Diabetes Risk Prediction with Machine Learning Models

Sundhar K A
Department of Computer Science Engineering
Vellore Institute of Technology
Tamil Nadu, India
s64kannan@gmail.com

Naveen Kumar V
Department of Computer Science Engineering
Vellore Institute of Technology
Tamil Nadu, India
naveenmadhan32@yahoo.in

Karthik V
Department of Computer Science Engineering
Vellore Institute of Technology
Tamil Nadu, India
karthikvijayakumar2004@gmail.com

Krishnamoorthy A
School of Computer Science & Engineering
Vellore Institute of Technology
Tamil Nadu, India
krishnamoorthyece@gmail.com

Abstract
The increasing prevalence of diabetes worldwide necessitates efficient and cost-effective methods for early diagnosis. This study investigates the performance of various machine learning algorithms in predicting diabetes using a publicly available dataset. The evaluated models include Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Support Vector Classifier (SVC), Gradient Boosting, and a Neural Network. Model performance was assessed using metrics such as accuracy, precision, recall, and F1-score. Among the models, the Decision Tree achieved the highest accuracy (79.22%), followed by Random Forest, KNN, and Neural Network (74–75%). These findings highlight the potential of machine learning in healthcare applications, particularly for early diabetes detection. The study also identifies strengths and limitations of each model, offering insights and recommendations for future research.

Keywords: Diabetes Prediction, Machine Learning, Logistic Regression, K-Nearest Neighbors, Random Forest, Decision Tree, Support Vector Classifier, Gradient Boosting, Neural Networks, Model Evaluation, Healthcare, Early Diagnosis.

1. Introduction

Diabetes is a widespread chronic disease affecting millions globally, with its prevalence posing a significant public health challenge. The long-term complications of diabetes, such as cardiovascular diseases, kidney failure, and neuropathy, place a substantial burden on healthcare systems. Early detection and diagnosis are essential to managing or preventing these complications. However, traditional diagnostic methods, while reliable, can often be time-consuming and costly.

In recent years, machine learning (ML) has emerged as a transformative tool in healthcare, including for diabetes prediction. ML algorithms can analyze large datasets, identify patterns, and make predictions based on features such as blood sugar levels, age, BMI, and family history. These capabilities offer a faster and potentially more accurate alternative to traditional diagnostic methods.

The purpose of this study is to assess how well different machine learning algorithms predict diabetes. We explore several widely used models, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Support Vector Classifier (SVC), Gradient Boosting, and Neural Networks. By analyzing their performance using metrics such as accuracy, precision, recall, and F1-score, we seek to identify the most suitable technique for diabetes prediction.

The findings of this study can contribute to developing efficient screening tools for early diabetes detection, assisting healthcare professionals in timely diagnosis and intervention. Additionally, this research highlights the potential of machine learning in healthcare applications while providing insights into the strengths and limitations of various predictive model.

2. Literature Survey

Prediction of Diabetic Condition Using Different Machine Learning Approaches and Different Datasets (Year: 2020) discusses various machine learning techniques for predicting diabetes, emphasizing the critical role of early detection to mitigate complications. It reviews algorithms such as Support Vector Machines (SVM), Random Forests, and Decision Trees, highlighting their effectiveness across different datasets. This is complemented by A Framework for Type-II Diabetes Prediction Using Machine Learning Approaches (Year: 2021), which introduces a methodology focused on preprocessing medical data with missing values and varying ranges. It evaluates the PIMA Indian Diabetes dataset and compares eight classification algorithms, including Naive Bayes and Logistic Regression, demonstrating that

effective data preprocessing is essential for enhancing model performance, with Naive Bayes achieving the highest accuracy.

