EVALUATING LOGISTIC REGRESSION AND RANDOM FOREST MODELS FOR PREDICTING DIABETES

Oluwaseun A. Adesina, Israel K. Adebayo, Taiwo J. Adejumo and Adedamola T. Bobade

[1,3&4]*Ladoke Akintola University of Technology, Ogbomoso, Nigeria*

[2] *Department of Mathematics and Statistics, Osun State College of Technology, Esa-Oke*

Department of Statistics

Corresponding Author: oaadesina26@lautech.edu.ng

RSS-NLG 2025 Conference Proceedings

Abstract:

Diabetes poses a significant global health challenge, necessitating effective predictive models for early diagnosis and intervention. This study evaluates the performance of logistic regression and random forest models using a dataset comprising information of 390 respondents which was extracted from the data. world to predict diabetes based on health biomarkers such as cholesterol levels, glucose concentrations, BMI, and blood pressure. Results indicate high performance for both models, with the logistic regression model achieving an Accuracy of 91%, Precision of 94%, Sensitivity of 95%, and Specificity of 75%. The random forest model yielded an Accuracy of 89%, Sensitivity of 92%, Precision of 93%, and a Similar Specificity of 75%. The logistic regression model outperforms the random forest in Accuracy, Precision, and Sensitivity, showing greater efficacy in distinguishing between diabetic and non-diabetic individuals.

Keywords: Diabetes, predictive models, logistic regression, Accuracy, random forest

1. INTRODUCTION

Diabetes mellitus is a chronic disease that affects millions of people of all ages. It is a condition that impairs the body's ability to produce or use insulin, resulting in high blood sugar levels. Azevedo & Alla (2008) emphasized that diabetes as a chronic disease affects millions of people worldwide, with type 2 diabetes being the most prevalent form. In Africa, the prevalence of diabetes has been increasing steadily since the mid-1980s. The majority of diabetes cases in Africa are of type 2, with genetic predisposition being one of the major contributing factors. Other factors include environmental factors, diet, lifestyle, and residence. Diabetes has raised a significant public health concern in Africa due to the disease's complications and associated morbidity and mortality.

Menezes et al. (2014) focused on sub-Saharan Africa, where diabetes is projected to impact over 20 million people by 2030. In Nigeria, prevalence rates range from 0.6% to 11.0%, highlighting the growing burden on patients, their families, and healthcare systems. Siegel et.al (2018), their study examined lifestyle behaviors among American adults without type 2 diabetes that are known to reduce the risk of developing the disease. The study found that only a small proportion of American adults engage in risk-reduction practices, with just 3.1% meeting most of the recommended guidelines. Additionally, younger individuals and those with lower education levels were associated with a lower likelihood of achieving these goals.

The exact cause of diabetes is not yet known, but several factors can contribute to its development. Age is a major factor, as the risk of developing diabetes increases with age. Family history is also important, as those with a family history of diabetes are more likely to develop the disease themselves. Pregnancy can also increase the risk of developing diabetes, especially in women who had gestational diabetes during pregnancy. Fluctuating glucose levels can also play a role, as repeated episodes of high blood sugar can damage the pancreas and reduce insulin production. High blood pressure is another factor that can increase the risk of developing diabetes. Managing diabetes requires careful monitoring of blood sugar levels, adherence to a healthy diet and exercise routine, and often medication or insulin therapy. It's important to stay informed about this condition and work closely with a healthcare professional to manage and prevent complications associated with diabetes (Ahamed et.al (2022).

Diabetes, spanning type 1 and type 2 variants and affecting all age groups which poses chronic tissue damage risks. Decision trees, random forests, and neural networks utilized physical examination data to predict diabetes, while methods such as Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) reduce data complexity. This approach highlights the importance of early detection and management of diabetes complications across different age groups, which can improve patient outcomes and lessen long-term health risks The study evaluates how well classifiers predict diabetes based on blood glucose levels (Zou et al., 2018). The exceeding prior predictions say that diabetes in the US is increasing rapidly, where a Markov model was used to project the prevalence of diagnosed diabetes through 2060 and found that the number of people with diabetes increased from 22.3 million in 2014 to 39.7 million in 2030 and on 60.6 million in 2060. Their estimates could help plan health services and public health programs to reduce the future burden of diabetes. Lin et.al (2018).

