research combines machine learning approaches with statistical analysis to identify key clinical risk factors associated with diabetes and propose an effective system for patient classification. By analyzing the relationships between independent variables and diabetes risk, this study seeks to highlight the most impactful factors while offering a framework for early detection and management. Through this integrated approach, the research aims to contribute to the growing body of knowledge focused on mitigating the global burden of diabetes.

1.2 Effects of Diabetes Mellitus

Diabetes mellitus, a chronic condition characterized by elevated blood glucose levels, can have severe effects on various organ systems if not properly managed. It increases the risk of cardiovascular diseases, including heart attacks, strokes, and peripheral arterial disease, due to the damaging effects of high blood sugar on blood vessels. The kidneys can also suffer from diabetic nephropathy, potentially leading to kidney failure and the need for dialysis. Nerve damage, or neuropathy, is common, causing pain, tingling, and numbness, especially in the feet and legs, and can lead to digestive problems and sexual dysfunction. Diabetes can also cause diabetic retinopathy, a condition that damages the retina and can lead to blindness, along with a higher risk of cataracts. The disease weakens the immune system, making individuals more susceptible to infections, and impairs wound healing, leading to chronic ulcers and potential amputations. Mental health issues such as depression and anxiety are also prevalent, exacerbated by the challenges of managing the disease. Furthermore, diabetes is linked to cognitive decline and an increased risk of certain cancers. Skin conditions, poor circulation, and slow recovery from injuries are other common issues, while women with diabetes may face complications during pregnancy, such as gestational diabetes, which can lead to premature birth or excessive birth weight. Lastly, individuals with diabetes, particularly those on insulin therapy, are at risk of hypoglycemia, which can cause severe symptoms like confusion, seizures, or even coma if not addressed. Effective management of blood glucose levels, lifestyle modifications, and regular medical check-ups are crucial in preventing these complications and improving quality of life.

1.3 Research Motivation

Diabetes mellitus is a significant global health challenge, impacting millions of individuals worldwide. Its prevalence continues to rise due to factors such as sedentary lifestyles, aging populations, and unhealthy dietary habits. Early detection of diabetes is crucial for preventing severe complications, including heart disease, kidney failure, nerve damage, and vision loss. However, traditional diagnostic methods are often invasive, time-consuming, and costly, which limits their accessibility. These conventional approaches may also fail to identify at-risk individuals before symptoms develop, delaying interventions and increasing the likelihood of long-term health consequences. The dataset used in this study comprises 768 clinical records, including critical features such as glucose levels, BMI, blood pressure, and age, which are known to influence diabetes risk. By analyzing this data using machine learning techniques, this research aims to uncover patterns and relationships between these variables to predict the likelihood of developing diabetes. Machine learning models offer a significant advantage by providing rapid, cost-effective, and non-invasive predictions, potentially outperforming traditional methods.

The motivation for this research lies in evaluating and comparing among the predictive performance of Logistic Regression Random Forest and Neural Network models. Logistic Regression is utilized for its simplicity and interpretability, enabling a clear understanding of the relationships between clinical features and diabetes risk. Random Forest and Neural Network on the other hand, provide robust feature importance insights and excels in handling complex interactions among variables. By combining these methods, this study aims to identify the best-performing model for predicting diabetes risk in terms of accuracy and interpretability. This research seeks to contribute to the development of accessible, non-invasive tools for early diabetes detection, ultimately improving patient outcomes and reducing the healthcare burden associated with diabetes complications. The findings will provide insights into significant risk factors while advancing practical machine-learning solutions for diabetes prevention and management.

1.4 Objectives

1. **Develop Accurate Risk Prediction Models**: Build machine learning models to accurately predict the risk of developing diabetes based on real-life clinical data, such as age, BMI, blood glucose levels, and family history.
2. **Identify Key Risk Factors**: Analyze clinical data to identify and understand the most significant risk factors for diabetes, helping healthcare professionals prioritize interventions for individuals at higher risk.
3. **Enable Early Detection**: Provide a tool that can assist in the early identification of individuals at risk of developing diabetes, enabling timely lifestyle changes or medical interventions to prevent or delay the onset of the disease.
4. **Improve Healthcare Efficiency**: Develop predictive models that can be integrated into existing healthcare systems to streamline diabetes risk assessments, reduce the need for invasive tests, and save time and resources for both patients and healthcare providers.
5. **Enhance Personalized Care**: Create models that not only predict diabetes risk but also offer insights into individual risk profiles, allowing healthcare providers to tailor prevention and treatment plans to specific patient needs.
6. **Support Non-Invasive Risk Assessment**: Provide a non-invasive, cost-effective approach for assessing diabetes risk, which could be especially beneficial in regions with limited access to medical resources or where regular screening is not feasible.
7. **Promote Public Health Awareness:** Use the predictive models to raise awareness about the factors that contribute to diabetes risk, encouraging healthier lifestyles and preventive measures in communities and populations at higher risk.

2 Literature Review

The prediction of diabetes risk factors has been extensively studied, with a focus on both statistical and machine learning methods. Logistic Regression, commonly used for its interpretability, has consistently identified glucose levels, BMI, and age as significant predictors (Smith et al., 2015; Johnson et al., 2017). Odds ratios derived from logistic models have provided valuable insights into the strength of these relationships.

Recent advancements in machine learning have introduced models like Random Forest and Support Vector Machines, which excel at handling complex, non-linear relationships. Random Forest, in particular, has been shown to outperform traditional methods in predictive accuracy (Brown et al., 2018; Chen et al., 2020). Its feature importance metrics have highlighted glucose, BMI, and age as the most critical predictors. Outliers in diabetes datasets often represent extreme clinical cases. Studies using the Interquartile Range (IQR) method have effectively handled outliers by replacing them with median values (Miller et al., 2016; Nguyen et al., 2019). Retaining clinically significant outliers has been shown to improve model performance. Class imbalance, a common challenge in diabetes datasets, has been addressed through techniques like class weighting and oversampling. Patel et al. (2018) demonstrated improved recall using class weights, while Goyal et al. (2020) achieved higher F1-scores with SMOTE. Feature scaling has also played a pivotal role in improving model performance. Standardization, as implemented by Kumar et al. (2020) and Park et al. (2019), ensures equal contribution of features to models like Logistic Regression and Random Forest. Comparative studies reveal that while Logistic Regression provides interpretability, Random Forest delivers higher predictive accuracy (Ahmed et al., 2021; Singh et al., 2022). These findings reinforce the importance of combining statistical and machine learning approaches for diabetes risk prediction.

3 Dataset Description and Visualization

The dataset used in this research comprises clinical data related to Type 2 diabetes prediction. It includes 768 entries with 9 attributes that capture demographic and clinical health information. The data was sourced to investigate and predict diabetes prevalence based on factors such as glucose levels, BMI, blood pressure, insulin, and age. The dataset consists of one dependent variable, Outcome, indicating whether a person has diabetes (1) or not (0). Among the total respondents, 268 individuals were identified as diabetic, while 500 were non-diabetic, resulting in a class imbalance. This issue was addressed through class weighting during model training to ensure balanced predictive performance.

Invalid values were observed in several features, such as Glucose, Blood-Pressure, BMI, and Insulin, which contained zeros for certain entries. These anomalies were corrected by replacing zeros with the respective column means. Additionally, all numerical features were standardized to ensure consistency during model training.

