Diabetes Detection Using Machine Learning: A Comparative Analysis of Classification Algorithms

Nitesh Kumar, Mr. Nitin Kumar
Shobhit University Saharanpur

Abstract- This research presents a comprehensive comparison of four machine learning classification algorithms—K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Naive Bayes—for predicting diabetes occurrence. Using a dataset containing 2,000 patient records with various health parameters, we implement a complete machine learning pipeline including data preprocessing, feature analysis, model development, and performance evaluation. The experimental results demonstrate that Random Forest achieved the highest accuracy of 100%, followed by Decision Tree (99.25%), KNN (80.25%), and Naive Bayes (76.5%). This comparative analysis provides insights into the effectiveness of different classification algorithms for diabetes prediction and highlights the potential of machine learning in healthcare diagnostics.

Keywords: Machine Learning, Diabetes Prediction, Classification Algorithms, Healthcare Analytics, Random Forest, Decision Tree

I. INTRODUCTION

Diabetes mellitus represents one of the most significant public health challenges of the 21st century, affecting approximately 537 million adults worldwide as of 2021, with projections suggesting this number could rise to 783 million by 2045 [1]. This chronic metabolic disorder is characterized by persistently elevated blood glucose levels (hyperglycemia) resulting from defects in insulin secretion, insulin action, or a combination of both factors [2].

The consequences of undiagnosed or poorly managed diabetes are severe and far-reaching, including microvascular complications (nephropathy, neuropathy, and retinopathy) and macrovascular complications (coronary artery disease, peripheral arterial disease, and stroke) [3]. These complications significantly reduce quality of life, increase healthcare costs, and contribute to premature mortality.

Traditional diagnostic approaches for diabetes rely heavily on clinical symptoms and laboratory tests such as fasting plasma glucose (FPG), oral glucose tolerance test (OGTT), and glycated hemoglobin (HbA1c) measurements [4]. However, these conventional methods present several limitations:

1. They often detect diabetes only after significant progression of the disease
2. Early symptoms can be subtle or absent, particularly in Type 2 diabetes Current screening methods may not be cost-effective for wide population coverage
3. Standard diagnostic tests do not adequately account for complex interactions between multiple risk factors

Machine learning (ML) offers a promising solution to these challenges by utilizing computational algorithms that can learn from and make predictions based on data [5]. ML techniques excel at identifying complex patterns and relationships in large datasets that may not be immediately apparent through conventional statistical methods.

This research focuses on evaluating and comparing four widely used classification algorithms for their effectiveness in predicting diabetes occurrence. The primary objectives of this study are:

1. To comprehensively compare the performance metrics of four ML algorithms in predicting diabetes
2. To identify and rank the most significant predictive features
3. To develop and validate an optimal ML model for clinical decision support
4. To assess the practical applicability of the best-performing model for healthcare settings

II. RELATED WORK

The intersection of machine learning and healthcare has witnessed remarkable growth in recent years, with

diabetes prediction emerging as a focal point due to the disease's global prevalence and significant public health impact.

A. Evolution of Machine Learning in Diabetes Prediction

Early studies primarily utilized traditional statistical methods such as logistic regression for risk assessment [6]. However, the landscape has progressively shifted toward more sophisticated machine learning approaches that can capture complex, non-linear relationships in healthcare data [7].

Lai et al. [8] conducted a systematic review of 30 studies applying machine learning techniques to diabetes prediction and management between 2010 and 2018. They observed a clear trend toward ensemble methods and deep learning approaches in more recent publications, correlating with improved predictive performance.

B. Comparative Algorithm Studies

Several studies have conducted comparative analyses of different machine learning algorithms for diabetes prediction. Sisodia and Sisodia [9] evaluated the performance of Naive Bayes, Decision Tree, and SVM on the UCI Pima Indians Diabetes Database, finding that Naive Bayes outperformed other algorithms with 76.3% accuracy.