RSS NIG 2025 Conference Proceedings

Lai et al. (2019) developed a predictive model to identify Canadian patients at risk for diabetes mellitus, a condition characterized by impaired glucose metabolism. They analyzed data from 13,309 patients aged 18 to 90, including demographics and lab results such as fasting blood glucose, BMI, HDL, triglycerides, blood pressure, and LDL. Using logistic regression and gradient boosting machine (GBM), the models were evaluated by AROC and sensitivity—with the GBM achieving an AROC of 84.7% (71.6% sensitivity) and logistic regression an AROC of 84.0% (73.4% sensitivity). Both models outperformed other techniques like Decision Trees and Random Forest, highlighting fasting blood glucose, BMI, HDL, and triglycerides as key predictors. The study supports integrating these models into online tools to help doctors predict diabetes risk, and its Canadian validation enhances its robustness compared to models developed for other populations.

Tan et al., (2021) shows a comprehensive systematic review of 32 studies, they examined the efficacy of machine learning (ML) models in predicting both microvascular and macrovascular complications in Type 2 diabetes patients. The review included examination of 87 ML models, with neural networks emerging as the most frequently utilized. Key predictors such as age, duration of diabetes, and body mass index were commonly integrated into these models. Notably, approximately 36% of the evaluated ML models demonstrated significant discrimination ability. Among the various ML algorithms, random forest exhibited the most promising overall performance. However, it's significant that the majority of the studies (31 out of 32) were found to have a high risk of bias, suggesting the imperative need for external validation studies. This highlights the need for thorough validation to ensure reliability and effectiveness before integrating ML-based prediction models for diabetes complications into clinical practice.

Wen et al. (2021) evaluated the effectiveness of machine learning in predicting recurrence risk among diabetic patients receiving team-based nursing care. Although long-term hyperglycemia in younger and middle-aged patients can lead to serious complications such as diabetic ketoacidosis, myocardial infarction, and infections, the study found that machine learning did not significantly improve diabetes knowledge, blood glucose management, or patients' psychological well-being. However, predictive models using random forest (RF) and logistic regression showed high accuracy, with RF at 81.46% and logistic regression at 80.21%. Despite the minimal impact on

RSS-NLG 2025 Conference Proceedings

patient outcomes, the findings suggest that machine learning could be integrated into clinical decision-making to personalize care and reduce long-term complications in this population.

Artificial intelligence, especially random forest (RF), has shown great potential in accurately predicting changes in glycated hemoglobin A1c (HbA1c) levels. RF models, which analyze health check-up data, outperform traditional methods in predictive accuracy. This approach is particularly effective in identifying key disease risk factors, underscoring its value in personalized medicine. Utilizing AI techniques like RF can enhance early intervention strategies for Type 2 diabetes, enabling customized treatments that slow disease progression and improve patient outcomes (Ooka et al., 2021).

A study approved by AJA University of Medical Sciences ethical committee, revealed a 3 lower prevalence of Type 2 Diabetes Mellitus (T2DM) compared to the general population. Incidence was notably higher among older individuals and staff members. T2DM prevalence correlated with obesity and high lipid profiles, particularly elevated total cholesterol, low-density lipoprotein cholesterol, and triglyceride levels. Key risk factors identified included age, body mass index, and lipid profile markers. The study highlights the importance of early identification and management strategies tailored to high-risk individuals (Sahebhonar et al. 2022).

Chen et al. (2023) conducted a comprehensive evaluation of machine learning models for predicting diabetic kidney disease (DKD). By systematically reviewing large databases, they compared various ML techniques. Logistic regression (LR) emerged as the most commonly used method, achieving a pooled AUROC of 0.83—comparable to that of non-LR models, which also reached a pooled AUROC of 0.83. Statistical analysis indicated no significant difference in predictive performance between LR and non-LR models, with all ML approaches showing satisfactory results (AUROC values above 0.7). Due to its simplicity and computational efficiency, LR stands out as a practical option. Overall, the study highlights the effectiveness of ML in forecasting DKD, with LR offering notable practical benefits.

Shojaee et al., (2024) their study aimed to develop a machine learning model that could predict the fasting blood glucose status of individuals. The data for the study was obtained from 3376 adults over 30 years old who had participated in a diabetes screening program in Tehran, Iran. The dataset included a range of variables, such as age, gender, waist-to-hip ratio, body mass index, systolic blood pressure, and other medical parameters. The study found that several factors were crucial in

RSS-NLG 2025 Conference Proceedings

predicting fasting blood glucose status, with age, waist-to-hip ratio, body mass index, and systolic blood pressure being the most important. They used various machine learning algorithms to develop the model, and the Cat Boost algorithm performed the best with an AUC of 0.737. This model can be used to help with diabetes management and prevention planning, as it can accurately predict the fasting blood glucose status of an individual. By identifying individuals who are at high risk of developing diabetes, healthcare providers can take proactive measures to help prevent the onset of the disease. Additionally, the model can be used to optimize diabetes management for those who have already been diagnosed, by providing personalized treatment plans based on their individual fasting blood glucose levels.

