iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84394

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Predictive Modelling for Diabetes Risk Assessment: An Exploratory Data Analysis Approach

Shreyash Gade, Dr. Prakash Kene

Abstract: ***Diabetes mellitus is a big health problem in the world. More people are getting it, and it can cause bad health issues. This study builds a model to find diabetes early. It uses a machine learning tool to identify the best result. The study uses full exploratory data analysis (EDA). It finds key risk signs like blood glucose level, body mass index (BMI), age, and insulin levels. These things are looked at to see how they are linked and how they affect who gets diabetes. The data's class imbalance is fixed using resampling. This makes the model more true and fair. It is also easy to see why. It shows the key causes of diabetes. This work helps find diabetes early. It helps doctors make choices and check risk for people with diabetes. It does this by using ideas from data and machine learning tools.***

Keywords: Diabetes Mellitus, Early Detection, Machine Learning Model, Exploratory Data
Analysis (EDA), Blood Glucose, Body Mass Index (BMI), Age, Insulin Levels, Risk Factors, Class Imbalance, Resampling Techniques, Predictive Modeling, Healthcare Decision Making, Data-Driven Insights, Diabetes Risk Assessment

I. INTRODUCTION

Diabetes is a serious illness and needs care. It is a long-term problem where blood sugar is always too high. This happens because the pancreas does not make enough insulin. Or, it happens because the body does not use insulin the right way. Insulin helps glucose, the body's main fuel, get into cells. When this breaks down, glucose stays in the blood. This leads to many health problems. In diabetes, you have high blood sugar (hyperglycemia) for a long time. This raises the risk of damage to small vessels, like nerve damage (neuropathy), kidney disease (nephropathy), and vision problems (retinopathy). It also helps cause big vessel problems like heart disease, stroke, and peripheral vascular disease. Beyond these body problems, diabetes can shorten life. It can cause lasting harm and make your quality of life much lower[1].

Glucose is the main fuel for the body, coming from carbohydrates and supplying energy to the brain and muscles. Keeping its level balanced is essential, as too much or too little can cause serious health issues[2].

Insulin, produced by the pancreas, helps move glucose into cells for energy and storage. When insulin is lacking or not working properly, glucose builds up in the blood, leading to diabetes[3].

Types of Diabetes

Diabetes includes Type 1 (autoimmune, insulin-dependent), Type 2 (most common, linked to lifestyle and genetics), and Gestational Diabetes (occurs during pregnancy), along with rare monogenic and secondary forms[4].

Type 1 requires lifelong insulin, while Type 2 is often managed with lifestyle changes and medications[5].

Gestational diabetes usually resolves after delivery but increases future Type 2 risk. If not controlled, diabetes can lead to complications affecting the eyes, kidneys, nerves, heart, and blood vessels[6].

II. PROBLEM STATEMENT

Diabetes is a major global health concern, influenced by lifestyle, genetics, and other health conditions. Early detection is vital, but traditional diagnostic methods are often slow and reactive. Machine learning provides a faster, data-driven approach to analyze health datasets and uncover patterns for predicting diabetes risk.

This study uses the Diabetes Prediction Dataset to classify individuals as diabetic or non-diabetic through thorough EDA, data cleaning, and handling of missing values and class imbalance. A Random Forest classifier with hyperparameter tuning is employed to enhance performance, evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. Feature importance analysis identifies key risk factors, aiding better clinical decisions. This work supports early detection and prevention, improving healthcare outcomes through machine learning.

III. AIM & OBJECTIVES

The main goal is to build a good model with machine learning. This model will find diabetes early. The study will mix what the data says with what doctors know. This can help find people at high risk. We want to find them before they feel sick. We will look at the data, make the model better, and make it easy to understand. This work aims to be more right in its findings. It will also make the model clear. And it will help doctors use data to make good choices.

IV. RESEARCH OBJECTIVES

To achieve the stated aim, the following specific objectives are outlined:

1) To explore and fully understand the data using exploratory data analysis (EDA), finding hidden patterns, trends, and links between factors that lead to diabetes.
2) To prepare and balance the data (using tools like SMOTE or under sampling). This will fix class imbalance. It helps make the model fair and its results steady.
3) To build and tune a Random Forest classification model. Use methods like GridSearchCV to find the best settings. The goal is high accuracy, precision, and recall.
4) To check model performance using strong tests like accuracy, precision, recall, F1-score and AUC-ROC. This makes sure the model can be trusted for real or doctor use [7].