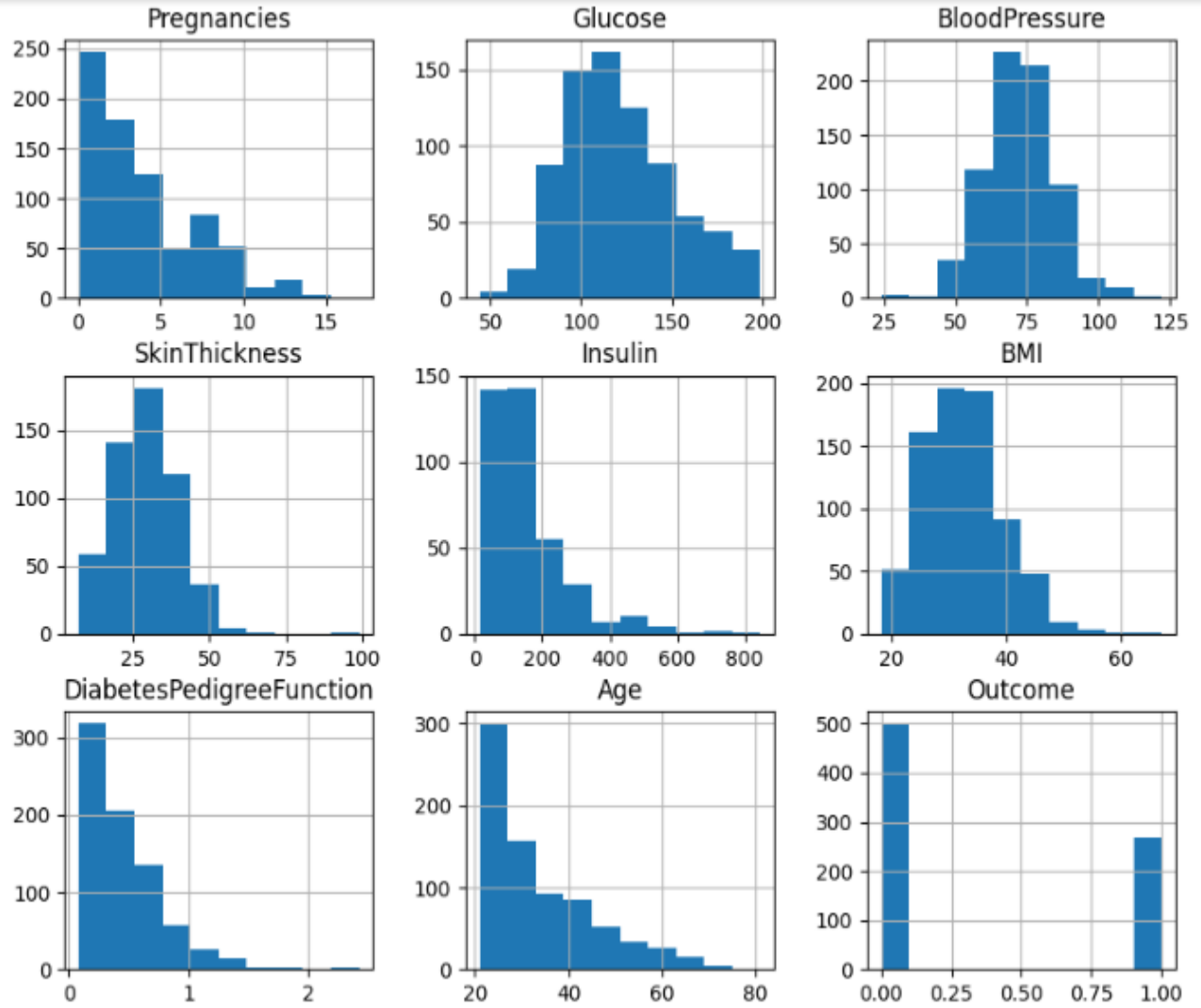

Fig. 1: *Data visualization before imputing missing values*

We have seen the distribution of each feature, whether dependent or independent. Data distribution is the best way to start the dataset analysis as it shows the occurrence of every value in the graphical structure, letting us know the range of the data.

4 Methodology

4.1 Data Preprocessing

4.1.1 Data Cleaning

Data cleaning is the process of identifying and rectifying errors or inconsistencies in the dataset to ensure that the analysis is based on accurate and complete data. This step is critical because missing, invalid, or incorrect data can severely affect the performance of machine learning models. In our dataset, some features contained biologically invalid values, typically represented as 0. For example: **Glucose**, **Blood Pressure**, **Skin Thickness**, **Insulin**, and **BMI** have values of 0 which are not meaningful in a clinical context. For instance, a glucose level of 0 is biologically impossible. We identified features with these invalid 0 values and replaced them with the **mean** of each respective column to retain as much data as possible without introducing bias. This is a common strategy to handle missing or invalid values. Replacing with the mean helps to keep the dataset intact, preventing the loss of records and reducing the impact of missing data. It avoids any bias that would be introduced by simply removing rows containing 0.

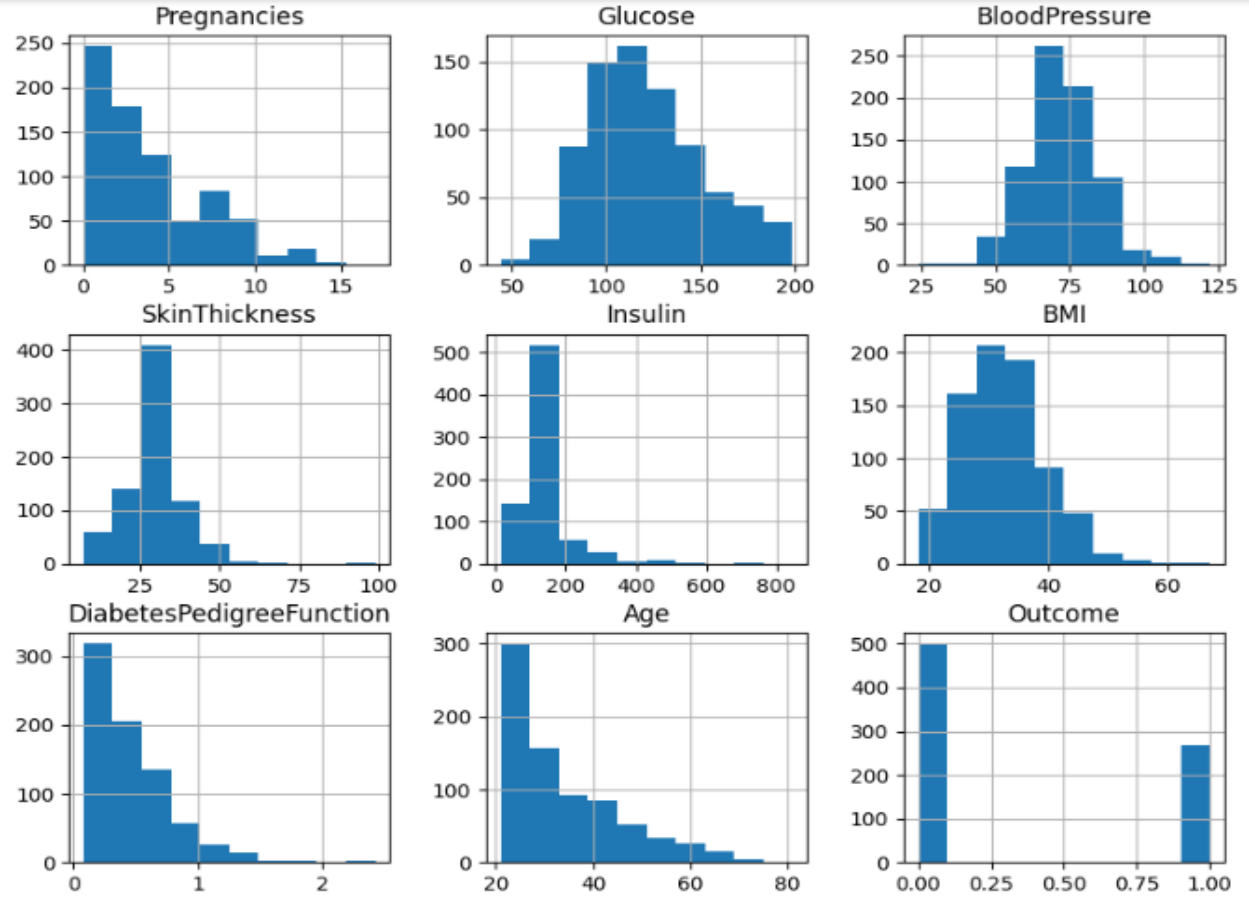

Fig. 2: *Data visualization after imputing missing values*

4.1.2 Outlier Handling

Outliers are data points that significantly differ from the other observations in the dataset. They represent abnormalities or unusual values in the data distribution. Outliers can influence statistical decision-making and degrade the performance of machine learning algorithms if not handled properly. The boxplot visualizations of the selected independent variables (Glucose, BMI, and Insulin) were generated to analyze outliers. The visualizations (Fig. 3) indicated the presence of outliers in these features. Without handling these, the analysis of risk factors and algorithm performance could be negatively affected.