This paper aimed to evaluate the performance of logistic regression and random forest models in predicting diabetes using health biomarkers, including cholesterol levels, glucose concentrations, BMI, and blood pressure.

The rest of this paper is organised as follows; section 2 discusses the materials and methodology used which comprises the Logistic regression model, theoretical concepts of the methods employed namely, Logistic Regression Odd Ratio, Odd Event, Random Forest Model, Classification Metrics such as sensitivity, specificity, precision and accuracy to evaluate the performance of the classifier models are briefly discussed, empirical illustrations, findings and discussions follow in Section 3 and Section 4 concludes the paper.

RSSNLG 2025 Conference Proceedings

2.0 MATERIALS AND METHOD

The data used is secondary data obtained from https://data.world/informatics-edu/diabetes-prediction/workspace/file?filename=Diabetes_Classification.xlsx. The explanatory variables (X) used in this study are cholesterol level, glucose concentration, BMI, and blood pressure while the dependent (Y) variable is the target (diabetes).

2.1 LOGISTIC REGRESSION MODEL

Logistic regression uses the natural logarithm of chances (logit) to express the relationship between the outcome variable and predictor factors (independent variables). In this scenario let X be a continuous predictor variable and y, be a dichotomous outcome variable with categories of “1” and “0”. Now, if we create a scatter plot, each outcome variable category will be represented by two parallel lines. Since the link does not exhibit a linear trend, a straightforward linear regression cannot be used to explain it. By applying a logit transformation to the result variable Y, logistic regression makes this scenario easier. The most straightforward logistic regression model is expressed as

$$\textbf{Logit(Y)} = \frac{\pi}{1-\pi} = \beta_0 + \beta_1 \qquad (1)$$

In the logistic regression equation, π represents the probability of the outcome Y, and π(1-π) represents the odds of success. The intercept and slope (regression coefficient) are denoted by β_0 and β_1 respectively. To estimate the probability of Y for a given value of predictor X, we can take antilog of both sides of the equation (1).

$$\boldsymbol{\pi} = (\boldsymbol{Y} \,/\, \boldsymbol{X} = \boldsymbol{x}) = \frac{e^{\beta_1+\beta_1 X_1}}{1+e^{\beta_1+\beta_1 X_1}} \qquad (2)$$

Both continuous and categorical predictor variables are possible for X. We can also expand the logistic model to include several predictors,

$$\textbf{Logit(Y) = In} \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \qquad (3)$$

2.2 LOGISTIC REGRESSION ODD RATIO

The generalized logistic regression model for p number of predictors is represented by equation (3). Either the weighted least squares approach or the maximum likelihood (ML) method can be used to estimate regression parameter β'*s*. Regression coefficient values between 1 and p show how X and Y's logit are related. A coefficient number greater than 0 implies that the logit of Y will increase as X increases, whereas a coefficient value less than 0 indicates that the logit of Y will fall as X increases. When the coefficient value is 0, it means that the logit of Y and the predictors X do not have a linear relationship. We often give the odds ratio along with the regression coefficient for ease of interpretation. This formula can be used to determine odds ratios:

$$\textbf{Odd ratio} = e^{\beta} \qquad (4)$$

2.3 Odd Event

In logistic regression analysis, Wald's test is commonly used to determine the statistical significance of the regression coefficient, while the likelihood ratio test or pseudo R2 test can be used to determine the significance of the entire model.

The odds of an event are the ratio of the probability that the event will occur to the probability that it will not occur. For example, if the probability of an event occurring is p, then the probability of the event not occurring is (1-P), and the corresponding odd value is given by:

$$\textbf{Odds of event} = \frac{p}{1-p} \qquad (5)$$

RSS-NIG 2025 Conference Proceedings

Logistic regression calculates the probability of an event occurring over the probability of an event not occurring. As a result, the impact of independent variables is often explained in terms of odds. In logistic regression, the relationship between the mean of the response variable p and an explanatory variable x is modeled using the equation:

$$\mathbf{p} = \alpha + \beta x \qquad (6)$$

However, this is a bad model since extreme x values will produce $\alpha + \beta x$ values that are not between 0 and 1. The odds are transformed using the natural logarithm in the logistic regression approach to this issue. We model the natural log chances using logistic regression as a linear function of the explanatory factor:

$$\textbf{Logit (p)} = \text{In(odds)} = \text{In}\left(\frac{p}{1-p}\right) = \alpha \qquad (7)$$

Where p is the probability of an interesting outcome and x is the explanatory variable. The parameters of the logistic regression are α and β. This is the simple logistic model.

