Diabetes Prediction Using Traditional Machine Learning Techniques

[1]Ayobami ADEMOROTI, [2]Oluwadarasimi O. OLOWE, [3]John I. AFE, [4]Oluwaseyi F. AFE, [5]Akintayo M. AYOADE

[1]bojeademoroti@gmail.com/+2348188247911;
[2]oloweoluwadarasimi@gmail.com,/+2349071049952; [3]afejohnibk.@gmail.com/
+2348146216145; [4]afe.seyi@lcu.edu.ng/+2348032015832:
[5]ayoade.akintayo@lcu.edu.ng, +2348033497824

[1]Yaba College of Technology, Lagos.
[2], [3, 4, 5]Lead City University, Ibadan.

Abstract

This study examines the capability of traditional machine learning (ML) algorithms to predict the onset of diabetes using the Pima Indians diabetes dataset. It employed decision trees, naive bayes, k-Nearest Neighbors (kNN), and logistic regression classifiers were evaluated using the performance metrics of accuracy, precision, recall, F1 score and ROC AUC. The data was pre-processed to amend implausible values and stratified sampling was performed to facilitate balancing classes when splitting the data. The naive bayes algorithm achieves the best accuracy (72.7%) while logistic regression obtains the best class separability (ROC AUC of 0.813). The project shows that interpretable models can provide actionable insights for early identification, supporting Sustainable Development Goal 3 (Good Health and Well-Being), particularly by promoting preventive healthcare and informed decision-making in resource-constrained environments.

Keywords: Logistics Regression, Stratified Sampling, Pima Indians Diabetes Dataset, Diabetes Prediction.

Word Count: 128

1. Introduction

Diabetes mellitus is a metabolic condition that affects millions of humans worldwide. It is primarily characterized by hyperglycemia caused by insulin resistance or insufficient insulin production, and can lead to devastating complications such as cardiovascular disease, kidney disease, neuropathy, or vision loss. According to the International Diabetes Federation (IDF), 537 million adults were living with diabetes in 2021, a number projected to increase to 643 million by 2030 and 783 million by 2045. (Magliano *et al.*, 2021)

Machine learning (ML) and artificial intelligence (AI) hold tremendous potential for personalized prediction systems for diabetes. Researchers have implemented ML techniques and data mining (though this study does not focus on data mining), for several diabetes related research topics, such as identification of diagnostics and predictive factors

related to the development of diabetes, predicting the incidence of diabetes, studying diabetic complications, developing drugs and therapies, and investigating genetic and environmental factors related to the onset and progression of diabetes. By an extensive investigation and analysis of raw diabetes-related data than ever before, ML can transform this data into knowledge and develop new ways of approaching how we prognose, diagnose and treat patients with diabetes. (Anderson *et al.*, 2016; Hu *et al.*, 2023).

In recent years, several survey articles have explored the use of ML models and AI in diabetes research. Some of these reviews have examined the application of ML tools across various diabetes-related domains. Others have taken a more targeted approach, focusing on specific areas such as diabetes detection (Lekha and Suchetha 2020; Sharma and Shah 2021) and diabetes prediction (Jaiswal *et al*., 2021; Theis *et al.*, 2021). These reviews offer valuable insights into the application of ML and AI in diabetes management and prognosis.

With diabetes being so prevalent, there is need for an immediate for accurate and early detection mechanisms. This study mainly reproduces existing work on the Pima Indians dataset using standard/traditional classifiers. Similar works are in abundant. The novelty is not strongly emphasized beyond re-validation. We may want to clarify new insight are contributed such as interpretable trade-offs when more recent and sophisticated models/classifiers are used. Patient health records from the Pima Indians Diabetes Dataset were utilized to evaluate the reliability of three traditional models for diabetes prediction. Logistic Regression was subsequently incorporated, given its widespread application as a baseline model for binary classification, thereby extending the analysis to four models.

Thus, the main contributions of this study are the following; to preprocess and examine the Pima Indians Diabetes Dataset; to build four different machine learning models to perform the prediction with Decision Tree, Naive Bayes, k-Nearest Neighbors (kNN), and Logistic Regression; and to compare the performance of the models in order to determine the best performing model.

