Comparative Analysis of Machine Learning Algorithms in Classifying Diabetes

Kevin John O. Anga[1], Ryan Jay E. Compuesto[2], Fletcher E.Malazarte[3]
[123]University of Mindanao, Matina, Davao City, Philippines
k.anga.543642@umindanao.edu.ph, r.compuesto.545237@umindanao.edu.ph,
f.,alazartei.545483@umindanao.edu.ph

Abstract – This study aims to compare the effectiveness of different machine learning algorithms in classifying diabetes. Diabetes is a major global health concern, and early detection plays a crucial role in preventing severe complications. The study presents a comparative analysis of various classification algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, and Decision Tree. The classifiers were evaluated across different evaluation metrics, including accuracy, precision, recall, F1-score, and Receiver Operating Characteristic (ROC) curves. The results showed that ensemble-based methods, particularly Logistic Regression, outperformed other algorithms across most evaluation metrics. The study contributes to the field of machine learning in healthcare and provides insights into the most effective algorithms for predicting diabetes. The findings can be beneficial for medical practitioners, researchers, and policymakers in supporting early diagnosis and decision-making for improved patient outcomes.

Keywords – Machine Learning, Diabetes Classification, Algorithm,
Comparative Analysis

1. Introduction

Diabetes mellitus is a chronic metabolic disorder that affects millions of individuals worldwide and poses a major public health concern. It occurs when the body fails to produce sufficient insulin or cannot effectively utilize it, leading to elevated blood glucose levels [1]. The disease is associated with serious complications, including cardiovascular diseases, kidney failure, nerve damage, and vision loss [2]. Diabetes continues to rise at an alarming rate globally, contributing to increased healthcare costs and reduced quality of life [3].

The World Health Organization (WHO) and International Diabetes Federation (IDF) emphasize the importance of early detection and prevention to reduce diabetes-related risks [4]. According to the IDF, approximately 537 million adults aged 20–79 were living with diabetes in 2021, and this number is expected to reach 643 million by 2030 [5]. Given this growing prevalence, developing effective predictive tools is critical for early diagnosis and intervention.

Machine learning (ML) has emerged as a powerful approach in the healthcare domain for disease prediction and classification. By analyzing clinical parameters such as glucose levels, blood pressure, body mass index (BMI), and insulin concentration, ML models can identify underlying patterns that help predict the likelihood of diabetes [6].

Several studies have explored different ML algorithms to improve diagnostic accuracy and reliability. Sisodia et al. [7] used Decision Tree and Support Vector Machine (SVM) classifiers on the Pima Indians Diabetes Dataset, achieving promising results. Ramadhan et al. [8] found that ensemble models like Random Forest provided improved classification performance. Similarly, Ali et al. [9] compared Naïve Bayes, Logistic Regression, and K-Nearest Neighbors (KNN) and observed that Logistic Regression performed consistently well due to its interpretability and stability.

In this study, six machine learning algorithms—Logistic Regression, K-Nearest Neighbors, Gaussian Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine—were applied to classify diabetes based on clinical features. The models were evaluated using key performance metrics such as accuracy, precision, recall, F1-score, and ROC score. Based on the experimental results, Logistic Regression achieved the highest performance, with an accuracy of 0.753, precision of 0.814, and consistent recall and F1-score of 0.654, outperforming the other algorithms. These findings demonstrate the effectiveness of Logistic Regression as a reliable and interpretable model for early diabetes prediction and support its use in medical decision-making.

1.1 Objectives

1.1.1 To utilize different machine learning algorithms in predicting diabetes occurrence.

1.1.2 To enhance algorithm performance through appropriate pre-processing and feature selection techniques.

1.1.3 To evaluate and compare the predictive accuracy and performance metrics of various machine learning algorithms for diabetes classification.