The odds ratio is a measure of the association between exposure and outcome. The odds of the outcome being present among individuals is defined as:

$$\frac{p(1)}{1-p(1)} \qquad (8)$$

Similarly, the odds of the outcome being present among individuals with x=0 is defined as:

$$\frac{p(0)}{1-p(0)} \qquad (9)$$

The odds ratio, denoted OR is defined as the ratio of the odds for to the odds for x=1 to the odds of x=0 and is given by the equation

$$\text{OR}=\frac{p(1)/[1-p(1)]}{p(0)/[1-p(0)]} \qquad (10)$$

2.4 RANDOM FOREST MODEL

Random Forest is a popular ensemble learning method used in machine learning. It is a combination of multiple decision trees that are trained on different subsets of the training data and using a random subset of the features for each tree. In a random forest, each decision tree is trained on a random subset of the training data, with replacement. This means that each tree has a slightly different view of the data and may learn different patterns. Additionally, at each split of the decision tree, only a random subset of the available features is considered, which helps to reduce overfitting and improve the generalization of the model. The final prediction of the random forest is then made by aggregating the predictions of all the individual trees. For classification problems,

RSS-NLG 2025 Conference Proceedings

this aggregation can be done by taking a majority vote, and for regression problems, the aggregation can be done by taking the average of the individual tree predictions. Random Forests are known for their ability to handle high-dimensional datasets, handle missing values, and avoid overfitting. They are widely used in various fields, such as finance, marketing, and healthcare, for tasks such as classification, regression, and feature selection. A predictor called a random forest is made up of M-randomized regression trees. The projected value at the query point x for the jth tree in the family is given as mn(x; j, Dn), where 1,..., and M are independent random variables that are distributed similarly to a generic random variable and independent of Dn. More exact definitions will be provided later. In practice, the variable is used to resample the training set before the growth of individual trees and to choose the subsequent directions for splitting. The jth tree estimate is expressed mathematically as:

$$\boldsymbol{m_n}\left(\boldsymbol{x};\boldsymbol{\theta_{j,}}\boldsymbol{D_n}\right) = {}_{i\epsilon D_n(\theta_j)}\Sigma \frac{x_j \epsilon \ A_n(x;\theta_j,D_n)Y_n}{N_n\ (x;\theta_{j,}D_n)} \qquad \text{`(11)}$$

where An(x;j, Dn) is the cell holding x, Dn*(j) is the collection of data points chosen before the tree is constructed, and Nn(x;j, Dn) is the number of (preselected) points that fall into An

(x;j, Dn).

The trees are joined at this point to get the (limited) forest estimate.

$$m_{M,n}\left(x;\theta_1,\dots,\theta_{M,}D_n\right) = \frac{1}{M}\sum_{j=1}^{M} m_n(x;\theta_{1,}D_n) \qquad (12)$$

The default setting for M (the number of trees in the forest) in the Random Forest R package is n-tree = 500. From a modelling perspective, it makes sense to let M go to infinity and take into account instead of (1) the (infinite) forest estimate since M may be selected to be as large as you like—the only restriction is the amount of processing power you have available.

$$m_{\infty,n}(x;D_n) = \mathrm{E}_\theta[m_n(x;\theta,D_n)] \qquad (13)$$

E denotes the expectation about the random parameter, conditional on Dn, in this definition. In fact, the rule of large numbers, which states that virtually certainly, subject to the condition that Dn,

$$\lim_{m\to\infty} M_{M,n}(x;\theta_1,\dots,\theta_{M,}D_n) = m_{\infty,n}(x;D_n) \qquad (14)$$

2.5 CLASSIFICATION METRICS

The original data set has some imbalance. Only using precision cannot effectively measure the performance of models. Therefore, in addition to precision, we also calculate other model

RSS-NIG 2025 Conference Proceedings

evaluation metrics such as sensitivity, specificity, precision and accuracy to evaluate the performance of the classifier models.