Further, Diabetes Prediction Using Python Machine Learning Techniques (Year: 2022) explores the application of various algorithms, such as Random Forest, SVM, and K-Nearest Neighbors (KNN), for predicting diabetes. This study emphasizes the significance of machine learning in healthcare, particularly for early diagnosis, which is crucial for effective diabetes management. Building on this, Diabetes Prediction Model Using Machine Learning Techniques (Year: 2023) presents an innovative model that employs a range of machine learning techniques, including ensemble methods like XGBoost and CatBoost. The findings indicate that ensemble learning significantly enhances prediction accuracy, with CatBoost achieving the highest accuracy and AUC-ROC score, suggesting its potential for clinical applications.
Additionally, Diabetes Prediction Using Different Machine Learning Classifiers (Year: 2021) investigates the effectiveness of various classifiers, including Decision Trees, SVM, and Naive Bayes, in predicting the onset of diabetes. It highlights the importance of tailored algorithm selection for optimizing predictive accuracy and reveals nuanced performance variations among classifiers, providing valuable insights for healthcare practitioners. Lastly, Diabetes Prediction Using Different Machine Learning Approaches (Year: 2022) further discusses the application of several machine learning algorithms to predict diabetes, stressing the significance of early detection. The study showcases the effectiveness of different algorithms and concludes that a combination of methods may yield the best predictive performance.

3. Methodology

3.1.Dataset

This study utilizes the *diabetes.csv* dataset, a widely recognized dataset for diabetes prediction research. It contains 768 entries and 9 attributes, capturing both medical indicators and demographic factors relevant to diabetes risk assessment. Key features include blood glucose levels, BMI (Body Mass Index), blood pressure, insulin levels, and skin thickness. Additionally, demographic factors such as age and family history of diabetes provide a broader understanding of predisposition to the condition. The target variable is binary, indicating whether an individual has diabetes (1) or not (0).

To address missing values and maintain data continuity, the forward fill method was applied. This approach propagates the last valid observation to fill gaps, ensuring data integrity while minimizing biases that could otherwise impact model performance.

3.2Preprocessing

To mitigate discrepancies in feature ranges and enhance model performance, feature scaling was performed. Continuous variables such as blood glucose levels and BMI were standardized using the StandardScaler, ensuring each feature had a mean of 0 and a standard deviation of 1. This step is particularly crucial for algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), which are sensitive to the scale of input features.

The dataset was split into training and testing sets in an 80:20 ratio to evaluate model performance. Stratified sampling was used to maintain the proportional representation of diabetic and non-diabetic cases in both subsets, a critical step for addressing class imbalance. This ensures that the models generalize effectively and avoid bias toward the majority class.

3.3 Machine Learning Models

The study utilized a variety of **machine learning algorithms**, both linear and non-linear, to predict the likelihood of diabetes. These include:

- **Logistic Regression**: A statistical model for binary classification that forecasts an outcome's likelihood is called logistic regression. It is a simple, interpretable model ideal for linear relationships between features and the target variable, often serving as a baseline for performance comparison.
- **K-Nearest Neighbors (KNN)**: A non-parametric, distance-based classification algorithm that assigns a class to a data point based on the majority class of its nearest neighbors. Key hyperparameters include the number of neighbors (n_neighbors) and the distance metric (Euclidean or Manhattan), which were optimized during the model's training.
- **Decision Tree**: A tree-based model that splits the dataset into branches based on feature thresholds to make predictions. Decision Trees can capture complex patterns without requiring feature scaling, and they offer the benefit of interpretability and visualization.
- **Random Forest**: An ensemble method that aggregates predictions from multiple decision trees. The model uses n_estimators (number of trees), max_depth (maximum depth of trees), and min_samples_split (minimum samples

needed to split a node) to reduce overfitting and improve accuracy.

- **Gradient Boosting**: An ensemble technique in which the errors made by the prior trees are corrected by successively building new trees. Important hyperparameters include learning_rate (which determines the contribution of each new tree), n_estimators (the number of trees), and max_depth (the depth of each tree). This model is designed to focus on improving areas where the previous models performed poorly.
- **Support Vector Classifier (SVC)**: A model that finds the optimal hyperplane to separate data points of different classes. Key hyperparameters include the kernel function (e.g., linear, RBF) and the C parameter, which controls the trade-off between a smooth decision boundary and classifying training points correctly.
- **Neural Network**: A deep learning model consisting of interconnected layers that learn non-linear relationships between input features and the target variable. Neural networks are more data-intensive and require larger datasets for training, often providing better performance in more complex tasks.

For hyperparameter tuning, **GridSearchCV** was employed to fine-tune models such as KNN, Random Forest, and Gradient Boosting by optimizing hyperparameters like the number of neighbors, the number of trees, and the learning rate.