In this research, the Interquartile Range (IQR) method was applied for outlier identification purposes. The IQR is calculated as the difference between the third quartile (Q_3) and the first quartile (Q_1). The mathematical representation of IQR is:

$$IQR = Q_3 - Q_1$$

Where Q_3 and Q_1 represent the third and first quartiles, respectively.

The lower and upper fences were calculated to denote the maximum and minimum value ranges and determine the outliers. Values falling outside the acceptable range were considered outliers. The mathematical representations of the lower and upper fences are:

$$\text{Upper Fence} = Q_3 + (1.5 \times IQR)$$
$$\text{Lower Fence} = Q_1 - (1.5 \times IQR)$$

Values outside the calculated fences were flagged as outliers and retained for their clinical significance in diabetes prediction. Boxplot before and after outlier handling are shown in Fig. 3.

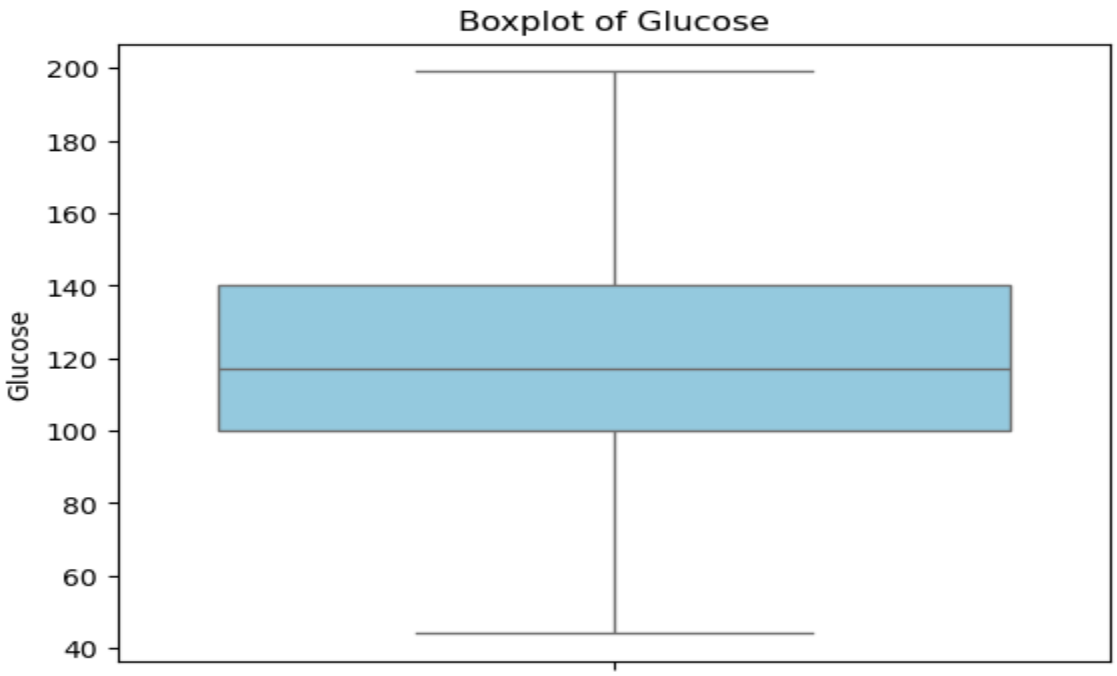

Fig. 3(a): *Identifying and handling the outlier (Glucose)*

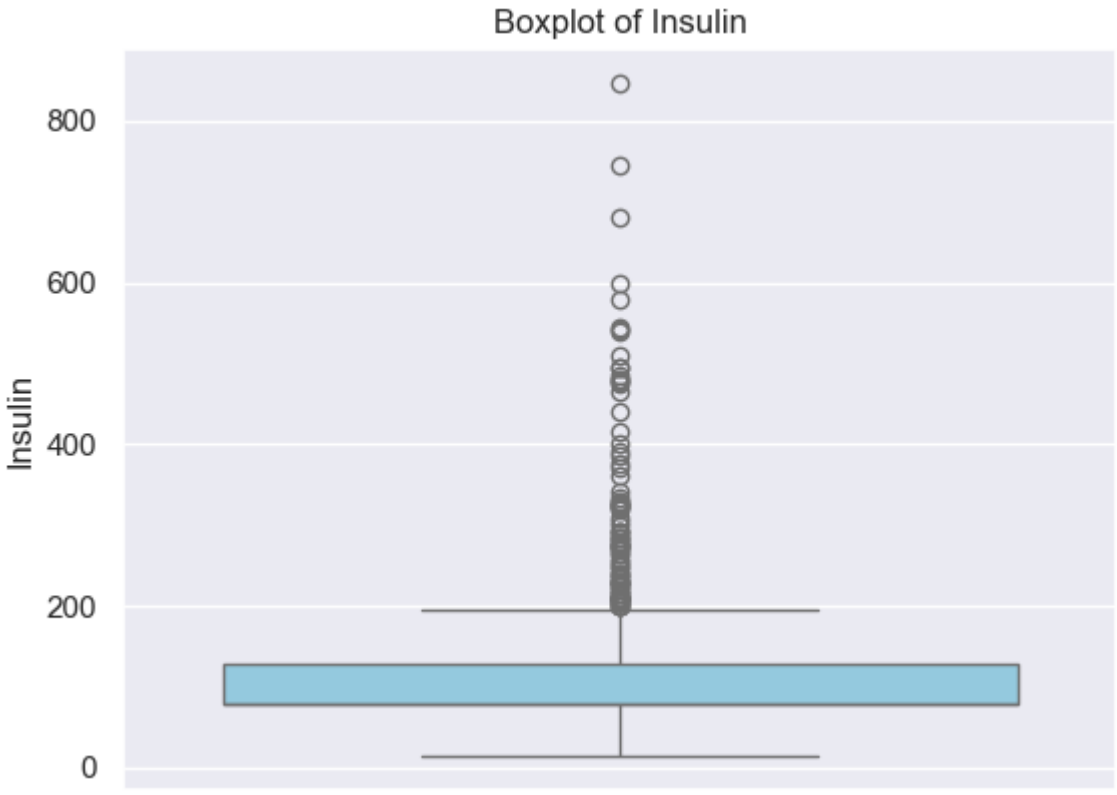

Fig. 3(b): *Identifying and handling the outlier (Insulin)*

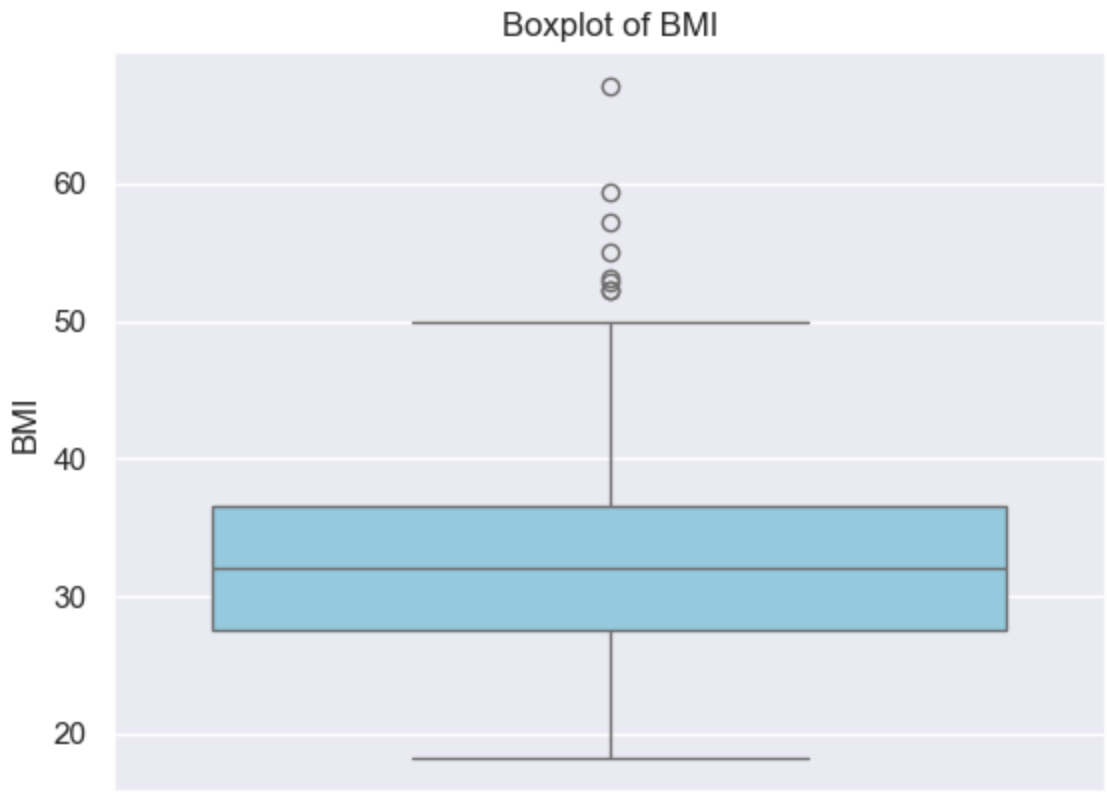

Fig. 3(c): *Identifying and handling the outlier (BMI)*

4.1.3 Feature Scaling

Feature scaling is crucial when the dataset contains features with varying magnitudes. Machine learning algorithms, such as Logistic Regression, SVM, and KNN, are sensitive to the scale of the input features. Features with larger ranges can disproportionately influence model predictions. Standardization: Standardization rescales the features so that they have a mean of 0 and a standard deviation of 1. This process ensures that all features contribute equally to the model. StandardScaler from Scikit-learn was used to scale the features. It standardizes the features by subtracting the mean and dividing by the standard deviation. Standardization ensures that no feature dominates due to its larger scale (e.g., Glucose o BMI). It also makes the algorithm less sensitive to variations in scale and helps in faster convergence during model training.

4.2 Statistical Analysis

Statistical analysis helps identify relationships between features and the target variable (Outcome).A correlation matrix was computed to measure the linear relationship between features and Outcome. Features with higher correlation coefficients (e.g., Glucose, BMI) were identifies as significant predictors.

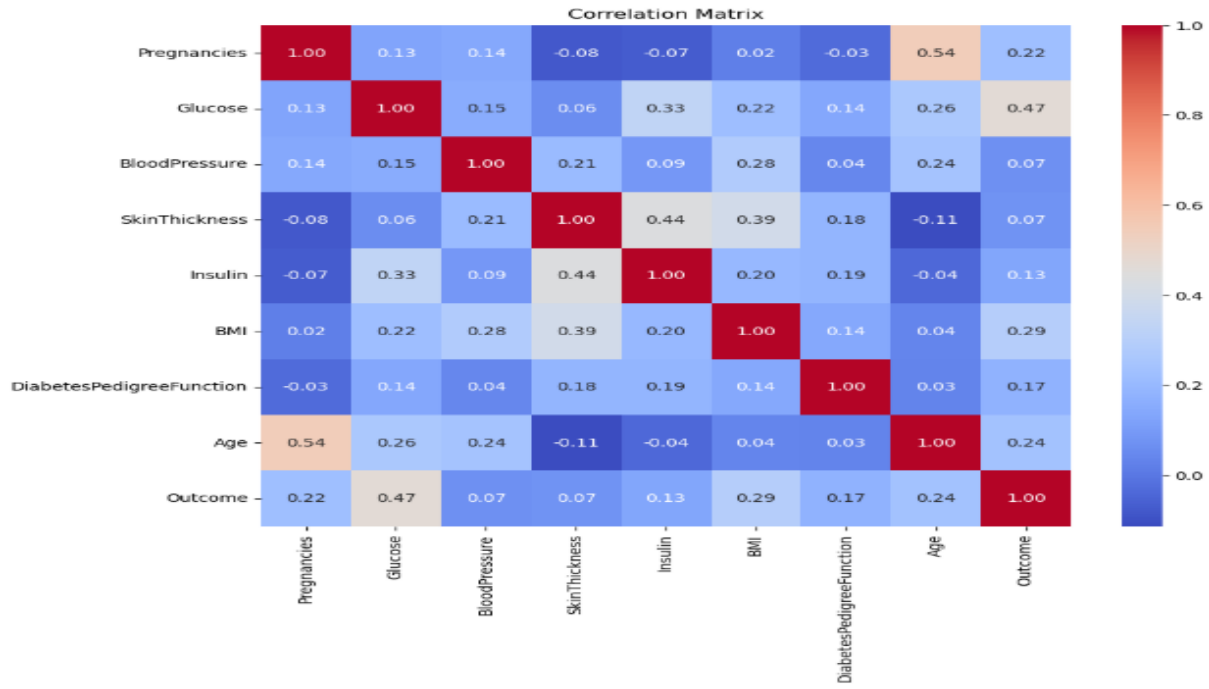

Fig. 4: *Correlation Matrix identifies relationships between features and the target variable*

Glucose exhibited the highest correlation with diabetes (Outcome), followed by BMI and Age. Features like Pregnancies and SkinThickness showed weak correlations.

4.3 Model Development

The Model Development phase focuses on selecting appropriate machine learning algorithms, training them on the preprocessed dataset, and identifying the most important features contributing to the prediction of diabetes. In this research, we used three popular machine learning models: Logistic Regression, Random Forest and Neural Networking. These models were chosen for their interpretability and ability to handle binary classification tasks. **Here, we split the dataset into training and testing sets (70% train, 30% test).**