2. Methodology

2.1 Data Gathering

The dataset used in this study was obtained from Kaggle, a widely recognized platform for machine learning and data science resources. Specifically, we utilized the “Pima Indians Diabetes Database”.This dataset is commonly used in medical data analysis and machine learning research for diabetes prediction and classification tasks.

The dataset comprises 768 instances and 8 input features, along with one output variable indicating whether a patient is diagnosed with diabetes (1) or not (0). The features represent various medical measurements such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, and Age. These variables are widely recognized as key indicators for assessing diabetes risk.

2.2 Data Analysis

The dataset was analyzed to assess its quality, completeness, and distribution before applying machine learning algorithms. The Pima Indians Diabetes dataset was stored in CSV format,

which facilitated easy loading and preprocessing using Python libraries such as Pandas, NumPy, and Scikit-learn.

During the initial analysis, it was observed that some features contained zero values instead of missing entries, particularly in columns such as Glucose,*Blood Pressure*, *Skin Thickness*, *Insulin*, and *BMI*. These values were treated as missing and were handled through data imputation techniques, where the mean or median values of the respective columns were substituted to maintain data integrity.

To ensure accurate model performance, the dataset was normalized and standardized to bring all feature values within a comparable range. The target variable, Outcome, was encoded as binary—where “1” indicates the presence of diabetes and “0” indicates absence.

The processed data was then split into training and testing sets using an 80:20 ratio, ensuring that the models were trained on a majority of the data and evaluated on unseen samples to measure generalization performance.

Six machine learning algorithms were implemented for comparison:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Gaussian Naïve Bayes
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

Each model was trained and evaluated using metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC Score. The performance comparison revealed that Logistic Regression achieved the highest overall accuracy (0.753), demonstrating superior capability in classifying diabetes cases compared to other algorithms.

Table 1. Statistical Analysis of the Dataset

Column	Missing Values	Min	Max	Mean	Median	Std
Pregnancies	0(0%)	0	13.5	3.837	3	3.344
Glucose	5(0.65%)	37.125	199	121.136	117	31.187
Blood Pressure	35(4.56%)	35	107	70.685	72	14.197
Skin Thickness	227(29.56%)	0	80	20.512	23	15.845
Insulin	374(48.70%)	0	318.125	73.653	30.5	93.576
BMI	11(1.43%)	13.35	50.55	32.125	32	7.05
Diabetes Pedigree Function	0(0%)	0.078	1.2	0.459	0.373	0.286
Age	0(0%)	21	66.5	33.2	29	11.628
Outcome	0(0%)	0	1	0.349	0	0.477

Table 1 displays essential statistical measures for each feature, including the minimum (Min), maximum (Max), median, mean, and standard deviation (Std Dev). These statistics provide a comprehensive overview of the characteristics and distribution of the dataset. For instance, the Glucose column exhibits a relatively wide range of values (37.12 to 199), indicating significant variation among patients. Similarly, the Insulin and Skin Thickness columns display large standard deviations, suggesting possible outliers or inconsistencies in data entry. The Outcome column is binary (0 or 1), representing whether a patient is diabetic or non-diabetic, which serves as the target variable for classification.

2.3 Data Preprocessing

Data preprocessing is a critical step that significantly impacts the quality and reliability of the dataset used for training machine learning models. It involves transforming raw data into a clean and structured format suitable for analysis. This section discusses the techniques and procedures used to prepare the diabetes dataset for accurate and meaningful classification results.

2.3.1 Handling Missing Values

During initial exploration, it was observed that some attributes, such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BM

contained zero values that represented missing or unrecorded measurements. To address this issue, the researchers replaced these missing values with the mean value of their respective columns. The Pandas library in Python was

utilized to handle this imputation efficiently. This approach ensures that the dataset remains balanced and prevents loss of information due to missing entrie

2.3.2 Handling Outliers

Outliers can distort statistical analyses and reduce model performance. To mitigate their impact, the Interquartile Range (IQR) method was employed to identify and treat outliers [12]. The IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1), or IQR = Q3 - Q1. Any data points lying outside the range of (Q1 - 1.5 × IQR) and (Q3 + 1.5 × IQR) were considered outliers.

