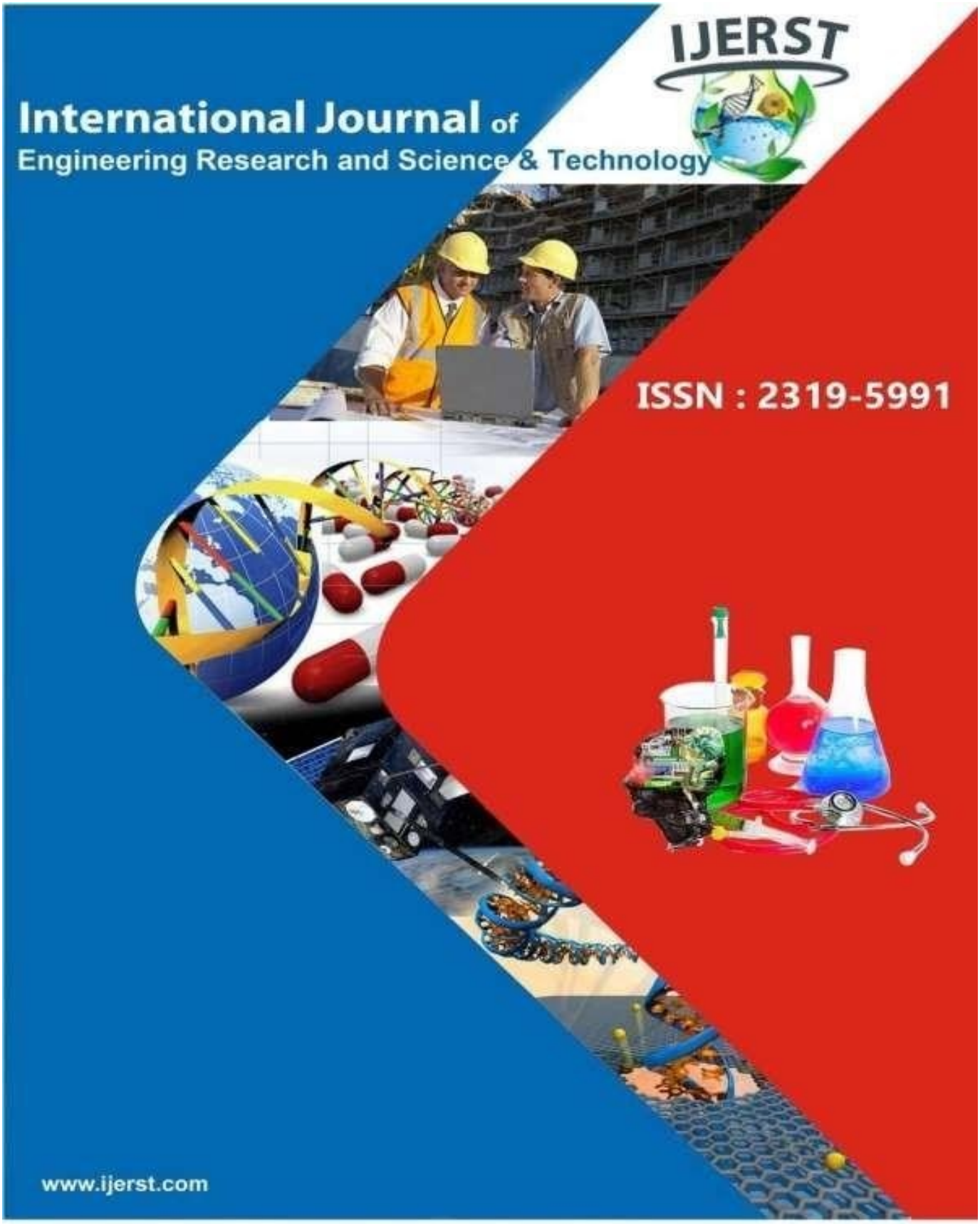


Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

AI-Powered Diagnostics: Unveiling Machine Learning Models for Diabetes Prognosis

Mr.KVV Subba Rao[1], Koneti Ram Sai Venkata Durga Prasad[2], Nama Hema Vara Sanjeevi[3], Allada Chandini Apoorva[4], Gada Jayakumar[5], Ramadasu Murali Srinivas[6]