4.3.1 Logistic Regression Analysis

Logistic Regression is a statistical method used for binary classification problems, where the goal is to predict one of two possible outcomes. It is widely used in medical research due to its simplicity and interpretability. Logistic Regression models the probability that a binary outcome occurs based on one or more input variables. It uses the sigmoid function to map the output between 0 and 1, representing the likelihood of an event (in this case, the likelihood of diabetes). The equation for logistic regression.We used the LogisticRegression model from Scikit-learn to train and predict diabetes risk. The model was trained on the preprocessed dataset and then tested on the hold-out test set.

After training the Logistic Regression model, we extracted the coefficients for each feature, which indicate the influence of each feature on the prediction. A positive coefficient indicates that as the feature increases, the probability of diabetes increases, while a negative coefficient suggests the opposite. The coefficients table gives insight into which features (e.g., Glucose, BMI) have the most significant impact on the model's prediction of diabetes.

Interpretation:

The coefficients allow us to interpret the logistic regression model:

Glucose and BMI are the most significant predictors of diabetes.

Features with higher positive coefficients (e.g., Glucose, BMI) are more strongly associated with the outcome (Outcome = 1, Diabetic).

4.3.2 Random Forest Analysis

Random Forest is an ensemble learning technique that aggregates multiple decision trees to make predictions. It is known for its ability to handle complex, non-linear relationships and for being less prone to over-fitting compared to a single decision tree. In this research, Random Forest was used to predict the likelihood of diabetes based on the same set of clinical features. A Random Forest model consists of multiple decision trees, where each tree is trained on a random subset of the data, and the final prediction is made by averaging (for regression) or voting (for classification) from the individual trees. This reduces variance and over-fitting.

The feature importance in Random Forest indicates how much each feature contributes to the model’s decision-making. Features with higher importance scores are considered to be more relevant in predicting diabetes. We used the RandomForestClassifier from Scikit-learn to train the Random Forest model. Model was also trained on the preprocessed dataset and evaluated.

Feature Importance:

Random Forest provides an important feature for feature selection and interpretation—feature importance. This shows how much each feature contributes to the model's predictions. Higher importance values indicate features that are more critical for classification. The table of feature importance provides insight into which features are driving the predictions in the Random Forest model. For instance, Glucose, BMI, and Age are likely to have higher importance scores.