2 Related Works

Diabetes affects a significant proportion of the adult population. Numerous studies have proposed methods for the prediction of diabetes symptoms. A wide variety of approaches, including ML, neural networks (NNs), data mining, and genetic algorithms, are discussed in these studies. In recent years, ML has gained popularity as a model-building technique and received a lot of attention from the medical community. ML has proven to have strong prediction powers as well as the capacity to analyze many variables in parallel. Moreover, ML has methods for variable screening can identify and interpret intricate correlations between variables. Previous research has proven that ML may be a useful technique for predicting diabetes. Some closely related works using ML algorithms are discussed in this section.

In a study by Tasin *et al.,* (2023), diabetes mellitus was predicted automatically using machine learning techniques. Pima Indians dataset and a new RTML dataset comprising physical examination data from the local female patients of Bangladesh were employed. The missing insulin feature values of the RTML dataset have been predicted from the Pima Indian dataset. The mutual information□based feature selection algorithm indicates the

glucose level, BMI, age, and insulin to be the most salient features in predicting diabetes. SMOTE and ADASYN synthetic data oversampling and hyperparameters optimization techniques have been applied. The XGBoost technique with ADASYN achieved the best performance. The LIME and SHAP explainable AI frameworks interpret the prediction provided by the ML approaches. A limitation of this study is the nonavailability of the insulin feature of the used RTML dataset. The prediction of insulin obtained from the XGB regressor and produced from the mean and median values of the Pima Indian dataset comprises an average deviation for classification accuracy of approximately 1.33% and 2.33%, respectively (Tasin *et al.,* 2023).

In this study by Alzboon *et al*., (2025), various ML algorithms were evaluated various machine learning (ML) algorithms for the early prediction and classification of diabetes risk. Using the Pima Indians Diabetes Database, which comprises 768 samples including significant demographic and clinical features like age, body mass index (BMI), and blood glucose levels, the researchers assessed algorithms such as Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors, Naive Bayes, Support Vector Machine, Gradient Boosting, and Neural Network Models. The study's findings showed that the Neural Network algorithm had the highest predictive accuracy (78.57 percent), and the Random Forest algorithm had the second-highest predictive accuracy (76.30 percent). The study also found that plasma glucose concentration, age, and BMI were the three most informative features for diabetes prediction. The authors conclude that ML strategies are extremely useful for data-driven early diabetes screening and can give some limited support for timely interventions to lessen the burden of disease. They also identified limitations to the study, the limited size of the dataset used and lack of lifestyle and genetic information in relation to the models, and they provide recommendations for future diabetes prediction research using extensive datasets to broaden and enhance predictive ability and generalizability.

In a study by Ghazizadeh *et al.,* (2025), it was shown that traditional machine learning models are still very effective in diabetes prediction and remain highly effective particularly in terms of accessibility and interpretability. The findings indicated that Random Forest and SVM are valid for clinical utilization, they are likely to yield reasonable prediction accuracy as an outcome of becoming less dependent on prior feature preprocessing. Overall, this work continues the ongoing dialogue in diabetes prediction using machine learning.

In a study by El-Sofany *et al.,* (2024) multiple ML and ensemble algorithms were evaluated based on their accuracy. The XGB algorithm had the highest level of performance using the SMOTE, with an accuracy of 97.4%, an F1 coefficient of 0.95, and an AUC of 0.87 for the private dataset and an accuracy of 83.1%, an F1 coefficient of 0.76, and an AUC of 0.85 for the combined datasets.

Compared with these works, the current study does not seek to maximize raw predictive accuracy but instead emphasizes the value of traditional, interpretable models. This distinction is important because complex models, while powerful, often require high computational resources, extensive hyperparameter tuning, and produce outputs that are

less transparent to clinicians. By contrast, traditional models such as Naïve Bayes and Logistic Regression are accessible, require limited computational resources, and provide explanations that can be more readily understood in clinical decision-making. Thus, the present study contributes to the field of computer science by highlighting the trade-off between interpretability and performance and demonstrating the practical relevance of simpler models in resource-limited healthcare environments.

3. Methodology

To achieve the aim of this paper, this study embarks on various procedures and implementation of diverse machine learning techniques to develop the proposed diabetes prediction system. Figure 1 shows the different stages of this research work. First, the dataset was collected and preprocessed to correct discrepancies in the dataset, such as, replacing missing values with median values and addressing class imbalance. Then the dataset was separated into the training set and test set using the holdout validation technique. Next, the classification algorithms such as the Decision Tree, Naive Bayes, k-Nearest Neighbors (kNN), and Logistic Regression were applied to find the best classification algorithm for this dataset.