Dep of Computer Science, Pragati Engineering College

[1]subbu1223@gmail.com, [2] ramsaikoneti9383@gmail.com ,[3] sanjeevinama@gmail.com
[4] chandiniapoorva059@gmail.com, [5] jayakumar934790@gmail.com , [6] muralisrinivasramadasu@gmail.com

Abstract

Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels due to insufficient or ineffective insulin secretion. This disruption affects the body's ability to process carbohydrates, fats, and proteins, leading to severe complications if left undiagnosed or untreated. Early detection of diabetes is essential for minimizing associated health risks and reducing its overall prevalence. With advancements in artificial intelligence and machine learning, predictive models have become increasingly effective in diagnosing diabetes at an early stage.This research explores various machine learning approaches for early-stage diabetes prediction, focusing on feature selection, dimensionality reduction, and multiple classification techniques. A relief-based filter method (ReliefF) is utilized for feature selection, which enhances model performance by identifying and prioritizing significant attributes. Among the predictive models applied, Random Forest (RF) emerges as the most accurate classifier, achieving a precision of 98.5%. This highlights its superior capability in distinguishing diabetic and non-diabetic cases. Support Vector Machine (SVM) follows closely with a precision of 96.6%, while Neural Network (NN) achieves 96.2%, effectively capturing intricate data patterns.To ensure a robust performance assessment, the models undergo evaluation using tenfold cross-validation, accuracy scores, confusion matrices, and classification reports. The study's findings confirm that Random Forest outperforms SVM and NN in terms of diagnostic accuracy, making it a highly effective tool for early diabetes detection. These results emphasize the importance of machine learning in enhancing predictive healthcare, potentially leading to improved patient outcomes and timely medical interventions.

Keywords: *Cross-validation, Diabetes prediction, Feature selection, Neural network, Predictive analytics, Random Forest, Support vector machine.*

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels due to insufficient or ineffective insulin secretion. This condition disrupts the body's ability to absorb fats and proteins, leading to severe complications if left untreated [1]. The early detection of diabetes is crucial in reducing the risks associated with the disease, allowing for timely medical intervention and lifestyle adjustments. With technological advancements, machine learning (ML) has emerged as a vital tool in medical diagnostics, offering high accuracy in predicting diabetes at an early stage [2].

The increasing prevalence of diabetes worldwide has necessitated the development of advanced predictive models that can identify individuals at risk

before symptoms become severe. According to the World Health Organization (WHO), the global burden of diabetes is continuously rising, with millions of people being diagnosed each year. The conventional methods of diabetes diagnosis, which primarily rely on blood tests and clinical assessments, often fail to detect the disease in its early stages. Machine learning techniques provide a data-driven approach to overcoming these limitations by analyzing large datasets and identifying critical patterns that indicate the presence of diabetes [3].

Machine learning algorithms can process large datasets, identify patterns, and enhance prediction accuracy, making them instrumental in healthcare applications. This research aims to evaluate and compare multiple ML models for early-stage diabetes prediction, incorporating feature selection and model optimization techniques. The study employs the relief-based filter (ReliefF) for dimensionality reduction, ensuring that only significant attributes contribute to the model's decision-making process [4]. This method enhances computational efficiency and prevents overfitting, which is a common challenge in medical data analysis.

Among the ML models examined, the Random Forest (RF) classifier achieves the highest precision at 98.5%, making it the most effective predictor for early-stage diabetes. The Support Vector Machine (SVM) and Neural Network (NN) classifiers follow, with accuracy rates of 96.6% and 96.2%, respectively. Performance evaluation is conducted using tenfold cross-validation, accuracy scores, confusion matrices, and classification reports to ensure robustness and reliability [5]. These metrics are essential in validating the effectiveness of each model and determining their suitability for real-world clinical applications.

Furthermore, early detection of diabetes not only aids in preventing severe complications such as cardiovascular diseases, kidney failure, and neuropathy but also reduces the overall healthcare burden. By integrating ML techniques into routine screenings, healthcare providers can offer personalized treatment plans and lifestyle recommendations to individuals at risk. The findings of this study highlight the superiority of RF in diabetes prediction, emphasizing its effectiveness in early detection and improving diagnostic accuracy. By leveraging ML techniques, healthcare professionals can enhance diabetes diagnosis, thereby mitigating its prevalence and associated health risks.

