IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICT DIABETES USING MACHINE LEARNING ALGORITHMS

Nischitha V P[1], Rashmi K A[2]

[1]Post Graduation Student Master of Technology, Department of Computer Science and Engineering, SJC Institute of Technology, Chikkaballapur, Karnataka, India

[2] Assistant Professor, Department of Computer Science and Engineering, SJC Institute of Technology, Chikkaballapur, Karnataka, India

Abstract

Diabetes mellitus is one of the most common chronic diseases affecting millions of people worldwide. Early prediction of diabetes helps healthcare professionals provide timely treatment and reduce severe complications. Machine Learning (ML) techniques have emerged as effective tools for disease prediction by analyzing medical datasets and identifying hidden patterns. This paper presents a diabetes prediction system using supervised machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The Pima Indians Diabetes Dataset is used for model training and evaluation. Data preprocessing techniques including missing value handling, feature scaling, and normalization are applied to improve model performance. The trained models are evaluated using accuracy, precision, recall, F1-score, and confusion matrix. Experimental analysis indicates that Random Forest provides better prediction accuracy than the other algorithms. The proposed system provides a simple, reliable, and efficient approach for assisting healthcare professionals in the early diagnosis of diabetes.

Index Terms— Diabetes Prediction, Machine Learning, Random Forest, Support Vector Machine, Logistic Regression, Healthcare Analytics.

I. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder caused by insufficient insulin production or ineffective utilization of insulin by the human body. According to the World Health Organization (WHO), diabetes has become one of the major global health concerns because of its rapidly increasing prevalence. Delayed diagnosis may result in severe complications including kidney failure, cardiovascular diseases, blindness, nerve damage, and stroke.

Traditional diabetes diagnosis mainly depends on laboratory examinations and physician analysis. Although these methods provide reliable results, they are often time-consuming and expensive. Machine

Learning provides an intelligent alternative by learning patterns from historical medical records and predicting whether a patient is likely to have diabetes based on clinical parameters.

The proposed system utilizes the Pima Indians Diabetes Dataset containing attributes such as glucose concentration, blood pressure, insulin level, body mass index (BMI), age, pregnancies, skin thickness, and diabetes pedigree function. Several supervised learning algorithms are implemented and compared to identify the most suitable model for diabetes prediction.

The major objective of this research is to develop an efficient and reliable machine learning-based system for the early prediction of diabetes. The study focuses on preprocessing the diabetes dataset to improve data quality and prediction accuracy, training multiple supervised machine learning classification algorithms, and comparing their performance using standard evaluation metrics. Based on the comparative analysis, the most effective prediction model is identified to achieve higher accuracy and reliability. The proposed system aims to assist healthcare professionals in the early diagnosis of diabetes, enabling timely medical intervention and improved patient care.

II. RELATED WORK

Several researchers have proposed Machine Learning-based diabetes prediction systems using different classification algorithms.

In recent years, Logistic Regression has been widely used because of its simplicity and effectiveness in binary classification problems. It provides interpretable prediction results and performs efficiently on structured healthcare datasets.

Decision Tree classifiers generate prediction rules that are easy for medical professionals to understand. However, individual decision trees are prone to overfitting when trained on complex datasets.

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Several research studies have reported that Random Forest achieves higher accuracy than traditional classifiers for diabetes prediction.

Support Vector Machine (SVM) has also shown promising performance in medical diagnosis because it effectively separates diabetic and non-diabetic samples using optimal hyperplanes.

K-Nearest Neighbors (KNN) predicts diabetes by comparing new patient records with previously classified medical records. Although simple to implement, its computational complexity increases with larger datasets.

Based on the literature, ensemble learning methods such as Random Forest generally outperform individual classifiers due to their robustness and better generalization capability.

III. RESEARCH METHODOLOGY

3.1 Dataset

The proposed system uses the **Pima Indians Diabetes Dataset**, one of the most widely used benchmark datasets for diabetes prediction. The dataset contains medical information collected from female patients aged above 21 years.