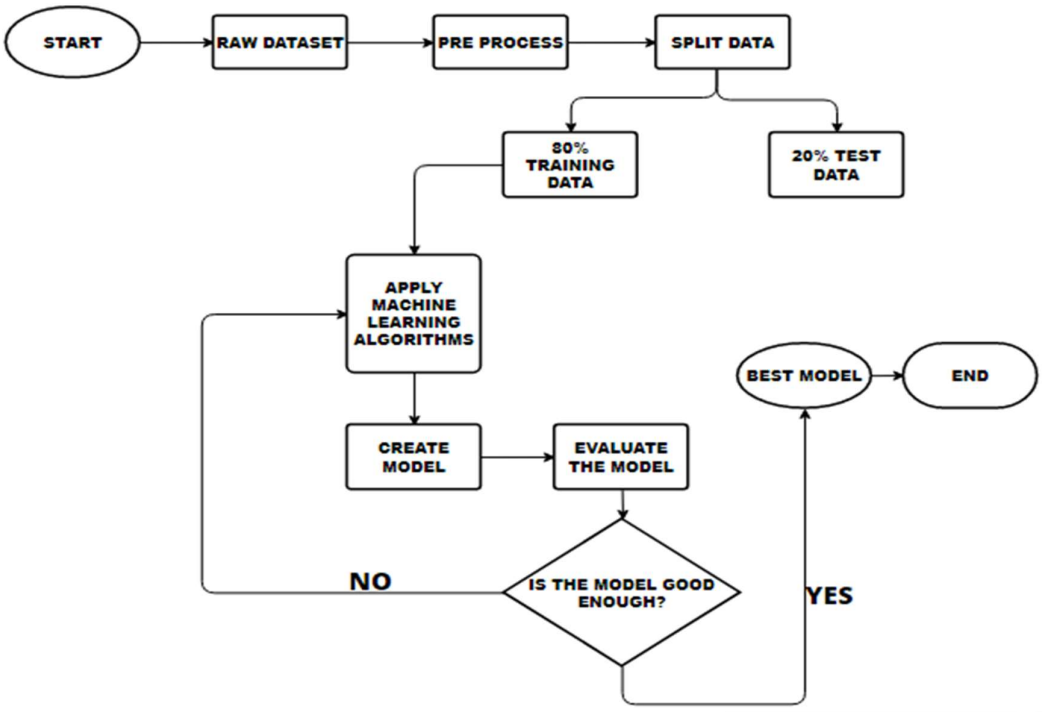

Figure 1: Workflow of the diabetes prediction model

3.1 Dataset

3.1.1 Dataset Components

The Pima Indians Diabetes Dataset (Pima Indians dataset is an open-source diabetes dataset that was initially gathered by the National Institute of Diabetes and Digestive and Kidney Diseases) UCI (2016) was obtained from the UCI Machine Learning Repository. The dataset

comprises data collected from female patients of Pima Indian heritage aged 21 years and above.

The dataset contains 768 observations with the following features:

- Pregnancies: Number of times pregnant.
- Glucose: Plasma glucose concentration.
- Blood Pressure: Diastolic blood pressure (mm Hg).
- Skin Thickness: Triceps skin fold thickness (mm).
- Insulin: 2-Hour serum insulin (mu U/ml).
- BMI: Body mass index (weight in kg/ ($\square eig \square t\ in\ m$)^2).
- Diabetes Pedigree Function: A function which scores likelihood of diabetes based on family history.
- Age: Age in years.
- Outcome: Class variable (0 for non-diabetic, 1 for diabetic).

Two derived features were created for exploratory analysis:

- Glucose_BMI: Product of Glucose and BMI values (to capture metabolic burden).
- Metabolic_Age: Product of Age and BMI (to assess age-adjusted metabolic risk). These features were not used in final models but informed mutual information analysis

This hyperlink contains the complete source code for data preprocessing, feature engineering, model training, and evaluation, along with instructions for replication.

Figure 2 shows the percentage of diabetes among Pima Indians participants. There are 768 records, and 268 of those individuals have been diagnosed with diabetes. Table 1 shows the eight features of the Pima Indians, and the comparison of the Pima Indians, including maximum, minimum, and average values, also shown in Table 1.