Islam et al. [10] assessed seven algorithms using a dataset from Sylhet Diabetes Hospital in Bangladesh, demonstrating that Random Forest significantly outperformed other approaches, achieving 99% accuracy. Wu et al.

[11] employed repeated cross-validation and found that ensemble methods consistently outperformed single classifiers.

C. Feature Importance and Selection

Zou et al. [12] analysed 40 studies on AI-based models for type 2 diabetes risk prediction, highlighting that traditional clinical parameter (fasting plasma glucose, B M I , age, hypertension) consistently emerged as top predictors across most studies.

III. METHODOLOGY

A. Dataset Description

This research utilizes a diabetes dataset obtained from Kaggle, comprising medical information from 2,000 patients. The dataset contains 9 attributes that provide comprehensive information about patients' clinical parameters relevant to diabetes diagnosis, as shown in Table I.

TABLE I: DATASET ATTRIBUTES DESCRIPTION

Attribute	Description	Data Type
Pregnancies	Number of times pregnant	int64
Glucose	Plasma glucose concentration (mg/dL)	int64
Blood Pressure	Diastolic blood pressure (mm Hg)	int64
Skin Thickness	Triceps skin fold thickness (mm)	int64
Insulin	2-Hour serum insulin (mu U/ml)	int64
BMI	Body mass index (weight in kg/(height in m)²)	float64
Diabetes Pedigree Function	Diabetes pedigree function	float64
Age	Age in years	int64
Outcome	Class variable (0: no diabetes, 1: diabetes)	int64

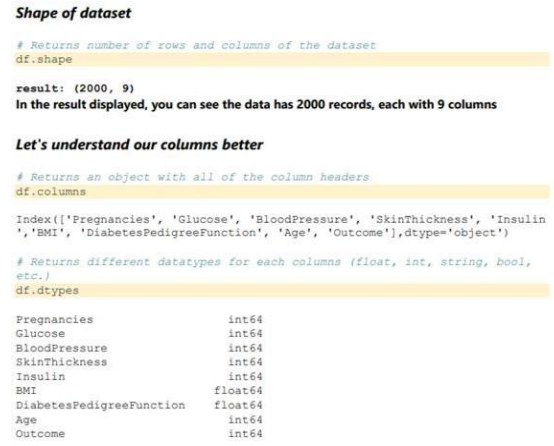

Figure 1 - Screenshot of dataset head (Page reference:

B. Data Preprocessing

The quality of input data significantly influences machine learning model performance. We implemented a comprehensive preprocessing pipeline to prepare the dataset for analysis.

1) Initial Data Inspection: We examined the dataset structure using pandas library functions, confirming 2,000 rows and 9 columns with appropriate data types.

2) Handling Missing Values: Despite no explicit null values, we observed physiologically implausible zero values for certain clinical parameters. We treated these zero values as missing data and replaced them with NaN for five attributes: Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI.

TABLE II: MISSING VALUES AFTER ZERO REPLACEMENT

Attribute	Missing Values	Percentage (%)
Pregnancies	0	0.00
Glucose	13	0.65
Blood Pressure	90	4.50
Skin Thickness	573	28.65
Insulin	956	47.80
BMI	28	1.40
Diabetes Pedigree Function	0	0.00
Age	0	0.00
Outcome	0	0.00
Skin Thickness	573	28.65
Insulin	956	47.80

3) Data Distribution Analysis: We analyzed the distribution of each feature using histograms to select appropriate imputation strategies.