1 Precision: Precision measures all samples that were predicted correctly, including positive and negative samples.

$$\textbf{Accuracy}=\frac{TN+TP}{TN+TP+FN+FP} \qquad (15)$$

Accuracy is a commonly used and easily understood evaluation index. However, its influence on positive and negative samples is the same. In the medical field, positive samples (such as CAD samples) are typically more important to doctors and patients than negative samples. When it comes to positive samples, the cost of missed diagnosis and misdiagnosis is different. In such cases, relying only on accuracy to assess classifier performance is insufficient. Additional evaluation metrics, such as precision, recall, and F1 score, may be necessary to more accurately assess the performance of a classifier.

2. Precision: Precision is used to measure the proportion of true positive samples in instances that are predicted to be positive.

$$\textbf{Precision} = \frac{TP}{TP+FP} \qquad (16)$$

Precision measures how likely it is that an instance predicted to be a positive sample is indeed a true positive sample. It indicates the accuracy of positive predictions made by the classifier.

3. Specificity: Specificity measures the proportion of true negative samples among all instances that are actually negative. It is particularly useful in binary classification tasks to evaluate the ability of a model to correctly identify negative samples.

Specificity can be calculated using the following formula:

$$\text{Specificity}=\frac{TN}{TN+FP} \qquad (17)$$

From the perspective of a negative sample, specificity measures how likely it is that an instance predicted to be negative is indeed a true negative sample. A higher specificity value indicates that the model is better at correctly identifying negative samples, minimizing false positive predictions.

4. Sensitivity: Sensitivity, also known as Recall, measures the proportion of true positive samples among all instances that are actually positive. It is a critical metric in classification tasks, especially in scenarios where identifying positive samples is of high importance, such as medical diagnosis.

Sensitivity can be calculated using the following formula:

RSS-NIG 2025 Conference Proceedings

Sensitivity=$\frac{TP}{TP+FN}$ (18)

3.0 ANALYSIS AND INTERPRETATION OF RESULTS

This chapter discusses the statistical analysis of the data collected. Logistic Regression and Random Forest, were used to carry out the study to provide answers to the objectives of this research work.

Table 1: Descriptive Analysis of the prioritized health biomarkers.

VARIABLES	**MINIMUM**	**1ST QUATILE**	**MEDIAN**	**MEAN**	**3RD QUATILE**	**MAXIMUM**
CHOLESTEROL LEVEL	78.0	98.25	195.50	195.50	292.75	390.00
GLUCOSE CONCENTRATION	78.0	179.0	203.0	207.2	229.0	443.O
BMI (BODY MASS INDEX)	15.20	24.10	27.80	28.78	32.27	55.80
SYSTOLIC BLOOD PRESSURE	90.0	122.0	136.0	137.1	148.O	250.0
DIASTOLIC BLOOD PRESSURE	48.00	75.00	82.00	83.00	90.00	124.00

Table 1 shows the frequencies of all the variables in the analysis having the minimum, maximum, mean, median,1STquartile and 3rdquartile

RSS-NLG 2025 Conference Proceedings

Table 2: Confusion Matrix for logistics regression model

Prediction	No diabetes	Diabetes
No diabetes	12	4
Diabetes	3	59

Accuracy: 0.9103

Table 3: Performance Measures for Logistic Regression Model

Accuracy	Sensitivity	Precision	Specificity
0.9103	0.9516	0.937	0.75

Table 3 shows the performance measures for Logistic Model, the model achieved an accuracy of 91% in identifying respondents with Diabetes and No diabetes, with a precision of 94%, Sensitivity of 95%, and Specificity of 75%.

Table 4: Odds Ratio

BMI	1.0457
Glucose	1.0366
Systolic_BP	1.0196
Cholesterol	1.0108
Diastolic_BP	0.9919

Table 4: Shows the odd ratio for each variable, we can see that the odd ratio for BMI is 1.0457, meaning that patients with a higher BMI are 1.0457 times less likely to have diabetes than patients with a lower BMI. The odd glucose index is 1.0366, which means that patients with the highest glucose index are 1.0366 times less likely to develop diabetes than patients with lower glucose levels. The odd index of systolic blood pressure is 1.0196, which means that patients with higher systolic blood pressure are 1.0196 times less likely to develop diabetes than patients with lower systolic blood pressure. The odd ratio of cholesterol is 1.0108, which means that patients with higher cholesterol levels are 1.0108 times more likely to develop diabetes than those with lower cholesterol levels. BMI, systolic, diastolic and cholesterol shows no relationship with the outcome

RSS-NIG 2025 Conference Proceedings

of diabetes, while the odds ratio Diastolic_BP is 0.9919, meaning that there is minimal to no relationship between the variable and the outcome of diabetes.