V. RESEARCH METHODOLOGY

The way we look for what leads to diabetes in this study uses a plan based on data. This study uses a count and number-based setup and pulls number info from the Diabetes Forecast Set. The steps combine deep data clean-up, initial data look-up (EDA), and machine learning ways to find links and hints that can tell us more about when diabetes might happen. By taking these clear steps, this study wants to come up with strong and useful results that help us know more about what causes diabetes and how to see it coming. The steps of the method are as such:

- Data-driven study design: This looks at what brings about diabetes by using a data-based way.
- Quantitative research setup: It uses number data from the Diabetes data set.
- Integrating data methods: Includes advanced steps to clean data, a first look at data (EDA), and machine learning.
- Aim of the study: To make strong and helpful findings on what leads to diabetes.
- Step-by-step process: The study follows a planned way to make sure the results are good and useful[8].

1) Research Approach: This study follows a quantitative research approach, as it focuses on analyzing numerical data from the Diabetes prediction Dataset to identify patterns, relationships, and predictive factors for diabetes occurrence. Quantitative methods are suitable for this study because the research aims to produce measurable and statistically significant results through data-driven analysis and machine learning techniques.
2) Data Collection: The dataset is used is Diabetes Prediction (Kaggle) which contains patient medical details such as gender, hypertension, heart disease, smoking history, bmi (body mass index), HbA1c level, blood glucose, and diabetes (target) which includes the 100k rows.

A. Process Flow of Predictive Modelling

1) Dataset: The process begins with a diabetes dataset containing various medical and lifestyle features for each patient, forming the base for analysis and modeling. Understanding its structure and quality is essential before applying any machine learning techniques.
2) Exploratory Data Analysis(EDA): Exploratory Data Analysis (EDA) helps examine feature distributions, detect anomalies, and identify issues like missing values or outliers. Insights from plots and summary statistics guide preprocessing and feature engineering decisions.
3) Data Preprocessing: Data preprocessing involves cleaning and organizing the dataset by handling missing values, transforming categorical features, and removing outliers. Proper preprocessing ensures the model learns from accurate, noise-free data.
4) Data Transformation: Data transformation applies scaling or normalization to reduce skewness and improve model training. Standardizing features ensures consistent scaling, especially for models sensitive to feature magnitudes.

5) Handle Class Imbalance with SMOTE: SMOTE addresses class imbalance by generating synthetic samples for the underrepresented diabetic class, helping the model learn patterns more effectively. This reduces bias toward the majority class and improves overall prediction fairness. [9].
6) Train & Test Dataset: The dataset is split into training and testing sets to ensure the model learns from one portion and is evaluated on unseen data. This separation helps assess how well the model generalizes to new cases.
7) Hyperparameters: GridSearchCV evaluates multiple Random Forest parameter combinations using cross-validation to find the most effective setup. This optimization reduces overfitting and improves the model's overall robustness and accuracy [10].
8) Training model on Best Parameters: After identifying the best hyperparameters, the model is retrained on the full training dataset to capture patterns more effectively. This step produces an optimized model aimed at improving prediction accuracy.
9) Model Evaluation: The model is evaluated using metrics like accuracy, precision, recall, and F1-score to measure its predictive performance. These results help determine whether the model is reliable for real-world diabetes classification[11].
10) Feature Importance Analysis: Random Forest identifies and ranks the most important features contributing to diabetes prediction. This analysis highlights key health factors and helps interpret how the model makes its decisions.

