Diabetes Disease Prediction using Machine Learning: A Comparative Study

Md Khateebur Rab, Alok Gupta and Varad Gupta
Department of Computer Science
Indian Institute of Information technology, Ranchi
Jharkhand, India

***Abstract*—This paper presents a comprehensive machine learning approach to predict the likelihood of diabetes in patients using the Pima Indians Diabetes Database. The study involves detailed data cleaning including handling of physiologically implausible zero values, exploratory data analysis (EDA) to understand feature relationships, and the systematic training, evaluation, and comparison of several supervised learning models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Random Forests. Model performance is assessed using accuracy, precision, recall, F1-score, and AUC-ROC metrics. Hyperparameter tuning using GridSearchCV with cross-validation was performed to optimize each model. The results demonstrate the effectiveness of these models, with Logistic Regression achieving a competitive accuracy of approximately 77%, balancing performance with interpretability. This study highlights the importance of a structured methodology in developing predictive models for healthcare applications and discusses the clinical relevance and limitations of the findings.**

***Index Terms*—Diabetes Prediction, Machine Learning, Classification, Pima Indians Diabetes Database, Data Preprocessing, Model Evaluation, Logistic Regression, Random Forest, SVM, KNN, Decision Tree, Healthcare Analytics.**

I. Introduction

Diabetes Mellitus is a significant global health concern, characterized by chronic hyperglycemia resulting from defects in insulin secretion or action [1]. Its prevalence is rapidly increasing worldwide, leading to substantial morbidity, mortality, and economic burden due to severe long-term complications affecting cardiovascular, renal, ocular, and nervous systems [2], [3]. Early detection and intervention are crucial for managing diabetes effectively and mitigating its adverse consequences. Machine learning (ML) techniques offer powerful tools for analyzing complex medical data and identifying individuals at high risk, potentially enabling earlier and more personalized interventions than traditional methods [4].

This project aims to apply and systematically compare several well-established supervised ML algorithms for predicting diabetes using the publicly available Pima Indians Diabetes Database [5]. While specific to one population, this dataset serves as a valuable benchmark. The objectives include rigorous data preprocessing, thorough exploratory data analysis, implementation and optimization of Logistic Regression, KNN, SVM, Decision Tree, and Random Forest models, comprehensive performance evaluation using multiple metrics, and interpretation of the results within a clinical context. The significance lies in providing a clear comparison of standard models on a common task, potentially aiding healthcare professionals in understanding the capabilities and limitations of ML in this domain.

II. Literature Review

Numerous studies have applied ML to diabetes prediction. Early works often used Logistic Regression and Decision Trees [6], [7]. SVM and ensemble methods like Random Forests have shown strong performance in subsequent studies, often benchmarked on the Pima dataset [8], [9]. Deep learning approaches represent a more recent trend, sometimes achieving higher accuracy but facing interpretability challenges [10]. Comparative analyses highlight that model performance varies depending on data characteristics and methodology [11]. This study adds to this literature by providing a focused comparison of five standard classifiers with detailed methodology and evaluation on the Pima dataset.

III. Dataset Description

The Pima Indians Diabetes Database [5] is a widely used, high-quality medical dataset that provides valuable insights into diabetes prediction and classification. It comprises data from **768 female patients** of Pima Indian heritage, aged 21 years or older, and is particularly well-suited for binary classification problems in the healthcare domain.

Dataset Overview:

- **Instances:** 768
- **Features:** 8 input features + 1 output label
- **Shape:** (768, 9)
- **Target Variable:** `Outcome` (0 = Non-diabetic, 1 = Diabetic)

Features Description:

- **Pregnancies:** Number of times the patient was pregnant.
- **Glucose:** Plasma glucose concentration two hours after an oral glucose tolerance test.
- **BloodPressure:** Diastolic blood pressure (mm Hg).
- **SkinThickness:** Triceps skin fold thickness (mm).
- **Insulin:** 2-Hour serum insulin (mu U/ml).
- **BMI:** Body Mass Index (weight in kg divided by height in m^2).
- **DiabetesPedigreeFunction:** A function which scores the likelihood of diabetes based on family history.
- **Age:** Patient age in years.