3.4 Evaluation Metrics

To comprehensively assess the performance of the machine learning models, multiple evaluation metrics were employed, each highlighting a distinct aspect of classification accuracy:

- **Accuracy**:
  Accuracy represents the proportion of correctly predicted instances out of the total instances. It provides a general measure of overall model performance across both positive and negative classes. However, it may not be reliable for imbalanced datasets, as it could be skewed by the majority class.
- **Precision**:
  Precision evaluates the quality of positive predictions by measuring the proportion of true positive instances among all predicted positive instances. It is especially critical in scenarios where false positives have significant consequences, such as misclassifying a non-diabetic individual as diabetic.
- **Recall (Sensitivity)**:
  Recall measures the model's ability to identify all actual positive cases, representing the proportion of true positive instances detected out of all actual positives. High recall is crucial in situations where missing positive cases (false negatives) could have severe repercussions, such as undiagnosed diabetes.
- **F1-Score**:
  The F1-score provides a balanced evaluation by taking the harmonic mean of precision and recall. It is particularly useful for imbalanced datasets, where considering both false positives and false negatives is essential.
- **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)**:
  The ROC-AUC metric evaluates the model's ability to distinguish between positive and negative classes across various decision thresholds. A higher AUC indicates better model performance, with a value of 1.0 representing perfect classification and 0.5 representing random guessing.
- **Confusion Matrix**:
  The confusion matrix visualizes the model's predictions by categorizing them into four groups: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It provides detailed insights into the types of errors the model makes, facilitating targeted optimizations.

4. Result

4.1 Performance Metrics

The performance of each model was evaluated using accuracy as the primary metric. The accuracies achieved by each machine learning model are as follows:

- **Logistic Regression**: 71.43%
- **K-Nearest Neighbor (KNN)**: 74.68%
- **Random Forest**: 74.68%
- **Decision Tree**: 79.22%
- **Neural Network**: 74.02%
- **Gradient Boosting**: 73.37%
- **SVC (Support Vector Classifier)**: 74.67%

For hyperparameter tuning, several models underwent optimisation. **K-Nearest Neighbor (KNN)** performed best with a k-value of 5, while **Random Forest** achieved improved performance with 100 trees. The **Neural Network** showed

notable improvements after 10 epochs, highlighting the model's potential for further refinement with additional training or deeper architectures.

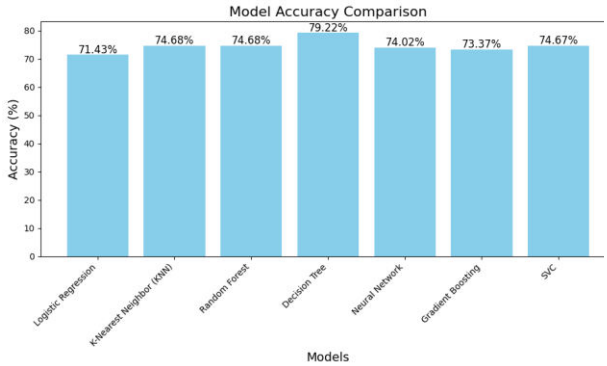

Fig.1 Accuracy Comparison

Confusion Matrix and Classification Reports

The confusion matrices for each model provide detailed insights into the performance, including true positives, false positives, true negatives, and false negatives. These matrices reveal the models' ability to classify positive and negative cases of diabetes correctly.