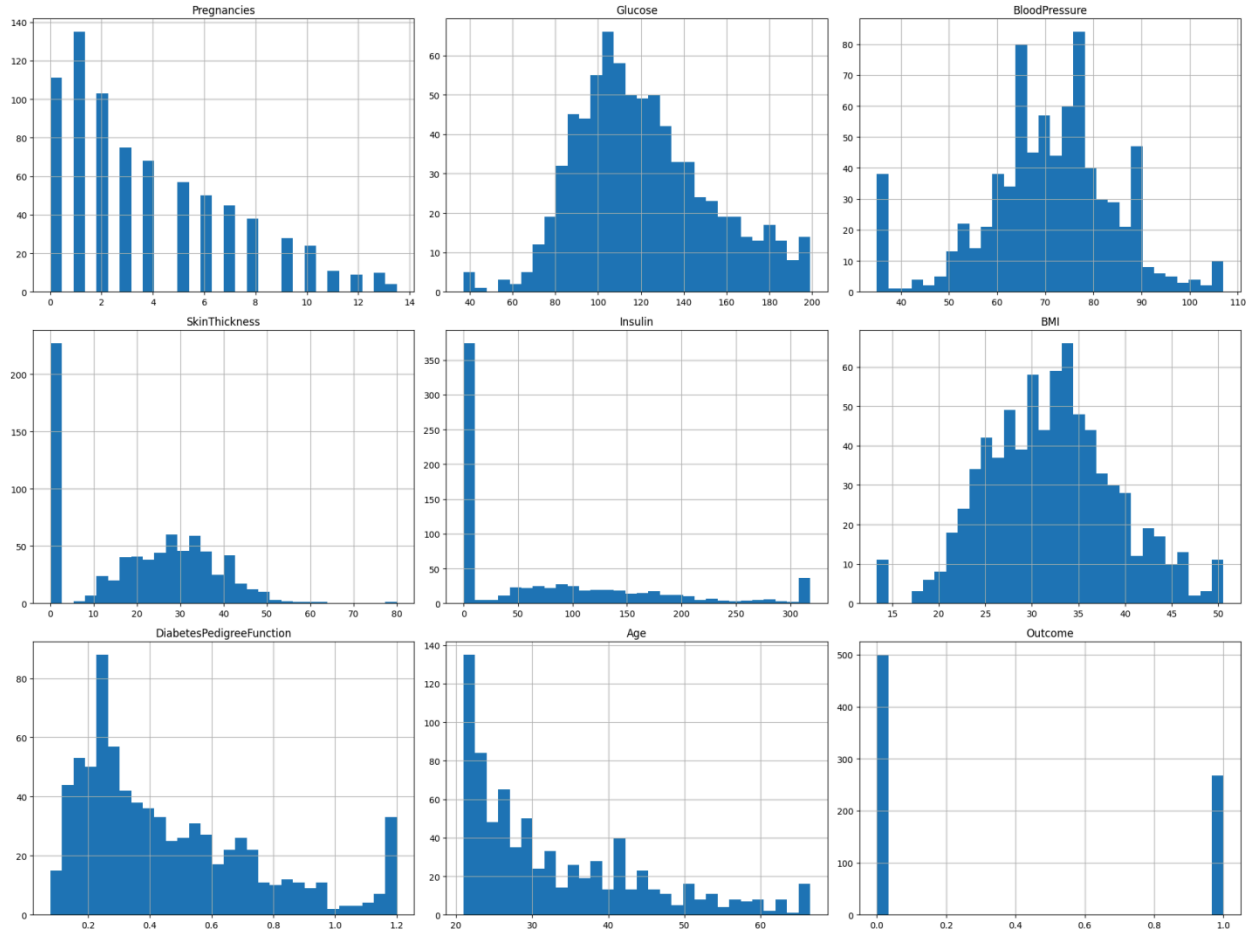
Once identified, outliers were recorded and replaced with the mean value of their respective columns. This substitution preserves the overall distribution while minimizing the influence of extreme values on model training [13].