The dataset consists of the following attributes:

Attribute	Description
Pregnancies	Number of pregnancies
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure
Skin Thickness	Triceps skin fold thickness
Insulin	Serum insulin level
BMI	Body Mass Index
Diabetes Pedigree Function	Family diabetes history
Age	Patient age
Outcome	Diabetes (0 = No, 1 = Yes)

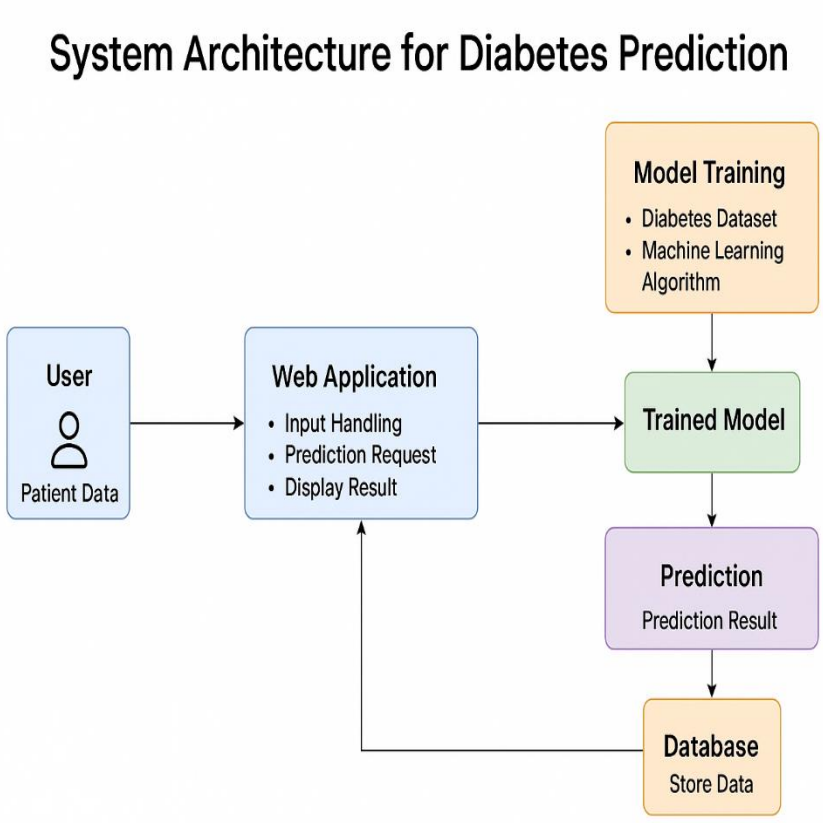


Figure 1. System Architecture of Diabetes Prediction System

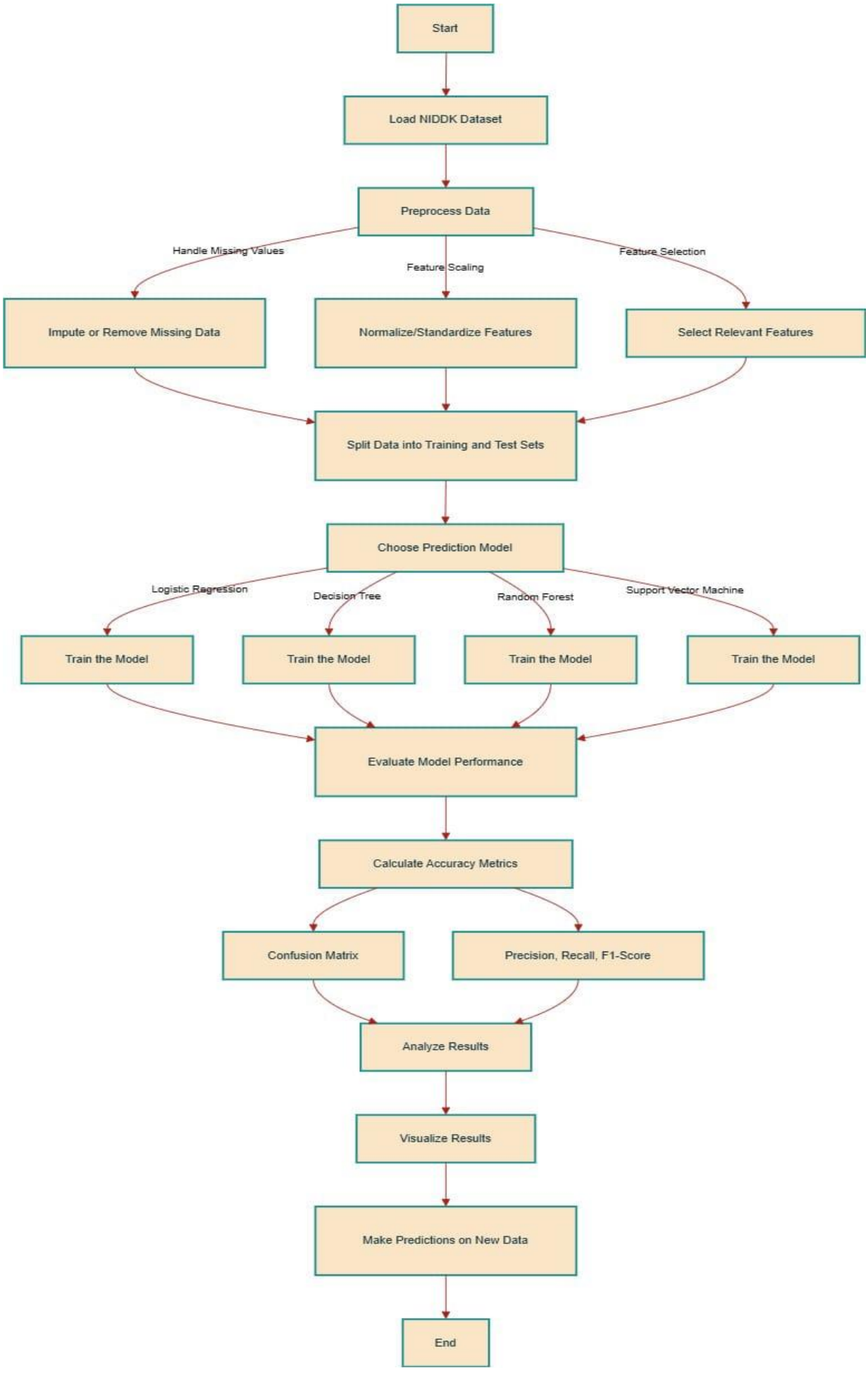
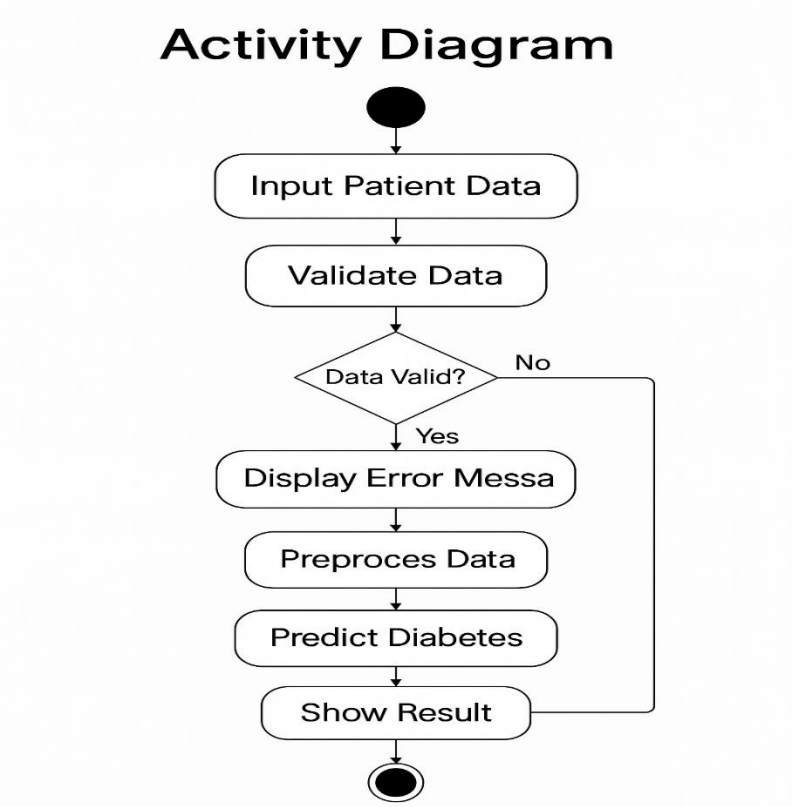


Figure 2 Flow Chart



3.2 Data Preprocessing

The collected dataset is preprocessed before training the Machine Learning models.

The preprocessing steps include:

- Handling missing values.
- Removing duplicate records.
- Feature normalization.
- Data scaling using Standard Scaler.
- Splitting the dataset into training and testing datasets (80:20 ratio).

These preprocessing techniques improve model accuracy and reduce bias.

3.3 Machine Learning Algorithms

The following supervised learning algorithms are implemented:

A. Logistic Regression

A statistical classification algorithm used for binary classification problems. It predicts the probability of diabetes occurrence based on patient attributes.

B. Decision Tree

A tree-based classifier that predicts diabetes using a sequence of decision rules generated from medical features.

C. Random Forest

Random Forest combines multiple decision trees using ensemble learning techniques to improve prediction accuracy and reduce overfitting.