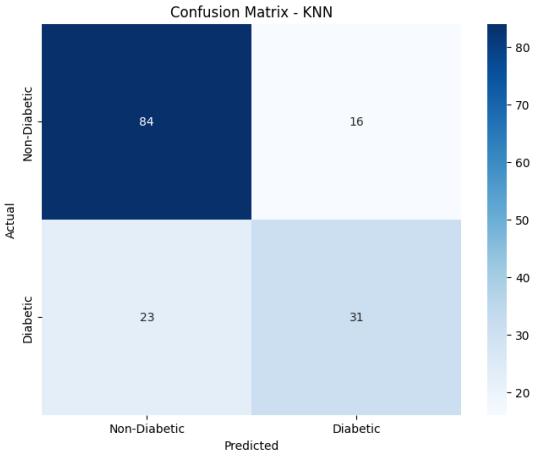

Fig.1. KNN confusion matrix

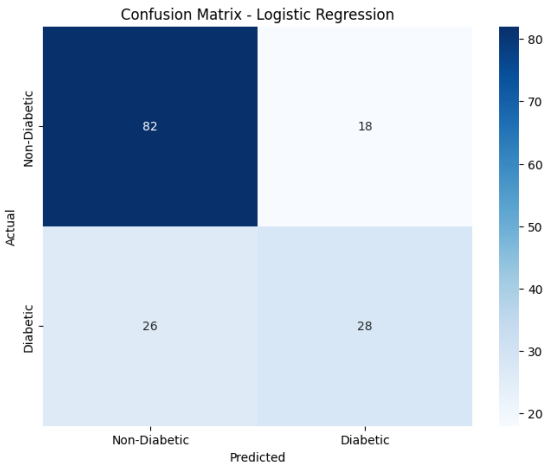

Fig.2. Logistic Regression confusion matrix

Fig.3. Decision Tree confusion matrix

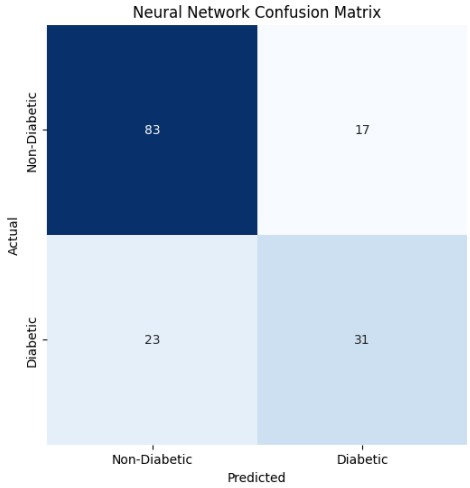

Fig.4. NN confusion matrix

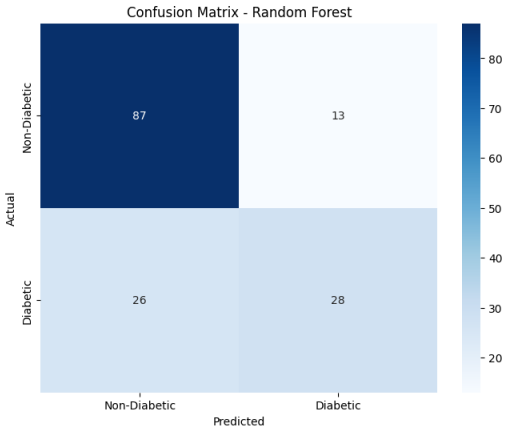

Fig.5 Random Forest confusion Matrix

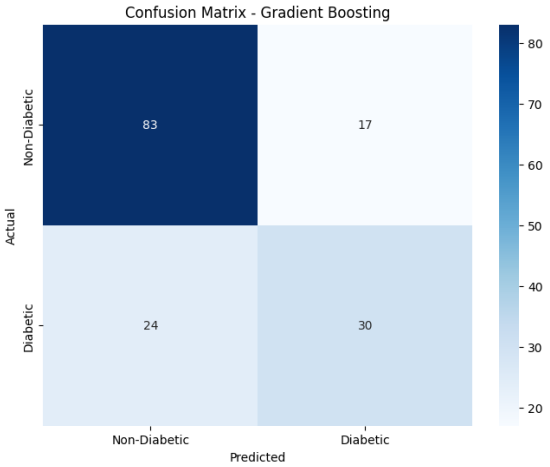

Fig.6 Gradient Boosting confusion matrix

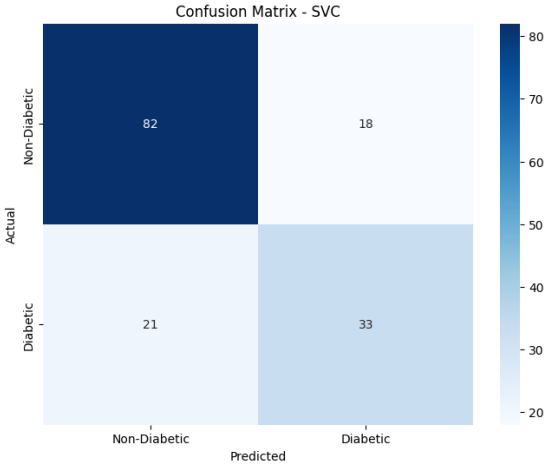

Fig.7. SVC confusion matrix

Key Observations

- The **Decision Tree** algorithm achieved the highest accuracy among traditional models, showcasing its ability to handle complex decision boundaries effectively.
- The **Neural Network** model showed consistent improvement in validation accuracy, indicating its potential for deeper models or further fine-tuning. While its initial accuracy was not as high as the Decision Tree, its learning capability may provide benefits as more data and training epochs are incorporated.
- **Random Forest** and **K-Nearest Neighbor (KNN)** demonstrated similar performance, with **Random Forest** showing slightly better results due to its ensemble nature.
- The **SVC** model performed similarly to **K-Nearest Neighbor (KNN)** and **Random Forest**, achieving an accuracy of 74.67%.
- **Gradient Boosting** achieved an accuracy of 73.37%, which is slightly lower than **KNN** and **Random Forest**, but still demonstrates its potential in handling complex data.