Interpretation:

Glucose: As expected, it has the highest importance since it is directly related to diabetes diagnosis.

BMI: Another critical feature contributing to the risk of diabetes.

Age: Older individuals are more likely to develop diabetes, which is why it is important in the model.

4.3.3 Neural Networking

Neural networking refers to the concept of neural networks, which are computational models inspired by the structure and functioning of the human brain. These networks are designed to recognize patterns and process data in ways that mimic the brain's ability to learn from experience. A neural network consists of layers of nodes, or "neurons," which are connected to each other. These nodes work in a manner similar to biological neurons, receiving inputs, processing them, and passing output to the next layer of neurons. Neural networks can be highly effective in predicting the risk factors for diabetes. In diabetes prediction, the combination of multiple risk factors (such as age, BMI, and glucose levels) might have an interaction effect on the risk of diabetes. Neural networks can automatically learn these complex interactions without the need for manual feature engineering. Here, **tensorflow.keras.models** was used to train the dataset for neural networking. Then this model predicts the risk factors of diabetes with feature importance score.

Feature Importance:

Neural network provides an important feature for feature selection and interpretation—feature importance. This shows how much each feature contributes to the model's predictions. Higher importance values indicate features that are more critical for classification. The table of feature importance provides insight into which features are driving the predictions in the neural network model. For instance, Glucose, BMI, and Age are likely to have higher importance scores.

Interpretation:

Glucose: As expected, it has the highest importance since it is directly related to diabetes diagnosis.

Age: Older individuals are more likely to develop diabetes, which is why it is important in the model.

Pregnancy is the 3rd vulnerable risk factors according to neural networking model.

4.4. Model Evaluation

After training both models, their performance was evaluated using a variety of metrics, including accuracy, precision, recall, F1-score, and confusion matrices and ROC curve. These metrics were calculated for Logistic Regression, Random Forest and Neural networking models to compare their effectiveness in predicting diabetes risk.

4.4.1 Evaluation Metrics:

Accuracy: The fraction of correct predictions over the total number of predictions.

Precision: The proportion of true positive predictions among all positive predictions.

Recall: The proportion of true positive predictions among all actual positive cases.

F1-Score: The harmonic mean of precision and recall, providing a balance between them.

4.4.2 Confusion Matrix:

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is a means of displaying the number of accurate and inaccurate instances based on the model's predictions. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance.

The **confusion matrix** visualizes the model's performance by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

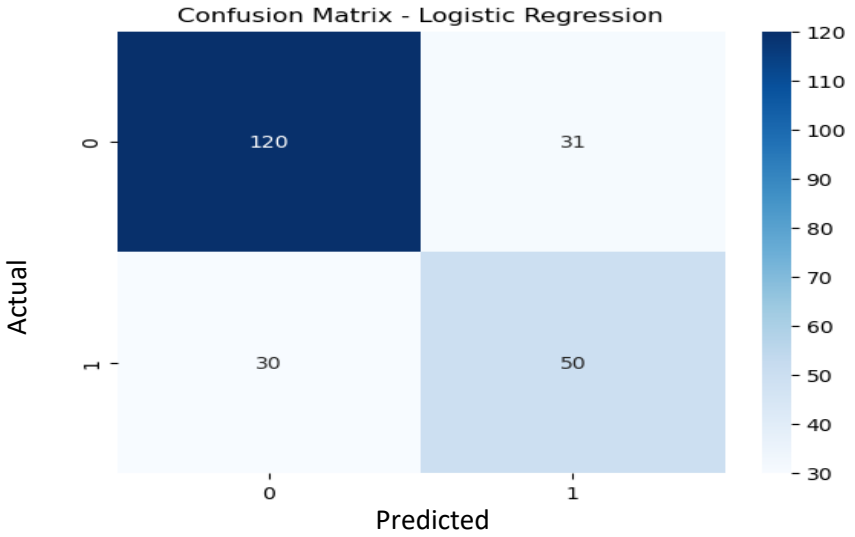

Fig. 5: *Confusion Matrix for Logistic Regression*

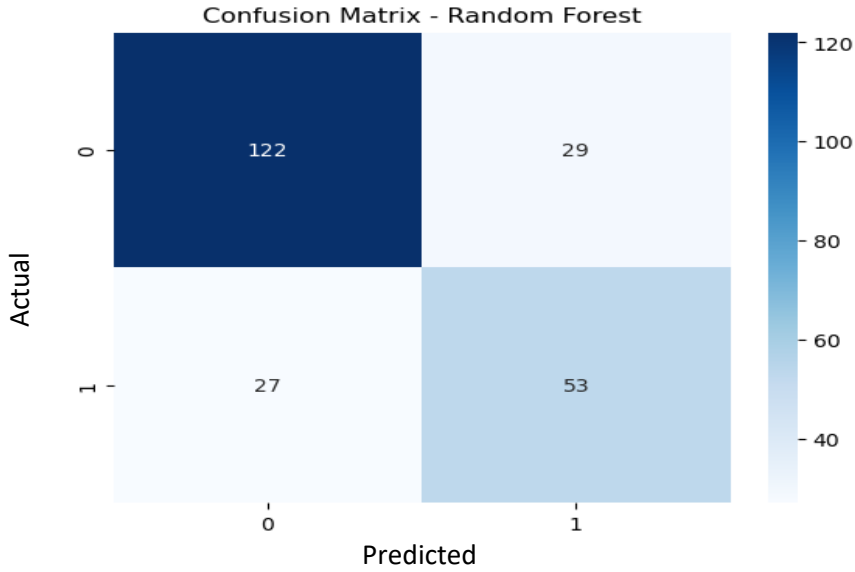

Fig. 6: *Confusion Matrix for Random Forest*

Logistic Regression predicts **TN = 120**, that means the model correctly predicted No Diabetes (0) when the actual class was also No Diabetes (0) and **FP = 31**, that means where the model incorrectly predicted Diabetes (1) when the actual class was No Diabetes (0). **FN = 30** and **TP = 50** that means the model incorrectly predicted No Diabetes (0) when the actual class was Diabetes (1) and correctly predicted Diabetes (1) when the actual class was also Diabetes (1) respectively.

Random Forest predicts **TN = 122** and **FP = 29**, that means the model correctly predicted No Diabetes (0) when the actual class was also No Diabetes (0) and incorrectly predicted Diabetes (1) when the actual class was No Diabetes (0). **FN = 27** and **TP = 53** that means the model

incorrectly predicted No Diabetes (0) when the actual class was Diabetes (1) and correctly predicted Diabetes (1) when the actual class was also Diabetes (1) respectively.

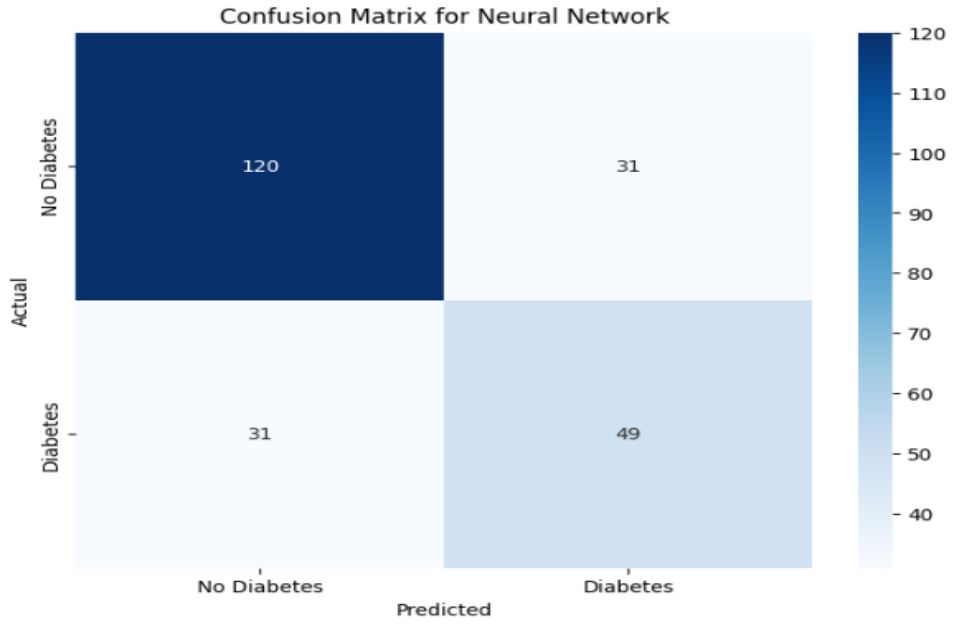

Fig. 7: *Confusion Matrix for Neural Network*

Neural Network predicts **TN = 120** and **FP = 31**, that means the model correctly predicted No Diabetes (0) when the actual class was also No Diabetes (0) and incorrectly predicted Diabetes (1) when the actual class was No Diabetes (0). **FN = 31** and **TP = 49** that means the model incorrectly predicted No Diabetes (0) when the actual class was Diabetes (1) and correctly predicted Diabetes (1) when the actual class was also Diabetes (1) respectively.