D. Support Vector Machine (SVM)

SVM constructs an optimal hyperplane that separates diabetic and non-diabetic patients with maximum margin.

E. K-Nearest Neighbors (KNN)

KNN predicts diabetes by comparing a patient's medical records with similar previously diagnosed patients.

3.4 Performance Evaluation

The prediction models are evaluated using the following performance metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

These evaluation metrics help identify the most reliable Machine Learning algorithm for diabetes prediction.

II.F Research Gap

Although numerous studies have applied Machine Learning techniques for diabetes prediction, several limitations remain. Many existing systems focus on improving prediction accuracy using a single algorithm without comparing the performance of multiple classification models. Some approaches require extensive feature engineering or large clinical datasets, making them difficult to implement in real-world healthcare environments. Additionally, several prediction models lack proper preprocessing techniques, resulting in reduced accuracy and poor generalization. Many existing systems also do not provide a simple and user-friendly interface that enables healthcare professionals to perform early diabetes prediction efficiently. Therefore, there is a need for an accurate, reliable, and easy-to-use diabetes prediction system that combines effective data preprocessing with multiple Machine Learning algorithms and comprehensive performance evaluation.

II.G Research Motivation

The increasing prevalence of diabetes across the world has created a significant need for early diagnosis and preventive healthcare. Delayed detection of diabetes can lead to severe complications such as heart disease, kidney failure, vision impairment, and nerve damage. Conventional diagnostic methods often require multiple laboratory tests and continuous medical supervision, which may not always be accessible or affordable. This research is motivated by the potential of Machine Learning to analyze patient health parameters and predict diabetes at an early stage. By comparing multiple supervised learning algorithms and selecting the most accurate prediction model, the proposed system aims to support healthcare professionals in making timely clinical decisions. The developed framework provides a simple, efficient, and cost-effective solution for early diabetes prediction, ultimately contributing to improved patient care and disease management.

IV. RESULTS AND DISCUSSION

The proposed diabetes prediction system was implemented using Python and the Scikit-learn library. After preprocessing the dataset, multiple supervised machine learning algorithms were trained and tested. The dataset was divided into training (80%) and testing (20%) sets to evaluate model performance.

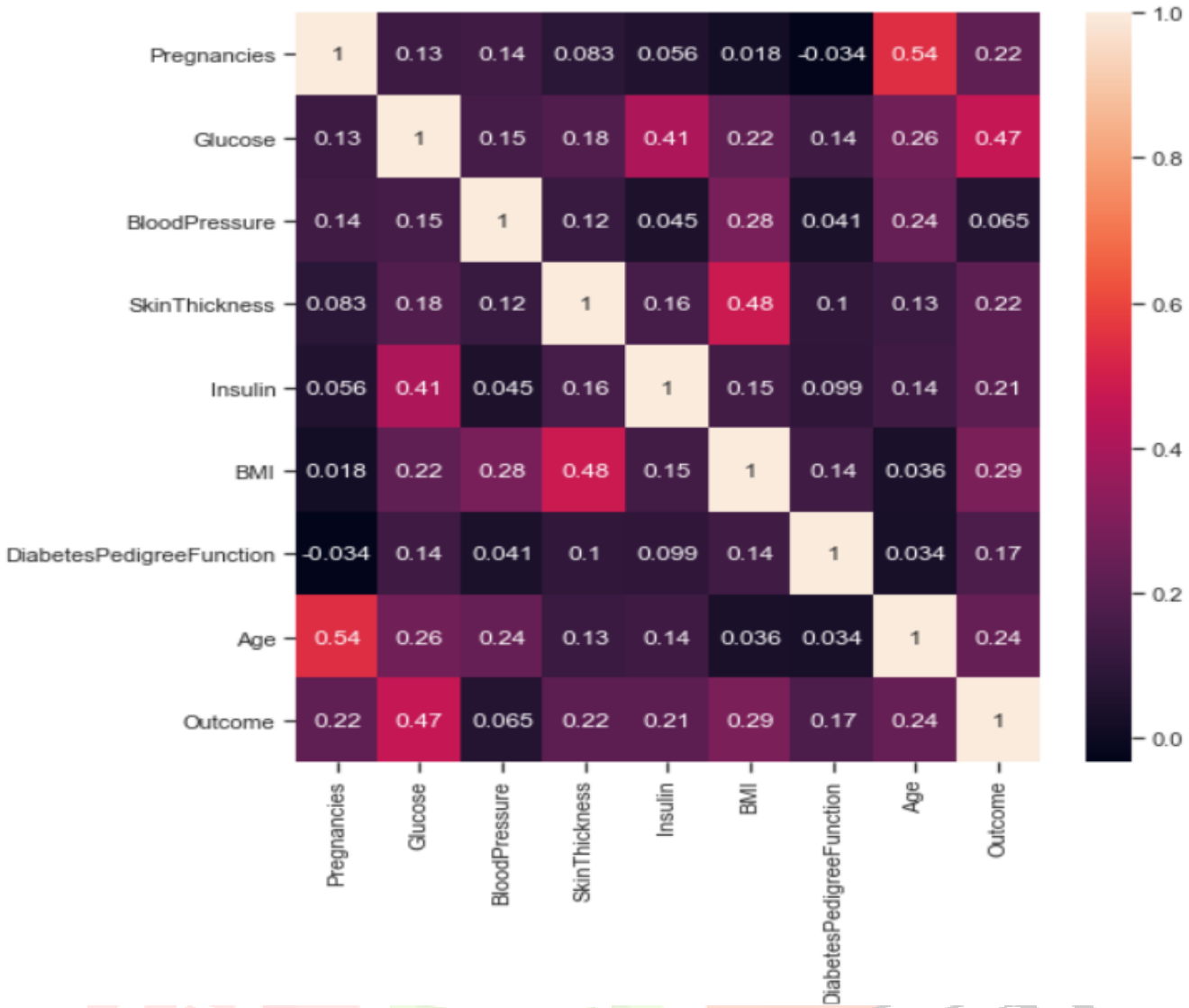
Table 1: Performance Comparison of Machine Learning Algorithms

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	79.20	0.79	0.78	0.78
Decision Tree	76.80	0.76	0.76	0.76
K-Nearest Neighbors	80.10	0.80	0.79	0.79
Support Vector Machine	82.60	0.82	0.82	0.82
Random Forest	**85.40**	**0.85**	**0.85**	**0.85**

The comparison shows that Random Forest achieved the highest prediction accuracy among all the implemented algorithms. The ensemble learning approach reduces overfitting and provides better generalization than individual classifiers.

Confusion Matrix

The confusion matrix was used to evaluate the prediction performance of the selected Random Forest model. It correctly classified most diabetic and non-diabetic patients while minimizing false predictions.



Actual / Predicted	Non-Diabetic	Diabetic
Non-Diabetic	91	7
Diabetic	9	47

Discussion

The experimental results demonstrate that proper data preprocessing significantly improves prediction performance. Random Forest achieved the best results because it combines multiple decision trees, making the model more robust and less sensitive to noisy data. Logistic Regression and SVM also produced satisfactory results, while Decision Tree showed relatively lower accuracy due to overfitting.

The proposed system can assist healthcare professionals in identifying individuals at risk of diabetes and supports early clinical decision-making.

V. CONCLUSION

This paper presented a machine learning-based diabetes prediction system using the Pima Indians Diabetes Dataset. Several supervised learning algorithms including Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, and Random Forest were implemented and evaluated.

Among all the models, Random Forest achieved the highest prediction accuracy and demonstrated superior classification performance. The system can serve as a reliable decision-support tool for the early detection of diabetes, enabling timely medical intervention and improved patient care.

Future work may include the use of larger clinical datasets, deep learning techniques, real-time prediction systems, and web or mobile-based healthcare applications to improve prediction accuracy and usability.

ACKNOWLEDGMENT

The authors sincerely thank **Rashmi K A**, Guide and Assistant Professor, Department of Computer Science and Engineering, SJC Institute of Technology, Chickballapur, Karnataka, for valuable guidance and continuous support. They also acknowledge the department's assistance and gratefully recognize the Kaggle dataset utilized, which significantly contributed to the successful completion of this research.

The authors also extend their sincere appreciation to the Department of Computer Science and Engineering, SJC Institute of Technology, Chickballapur, for providing the necessary facilities and a supportive research environment.

Furthermore, the authors gratefully acknowledge **Kaggle** for providing the publicly available dataset utilized in this study, which significantly contributed to the successful completion of the research.

REFERENCES

[1] UCI Machine Learning Repository, Pima Indians Diabetes Database, University of California, Irvine.

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[3] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[4] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[5] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2nd Edition, 2009.

[6] I. H. Witten, E. Frank, and M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2017.

[7] World Health Organization (WHO), Diabetes Fact Sheet, Geneva, Switzerland.

[8] American Diabetes Association, "Standards of Medical Care in Diabetes," Diabetes Care, Annual Edition.