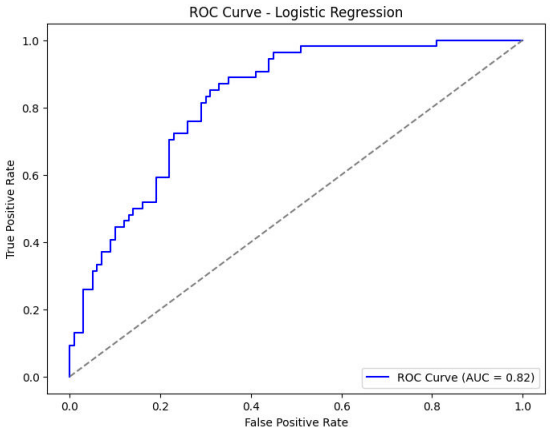

Fig.8.Logistic Regression ROC curve

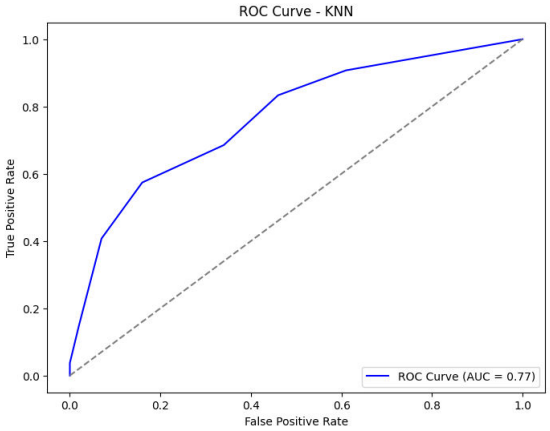

Fig.9. KNN ROC curve

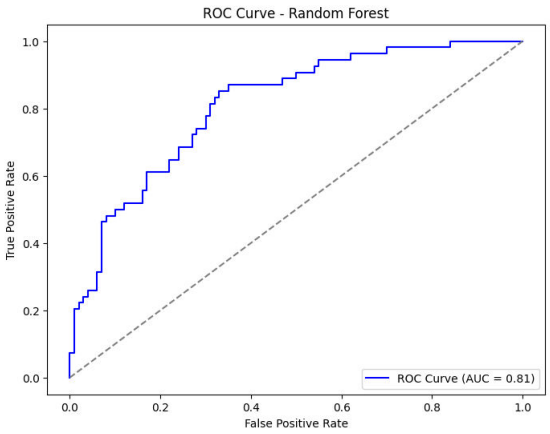

Fig.10. Random Forest ROC curve

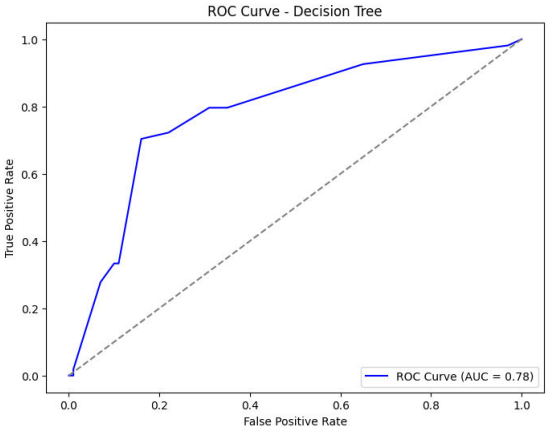

Fig.11. Decision Tree ROC curve

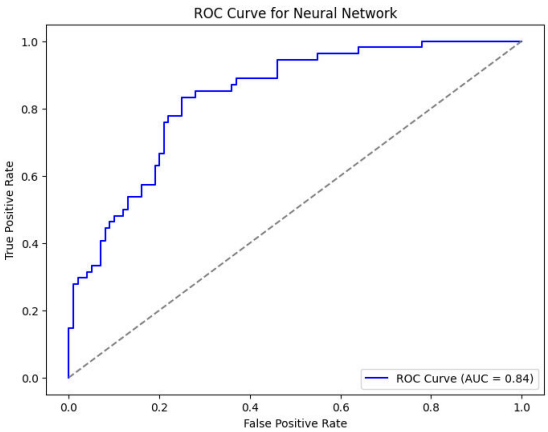

Fig.12. NN ROC curve

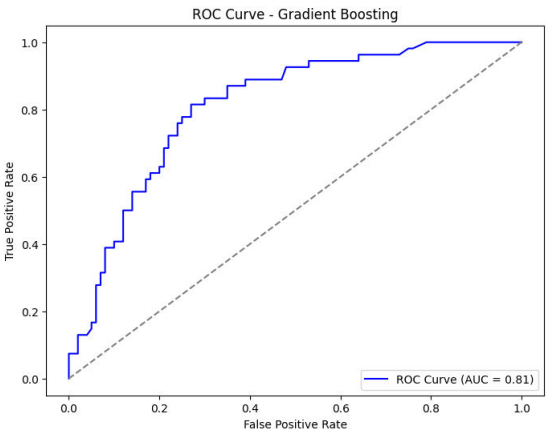

Fig.13. Gradient Boosting ROC curve

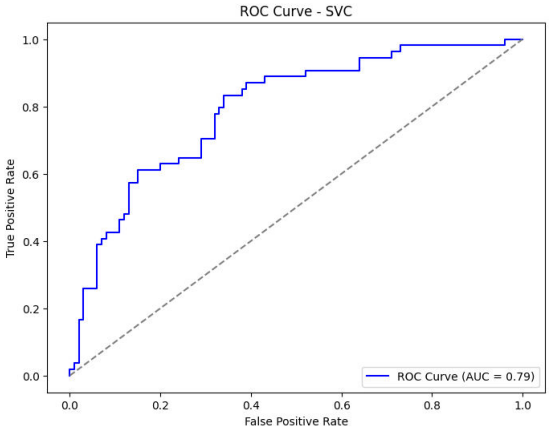

Fig.14. SVC ROC curve

5. Discussion

This research aimed to develop an effective machine learning model for diabetes prediction, using a range of algorithms to assess their ability to classify individuals as diabetic or non-diabetic based on medical and demographic features. The models tested included Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting, Support Vector Classifier (SVC), and Neural Networks.

Among these, Decision Tree and Random Forest performed the best, providing strong accuracy and interpretability. Random Forest, being an ensemble method, reduced the risk of overfitting, while Decision Tree offered clear insights into the feature importance and classification process, making it potentially more interpretable and actionable for healthcare professionals. Gradient Boosting, which builds trees sequentially to correct errors from previous trees, also showed promising results, though it required careful tuning to optimize performance. SVC showed competitive results but was sensitive to kernel selection and parameter tuning. Neural Networks, while capable of learning complex patterns, did not outperform the tree-based methods, likely due to the dataset's size and the simplicity of the network used.