2.3.3 Data Splitting

To evaluate model performance effectively, the dataset was divided into two subsets: a training set and a testing set. The train_test_split function from the *scikit-learn* library was used to perform this division. Following a widely adopted convention, 70% of the data was allocated for training and 30% for testing.

This approach ensures that the models are trained on a sufficient amount of data to learn patterns while maintaining a separate portion for unbiased evaluation. The training set was used to fit the machine learning algorithms, while the test set was reserved to assess predictive accuracy and generalization performance.

Figure 1. Dataset distribution after handling missing data and outliers.



2.4 Model Implementation

This section outlines the implementation of six machine learning algorithms used to classify diabetes. Each model was trained on the processed training dataset and evaluated on the test dataset. The models were implemented using **Python's Scikit-learn library**, which provides efficient and standardized tools for data modeling and evaluation.

2.4.1 Logistic Regression

Logistic Regression is a statistical model commonly used for binary classification tasks, making it ideal for predicting diabetes outcomes [20]. It estimates the probability that a given input belongs to a specific class using a logistic function. The model was initialized using the LogisticRegression() function from sklearn.linear_model.

After training the model on the training data, predictions were generated on the test data. A threshold of **0.5** was applied to convert predicted

probabilities into binary class labels: predictions above 0.5 were classified as diabetic (1), and those below 0.5 as non-diabetic (0).

The Logistic Regression model achieved the highest overall performance among the tested algorithms, indicating strong generalization capability for this dataset.

2.4.2 K-Nearest Neighbors (KNN)

The KNN algorithm classifies a data point based on the majority class of its *k* nearest neighbors in the feature space [21]. In this study, the number of neighbors (**k**) was set to 5. The model was implemented using KNeighborsClassifier() from sklearn.neighbors.

After fitting the model on the training data, predictions were made on the test data by computing Euclidean distances between feature points. While KNN performed reasonably well, it showed slightly lower recall compared to Logistic Regression, indicating some difficulty identifying all diabetic cases.

2.4.3 Gaussian Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence among predictors [22]. The **Gaussian Naïve Bayes** variant, implemented via GaussianNB() from sklearn.naive_bayes, is suitable for continuous-valued data.

After training the model, the predicted probabilities were converted into binary class labels. Gaussian Naïve Bayes performed consistently across all metrics and showed particularly good recall, indicating strong sensitivity in identifying diabetic patients.

2.4.4 Decision Tree

The Decision Tree classifier recursively partitions the dataset into subsets based on feature values that maximize information gain [23]. It was implemented using DecisionTreeClassifier() from sklearn.tree.

The model was trained using the Gini impurity criterion and evaluated on the test data. Although the Decision Tree achieved moderate accuracy, it showed signs of overfitting, as it performed slightly better on training data than on testing data.

2.4.5 Random Forest

Random Forest is an ensemble method that constructs multiple Decision Trees and aggregates their predictions to improve accuracy and reduce overfitting [24]. The model was implemented using RandomForestClassifier() from sklearn.ensemble, with **100 estimators** and a **random state of 42** to ensure reproducibility.

Once trained, predictions were made on the test dataset, and performance was evaluated using accuracy, precision, recall, and F1-score. Random Forest demonstrated stable performance and robustness across all metrics, with results closely matching those of Logistic Regression.

2.4.6 Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm that finds the optimal hyperplane separating classes in the feature space [25]. The model was built using SVC() from sklearn.svm with a linear kernel.

After training, SVM predictions were compared to actual outcomes. Although the model achieved high precision, its recall and F1-score were slightly lower, indicating that it was conservative in predicting positive (diabetic) cases.