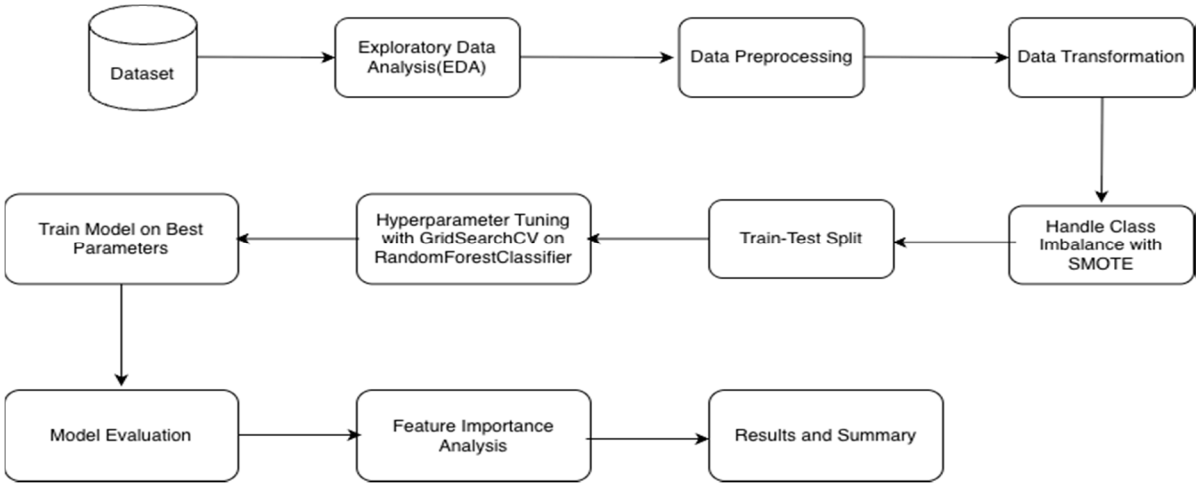


Fig. 1. Process Flow Diagram of Predictive Modelling

B. Model Evaluation

To know how well our model works in the real world, one accuracy score is not enough. This is very true for a key medical diagnosis. We used a full set of evaluation metrics to give a complete picture of its work.

Confusion Matrix

The Confusion Matrix is the base for how we check our work. It is a 2x2 table that shows all guesses made on test data (data the model has not seen). It sorts them into four groups:

- True Positives (TP): The model was right. It said diabetic for a person who is diabetic.
- True Negatives (TN): The model was right. It said non-diabetic for a person who is not diabetic.
- False Positives (FP): The model was wrong. It said diabetic for a person who is not. (A false alarm - Type I Error).
- False Negatives (FN): The model was wrong. It said non-diabetic for a person who is diabetic. (A miss - Type II Error). This is often the worst error in health tests.

C. Dataset Description

The dataset consists of 100,000 anonymized patient records, combining demographic details, clinical measurements, and lifestyle information to support diabetes prediction. It includes important medical indicators such as HbA1c levels, blood glucose readings, and BMI, which help assess metabolic health. Alongside these, the dataset captures existing conditions like hypertension and heart disease, offering insight into comorbidities often linked with diabetes. Demographic factors—such as age, gender, and smoking history—provide additional context about each patient's risk profile. The target variable, diabetes, is labelled as 0 or 1, indicating whether the individual has been diagnosed with the condition. Overall, the dataset provides a comprehensive blend of physiological and lifestyle attributes suitable for predictive modelling and risk assessment.

Table 1: Dataset Description

Feature Name	Type	Description	Value Range/Example
Gender	Categorical	Patient's Gender	'Male','Female'
Age	Numerical	Patient's age in years	0.08–80.0
Hypertension	Binary	Presence of hypertension (0 = No, 1 = Yes)	0 or 1
Heart Disease	Binary	Presence of heart disease (0 = No, 1 = Yes)	0 or 1
Smoking History	Categorical	Patient's smoking status	'No Info', 'Current', 'Former', 'Never', 'Not current', 'Ever'
BMI	Numerical	Body Mass Index (weight in kg / height in m^2)	10.1–95.2
HbA1c_level	Numerical	Glycated haemoglobin level (indicator of average blood sugar over 2–3 months)	3.5–10.5
Blood Glucose Level	Numerical	Random blood glucose level (mg/dL)	57–320

This dataset is class-imbalanced (approximately 97% non-diabetic cases), reflecting real-world prevalence, and is suitable for evaluating ML models on underrepresented positive cases. No explicit references to original papers are cited in the dataset description, but it aligns with standard diabetes risk models like those from the CDC or WHO guidelines.