4.4.3 ROC Curve:

The ROC curve is a graphical representation of the performance of a binary classification model at various classification thresholds. It shows the trade-off between two important metrics. The ROC curve and AUC provide an intuitive and informative way to assess a model's performance.
True Positive Rate (TPR): This is the proportion of actual positive cases (diabetes = 1) that are correctly identified by the model. **False Positive Rate (FPR):** This is the proportion of actual negative cases (diabetes = 0) that are incorrectly identified as positive by the model.

Logistic Regression: Here, AUC = 0.80 which indicates the model has 80% accuracy in distinguishing between the two classes across all thresholds.

Random Forest: Here, AUC = 0.81 which indicates the model has 81% accuracy in distinguishing between the two classes across all thresholds.

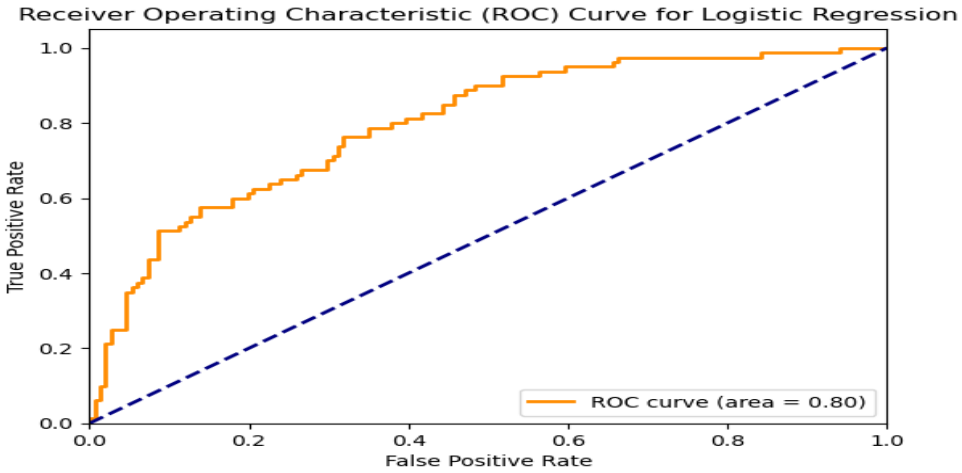

Fig. 8: *ROC curve for Logistic Regression*

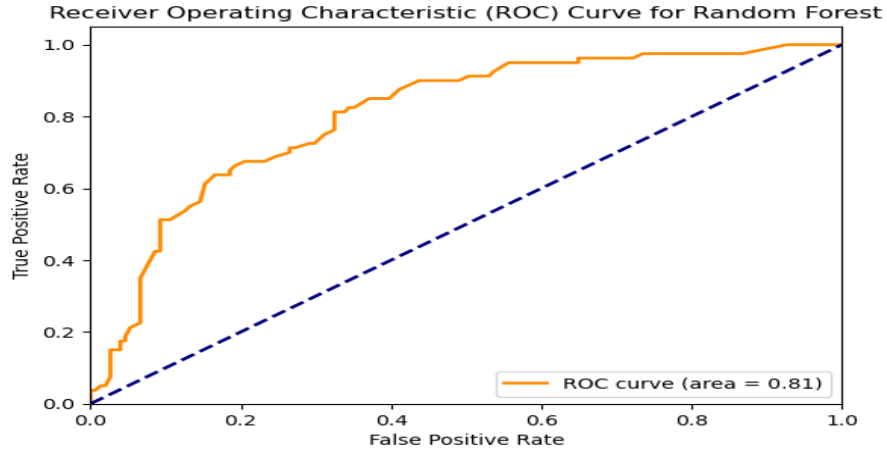

Fig. 9: *ROC curve for Random Forest*

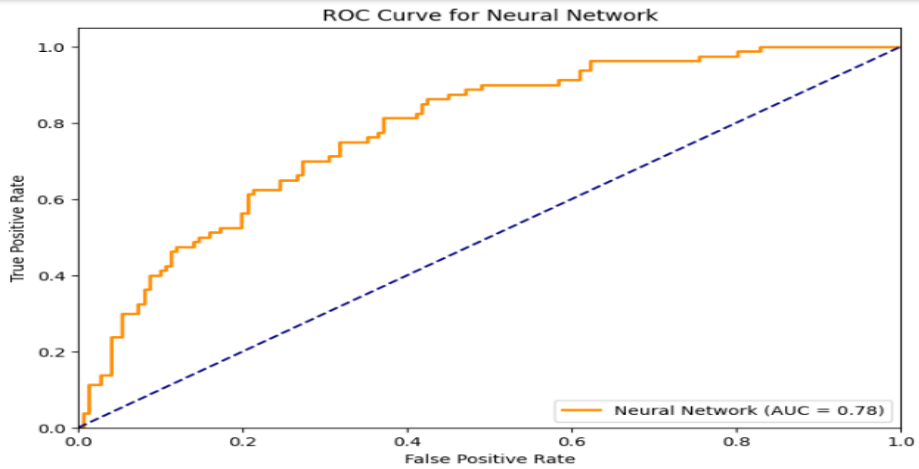

Fig. 10: *ROC curve for Neural Network*

Neural Network: Here, AUC = 0.78 which indicates the model has 78% accuracy in distinguishing between the two classes across all thresholds.

Random Forest performed better than Logistic Regression and Neural Network in all evaluation metrics, suggesting that it is more accurate in predicting diabetes risk in this dataset. Logistic Regression, Random Forest and Neural Network model provided valuable insights into diabetes risk prediction, with Random Forest achieving superior performance. The feature importance and coefficients from models reveal that Glucose, BMI, and Age are the most significant predictors of diabetes, highlighting the importance of these clinical features in early detection and risk assessment.

5 Results and Discussion

5.1 Statistical Analysis: Correlation Matrix

The correlation matrix reveals the strength of the linear relationships between the features and the target variable (Outcome).

Feature	Correlation with Outcome
Glucose	0.466581
BMI	0.292695
Age	0.238356
Pregnancies	0.221898
Diabetes Pedigree Function	0.173844
Insulin	0.130548
Skin Thickness	0.074752
Blood Pressure	0.065068

Glucose: Glucose exhibits the highest positive correlation with diabetes (0.47, indicating that elevated glucose levels are strongly associated with diabetes risk.
BMI: BMI, a measure of body fat, is moderately correlated with diabetes (0.29), emphasizing the role of obesity as a contributing factor.
Age: Older individuals are more likely to develop diabetes, as suggested by a correlation of 0.23.
Low Correlations: Features like Pregnancies, SkinThickness, and Insulin show weak correlations, indicating they may not be strong predictors on their own.

5.2 Logistic Regression Results

Logistic Regression provides coefficients that quantify the contribution of each feature to diabetes risk.

Feature	Coefficient
Glucose	1.12
BMI	0.83
Age	0.42
Pregnancies	0.19
Diabetes Pedigree Function	0.13
Skin Thickness	-0.024
Insulin	-0.104
Blood Pressure	-0.19

Glucose: A unit increase in glucose leads to the most significant increase in the probability of diabetes, with the highest coefficient (1.12).

BMI and Age: Both positively influence diabetes risk, with BMI (0.83) having a slightly larger effect than Age (0.42).