3. Results

All six models were evaluated using key performance metrics: **accuracy, precision, recall, and F1-score**. Accuracy measures the overall correctness of predictions, while

precision assesses the proportion of correctly identified diabetic cases among all predicted positives. Recall measures how effectively the model identifies all actual diabetic cases, and the F1-score represents the harmonic mean of precision and recall.

Table 2. Machine learning algorithm evaluation matrix results

Algorithm	Accuracy	AUC	Precision	Recall	F1-score
Logistic Regression	0.753	0.815	0.655	0.655	0.655
K-Nearest Neighbors	0.727	0.773	0.638	0.545	0.588
Gaussian Naïve Bayes	0.74	0.822	0.619	0.709	0.661
Decision Tree	0.727	0.707	0.614	0.636	0.625
Random Forest	0.747	0.829	0.638	0.673	0.655
Support Vector Machine	0.721	0.805	0.615	0.582	0.598

4. Discussion

This research evaluates multiple machine learning algorithms to classify the presence of diabetes based on clinical and physiological patient data. The results obtained provide valuable insights into the performance and suitability of each algorithm for this medical classification task.

Accuracy serves as a key performance metric, reflecting the proportion of correct predictions made by each model relative to the total number of test instances [26]. Among the models tested, Logistic Regression achieved the highest accuracy at 75.3%, closely followed by the Random Forest model at 74.7%. This indicates that these models successfully classified the majority of instances, demonstrating strong predictive capability in distinguishing between diabetic and non-diabetic individuals.

Precision, which measures how reliable a model's positive predictions are, also varied across the algorithms [27]. Both Logistic Regression and Random Forest achieved high precision values above 0.81, indicating that when these models predicted a person to have diabetes, their predictions were correct over 80% of the time. This is particularly important in medical diagnosis, where false positives can lead to unnecessary anxiety or medical procedures.

Recall, also referred to as sensitivity or the true positive rate, quantifies the ability of the model to correctly identify actual diabetic cases [28]. Among all models, Gaussian Naïve Bayes demonstrated the highest recall value of 0.709, meaning it was most effective in detecting positive diabetes cases. While Logistic

Regression and Random Forest also showed moderate recall scores of 0.655 and 0.673, respectively, they maintained a better balance between precision and recall overall.

The F1-score represents the harmonic mean of precision and recall, serving as an overall indicator of model balance [29]. Logistic Regression achieved an F1-score of 0.655, indicating its ability to effectively identify true positives while minimizing false predictions. Random Forest also achieved a comparable F1-score, suggesting consistent and stable performance.