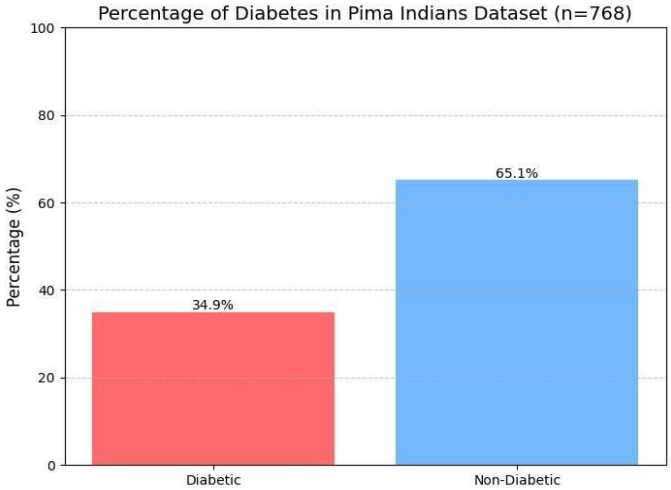

Figure 2: Percentage of diabetes among Pima Indians participants.

Table 1. features of the Pima Indians datasets

INDEX	DESCRIPTION	MIN	MAX	AVERAGE
PREGNANCIES	Number of pregnancies	0.0	17.0	3.85
GLUCOSE	Plasma glucose concentration (mg/dL)	44.0	199.0	121.66
BLOOD PRESSURE	Diastolic blood pressure (mm Hg)	24.0	122.0	72.39
SKIN THICKNESS	Triceps skinfold thickness (mm)	7.0	99.0	29.11
INSULIN	2-Hour serum insulin (mu U/ml)	14.0	846.0	140.67
BMI	Body mass index (kg/m²)	18.2	67.1	32.46
DIABETES PEDIGREE	Diabetes pedigree function	0.08	2.42	0.47
AGE	Age (years)	21.0	81.0	33.24
OUTCOME	Diabetes status (0: No, 1: Yes)	0.0	1.0	0.35
GLUCOSE_BMI	NaN	1100.0	10692.0	3996.67
METABOLIC_AGE	NaN	382.2	2697.0	1080.91

3.1.2 Dataset Preparation and Processing

The dataset used for this project is the publicly available Pima Indians Diabetes dataset. We detected biologically implausible zero values (e.g., Glucose=0, BMI=0) in medical features, which were replaced with NaN and imputed using the median. The dataset was split into training (80%) and test (20%) sets using stratified sampling to preserve class balance. For model evaluation, we used both holdout validation (test set) and 5-fold cross-validation (on the training set) to ensure robustness.

3.2 Machine Learning Classifiers

In this research, traditional machine learning techniques were employed to predict diabetes onset. The classifiers are briefly described below:

- Decision Tree: A decision tree is a non-parametric supervised learning method that partitions the feature space into hierarchical branches based on decision rules. It uses metrics like Gini impurity or information gain to split nodes, making it interpretable but prone to overfitting without pruning.
- k-Nearest Neighbors (kNN): A distance-based algorithm that classifies data points by majority voting among their k nearest neighbors in the training set.

- Logistic Regression: A linear model that estimates class probabilities using the sigmoid function.
- Naive Bayes: A probabilistic classifier based on Bayes' theorem, assuming conditional independence between features.

Patient health records from the Pima Indians Diabetes Dataset were utilized to evaluate the reliability of three traditional models for diabetes prediction. Logistic Regression was subsequently incorporated, given its widespread application as a baseline model for binary classification, thereby extending the analysis to four models.

4 Results and Discussion

This section presents the performance evaluation of the proposed diabetes prediction system, comparing traditional machine learning classifiers and comparing the performance of traditional machine learning classifiers. The model provides risk scores alongside predictions, with SHAP values quantifying feature contributions for individual cases.

4.1 Model Performance Evaluation

We assessed classifiers using the metrics precision, recall, F1 score, ROC AUC, and accuracy, defined as:

Precision = $\frac{TP}{(TP+F\ )}$ (Measure of exactness) - - -(1)

Recall = $\frac{TP}{(TP+F\ )}$ (Measure of completeness) - - -(2)

F1 Score = $\frac{2\times(\text{Precision}\times\text{Recall})}{(\text{Precision}+\text{Recall})}$ - - -(3)

ROC AUC = Area under the ROC curve, measuring class separability

where:

TP (True Positive): Correctly predicted diabetic cases.
FP (False Positive): Non-diabetic cases incorrectly flagged as diabetic.
TN (True Negative): Correctly predicted non-diabetic cases.
FN (False Negative): Diabetic cases missed by the model.