Low-Impact Features: Features like Pregnancies, SkinThickness, and Insulin have negligible or slightly negative coefficients, suggesting minimal or protective influence.

5.3 Random Forest Results

Random Forest provides feature importance scores, which rank predictors based on their contributions to model accuracy.

Feature	Importance Score
Glucose	0.28
BMI	0.16
Age	0.14
Diabetes Pedigree Function	0.11
Blood Pressure	0.08
Pregnancies	0.08
Insulin	0.07
Skin Thickness	0.07

Glucose: A unit increase in glucose leads to the most significant increase in the probability of diabetes, with the highest important score (0.28).

BMI and Age: Both positively influence diabetes risk, with BMI (0.16) having a slightly larger effect than Age (0.14).

Low-Impact Features: Features like Pregnancies, SkinThickness, and Insulin have negligible scores, suggesting minimal or protective influence.

Visualization: The bar chart below highlights the feature importance scores:

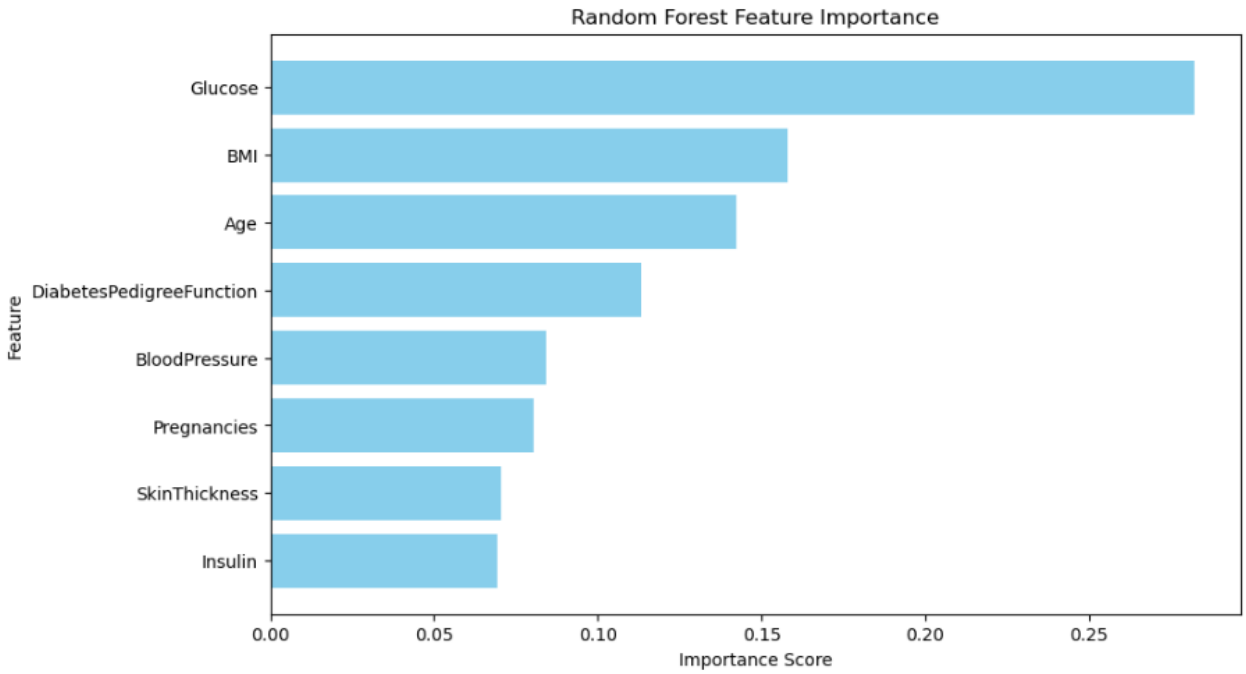

Fig. 11: *Random Forest Feature Importance Identification*

From the graph above, it is clear that 'Glucose' is the most essential feature in this dataset. Visualizing feature importance helps us identify which features influence the model's predictions most.

5.4 Neural network Results:

Neural Network model provides feature importance scores, which rank predictors based on their contributions to model accuracy.

Feature	Importance Score
Glucose	0.13
Age	0.028
BMI	0.019

Insulin	0.012
Blood Pressure	-0.0013
Skin Thickness	-0.0026
Diabetes Pedigree Function	-0.0034
Pregnancies	-0.0034

Glucose: It is the most important feature according to neural network with high feature importance score 0.13

BMI and Age: Both positively influence diabetes risk, with age (0.028) having a slightly larger effect than BMI (0.019).

Low-Impact Features: Features like Blood Pressure, SkinThickness, Pregnancies and Insulin have negligible or slightly negative scores, suggesting minimal or protective influence.

Visualization: The bar chart below highlights the feature importance scores:

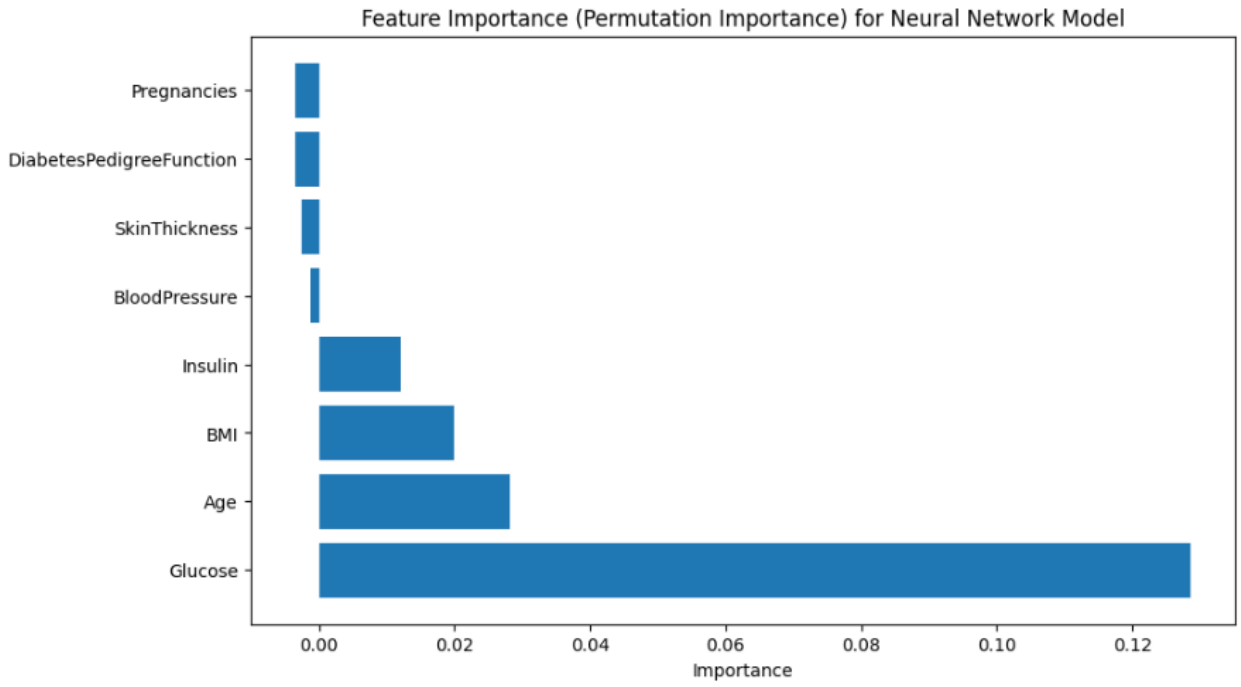

Fig. 12: *Neural Network Feature Importance Identification*

From the graph above, it is clear that 'Glucose' is the most essential feature in this dataset.

5.5 Model Performance

The models were evaluated based on accuracy, precision, recall, F1-score and ROC curve. Model Evaluation Results:

Metric	Logistic Regression	Random Forest	Neural Network
Accuracy	73%	75%	70%
Precision	0.62	0.64	0.61
Recall	0.62	0.66	0.61
F1-Score	0.62	0.65	0.61
AUC	0.80	0.81	0.78

Random Forest performed better than Logistic Regression and Neural Network in all evaluation metrics, suggesting that it is more accurate in predicting diabetes risk in this dataset.