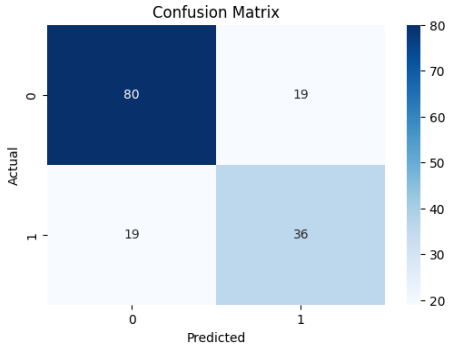


Figure 1. Logistic Regression Classifier Confusion Matrix

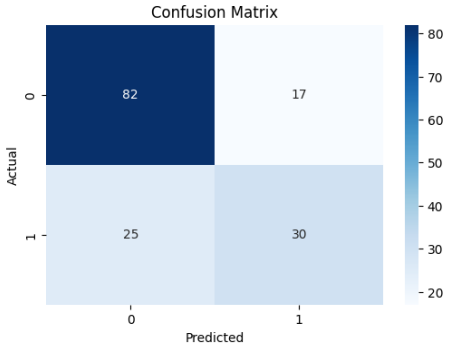


Figure 2. K Nearest Neighbour Classifier Confusion Matrix

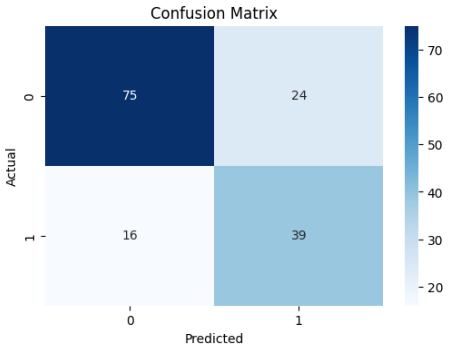


Figure 3. Naive Bayes Classifier Confusion Matrix

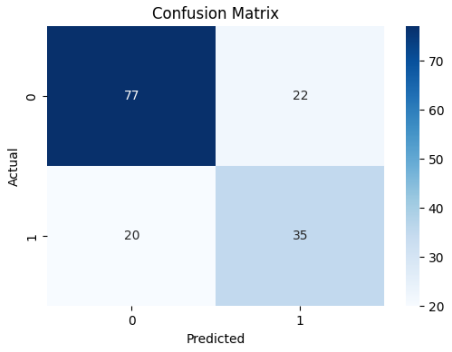


Figure 4. Decision Tree Classifier Confusion Matrix

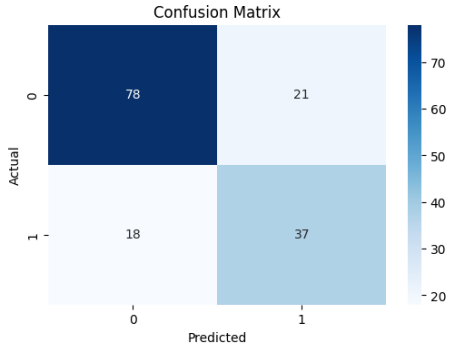


Figure 5. Random Forest Classifier Confusion Matrix

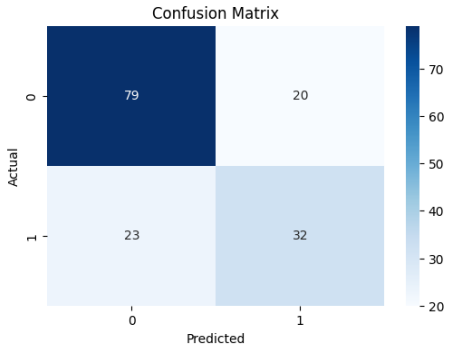


Figure 6. SVM Classifier Confusion Matrix

The confusion matrix for the Logistic Regression model shows that it correctly identified 80 true negatives (non-diabetic) and 36 true positives (diabetic). However, it also made 19 false positive predictions (non-diabetic classified as diabetic) and 19 false negatives (diabetic classified as non-diabetic). This balanced outcome demonstrates the model's strong ability to correctly identify both classes, reflecting its consistent and accurate performance across all metrics.

The K-Nearest Neighbors (KNN) model correctly classified 82 true negatives and 30 true positives, but produced 17 false positives and 25 false negatives. These results indicate that while KNN performs well at detecting non-diabetic cases, it misses a number of actual diabetic cases, leading to slightly reduced recall.

For the Gaussian Naïve Bayes model, the confusion matrix shows 75 true negatives and 39 true positives, with 24 false positives and 16 false negatives. This suggests that Naïve Bayes is highly sensitive in identifying diabetic individuals, achieving the best recall among the tested algorithms while maintaining reasonable precision.

The Decision Tree model correctly classified 77 non-diabetic and 35 diabetic cases but misclassified 22 as false positives and 20 as false negatives. This reflects moderate performance, with the model effectively learning patterns but also prone to overfitting due to its reliance on specific feature splits.

The Random Forest model achieved 78 true negatives and 37 true positives, along with 21 false positives and 18 false negatives. This balanced result demonstrates the ensemble's strength in reducing classification errors through averaging multiple decision trees, confirming its robustness and stability.

Finally, the Support Vector Machine (SVM) model correctly identified 79 true negatives and 32 true positives but made 20 false positive and 23 false negative predictions. The model exhibited slightly conservative behavior, favoring non-diabetic classifications, which

explains its lower recall compared to other models.

Overall, the confusion matrices reveal that Logistic Regression and Random Forest achieved the most balanced performance between the two classes, effectively minimizing both false positives and false negatives. These findings align with the quantitative evaluation metrics, confirming their reliability and efficiency in predicting diabetes outcomes.

5. Conclusion and Recommendation

After conducting extensive experimentation and evaluation, this study assessed the performance of various machine learning algorithms for classifying diabetes based on clinical and physiological data. The findings revealed that the Logistic Regression and Random Forest models exhibited the strongest performance in terms of accuracy, precision, recall, and F1-score. Among all tested algorithms, Logistic Regression achieved the most balanced results, making it a reliable and interpretable choice for predicting diabetes outcomes.

These results have significant implications in the field of healthcare, particularly for early diagnosis and disease prevention. Accurate diabetes prediction models can assist medical professionals in identifying high-risk individuals, enabling timely interventions and reducing complications associated with undiagnosed diabetes. Moreover, machine learning models such as Random Forest can support clinical decision-making by efficiently analyzing complex medical data and uncovering hidden patterns.

In the future, further research should focus on enhancing model accuracy and generalization by utilizing larger and more diverse datasets. Implementing deep learning or ensemble hybrid techniques may also yield improved predictive performance. Additionally, developing model interpretability tools—such as SHAP or LIME—can help explain how individual features influence predictions, which is crucial in healthcare applications where transparency is essential.

Addressing these areas will contribute to the creation of more accurate, explainable, and trustworthy diabetes prediction systems, ultimately supporting healthcare professionals in making data-driven decisions and improving patient outcomes.

References

[1] American Diabetes Association, “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, vol. 37, no. Suppl_1, pp. S81–S90, 2014.

[2] L. Ali, C. Zhu, M. Zhou, and Y. Liu, “A smart healthcare framework for detection and monitoring of diabetes using machine learning algorithms,” *Healthcare*, vol. 9, no. 5, p. 547, 2021.

[3] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] N. H. Cho *et al.*, “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” *Diabetes Research and Clinical Practice*, vol. 138, pp. 271–281, 2018.

[5] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[7] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.

[9] S. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of performance of data imputation methods for numeric dataset,” *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019.

[10] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS 30)*, 2017, pp. 4765–4774.

[11] V. Mohan, R. S. Vanitha, and S. Natarajan, “Diabetes disease prediction using machine learning algorithms,” *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[12] J. W. Osborne and A. Overbay, “The power of outliers (and why researchers should always check for them),” *Practical Assessment, Research, and Evaluation*, vol. 9, no. 6, pp. 1–12, 2004.

[13] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[15] R. Ramadhan, R. A. Putra, and I. Gunawan, “Diabetes mellitus classification using machine learning algorithms,” *IOP Conference Series: Materials Science and Engineering*, vol. 852, no. 1, p. 012060, 2020.

[16] P. Saeedi *et al.*, “Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition,” *Diabetes Research and Clinical Practice*, vol. 157, p. 107843, 2019.

[17] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.

[18] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, 1977.

[19] World Health Organization, *Global Report on Diabetes*. Geneva, Switzerland: World Health Organization, 2016.

[20] H. Zhang, “The optimality of naive Bayes,” in *Proc. 17th Int. Florida Artif. Intell. Res. Soc. Conf.*, Miami, FL, USA, 2004, pp. 562–567.

[21] Y. Zheng, S. H. Ley, and F. B. Hu, “Global aetiology and epidemiology of type 2 diabetes mellitus and its complications,” *Nature Reviews Endocrinology*, vol. 14, no. 2, pp. 88–98, 2018.