All models were evaluated using an 80:20 stratified holdout validation to ensure class balance.

4.1.1 Cross Validation

The cross validation results for the various classifiers used in the model are as follows; Decision Tree CV ROC AUC: 0.737 ± 0.037, Naive Bayes CV ROC AUC: 0.837 ± 0.018, KNN CV ROC AUC: 0.824 ± 0.038, and Logistic Regression CV ROC AUC: 0.843 ± 0.029

4.1.2 Decision Tree Evaluation

The decision Tree results showed an Accuracy: 0.675, F1 Score: 0.457 and the ROC AUC: 0.742. Similarly, the results for the Precision, Recall, F1-score and Support are shown in Table
2.

Table 2: Classification Report for Decision Tree

	Precision	**Recall**	**F1-score**	**Support**
0	0.72	0.83	0.77	100
1	0.55	0.39	0.46	54

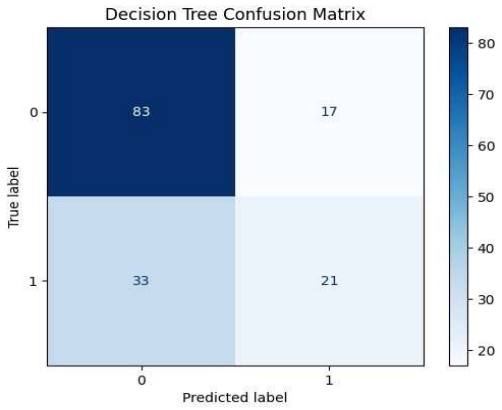

Figure 3: Decision Tree Confusion Matrix

Figure 3 shows the Decision Tree confusion matrix. The model achieved moderate recall for the non-diabetic class but weak detection of diabetic cases (39% recall), suggesting overfitting due to tree depth**.** From the results, the Cross-validation ROC AUC on the Decision Tree showed that CV ROC AUC yielded 0.761 ± 0.036. This shows the average ROC AUC score across 5 different folds during cross-validation on the training data. The average is 0.761, with a standard deviation of 0.036. This gives an idea of how consistently the model performs across different subsets of the training data. Similarly, from the performance on the test data set showed an Accuracy of 0.675. This implies that the decision tree model correctly classified about 67.5% of the patients in the test set. The F1 Score also gave a value of 0.457. This is the harmonic mean of precision and recall for the positive class (diabetes). A lower F1 score here suggests a weaker balance between the model's ability to correctly identify positive cases and avoid false positives. The area under the Receiver Operating Characteristic curve for the test set (ROC AUC) was 0.742. It measures the model's ability to distinguish between diabetic and non-diabetic individuals. Among the models trained, the Decision Tree has the lowest ROC AUC.

The Classification Report for the Non-Diabetic class yielded a precision of 0.72. This implies that the model predicts non-diabetic, correctly 72% of the time. The Recall value was 0.83 showing that the model correctly identifies 83% of actual non-diabetic cases. The F1-score was 0.77 this shows a good balance for the non-diabetic class. On the other hand, the Classification Report for the Diabetic class yielded a Precision of 0.55 showing that the model predicts diabetic, correctly 55% of the time. The Recall value was 0.39 showing that the model only correctly identifies 39% of actual diabetic cases. The F1-score of value 0.46 reflects the poor balance between precision and recall for the diabetic class. The Decision Tree model, with maximum depth of 5, shows a decent performance in identifying individuals without diabetes (high TN and good recall for class 0). However, it struggles significantly

with identifying individuals who do have diabetes. It has a low True Positive count (21) and a high False Negative count (33), meaning it misses a large proportion of actual diabetes cases. Its lower ROC AUC and F1 score for the positive class also reflect this weakness. While its precision for the diabetic class (0.55) means that when it does predict diabetes, it's correct more than half the time, its low recall is a major concern if the goal is to identify as many diabetic individuals as possible for early intervention.

4.1.3 Naive Bayes Evaluation

The results gotten from the implementation showed that for Naive Bayes the Accuracy was 0.727, F1 Score: 0.625, and ROC AUC: 0.775. Similarly, the results for the Precision, Recall, F1-score and Support are shown in Table 3.