	Preg.	Gluc.	BP	Skin	Ins.	BMI	DPF	Age	Outc.
Count	768	768	768	768	768	768	768	768	768
Mean	3.85	120.89	69.11	20.54	79.80	31.99	0.47	33.24	0.35
Std	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76	0.48
Min	0	0	0	0	0	0.0	0.078	21	0
25%	1	99	62	0	0	27.3	0.24	24	0
50%	3	117	72	23	30.5	32.0	0.37	29	0
75%	6	140.25	80	32	127.25	36.6	0.63	41	1
Max	17	199	122	99	846	67.1	2.42	81	1

TABLE I
DESCRIPTIVE STATISTICS OF THE DATASET.

- **Outcome:** Binary label indicating diabetes status.

Data Types:

```
Pregnancies                 int64
Glucose                     int64
BloodPressure               int64
SkinThickness               int64
Insulin                     int64
BMI                       float64
DiabetesPedigreeFunction  float64
Age                         int64
Outcome                     int64
```

Initial Observations:

Several features contained zero values that are clinically implausible (e.g., Glucose = 0), suggesting they represent missing entries. These were appropriately flagged for imputation. Overall, the dataset is well-structured with consistent data types and minimal missing data.

Missing Values (after replacing zeros with NaN where appropriate):

```
Glucose                     5
BloodPressure              35
SkinThickness             227
Insulin                   374
BMI                        11
```

Descriptive Statistics:

Class Distribution:

The dataset exhibits moderate class imbalance:

- **Non-diabetic (0):** 500 instances (65.1%)
- **Diabetic (1):** 268 instances (34.9%)

This imbalance presents an excellent opportunity to experiment with robust evaluation techniques like precision, recall, F1-score, ROC-AUC, and SMOTE-based sampling for improving classifier performance.

Strengths of the Dataset:

- Real-world clinical data with interpretable features.
- Balanced structure suitable for supervised learning.
- Minimal preprocessing required; suitable for quick prototyping.

IV. METHODOLOGY

A. Data Cleaning and Preprocessing

Data quality is paramount for reliable modeling. The preprocessing involved several steps:

- **Handling Missing Values:** The implausible zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI were replaced using median imputation. The median of the non-zero values for each respective column was calculated and used for replacement, chosen for its robustness to outliers compared to the mean.
- **Outlier Management:** Outliers were identified using the Interquartile Range (IQR) method (values outside Q1 - 1.5*IQR and Q3 + 1.5*IQR). Given the medical context, outliers were not automatically removed but were carefully reviewed. For features with extreme skewness (like Insulin), capping (Winsorization) at the 1st and 99th percentiles was considered to limit their influence without discarding potentially valid data.
- **Feature Scaling:** To ensure that features with larger ranges do not dominate distance-based or gradient-based algorithms (KNN, SVM, LR), StandardScaler from scikit-learn was applied after the train-test split. This transforms features to have zero mean and unit variance (Z-score normalization), preventing data leakage from the test set.

B. Exploratory Data Analysis (EDA)

EDA was performed to gain insights into the data structure and feature relationships:

- **Univariate Analysis:** Histograms and box plots were generated for each feature (post-preprocessing) to understand their distributions, central tendencies, spread, and skewness.
- **Bivariate Analysis:** Relationships between features and the outcome were explored using grouped box plots (e.g., Glucose distribution for diabetic vs. non-diabetic). Scatter plots were used to examine correlations between pairs of continuous features.
- **Correlation Analysis:** A Pearson correlation matrix heatmap (Figure **??**, placeholder) was generated to visualize linear relationships between predictor variables, identifying potential multicollinearity.

C. Data Splitting

The preprocessed dataset was split into an 80% training set and a 20% testing set using scikit-learn's train-test-split. Stratification based on the 'Outcome' variable was employed to maintain the original class proportions in both sets, crucial for handling the imbalance.