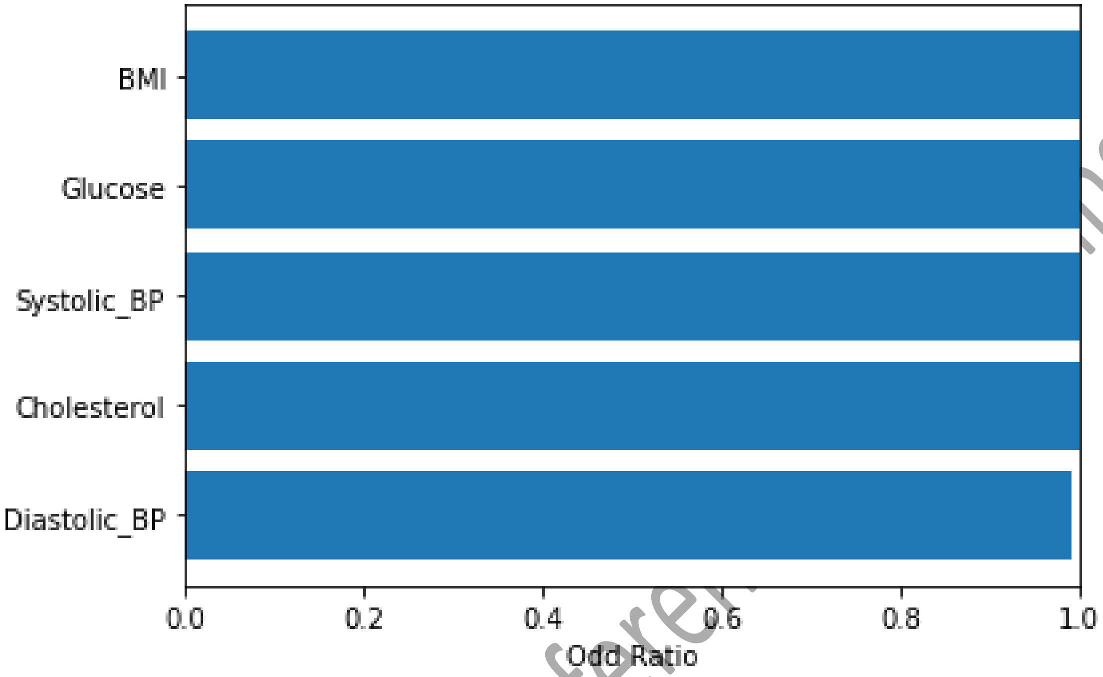

Figure 1: The Pictorial representation of odds ratio.

Figure 1 shows the odds ratios for each health biomarker (BMI, Glucose, Systolic BP, Cholesterol, Diastolic BP) used in predicting diabetes.

Table 5: Confusion matrix for Random Forest Model

Prediction	NO Diabetes	Diabetes
Diabetes	12	4
No Diabetes	5	57

Accuracy: 0.885

Table 5 Shows the Confusion matrix of No diabetes and Diabetes for the Random Forest Model.

RSS-NLG 2025 Conference Proceedings

Table 6: Performance measures for the Random Forest regression model

	Accuracy	Sensitivity	Precision	Specificity
	0.885	0.919	0.934	0.75

Table 6 shows the performance measures for Random Forest Model, the model achieved an accuracy of 89% in identifying respondents with Diabetes and No diabetes, with a precision of 93%, Sensitivity of 92%, and Specificity of 75%.