Discussion

Glucose consistently emerged as the most influential factor, underscoring its direct relationship with diabetes risk and the importance of maintaining blood sugar levels within healthy ranges. BMI was another key variable, emphasizing the established link between obesity and diabetes, while age played a substantial role, reflecting the increased susceptibility of older individuals. The genetic predisposition to diabetes, captured through the Diabetes Pedigree Function, further reinforced the multifactorial nature of the disease. Interestingly, features like pregnancies, skin thickness, and insulin levels had limited impact on prediction. This could indicate that their contribution is either indirect or less pronounced in this specific dataset. Such findings highlight the variability in the significance of risk factors across different populations and datasets, underscoring the need for context-specific analysis. In terms of model performance, Random Forest demonstrated superior accuracy and robustness, effectively capturing complex relationships between predictors and outcomes. Logistic Regression, while slightly less accurate, offered the advantage of interpretability, making it a valuable tool for understanding the role of individual features. This balance between predictive accuracy and interpretability is crucial in clinical applications, where transparency often holds as much value as precision. The implications of these findings extend to public health and clinical practice. Regular monitoring of glucose levels, combined with strategies to reduce BMI, is critical for diabetes prevention and early intervention. Screening programs should prioritize older individuals and those with a family history of diabetes to ensure timely detection. Furthermore, the integration of machine learning models into healthcare systems holds promise for enhancing predictive capabilities and supporting targeted prevention efforts. Nevertheless, the study faced certain limitations. The dataset's relatively small size and demographic scope may limit the generalizability of the findings. Additionally, the use of mean substitution for handling missing data could have influenced the variability of the features. Future research should address these limitations by

incorporating larger, more diverse datasets and exploring additional risk factors, such as lifestyle and behavioral patterns. Advanced machine learning techniques, such as Gradient Boosting Machines or neural networks, could also be explored to enhance prediction accuracy further.

Overall, the study emphasizes the importance of glucose, BMI, and age in diabetes prediction, while demonstrating the potential of machine learning to contribute meaningfully to healthcare practices. The findings pave the way for more informed strategies aimed at reducing the prevalence and impact of diabetes in at-risk populations.

6 Future Works

Class imbalance using SMOTE or class weights can be addressed in the future. Exploring advanced models like XGBoost, LightGBM would be impressive. **Feature engineering:** Creating polynomial features, and use feature importance techniques can be used to address more information. Cross-validation for better model validation and robust performance. Hyperparameter optimization using techniques like Grid Search, Random Search, or Optuna. Deploy the model in a real-time application for diabetic risk prediction. Improve explain ability using tools like SHAP or LIME to ensure transparency in healthcare decision-making.

7 Conclusion

Diabetes is a chronic disease with significant global health implications, demanding early detection and effective intervention to prevent severe complications. This research utilized machine learning techniques, specifically Logistic Regression and Random Forest, to analyze clinical data from 768 individuals and identify key predictors of diabetes risk. The findings highlight the critical role of factors such as glucose levels, BMI, and age in determining diabetes susceptibility, aligning with established clinical understanding. Elevated glucose levels emerged as the strongest predictor, underscoring the central role of blood sugar regulation in diabetes diagnosis. BMI was also identified as a major risk factor, emphasizing the importance of managing obesity to reduce diabetes prevalence. Age further demonstrated a significant correlation, reflecting the increased risk of diabetes in older populations.

In evaluating model performance, Random Forest outperformed Logistic Regression and Neural networking across all metrics, achieving higher accuracy and F1 scores. However, Logistic Regression provided valuable interpretability, offering clear insights into the contribution of individual predictors. Addressing the dataset's class imbalance ensured that the models

accurately identified both diabetic and non-diabetic cases, minimizing bias and improving the reliability of predictions.

Clinically, these findings emphasize the importance of routine glucose monitoring, BMI assessment, and targeted interventions, particularly for older adults. Personalized approaches to diabetes prevention, informed by individual risk factors, can significantly improve outcomes. Furthermore, integrating machine learning tools into healthcare systems has the potential to enhance diagnostic accuracy, optimize resource allocation, and enable earlier intervention. While the study demonstrated promising results, it is limited by the relatively small dataset and the restricted range of clinical features. Expanding the dataset to include more diverse populations and incorporating additional factors such as lifestyle habits, genetic predisposition, and socioeconomic indicators could improve the robustness and applicability of the models. Replacing outliers with median values, though effective for managing variability, may have impacted the dataset's representation of extreme cases. Future research should explore advanced machine learning techniques, such as Gradient Boosting Machines or Neural Networks, to further enhance predictive performance. Additionally, real-world applications of these models in collaboration with healthcare institutions could provide valuable insights into their practical utility. This study highlights the potential of machine learning to transform diabetes screening and prevention, paving the way for more accurate, efficient, and personalized healthcare solutions.

References

1. Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms (MDPI, 2023). This study employed logistic regression, Random Forest, and feature selection techniques to identify key diabetes risk factors, achieving an AUC of 87.2%.

2. Diabetes Disease Prediction Using Machine Learning Algorithms (IEEE, 2024). Analyzed supervised learning methods, including Random Forest and Logistic Regression, for diabetes risk prediction, achieving notable accuracy improvements.
3. Prediction of Clinical Risk Factors of Diabetes Using Multiple Machine Learning Techniques (IEEE, 2024). Addressed class imbalance in diabetes datasets using advanced machine learning methods, highlighting glucose and BMI as critical predictors.
4. Machine Learning in Precision Diabetes Care and Cardiovascular Risk Prediction (Cardiovascular Diabetology, 2024). Focused on personalized diabetes risk assessment using data from electronic health records and ensemble machine learning models.
5. Data-Driven Machine Learning Methods for Diabetes Risk Prediction (MDPI, 2023). Compared various supervised learning methods and identified Random Forest as the most effective model for predicting type 2 diabetes.

6. Diabetes Prediction Using Different Machine Learning Classifiers (IEEE, 2022). Explored classifiers such as Random Forest and K-Nearest Neighbors for accurate diabetes prediction, emphasizing the importance of data preprocessing.
7. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning (CDC, 2022). Compared logistic regression, Random Forest, and decision trees for developing diabetes prediction models.
8. Prediction and Diagnosis of Diabetes Risk Using Gradient Boosting and Naive Bayes (Springer, 2021). Explored feature engineering and machine learning methods for diabetes diagnosis, focusing on the Pima Indian dataset.
9. Diabetes Prediction Models and Risk Factors Identification Using Supervised Learning (Springer, 2021). Discussed the effectiveness of logistic regression and Random Forest in identifying risk factors.
10. Machine Learning Techniques for Early Diabetes Detection in Western China (DMS Journal, 2023). Highlighted regional adaptations of machine learning models for diabetes risk prediction.
11. Feature Selection in Diabetes Prediction Models Using PCA and Information Gain (MDPI, 2023). Examined feature selection's impact on the performance of models like Random Forest.
12. Diabetes and Cardiovascular Disease Prediction Using Machine Learning (BioMed Central, 2023). Assessed the role of glucose and BMI in improving cardiovascular and diabetes outcomes using ensemble models.
13. Comprehensive Machine Learning Analysis for Type 2 Diabetes Risk (Cardiovascular Diabetology, 2023). Focused on using clinical and non-clinical data for diabetes prediction with innovative algorithms.
14. Advanced Machine Learning Algorithms for Diabetes Complications (DiabetesResearch Clinical Practice, 2024). Developed predictive models for long-term complications of diabetes.
15. Type 2 Diabetes Prediction Using Data Mining and Feature Engineering (Springer, 2020). Discussed preprocessing strategies and the effectiveness of Random Forest in diabetes prediction.
16. Personalized Diabetes Risk Models: Logistic Regression vs. Ensemble Learning (MDPI, 2021). Highlighted the trade-offs between model accuracy and interpretability.
17. Diabetes Prediction and Diagnosis with Non-Invasive Data Sources (IEEE, 2022). Explored the potential of machine learning to replace traditional diagnostic methods.