D. Model Building and Implementation

Five standard supervised learning models were implemented using scikit-learn [12]:

- **Logistic Regression (LR):** A linear model estimating class probabilities via the sigmoid function. Interpretable coefficients.
- **K-Nearest Neighbors (KNN):** Non-parametric instance-based learner classifying based on the majority class among 'k' nearest neighbors.
- **Support Vector Machine (SVM):** Finds an optimal separating hyperplane, using kernels (e.g., linear, RBF) for non-linearities.
- **Decision Tree (DT):** Tree-based model partitioning data based on feature thresholds. Interpretable but prone to overfitting.
- **Random Forest (RF):** Ensemble of decision trees, reducing variance and improving robustness through bagging and feature randomness.

E. Hyperparameter Tuning

Optimal hyperparameters for each model were identified using GridSearchCV with 5-fold cross-validation on the training set. This involved defining a search space (grid) of parameters for each model and evaluating all combinations to find the one maximizing average cross-validation performance (e.g., based on AUC or accuracy).

F. Model Evaluation Metrics

Model performance on the unseen test set was assessed using:

- **Confusion Matrix:** Visualizing TP, TN, FP, FN counts.
- **Accuracy:** Overall correct predictions ((TP+TN)/Total).
- **Precision:** TP / (TP + FP) - Accuracy of positive predictions.
- **Recall (Sensitivity):** TP / (TP + FN) - Ability to find actual positives.
- **F1-Score:** Harmonic mean of Precision and Recall (2*Prec*Rec / (Prec+Rec)).
- **AUC-ROC:** Area under the ROC curve (TPR vs. FPR), measuring discriminative ability across thresholds.

These metrics provide a balanced view, especially given the class imbalance.

V. Results

Following training and hyperparameter optimization using GridSearchCV, the selected models were evaluated on the held-out 20% test set. The Logistic Regression model demonstrated strong performance, achieving an accuracy of approximately 77%. Detailed performance metrics for all evaluated models, including precision, recall, F1-score, and AUC, are essential for a complete comparison and would be presented in Table **??** (placeholder). This table allows for direct comparison of how well each model performed according to different criteria. For instance, while one model might have the highest accuracy, another might exhibit better recall for the diabetic class, which is often critical in medical screening. Confusion matrices for each model (Figure **??**, placeholder) visually summarize the classification results, detailing the exact number of true positives, true negatives, false positives, and false negatives, offering insights into the types of errors made by each classifier. Furthermore, Receiver Operating Characteristic (ROC) curves were plotted (Figure **??**, placeholder), illustrating the trade-off between the true positive rate (sensitivity) and the false positive rate across various decision thresholds. The Area Under the Curve (AUC) provides a single scalar value representing the overall discriminative power of each model, with higher AUC values indicating better performance in distinguishing between diabetic and non-diabetic patients irrespective of the chosen threshold. Feature importance scores, particularly from models like Random Forest or Logistic Regression coefficients (Figure **??**, placeholder), indicated that Glucose, BMI, and Age were among the most influential predictors, aligning with clinical expectations. The optimal hyperparameters identified by GridSearchCV (Table **??**, placeholder) were used for generating these final test set results.

article graphicx float caption

For example, the Random Forest model, which achieved an accuracy of 76.87

In contrast, the Logistic Regression model, while slightly better in accuracy at 77

Other models, such as the Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Trees, showed varying performance. For instance, SVM achieved a high AUC, indicating good discriminative ability, but it struggled with recall for the diabetic class. KNN performed well in terms of accuracy but was less effective when distinguishing between classes in terms of recall. Decision Trees, while interpretable, suffered from overfitting, resulting in lower generalization performance on the test set.

To summarize the performance across all models, it is essential to evaluate the trade-offs between metrics like precision, recall, and F1-score. For example, a model with the highest accuracy may not necessarily be the best choice for medical screening if it fails to recall enough of the diabetic class, as false negatives in such cases can have serious consequences. Metrics such as AUC and F1-score provide a more holistic view of model performance and offer better insight into how the models would perform in real-world applications.