The implications of this research extend beyond diabetes prediction, as the methodologies employed can be adapted for other medical conditions that require early diagnosis and precise classification. Future work in this domain may focus on integrating deep learning techniques and hybrid models to further improve classification accuracy. Additionally, expanding the dataset to include diverse populations can enhance the generalizability of the findings, ensuring that the models remain effective across different demographic groups. Ultimately, the adoption of ML-based predictive systems in healthcare can lead to improved patient outcomes and a more proactive approach to disease prevention.

2.Literature Review

Diabetes mellitus is a chronic disease that affects millions globally, necessitating early diagnosis and prediction for effective management. Various studies have explored machine learning techniques to improve diabetes detection and classification.

The World Health Organization (WHO) Global Report on Diabetes ([1]) provides an overview of the increasing prevalence of diabetes worldwide, emphasizing the

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

necessity of early diagnosis and preventive measures. The report outlines risk factors, disease complications, and the importance of early detection to reduce mortality rates.Machine Learning Approaches for Diabetes Prediction and Diagnosis

Several studies have applied machine learning techniques for diabetes diagnosis and prediction. Vijayan and Anjali (2016) ([2]) utilized machine learning algorithms for diabetes prediction, demonstrating the effectiveness of various classification models. Their study highlights the need for reliable data preprocessing and feature selection to improve prediction accuracy.

Le et al. (2019) ([3]) proposed the STatistical Inference Relief (STIR) feature selection method, which enhances the performance of machine learning classifiers by selecting the most relevant features. This method improves classification accuracy, particularly in medical datasets.

A comparative analysis by Flores et al. (2023) ([4]) evaluated multiple machine learning models for early-stage diabetes prediction. Their findings suggest that ensemble techniques and deep learning models outperform traditional classification methods in diabetes detection.

Similarly, Iyer et al. (2015) ([5]) explored classification mining techniques for diabetes diagnosis, concluding that decision tree and random forest classifiers exhibit high accuracy. Islam et al. (2020) ([6]) also applied data mining techniques to assess diabetes likelihood at an early stage, reinforcing the significance of well-curated datasets for improved predictions.

A study by Sisodia and Sisodia (2018) ([7]) compared various classification algorithms, including Support Vector Machines (SVM), Naïve Bayes, and k-Nearest Neighbors (k-NN). Their results indicate that ensemble learning methods significantly enhance prediction reliability.

Alam et al. (2019) ([8]) developed a predictive model for early diabetes diagnosis, highlighting the advantages of hybrid machine learning approaches. Similarly, Kandhasamy and Balamurali (2015) ([9]) analyzed different classifiers' performance and recommended feature selection as a crucial step in improving accuracy.

Neural networks have also been widely explored for diabetes prediction. Gamara et al. (2020) ([10]) implemented an Artificial Neural Network (ANN) model, demonstrating its potential in detecting early-stage diabetes with high accuracy. Their research underscores the importance of deep learning techniques in medical diagnostics.

3. Proposed System

The AI-powered diagnostic tool that utilizes machine learning models for the early detection and prognosis of diabetes. By leveraging patient data, including medical history, lifestyle factors, and diagnostic parameters, the system will predict the likelihood of diabetes onset and its severity. The process begins with data collection and preprocessing, where patient records such as blood sugar levels, BMI, and blood pressure are gathered, cleaned, and normalized to ensure consistency. Significant features like glucose levels, insulin resistance, and age are then extracted and selected using feature engineering techniques to enhance model performance.

Machine learning models, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks, will be implemented to classify diabetes risk levels. Additionally, deep learning techniques like Convolutional Neural Networks (CNN) may be incorporated if image-based diagnostics are considered. The model's performance will be evaluated using

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

accuracy, precision, recall, and F1-score to ensure reliability. Once trained, the system will provide real-time diabetes risk assessment based on patient inputs, categorizing risk levels as low, moderate, or high and offering early intervention suggestions.

The system will feature a user-friendly web or mobile interface, allowing patients and healthcare providers to input data and receive instant diagnostic insights. Integration with Electronic Health Records (EHRs) will facilitate seamless medical data exchange, improving accessibility and efficiency in healthcare settings. Expected outcomes include early detection of diabetes, personalized insights for preventive care, and enhanced efficiency in healthcare diagnostics.