Evaluation metrics like precision, recall, F1-score, and ROC-AUC helped assess the model's ability to handle class imbalance, a common challenge in diabetes datasets. In healthcare contexts, recall is particularly critical, as false negatives could mean missing a diabetes diagnosis, potentially delaying treatment. Stratified sampling ensured that both diabetic and non-diabetic cases were represented in the train-test split, but models still showed room for improvement, particularly in reducing false negatives. These findings suggest that Decision Trees could be especially useful in healthcare due to

their interpretability, which allows practitioners to better understand and trust the model's predictions. Future work should focus on optimizing models to minimize false negatives and improve recall, ensuring more reliable detection of diabetic cases.

6. Conclusion

In conclusion, this study explored the effectiveness of various machine learning algorithms for predicting diabetes based on medical and demographic features. Among the models tested, Decision Tree and Random Forest achieved the highest performance, offering a good balance between accuracy and interpretability. Random Forest, in particular, provided robust predictions due to its ensemble nature, which helps reduce overfitting. Other models, such as KNN, Gradient Boosting, and SVC, also showed promising results, with fine-tuning improving their performance.

The evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, were essential in understanding the models' strengths and limitations, especially when handling imbalanced datasets. While the models performed well overall, there is still room for improvement, particularly in reducing false negatives and increasing recall for the diabetic class.

This research highlights the potential of machine learning in healthcare, particularly for early diabetes prediction. Further work could focus on incorporating more features, addressing class imbalance, and exploring more advanced models to improve prediction accuracy.

References:

[1] P. Thakral and J. K. Sandhu, "Prediction of Diabetic condition using Different Machine Learning Approaches and Different Datasets," in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, pp. 1383-1388, Aug. 2023, IEEE.

[2] A. Kiran, M. Khan, J. C. Babu, B. S. Kumar, S. J. Ahmed, and Z. A. Khan, "Diabetes Prediction Using Python Machine Learning Techniques," in *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, vol. 1, pp. 1-8, Oct. 2023, IEEE.

[3] S. K. S. Modak and V. K. Jha, "Diabetes prediction model using machine learning techniques," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 38523-38549, 2024.

[4] M. F. U. Bhuiyan, M. T. Rahman, M. A. Anik, and M. Khan, "A Framework for Type-II Diabetes Prediction Using Machine Learning Approaches," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-6, Jul. 2021, IEEE.

[5] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 367-371, Mar. 2019, IEEE.

[6] Anonymous, "Diabetes Prediction Using Different Machine Learning Approaches," Year: 2022.

[7] A.A. Aljumah, M.G. Ahamad, and M.K. Siddiqui, "Application of data mining: Diabetes health care in young and old patients," *Journal of King Saud University - Computer and Information Sciences*, vol. 25, pp. 127–136, 2013. doi:10.1016/j.jksuci.2012.10.003.

[8] R. Arora and Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," *International Journal of Computer Applications*, vol. 54, pp. 21–25, 2012. doi:10.5120/8626-2492.

[9] M.P. Bamnote and G.R., "Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming," *Advances in Intelligent Systems and Computing*, vol. 1, pp. 763–770, 2014. doi:10.1007/978-3-319-11933-5.

[10] D.K. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima Indian Diabetes Dataset Using Naive Bayes with Genetic Algorithm as an Attribute Selection," in *Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016)*, pp. 451–455, 2017.

[11] Kanchan B. Dhomse and M.K.M., "Study of Machine Learning Algorithms for Special Disease Prediction Using Principal of Component Analysis," in *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication*, IEEE, pp. 5–10, 2016.

[12] A.A. Sharief and A. Sheta, "Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming," *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, vol. 3, pp. 54–59, 2014. doi:10.14569/IJARAI.2014.031007.

[13] D. Sisodia, S.K. Shrivastava, and R.C. Jain, "ISVM for Face Recognition," in *Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010*, pp. 554–559, 2010. doi:10.1109/CICN.2010.109.

[14] D. Sisodia, L. Singh, and S. Sisodia, "Fast and Accurate Face Recognition Using SVM and DCT," in *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, Springer, pp. 1027–1038, 2014.

[15] https://www.kaggle.com/johndasilva/diabetes

[16] A.S. Rani and S. Jyothi, "Performance Analysis of Classification Algorithms Under Different Datasets," in *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, IEEE, pp. 1584–1589, March 2016.