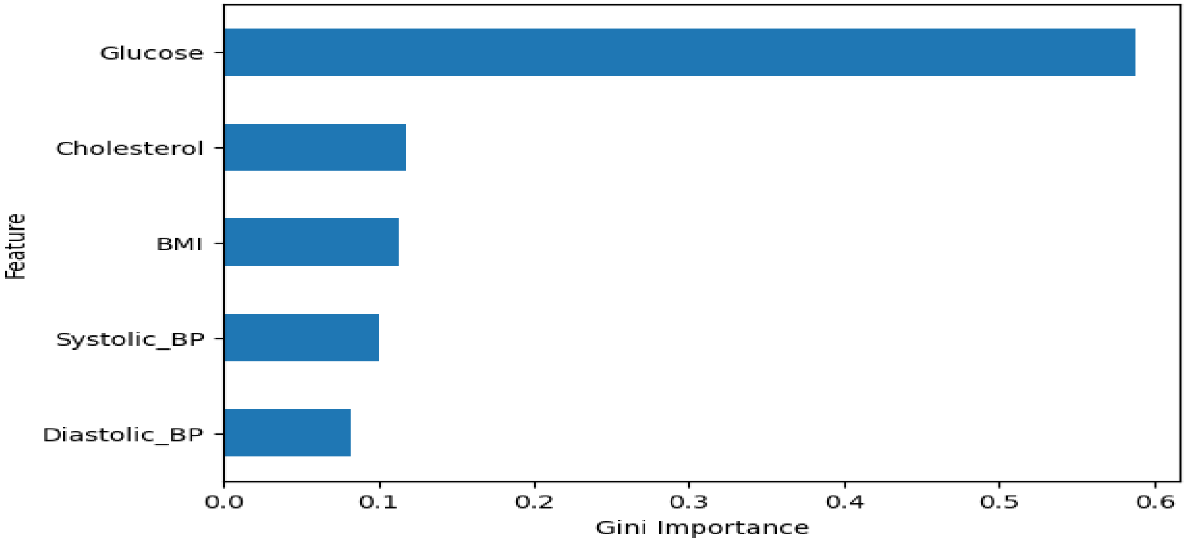

Figure 2**:** The graphical representation of Random Forest model

Figure 2 shows the relative significance of each variable as identified by the Random Forest model. Glucose is highlighted as the most important variable, followed by Cholesterol, BMI, Systolic_BP, and Diastolic_BP. This order indicates that variations in Glucose levels have the strongest effect on the model's ability to accurately predict diabetes outcomes.

RSS-NG 2025 Proceedings

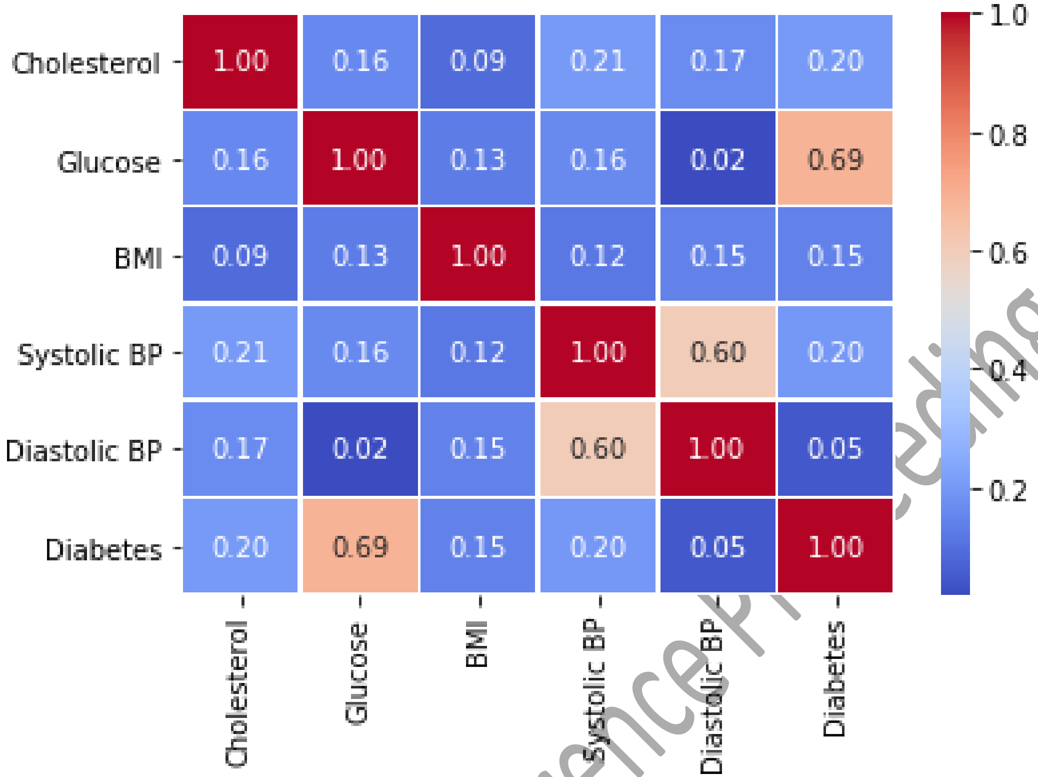

Figure 3. The Heatmap Correlations Matrix Among Health Metrics.

Figure 3 shows the relationships among various health metrics using a heatmap. It reveals a strong positive correlation between diabetes and glucose levels (0.69), moderate correlations between systolic and diastolic blood pressure (0.60) and between cholesterol and systolic blood pressure (0.21), while other features exhibit weaker linear associations, indicating limited linear relationships. Glucose is a key metric for Diabetes.