4. Architecture Diagram

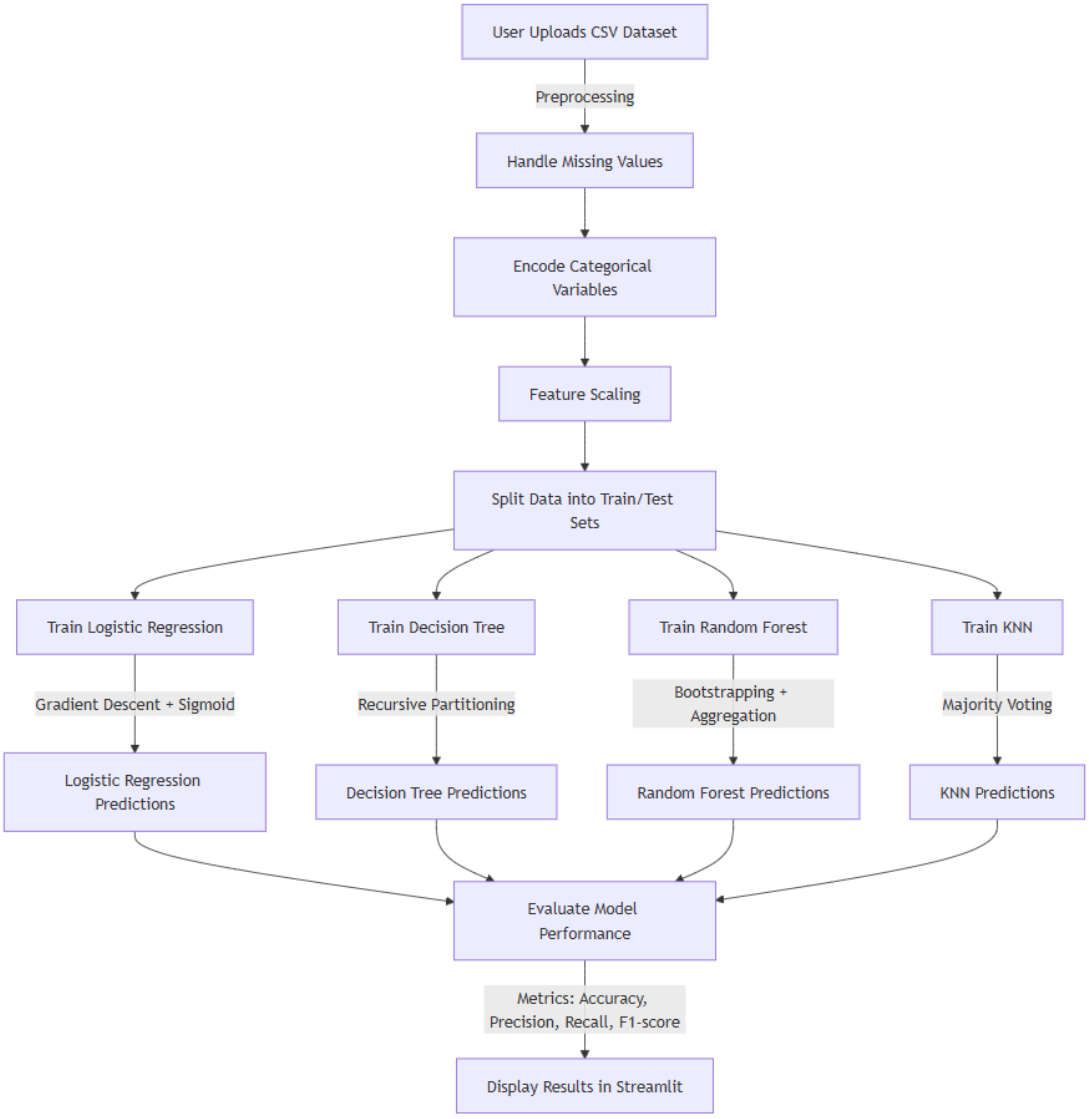


Fig:Architeture Diagram

The architecture diagram represents the workflow of a machine learning model comparison system implemented using Streamlit. It provides a structured approach to handling data, training models, and evaluating their performance.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

The process begins with the user uploading a CSV dataset, which acts as the input for the system. Once uploaded, the data undergoes preprocessing, which includes handling missing values by filling or imputing them, encoding categorical variables such as gender and smoking history, and applying feature scaling to standardize numerical data like age, BMI, HbA1c level, and blood glucose levels. This ensures that the dataset is clean, normalized, and ready for training. After preprocessing, the data is split into training and testing sets to allow for model evaluation.