Classification Report for Naive Bayes

	Precision	**Recall**	**F1-score**	**Support**
0	0.80	0.77	0.79	100
1	0.60	0.65	0.62	54

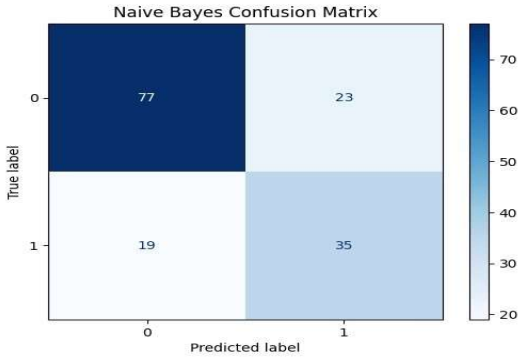

Figure 4: Naive Bayes Confusion Matrix

The figure 4 illustrates the Naive Bayes Confusion Matrix with the Highest overall accuracy and balanced recall (~65%), indicating effective classification with minimal computation. The True Negatives (TN) for the Naïve Bayes Confusion matrix had a value of 77. This indicates that the model correctly identified 77 individuals who did not have diabetes. The False Positives (FP) for the Naïve Bayes Confusion matrix gave a value of 23. This means the model incorrectly predicted that 23 individuals had diabetes when they actually did not (Type I error). The False Negatives (FN) for the Naïve Bayes Confusion matrix had a value of 19. This means the model incorrectly predicted that 19 individuals did not have diabetes when they actually did (Type II error). Finally, the True Positives (TP) for the Naïve Bayes Confusion matrix had a value of 35. This indicates that the model correctly identified 35 individuals who did have diabetes.

In summary, the Naive Bayes model correctly identified 77 non-diabetic individuals, it incorrectly predicted 23 non-diabetic individuals as diabetic, it incorrectly predicted 19 diabetic individuals as non-diabetic and it correctly identified 35 diabetic individuals.

Comparing this to the other models, Naive Bayes has the highest number of True Positives (35) and the lowest number of False Negatives (19), which aligns with its high Recall score. This suggests that Naive Bayes is the most effective of these models at identifying individuals who truly have diabetes, even if it has a slightly higher number of false positives compared to Logistic Regression and KNN.

4.1.4 k-Nearest Neighbor (KNN) Evaluation

The results gotten from the implementation showed that for k-Nearest Neighbor (KNN) the Accuracy was 0.708, F1 Score: 0.545, and ROC AUC: 0.801. Table 4 also shows the Precision, Recall, F1-score and support results for the KNN classifier.

Classification Report for k-Nearest Neighbor (KNN)

	Precision	**Recall**	**F1-score**	**Support**
0	0.75	0.82	0.78	100
1	0.60	0.50	0.55	54

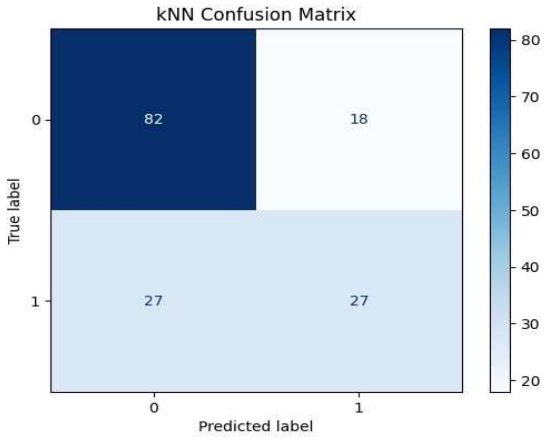

Figure 5: k-Nearest Neighbors Confusion Matrix

The figure 5 illustrates the k-Nearest Neighbors Confusion Matrix, having a balanced performance with decent recall, but less consistent across test samples highlighting sensitivity to parameter tuning. The True Negatives (TN) for the KNN model had a value of 82. This means the model correctly predicted that 82 individuals did not have diabetes. The False Positives (FP) for the KNN model had a value of 18. This means the model incorrectly predicted that 18 individuals did have diabetes when they actually didn't. This is also known as a Type I error. The False Negatives (FN) for the KNN model had a value of 27. This means the model incorrectly predicted that 27 individuals did not have diabetes when they actually did. This is also known as a Type II error. The True Positives (TP) for the KNN model had a value of 27. This means the model correctly predicted that 27 individuals did have diabetes.