4.0 CONCLUSION

The variables in the equation show that it is important to have glucose. While BMI (Body Mass Index), systolic blood pressure, diastolic blood pressure and cholesterol are insignificant.

The logistic model and the random forest model have the same specificity of 75%, while the random forest has a minimum precision of 89%, a sensitivity of 92% and a precision of 93%. Therefore, logistic regression model is the best machine learning model to predict the presence of diabetes in this context.

REFERENCE

Ahamed B. S, Arya M. S and Nancy V A. O (2022) Prediction of Type-2 Diabetes Mellitus

RSS-NIG 2025 Conference Proceedings

Disease Using Machine Learning Classifiers and Techniques. Frontiers in Computer Science vol (4) https://doi.org/10.3389/fcomp.2022.835242.

Azevedo, M., and Alla, S. (2008). Diabetes in Sub-Saharan Africa: Kenya, Mali, Mozambique, Nigeria, South Africa and Zambia. International Journal of Diabetes in Developing Countries, 28(4), 101-108. https://doi.org/10.4103/0973-3930.45268

Chen, L., Shao, X. and Yu, P. (2023). Machine learning prediction models for diabetic kidney disease: systematic review and meta-analysis. *Endocrine* . https://doi.org/10.1007/s12020-023-03637-8

Siegel K. R., Bullard K. M., Imperatore G., Mohammed K., Albright A., Mercado C. I., Li R. and Gregg E. W. (2018). Prevalence of Major Behavioral Risk Factors for Type 2 Diabetes. *Diabetes Care*, 41 (5): 1032–1039. https://doi.org/10.2337/dc17-1775

Lai H., Huang H., Keshavjee K., Guergachi A. and Gao X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord 19, 101. https://doi.org/10.1186/s12902-019-0436-6

Lin J., Thompson T. J., Cheng Y. J., Zhuo X., Zhang P., Gregg E. and Rolka D. B. (2018). Projection of the future diabetes burden in the United States through 2060. Population Health Metrics 16 (9) https://doi.org/10.1186/s12963-018-0166-4

Menezes C. N., Crowther N. J., Duarte R., Amsterdam D. V., Evans D., Dickens C., Dix-Peek T., Rassoo M., Prinsloo A., Raal F. and Sanne I..(2014). A randomized clinical trial comparing metabolic parameters after 48 weeks of standard- and low-dose stavudine therapy and tenofovir disoproxil fumarate therapy in HIV-infected south African patients HIV Med. 2014;15(1):3-12

Ooka T., Johno H., Nakamoto K., Yoda Y., Yokomichi H. and Yamagata Z. (2021). Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health checkup data in Japan. BMJ Nutrition, Prevention and Health 2021;4:e000200. doi:10.1136/ bmjnph-2020-000200

Sahebhonar M, Gholampour D. M., Kazemi-Galougahi M. H. and Soleiman-Meigooni S. (2022). A Comparison of Three Research Methods: Logistic Regression, Decision Tree, and Random Forest to Reveal Association of Type 2 Diabetes with Risk Factors and Classify

RSS-NLG 2025 Conference Proceedings

Subjects in a Military Population. J Arch Mil Med. 2022;10(2):e118525. https://doi.org/10.5812/jamm-118525.

Shojaee-Mend H, Velayati F, Tayefi B. and Babaee E. (2024). Prediction of Diabetes Using Data Mining and Machine Learning Algorithms: A Cross-Sectional Study. Healthc Inform Res. 2024 Jan;30(1):73-82. doi: 10.4258/hir.2024.30.1.73.

Tan, K. R., Benjamin Seng, J. J., Kwan, Y. H., Chen, Y. J., Zainudin, S. B., Fang Loh, D. H., Liu, N., and Low, L. L. (2021). Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review. Journal of Diabetes Science and Technology. https://doi.org/10.1177/19322968211056917

Wen R., Zheng K., Zhang Q., Zhou L., Liu Q., Yu G.,Gao X., Hao L., Lou Z. and Zhang W. (2021). Machine learning-based random forest predicts anastomotic leakage after anterior resection for rectal cancer. Journal of Gastrointestinal Oncology 12 (3), 921–932. https://doi.org/10.21037/jgo-21-268

Zou Q., Qu K., Luo Y., Yin D., Ju Y. and Tang H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 9:515. doi: 10.3389/fgene.2018.00515

RSS-NLG 2025 Conference Proceedings