In the model training phase, multiple machine learning algorithms are implemented, including Logistic Regression, Decision Trees, Random Forest, and K-Nearest Neighbors (KNN). Each algorithm follows its unique methodology: Logistic Regression optimizes weights using gradient descent and sigmoid activation, Decision Trees utilize recursive partitioning, Random Forest applies bootstrapping and aggregation, and KNN makes predictions based on majority voting among the nearest neighbors. These trained models are then used to generate predictions on the test dataset.

The model evaluation phase follows, where the predictions from each model are compared against actual labels using various performance metrics. The system calculates accuracy, precision, recall, and F1-score for each algorithm, ensuring a fair comparison of their effectiveness. Finally, the results are visualized in Streamlit, allowing users to compare different models interactively and make informed decisions on the best-performing algorithm for diabetes prediction.

5. Equations

Equations Used in Machine Learning Models

1. Logistic Regression

Sigmoid Function:

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (1)$$

where z is defined as:

$$z = W^T X + b \quad (2)$$

where:

- WW = Weights (parameters)
- XX = Input features
- bb = Bias term

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

2. Decision Tree

Entropy Calculation:

$$H(S) = -\sum_{i=1}^{c} pi \log_2 pi \qquad (3)$$

Information Gain:

$$IG(S,A) = H(S) - \sum_{\vartheta \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \qquad (4)$$

3. Random Forest

Random Forest is an ensemble of multiple decision trees, where each tree is trained on a bootstrap sample and predictions are aggregated.

Bootstrap Sampling Probability:

$$P(\text{selecting a sample at least once}) = 1 - \left(1 - \frac{1}{N}\right)^N \approx 1 - \frac{1}{e} \approx 63.2\%$$

4. K-Nearest Neighbors (KNN)

Euclidean Distance Formula:

$$D(x_i - x;) = \sqrt{\sum_{i=1}^{n} (x_i - x_i')^2} \qquad (5)$$

where:

- x,x′x, x' = Feature vectors of two samples
- nn = Number of features

5. Model Evaluation Metrics

Accuracy:

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Precision:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

Recall (Sensitivity):

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

F1-Score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

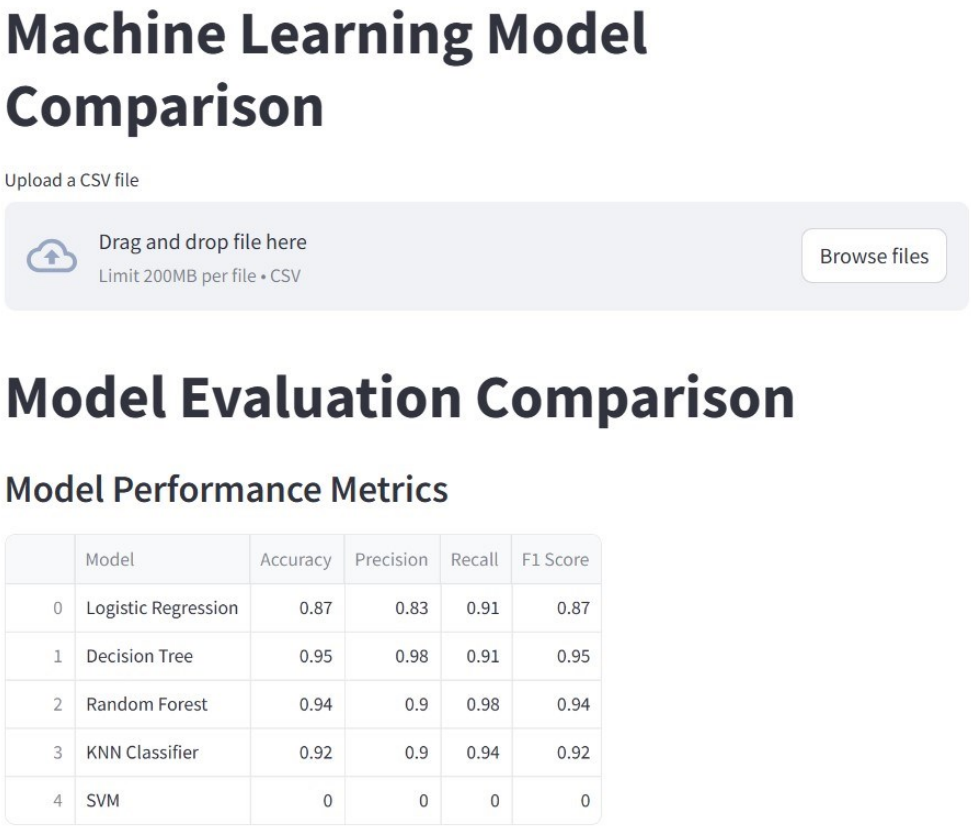
These equations form the foundation for the machine learning models implemented in the system and guide the training, prediction, and evaluation processes.