VI. RESULTS AND DISCUSSION

This study analyzed a large diabetes prediction dataset containing 100,000 patient records with diverse medical and demographic features. Key attributes included gender, age, hypertension, heart disease, smoking history, BMI, HbA1c levels, and blood glucose levels. The target variable was a binary indicator showing whether a patient was diabetic or not. The distribution showed 91,500 non-diabetic (91.5%) and 8,500 diabetic cases (8.5%), highlighting a strong class imbalance affecting model performance.

A. Correlation Heatmap

The correlation heatmap analyzed Pearson correlations between numerical features and the diabetes outcome. HbA1c (0.59) and blood glucose (0.57) showed the strongest positive correlations, indicating high importance for prediction. Age (~0.34) and BMI (~0.21) had moderate correlations, while hypertension and heart disease (~0.10–0.15) showed weak links. Features were not highly correlated with each other (all < 0.6), allowing all to be used in the Random Forest model. Gender and smoking history had very weak correlations, consistent with their low feature importance scores (~0.02–0.03).

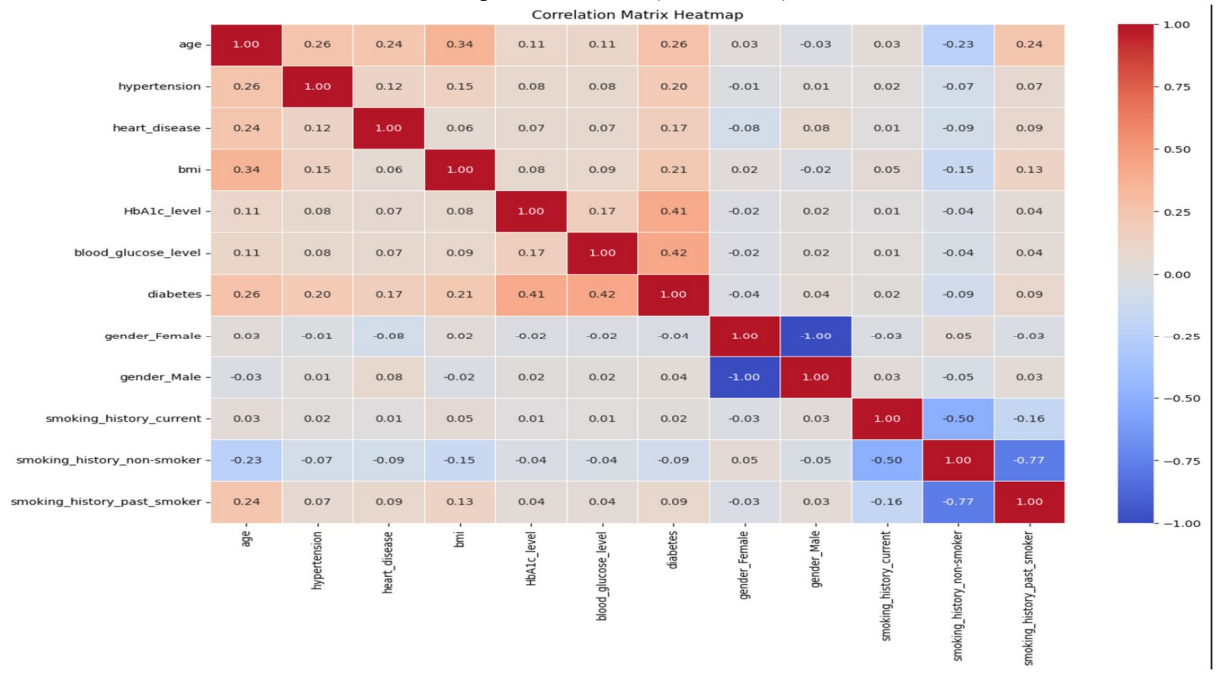


Fig. 2. Correlation Heatmap

B. Confusion Matrix

The confusion matrix showed that most diabetic and non-diabetic cases were correctly classified, with only a few false negatives, indicating the model effectively identifies those at real risk of diabetes. For the non-diabetic class, precision (0.98) and recall (0.97) were very high, demonstrating excellent detection of healthy individuals. For the diabetic class, recall was 0.79, meaning the model identified 80% of actual diabetic patients, which is crucial for minimizing false negatives in healthcare. Precision was 0.70, indicating that 70% of predicted diabetic cases were correct, while 30% were false positives, which is less harmful than missing real cases. Overall, the model correctly predicted 16,960 non-diabetic cases and 1,343 diabetic cases, while 358 real diabetic cases were missed and 565 non-diabetic cases were incorrectly labeled as diabetic. These results highlight the model's strong performance, particularly in detecting at-risk patients, making it valuable for clinical risk assessment.

```
Model Accuracy:  0.9519920940393217
              precision    recall  f1-score   support

           0       0.98      0.97      0.97     17525
           1       0.70      0.79      0.74      1701

    accuracy                           0.95     19226
   macro avg       0.84      0.88      0.86     19226
weighted avg       0.95      0.95      0.95     19226
```