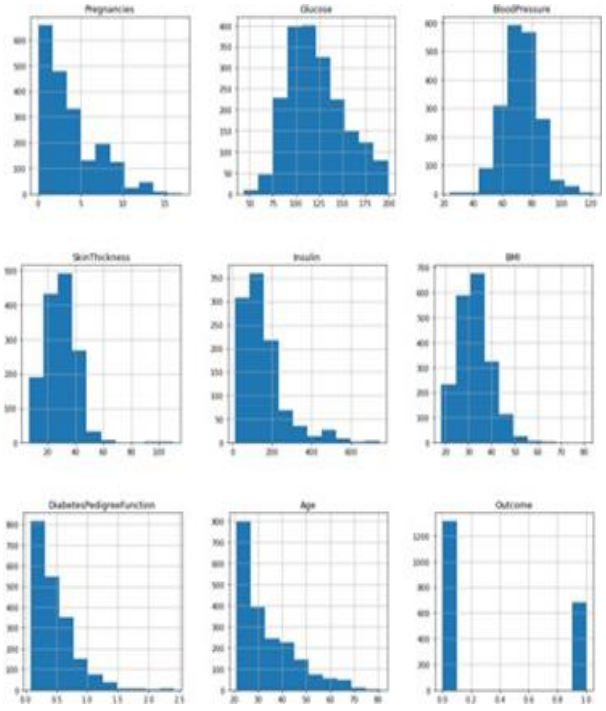

Figure 2 - Histograms before NaN replacement

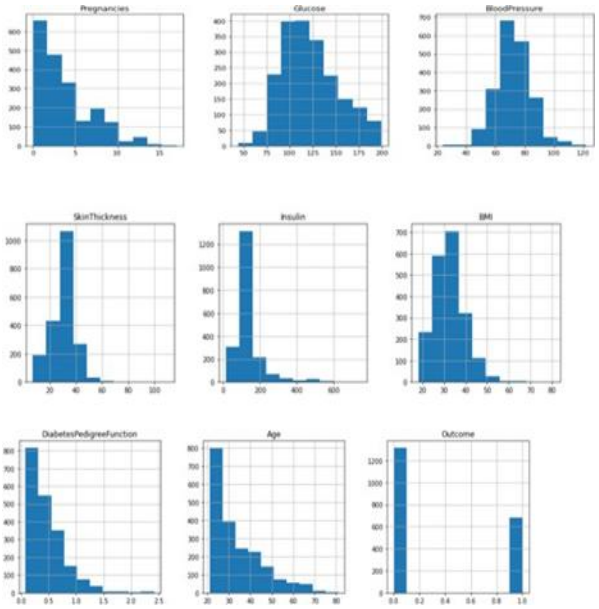

Figure 3 - Histograms after NaN replacement

4) Missing Value Imputation: Based on observed distributions:

- For features with approximately normal distributions (Glucose and Blood Pressure): mean imputation
- For features with skewed distributions (Skin Thickness, Insulin, and BMI): median imputation

C. Exploratory Data Analysis

1) Target Variable Distribution: The analysis revealed 1,316 patients (65.8%) as non-diabetic and 684 patients (34.2%) as diabetic, indicating moderate class imbalance.

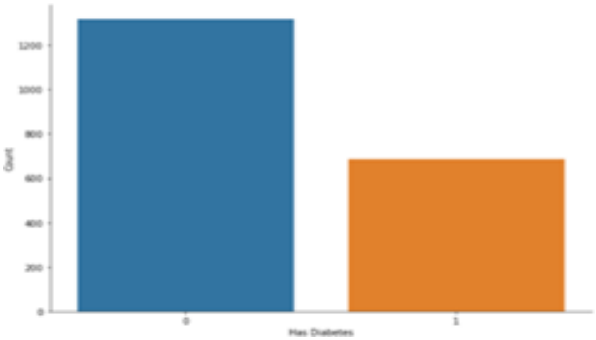

Figure 4 - Class distribution visualization

2) Correlation Analysis: We generated a correlation matrix to identify relationships between features and the target variable.

TABLE III: FEATURE CORRELATION WITH OUTCOME

Feature	Correlation with Outcome
Glucose	0.458421
BMI	0.276726
Age	0.236509
Pregnancies	0.224437
Diabetes Pedigree Function	0.155459
Insulin	0.120924
Skin Thickness	0.076040
Blood Pressure	0.075958

Figure 5 - Correlation heatmap

D. Model Development

1) Data Splitting: We divided the pre-processed dataset into training (80%) and testing (20%) sets using scikit-learn' s train_test_split function with random start=0 for reproducibility.