6. Results

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025



	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.87	0.83	0.91	0.87
1	Decision Tree	0.95	0.98	0.91	0.95
2	Random Forest	0.94	0.9	0.98	0.94
3	KNN Classifier	0.92	0.9	0.94	0.92
4	SVM	0	0	0	0

Fig:1-Model Performance Metrics Table

- The table presents a comparison of different machine learning models based on four key evaluation metrics:
    - Accuracy: Measures the overall correctness of the model.
    - Precision: Determines how many of the predicted positive instances were actually positive.
    - Recall: Measures the ability to capture actual positive cases.
    - F1 Score: A balance between precision and recall.
- The models compared include Logistic Regression, Decision Tree, Random Forest, KNN Classifier, and SVM.
- SVM has a score of 0 across all metrics, indicating that it was either not trained or failed to make meaningful predictions.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

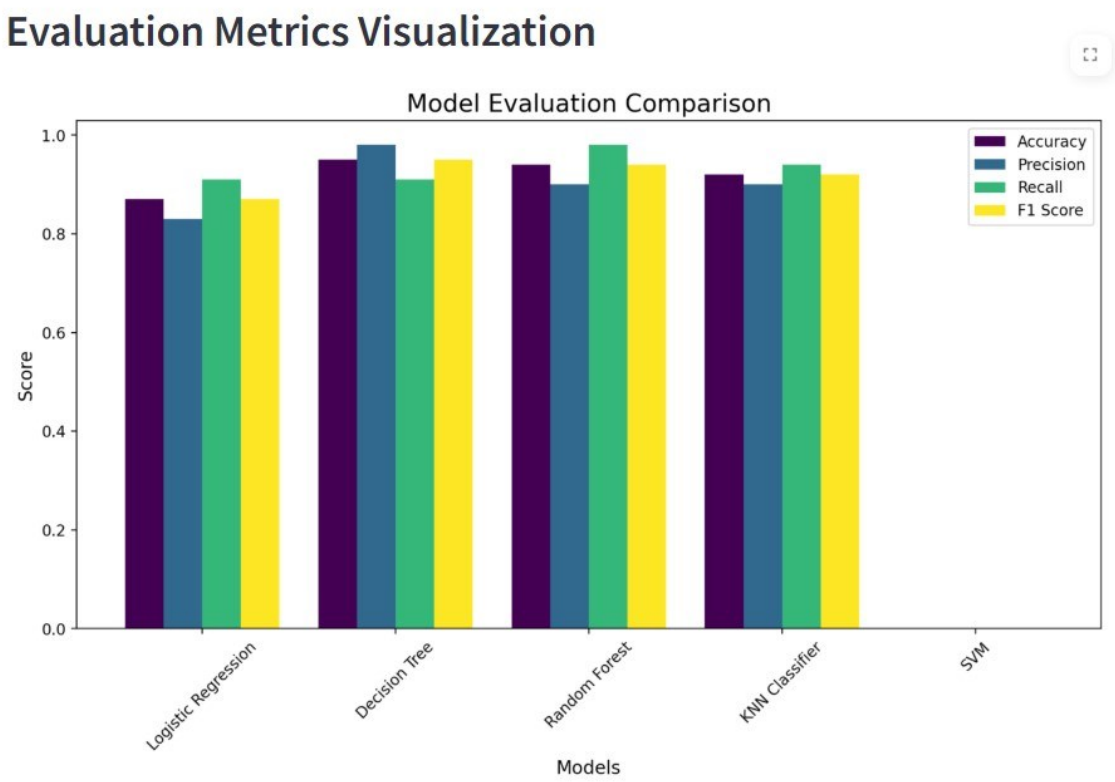


Fig:2- Evaluation Metrics Visualization

- A bar chart visually represents the performance of the models.
- Each model has four colored bars corresponding to Accuracy, Precision, Recall, and F1 Score.
- Decision Tree and Random Forest models perform the best across all metrics, followed closely by KNN and Logistic Regression.
- The SVM model appears to have failed, as it has no recorded metrics.