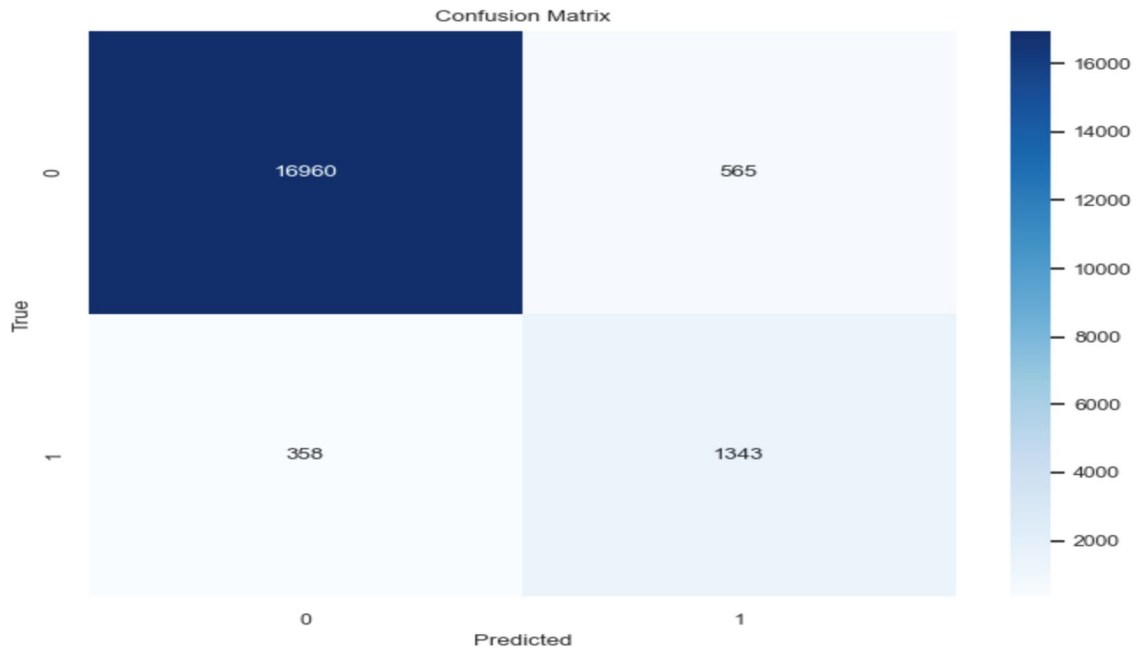


Fig. 3. Confusion matrix visualization

Table 2: Model Performance Metrics for Diabetes Prediction

Metrics	Description	Value
Accuracy	Overall correctness of predictions	0.95
Precision	Correct diabetic predictions among all predicted diabetic cases	0.70
Recall	Correctly identified diabetic patients among all actual diabetic cases	0.79
F1-Score	Harmonic mean of precision and recall	0.74

But, some false positives were found. These could lead to extra doctor visits that are not needed. Still, in health care, a high recall (sensitivity) is more important. This is because missing a patient with diabetes can lead to big problems.

C. *Tuned Random Forest Model Performance*

To get better results and stop overfitting, we tuned the Random Forest model. We used the GridSearchCV (Grid Search Cross-Validation) method. This process checked how well the model worked across a set grid of key settings. It tested things like the number of trees (n_estimators), the max depth of each tree (max_depth), the smallest number of samples for a split (min_samples_split), and the rule for how to split (criterion).

The goal was to find the best mix of parameters for the best results. We used a measure like the F1-score during cross-validation. This works better than plain accuracy when there is class imbalance. The final model was tested on the hold-out test set. It had a great overall accuracy of 94.9%.