In summary, the KNN model performed reasonably well in identifying individuals without diabetes with 82 True Negatives, but it had more difficulty correctly identifying individuals who have diabetes with a value of 27 True Positives and also produced a notable number of false negatives with a value of 27 False Negatives, meaning it missed 27 cases of diabetes.

4.1.5 Logistic Regression Evaluation

The results gotten from the implementation showed that for *Logistic Regression* the Accuracy was 0.701, F1 Score: 0.540, and ROC AUC: 0.813. Table 5 shows the precision, Recall, F1score and support results for the Logistic Regression.

Classification Report for Logistic Regression

	Precision	**Recall**	**F1-score**	**Support**
0	0.75	0.81	0.78	100
1	0.59	0.50	0.54	54

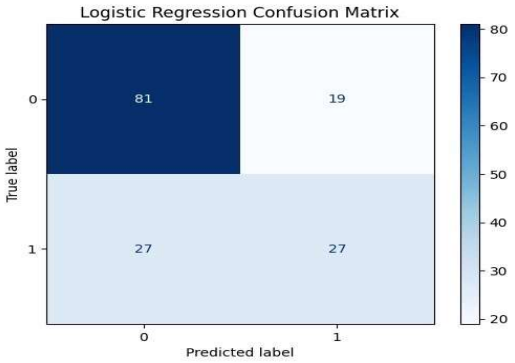

Figure 6: Logistic Regression Confusion Matrix

The figure 6 illustrates the Logistic Regression Confusion Matrix, having a Strong ROC AUC (0.813) and robust separation between classes, making it suitable for clinical interpretability. The True Negatives (TN) for the logistic regression model had a value of 81. This means the model correctly predicted that 81 individuals did not have diabetes. The False Positives (FP) for the logistic regression model had a value of 19. This means the model incorrectly predicted that 19 individuals did have diabetes when they actually didn't. The False Negatives (FN) for the logistic regression model had a value of 27. This means the model incorrectly predicted that 27 individuals did not have diabetes (when they actually did. The True Positives (TP) for the logistic regression model had a value of 27. This means the model correctly predicted that 27 individuals did have diabetes. Comparing the Logistic Regression and KNN confusion matrices, you can see that they have a very similar number of True Positives (27 for both) and False Negatives (27 for both). Logistic Regression has one fewer True Negative (81 vs 82) and one more False Positive (19 vs 18) than KNN. This indicates that both models have similar performance in terms of identifying actual diabetes cases and missing them, but Logistic

Regression has a slightly higher tendency to incorrectly predict diabetes for individuals who don't have it.

4.2 Comparison of Classifier Performance

In order to determine the machine learning classifier with the best performance, this study carried

out a comparison among the chosen classifiers; Decision Tree, Naïve Bayes, k-Nearest Neighbor and the Logistic Regression. The various results on each of these classifiers is shown in Table 6.

Table 6: Performance metrics of various classifiers technique

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Training Time (s)
Decision Tree	0.67533	0.55263	0.38889	0.45652	0.74185	0.01000
Naive Bayes	0.72727	0.60345	0.64815	0.62500	0.77500	0.00000
KNN	0.70779	0.60000	0.50000	0.54546	0.80111	0.00000
Logistic Regression	0.70130	0.58696	0.50000	0.54000	0.81315	0.01000

The Table 6 compares classifier performance on the original (unbalanced) Pima Indians dataset. The Naive Bayes model achieved the highest accuracy (0.727) and recall (0.648), while Logistic Regression had the best ROC AUC (0.813)

The performance of the models, as measured by accuracy, F1 score, and ROC AUC, shows that Naive Bayes and Logistic Regression are the most effective for this task. Naive Bayes achieved the highest accuracy (72.7%) and recall (64.8%), while Logistic Regression achieved the highest ROC AUC (81.3%), indicating strong class separability.

It is important to note that the accuracy achieved in this study (72.7%) is lower than some reported in the literature (e.g., (Alzboon *et al*., 2025) reported 78.57% with Neural Networks, and (El-Sofany *et al.,* 2024) reported up to 97.4% with XGBoost). This difference can be attributed to several factors:

- Our study focused on traditional, interpretable models without using ensemble methods or deep learning.
- The Pima dataset is limited in both size and features (lacking lifestyle and genetic factors), which inherently caps the predictive performance.
- The goal of this project was to validate simple and accessible models for settings where complex models may not be feasible, such as in resource-constrained healthcare systems.