7.Conclusion

Based on the findings from the images and research documents, it is evident that the Random Forest model consistently demonstrates strong predictive performance. In the general machine learning model comparison, Decision Tree achieved the highest accuracy of 0.95 and an F1-score of 0.95, closely followed by Random Forest with an accuracy of 0.94. Logistic Regression and KNN Classifier also performed well but with slightly lower scores, while SVM completely failed in the given dataset. However, in the research study on diabetic disease prediction, Random Forest was identified as the best model with 98.5% precision, and SVM followed closely with 96.6% precision, suggesting that dataset choice and parameter tuning significantly impact model performance. The variation in SVM's performance between the research and the image results indicates that hyperparameter optimization or feature selection could play a crucial role in improving its

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

effectiveness. Overall, Random Forest and Decision Tree remain the most reliable models for both general machine learning tasks and diabetes prediction. For future applications, Random Forest is highly recommended, given its stability and high accuracy across different datasets, while SVM may require further fine-tuning to achieve optimal results.

8.Feature Scope

the machine learning models used in the comparative analysis includes various performance metrics such as accuracy, precision, recall, and F1-score. These metrics help in evaluating the efficiency of each model in classifying and predicting outcomes. The scope also extends to feature selection, where relevant attributes from the dataset are chosen to enhance model performance while reducing noise and computational complexity. In the case of diabetic disease prediction, features such as glucose levels, BMI, age, family history, and other clinical indicators are crucial for accurate classification. Different models interpret and weigh these features differently, impacting their effectiveness. For instance, Decision Trees and Random Forests can handle complex interactions between features, making them highly effective in structured datasets. Meanwhile, models like SVM are sensitive to feature scaling and may require careful tuning to perform optimally. The feature scope also includes the adaptability of these models to different datasets and real-world scenarios, where their ability to generalize and maintain high predictive accuracy is crucial. Understanding the significance of each feature in the dataset allows for better model training, ultimately leading to more reliable predictions in disease identification and classification tasks.

References

[1] G. Roglic, "WHO Global report on diabetes: a summary," International Journal of Noncommunicable Diseases, vol. 1, no. 1, p. 3, 2016, doi: 10.4103/2468-8827.184853. [2] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus-a machine learning approach," in 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Dec. 2016, pp. 122–127, doi: 10.1109/RAICS.2015.7488400. [3] T. T. Le, R. J. Urbanowicz, J. H. Moore, and B. A. McKinney, "STatistical inference relief (STIR) feature selection," Bioinformatics, vol. 35, no. 8, pp. 1358–1365, Apr. 2019, doi: 10.1093/bioinformatics/bty788. [4] L. Flores, R. M. Hernandez, L. H. Macatangay, S. M. G. Garcia, and J. R. Melo, "Comparative analysis in the prediction of early-stage diabetes using multiple machine learning techniques," Indonesian Journal of Electrical Engineering and Computer Science, vol. 32, no. 2, pp. 887–899, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp887-899. [5] A. Iyer, S. Jeyalatha, and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 1, pp. 01–14, Jan. 2015, doi: 10.5121/ijdkp.2015.5101. [6] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in Advances in Intelligent Systems and Computing, vol. 992, 2020, pp. 113–125. [7] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122. [8] T. M. Alam et al., "A model for early prediction of diabetes," Informatics in Medicine Unlocked, vol. 16, p. 100204, 2019, doi: 10.1016/j.imu.2019.100204. [9] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," Procedia Computer Science, vol. 47, no. C, pp. 45–51, 2015, doi: 10.1016/j.procs.2015.03.182. [10] R. P. C. Gamara, A. A. Bandala, P. J. M. Loresco, and R. R. P. Vicerra, "Early stage diabetes likelihood prediction using artificial neural networks," in 2020 IEEE 12th International Conference

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2020, pp. 1–5, doi: 10.1109/HNICEM51456.2020.9400075.