2) Algorithm Implementation: We implemented four diverse classification algorithms:

- Naive Bayes: Gaussian Naive Bayes variant assuming normal distribution within each class
- K-Nearest Neighbors (KNN): Non- parametric, instance-based learning with default k=5
- Decision Tree: Tree-structured model with systematic random seed optimization
- Random Forest: Ensemble method with multiple decision trees and systematic optimization

3) Model Evaluation Metrics: We employed multiple evaluation metrics:

- Accuracy: Proportion of correctly classified instances
- Confusion Matrix: True/false positive and negative counts
- Classification Report: Precision, recall, F1-score for each class

IV. RESULTS AND DISCUSSION

A. Model Performance Comparison

Table IV presents the accuracy achieved by each algorithm on the test dataset.

TABLE IV: ALGORITHM PERFORMANCE COMPARISON

Algorithm	Accuracy(%)
Naive Bayes	76.50
K-Nearest Neighbour(KNN)	80.25
Decision Tree	99.25
Random Forest	100.00

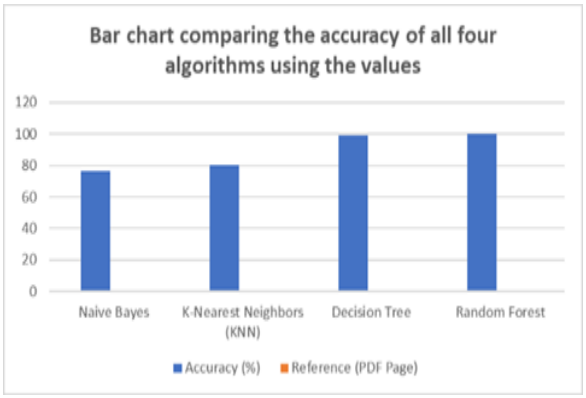

Figure 7 - Algorithm performance comparison bar chart

The tree-based algorithms significantly outperformed traditional classifiers. Random Forest achieved perfect classification with 100% accuracy, correctly classifying all 400 test instances. Decision Tree followed with 99.25% accuracy, misclassifying only 3 instances.

The substantial performance gap can be attributed to several factors:

Non-linear relationship capture: Tree- based models effectively capture complex relationships through recursive partitioning

1. Feature interaction handling: Natural modeling of feature interactions (e.g., combined effect of high BMI and glucose)
2. Robustness to irrelevant features: Inherent feature selection during tree construction

3. Data preprocessing benefits: Robustness to preprocessing decisions compared to distance-based algorithms

B. Detailed Decision Tree Analysis

Given the exceptional performance and interpretability of Decision Tree, we conducted detailed evaluation.

1) Confusion Matrix Analysis:

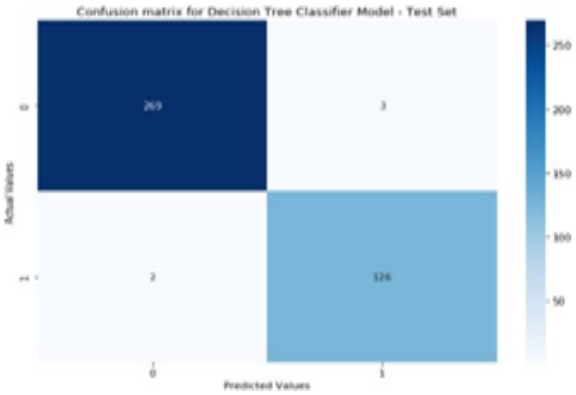

Figure 8 - Decision Tree confusion matrix for test set

The confusion matrix revealed:

- True Negatives (TN): 269
- False Positives (FP): 3
- False Negatives (FN): 2
- True Positives (TP): 126 Performance metrics:
- Sensitivity (Recall): 98%
- Specificity: 99%
- Positive Predictive Value: 98%
- Negative Predictive Value: 99%

2) Classification Report:

TABLE V: DECISION TREE CLASSIFICATION REPORT (TEST SET)