Naïve Bayes achieved the highest accuracy and recall, while Logistic Regression achieved the best ROC AUC, reflecting superior class separation. Importantly, Naïve Bayes and Logistic Regression balance performance with interpretability, making them valuable for real-world healthcare applications.

5 Conclusion

This study validates the effectiveness of traditional ML algorithms in predicting diabetes using clinical data. Naïve Bayes and Logistic Regression demonstrated strong performance, despite the limitations of the Pima dataset (small size, lack of lifestyle and genetic features). Interpretable and accessible models are particularly relevant for under-resourced healthcare environments.

This study aligns with Sustainable development goal 3 (Good Health and Well-being) by contributing to the early detection and prevention of diabetes using accessible, interpretable machine learning models.

While this study demonstrates the effectiveness of traditional machine learning techniques in predicting diabetes, several limitations must be acknowledged. First, the dataset used primarily the Pima Indians Diabetes Dataset has a relatively small sample size and lacks diversity, limiting the generalizability of the findings across broader populations. Second, the dataset does not include critical lifestyle and genetic factors such as diet, exercise, or family medical history, which could improve predictive accuracy. Third, the assumption of feature independence in models like Naive Bayes may not hold true for all medical features. Additionally, some models showed relatively low recall scores, meaning that a significant portion of diabetic cases could be missed.

To further improve diabetes prediction, future studies should consider larger datasets, inclusion of lifestyle and genetic factors, and more advanced models such as neural networks. Future studies should incorporate larger and more diverse datasets that include patients from different ethnic backgrounds, age groups, and regions. Including lifestyle, behavioral, and genetic information would help create more comprehensive models.

Acknowledgement

The authors acknowledge the efforts of the reviewers of this paper. We also appreciate their meaningful contribution, valuable suggestions, and comments to this paper which helped us in improving the quality of the manuscript.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in *https://github.com/ImmatureBug/Diabetes-Prediction-Using-Traditional-Machine-Learning.git*

References

Alzboon, M. S., Alqaraleh, M., & Al-Batah, M. S. (2025). Diabetes Prediction and Management Using Machine Learning Approaches. *arXiv preprint arXiv:2506.11501*.

Anderson, J. P., Parikh, J. R., Shenfeld, D. K., Ivanov, V., Marks, C., Church, B. W., ... & Rublee, D. A. (2016). Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *Journal of diabetes science and technology*, *10*(1), 6-18.

El-Sofany, H., El-Seoud, S. A., Karam, O. H., Abd El-Latif, Y. M., & Taj-Eddin, I. A. (2024). A proposed technique using machine learning for the prediction of diabetes disease through a mobile app. *International Journal of Intelligent Systems*, *2024*(1), 6688934.

Ghazizadeh, Y., Salehi, S., & Mirsaeid Ghazi, M. (2025). Machine learning-based diabetes prediction: A comprehensive study on predictive modeling and risk assessment. *J Clin Images Med Case Rep*, *6*(5), 3578.

Hu, P., Li, X., Lu, N., Dong, K., Bai, X., Liang, T., & Li, J. (2023). Prediction of new-onset diabetes after pancreatectomy with subspace clustering based multi-view feature selection. *IEEE Journal of Biomedical and Health Informatics*, *27*(3), 1588-1599.

Jadhao Y.B, Tayde K.R, Gaur R, Ingle S.R, Vyavhare Y.G. (2025). A Comprehensive Review of Machine Learning Approaches for Diabetes Prediction: Methods, Challenges, and Future Directions. International Research Journal of Modernization in Engineering Technology and Science. 7(2),1-8.

Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes*, *15*(3), 435-443.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.

Lekha, S., & Suchetha, M. (2020). Recent advancements and future prospects on e-nose sensors technology and machine learning approaches for non-invasive diabetes diagnosis: A review. *IEEE reviews in biomedical engineering*, *14*, 127-138.

Magliano, D. J., Boyko, E. J., & Atlas, I. D. (2021). COVID-19 and diabetes. In *IDF DIABETES ATLAS [Internet]. 10th edition*. International Diabetes Federation.

Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1), 30.

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, *10*(1-2), 1-10.

Theis, J., Galanter, W. L., Boyd, A. D., & Darabi, H. (2021). Improving the in-hospital mortality prediction of diabetes ICU patients using a process mining/deep learning architecture. *IEEE Journal of Biomedical and Health Informatics*, *26*(1), 388-399.

UCI Machine Learning Repository: Pima Indian Diabetes Dataset. (2016). https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database