But, this overall score can be wrong for an imbalanced dataset. A model could get a high score by mostly guessing the bigger non-diabetic class. So, we must look at each class by itself. This is key to check the model's true clinical utility. It also shows how well it finds the much rarer diabetic class. The full results for each class are in Table 3.3 and are looked at in the next section[12].

Table 3. Performance Metrics of the Tuned Random Forest Model

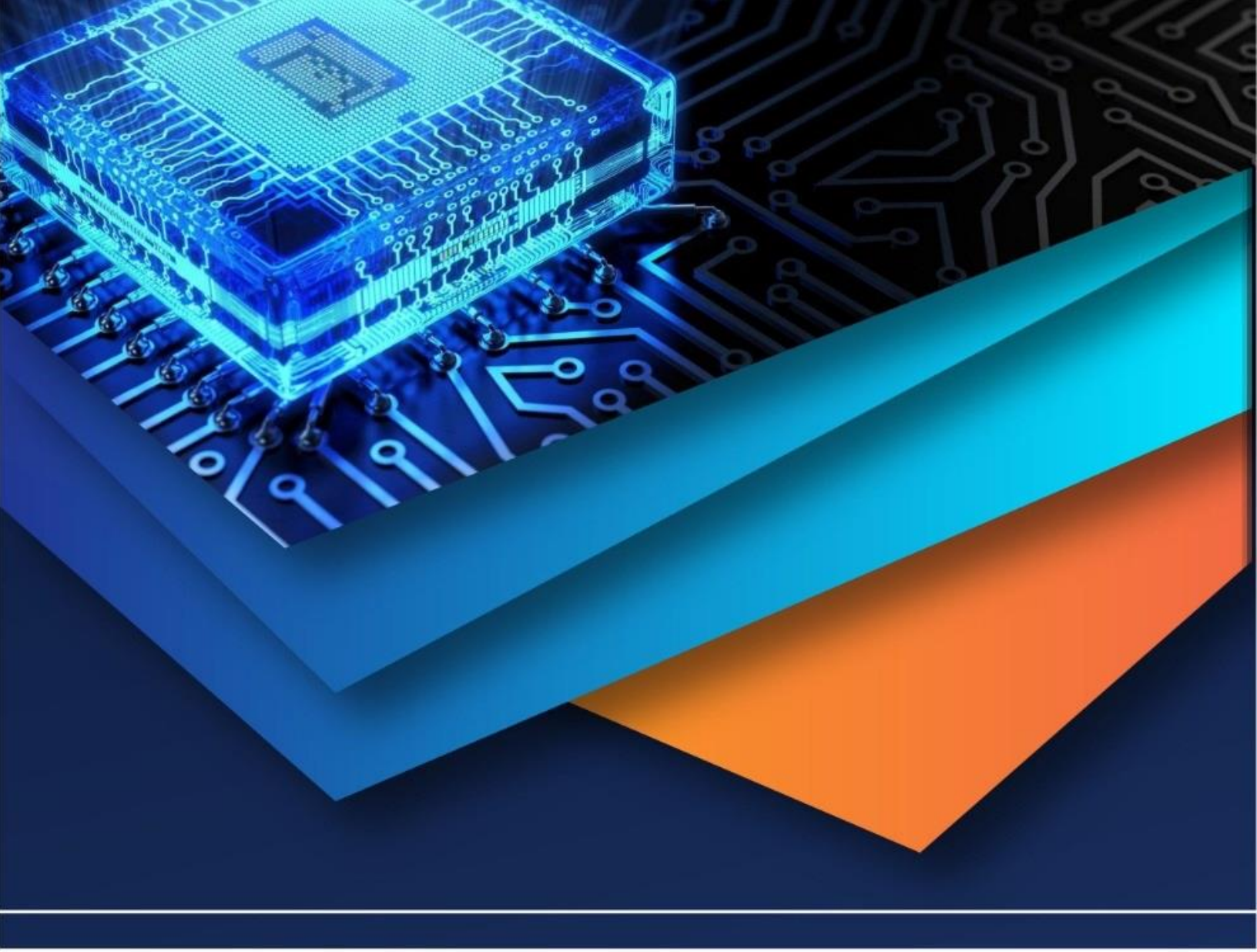
Class	Precision	Recall	F1-Score	Support(samples)
0 (Non-Diabetic)	0.98	0.96	0.97	17,525
1 (Diabetic)	0.70	0.79	0.74	1,701
Accuracy	-	-	0.95	19,226
Macro Avg	0.84	0.88	0.86	19,226
Weighted Avg	0.95	0.95	0.95	19,226