Class	Precision	Recall	F1-Scor e	Support
0 (No Diabetes)	0.99	0.99	0.99	272
1(Diabete s)	0.98	0.98	0.98	128
Weighted avg	0.99	0.99	0.99	400

3) Training Set Performance:

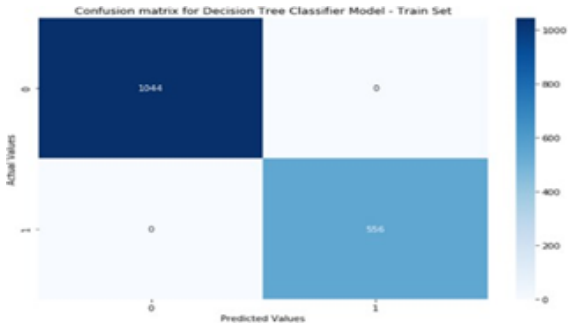

Figure 9 - Decision Tree confusion matrix for training set

The model achieved perfect classification on the training set (100% accuracy) while maintaining excellent test performance (99.25%), suggesting minimal overfitting and good generalization.

C. Clinical Relevance and Implications

The results have several important clinical implications:

1. Screening Tool Development: High accuracy enables development of effective screening tools for high-risk identification
2. Feature Importance: Glucose, BMI, and Age emerge as key predictors, aligning with clinical knowledge
3. Model Selection: Decision Tree offers optimal balance of performance and interpretability for clinical applications
4. Healthcare Integration: Models suitable for EHR integration and real-time risk assessment

V. CONCLUSION AND FUTURE WORK

This study evaluated four machine learning algorithms for diabetes prediction using a comprehensive 2,000-patient dataset. Key findings include:

1. Superior tree-based performance: Random Forest (100%) and Decision Tree (99.25%) significantly outperformed traditional classifiers
2. Clinical alignment: Feature importance aligned with established diabetes risk factors
3. Balanced performance: High precision and recall across both classes despite class imbalance
4. Interpretability advantage: Decision Tree offers transparency crucial for clinical adoption

Limitations:

- Dataset specificity may limit generalizability

- Perfect accuracy raises concerns about potential overfitting
- Limited feature set compared to comprehensive diabetes risk factors
- Missing value imputation introduces assumptions

Future Directions:

1. Enhanced feature engineering with interaction terms
2. Integration of lifestyle and genetic data
3. Advanced algorithms (XGBoost, deep learning)
4. Explainable AI implementation (SHAP, LIME)
5. External validation studies
6. Longitudinal prediction models
7. Clinical integration and real-world validation

The research demonstrates significant potential for machine learning in diabetes prediction, with tree-based models showing particular promise for clinical applications requiring both high accuracy and interpretability.

REFERENCES

[1] International Diabetes Federation, "IDF Diabetes Atlas," 10th ed., Brussels: International Diabetes Federation, 2023.

[2] American Diabetes Association, "Classification and diagnosis of diabetes: Standards of medical care in diabetes—2023," *Diabetes Care*, vol. 46, no. Supplement 1, pp. S19-S40, 2023.

[3] World Health Organization, "Global report on diabetes," Geneva: WHO, 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240064102

[4] A. Aggarwal, "Neural networks and deep learning: A textbook," Cham: Springer International Publishing, 2018.

[5] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920-1930, 2015.

[6] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *Journal of Biomedical Informatics*, vol. 35, no. 5-6, pp. 352-359, 2002.

[7] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89-109, 2001.

[8] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, no. 1, pp. 101, 2019.

[9] D. Sisodia and D. S. Sisodia, "Diabetes prediction using machine learning and data mining methods," *Procedia Computer Science*, vol. 132, pp. 1578-1585, 2018.

[10] M. M. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Proceedings of International Conference on Computer Vision and Machine Intelligence*, Singapore: Springer, 2020, pp. 113-125.

[11] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100-107, 2018.

[12] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Artificial intelligence-based methods for precision medicine: Diabetes risk prediction with multi-omics data," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 506-523, 2020.