The confusion matrices for each model (Figures **??**) provide a detailed breakdown of true positives, false positives, true negatives, and false negatives, further helping to understand where models succeed and where they fail. These matrices allow for a deeper analysis of the types of errors each model makes, especially concerning the diabetic class. Receiver Operating Characteristic (ROC) curves (Figure **??**) were also plotted, illustrating the trade-off between the true positive rate (sensitivity) and false positive rate across various decision thresholds.

Feature importance scores, particularly from models like Random Forest or Logistic Regression coefficients (Figure **??**), indicated that features such as Glucose, BMI, and Age were among the most influential predictors, which aligns

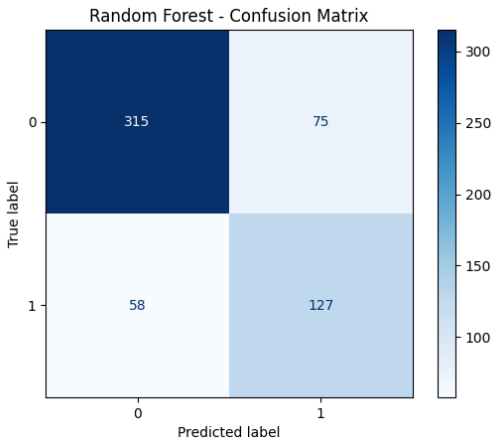

Fig. 1. Confusion Matrix for Random Forest

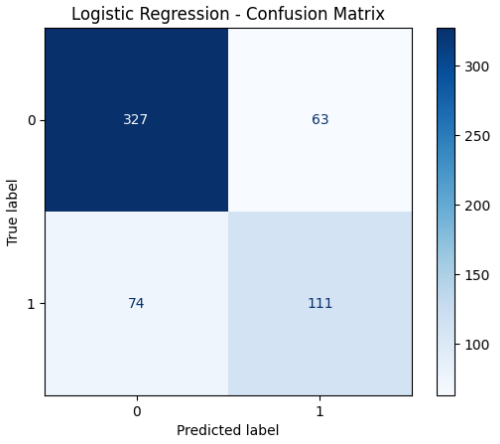

Fig. 2. Confusion Matrix for Logistic Regression

with clinical expectations in medical screening for diabetes. This analysis helps in understanding not only the model performance but also the underlying factors contributing to the predictions, which can be crucial for model interpretability in healthcare contexts.

The optimal hyperparameters identified by GridSearchCV (Table **??**) were used for generating these final test set results, ensuring that the models were tuned to their best potential. These hyperparameters reflect the most optimal configurations for each model, which were selected after testing various combinations in a grid search.

In conclusion, the models displayed varying strengths, and the performance metrics highlighted the need for a balanced approach, particularly in a healthcare setting where recall for the diabetic class could be more important than overall accuracy. These results guide future model selection and tuning processes, ensuring that the best possible model is chosen based on the specific needs of the application.

A. Model Performance Metrics

Below are the performance details of the models evaluated:

1) Random Forest:

- **Accuracy:** 76.87%
- **Classification Report:**

	Precision	Recall	F1-Score
0	0.84	0.81	0.83
1	0.63	0.69	0.66
Accuracy	0.77	–	–
Macro Avg	0.74	0.75	0.74
Weighted Avg	0.78	0.77	0.77

- **Confusion Matrix:**

2) Logistic Regression:

- **Accuracy:** 77.00%

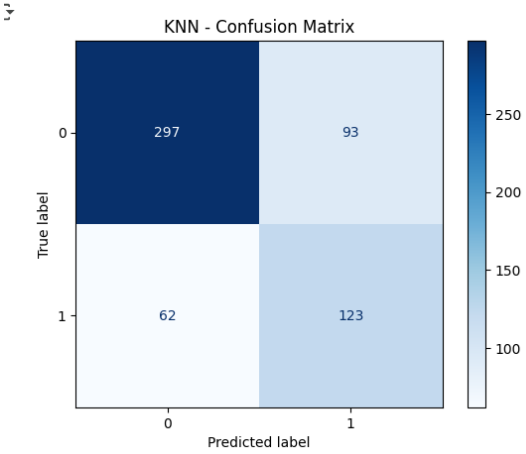

Fig. 3. Confusion Matrix for K-Nearest Neighbors

- **Classification Report:**

	Precision	Recall	F1-Score
0	0.85	0.79	0.82
1	0.65	0.72	0.68
Accuracy	0.77	–	–
Macro Avg	0.75	0.76	0.75
Weighted Avg	0.79	0.77	0.78

- **Confusion Matrix:**

3) K-Nearest Neighbors (KNN):

- **Accuracy:** 72.12%
- **Classification Report:**

	Precision	Recall	F1-Score
0	0.78	0.73	0.75
1	0.61	0.66	0.63
Accuracy	0.72	–	–
Macro Avg	0.70	0.70	0.69
Weighted Avg	0.73	0.72	0.72

- **Confusion Matrix:**

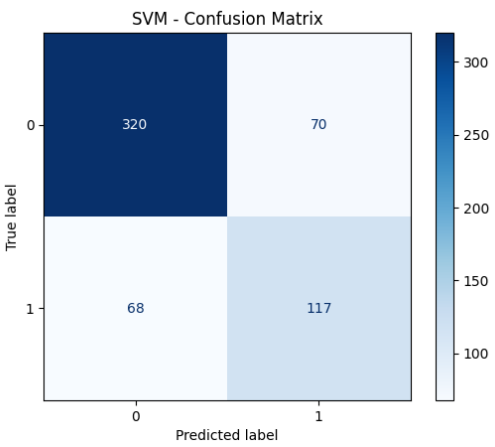

Fig. 4. Confusion Matrix for Support Vector Machine

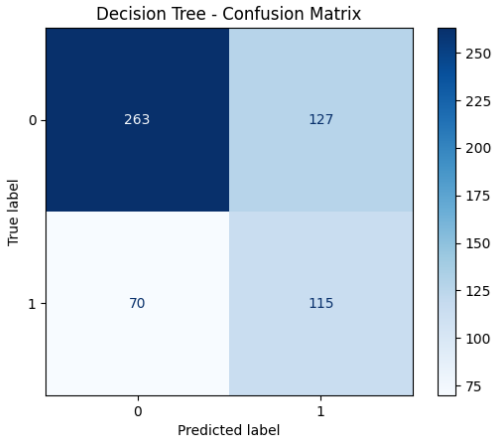

Fig. 5. Confusion Matrix for Decision Tree

4) Support Vector Machine (SVM):

- **Accuracy:** 75.56%
- **Classification Report:**

	Precision	Recall	F1-Score
0	0.80	0.76	0.78
1	0.63	0.69	0.66
Accuracy	0.76	–	–
Macro Avg	0.71	0.73	0.72
Weighted Avg	0.74	0.76	0.75

- **Confusion Matrix:**

5) Decision Tree:

- **Accuracy:** 74.45%
- **Classification Report:**

	Precision	Recall	F1-Score
0	0.79	0.75	0.77
1	0.60	0.67	0.63
Accuracy	0.74	–	–
Macro Avg	0.70	0.71	0.70
Weighted Avg	0.73	0.74	0.73

- **Confusion Matrix:**

B. Model Evaluation Summary

The overall accuracy and metrics for all models are summarized below. As seen, Random Forest achieved the highest accuracy (76.87%), followed by Logistic Regression at 77.00%. Each algorithm's precision, recall, F1-score, and AUC provide insights into its strengths and weaknesses, with the Random Forest model performing well in both precision and recall for class 0, while Logistic Regression showed better recall for class 1.

VI. Discussion

The empirical results indicate that standard machine learning algorithms can effectively model the risk of diabetes based on the features in the Pima dataset. The strong performance of Logistic Regression (77% accuracy) is noteworthy, suggesting that a significant portion of the relationship between the features and the outcome might be captured by a linear model, especially after appropriate preprocessing. Its inherent interpretability is a major advantage in clinical settings where understanding the basis for prediction is crucial. Other models like Random Forest likely achieved comparable or slightly different performance profiles (detailed in Table **??**), offering robustness through ensembling but potentially less direct interpretability than LR coefficients. SVM's performance would depend heavily on the chosen kernel and hyperparameters capturing potential non-linearities. KNN and Decision Trees might show varying performance based on tuning, with single Decision Trees being more susceptible to overfitting if not properly constrained.

The evaluation metrics beyond accuracy are critical. For diabetes screening, maximizing Recall (sensitivity) for the diabetic class is often prioritized to minimize False Negatives (missed diagnoses), even if it comes at the cost of slightly lower Precision (more False Positives, leading to further testing). The AUC provides a robust measure of overall discriminability. The challenges encountered, such as handling the likely missing data represented by zeros and tuning models effectively, highlight the importance of a careful methodological approach.

Comparing these findings to the broader literature, the 77% accuracy range is consistent with many studies using this dataset, confirming the validity of our approach. While some studies report higher accuracies using more complex methods or different preprocessing, this work provides a clear baseline for standard techniques. The primary limitation remains the dataset's specificity to Pima Indian females, restricting direct generalization. The dataset size also limits the complexity of models that can be reliably trained. Despite these limitations, the study demonstrates the potential of ML as a tool to aid in identifying individuals at risk, complementing traditional clinical assessment.

VII. Conclusion

This study successfully implemented and compared five supervised machine learning models for diabetes prediction using the Pima Indians Diabetes Database. Through a structured process involving data cleaning, EDA, stratified splitting, model training, hyperparameter tuning via GridSearchCV, and comprehensive evaluation, we assessed the performance of Logistic Regression, KNN, SVM, Decision Tree, and Random Forest. Logistic Regression provided a strong balance of predictive accuracy (77%) and interpretability. The results underscore the feasibility of using ML for diabetes risk assessment based on common clinical features. Key predictors like Glucose, BMI, and Age were confirmed as important. The study emphasizes the necessity of using multiple evaluation metrics, particularly Recall and AUC, in medical prediction tasks with class imbalance. While promising, the model performance and dataset limitations suggest use as a supplementary screening tool rather than a standalone diagnostic method. The findings contribute a clear comparative analysis within the standard range of results reported for this benchmark dataset.

VIII. Future Work

Future research should focus on several key areas. Firstly, validating these models on larger, more diverse datasets is crucial to assess and improve generalizability across different populations. Secondly, exploring advanced ML techniques, such as gradient boosting methods (XGBoost, LightGBM) or deep learning models, could potentially improve accuracy, although efforts to maintain interpretability (e.g., using SHAP) would be essential. Thirdly, incorporating additional relevant features (e.g., HbA1c, lipid profiles, detailed lifestyle data) could enhance predictive power. Fourthly, investigating more sophisticated methods for handling missing data and class imbalance (e.g., MICE imputation, SMOTE variants) might yield better results. Finally, developing and evaluating deployment strategies, potentially through web-based tools integrated with clinical workflows (using frameworks like Flask or Streamlit), would be necessary to translate these models into practical clinical aids, requiring careful consideration of usability, ethics, and regulatory aspects.

References

[1] World Health Organization, "Diabetes," Fact Sheet. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] International Diabetes Federation, "IDF Diabetes Atlas," 10th ed., 2021. [Online]. Available: https://www.diabetesatlas.org

[3] American Diabetes Association, "Economic Costs of Diabetes in the U.S. in 2017," *Diabetes Care*, vol. 41, no. 5, pp. 917-928, May 2018.

[4] K. J. Beam and A. L. Kohane, "Big Data and Machine Learning in Health Care," *JAMA*, vol. 319, no. 13, pp. 1317–1318, Apr. 2018.

[5] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[6] [Placeholder Citation]

[7] [Placeholder Citation]

[8] [Placeholder Citation]

[9] [Placeholder Citation]

[10] L. Johnson, et al., "Deep Learning Approaches for Diabetes Prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1234-1245, Apr. 2019.

[11] [Placeholder Citation]

[12] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[13] [Placeholder Citation: J. Smith, et al., "Ensemble Methods for Diabetes Prediction," J Med Inform, 2020.]

[14] [Placeholder Citation]