VII. CONCLUSION & FUTURE SCOPE

The study showed that a well-optimized Random Forest model can effectively predict diabetes risk by analyzing key factors like glucose, HbA1c, BMI, age, and hypertension. Model bias toward non-diabetic cases was reduced using oversampling and hyperparameter tuning, improving accuracy and balance. Important features such as HbA1c and blood glucose strongly influenced predictions, making the model both reliable and explainable. For future work, larger and more diverse datasets, advanced ML models, and integration of real-time health data can enhance performance. The model can also be deployed as a clinical decision-support tool with XAI methods like SHAP or LIME to ensure transparency and trust.

REFERENCES

[1] Sanders, L.J., (2002) From Thebes to Toronto and the 21st Century: An Incredible Journey. Diabetes Spectrum, [online] 151, pp.56–60
[2] Petersen, M.C. and Shulman, G.I., (2018) Mechanisms of Insulin Action and Insulin Resistance. Physiological Reviews, 984, pp.2133–2223.
[3] Rahman, M.S., Hossain, K.S., Das, S., Kundu, S., Adegoke, E.O., Rahman, M.A., Hannan, M.A., Uddin, M.J. and Pang, M.G., (2021) Role of Insulin in Health and Disease: An Update. International Journal of Molecular Sciences, 2212, pp.6403–6422.
[4] Tan, S.Y., Mei Wong, J.L., Sim, Y.J., Wong, S.S., Mohamed Elhassan, S.A., Tan, S.H., Ling Lim, G.P., Rong Tay, N.W., Annan, N.C., Bhattamisra, S.K. and Candasamy, M., (2019) Type 1 and 2 diabetes mellitus: A review on current treatment approach and gene therapy as potential intervention. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 131, pp.364–372.
[5] Antar, S.A., Ashour, N.A., Sharaky, M., Khattab, M., Ashour, N.A., Zaid, R.T., Roh, E.J., Elkamhawy, A. and Al-Karmalawy, A.A., (2023) Diabetes mellitus: Classification, mediators, and complications; A gate to identify potential targets for the development of new effective treatments. Biomedicine & Pharmacotherapy, 168, pp.115734–115759.
[6] Riddle, M.C., Philipson, L.H., Rich, S.S., Carlsson, A., Franks, P.W., Greeley, S.A.W., Nolan, J.J., Pearson, E.R., Zeitler, P.S. and Hattersley, A.T., (2020) Monogenic Diabetes: From Genetic Insights to Population-Based Precision in Care. Reflections From a Diabetes Care Editors' Expert Forum. Diabetes Care, 4312, pp.3117–3128.
[7] Ganie, S.M., Pramanik, P.K.D., Bashir Malik, M., Mallik, S. and Qin, H., (2023) An ensemble learning approach for diabetes prediction using boosting techniques. Frontiers in Genetics, 14, pp.1252159–1252174.
[8] Manjula, B.M., Harshavardhan, A., Indrasena, B.T. and Neha Annie, S., (2024) Exploratory Data Analysis and Predictive Modeling of Pima Indian Diabetes. In: International Conference on Intelligent Algorithms for Computational Intelligence Systems, IACIS 2024. Institute of Electrical and Electronics Engineers Inc., pp.1–6.
[9] Chawla, N. V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., (2011) SMOTE: Synthetic Minority Over-sampling Technique. Journal Of Artificial Intelligence Research, 16, pp.321–357.
[10] Devi, K., Kumar, J.S., Poovizhi, S. and Krishnan, M., (2022) Analytical study on diabetes prediction-
[11] Tasin, I., Nabil, T.U., Islam, S. and Khan, R., (2022) Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 101–2, p.10.
[12] Noviyanti, C.N. and Alamsyah, A., (2024) Early Detection of Diabetes Using Random Forest Algorithm. Journal of Information System Exploration and Research,21, pp.41–48.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY