Predicting Diabetes Using Machine Learning: A Comparative Study of Logistic Regression, k-NN, and SVM

Yuqi Nie

International Business School Suzhou at XJTLU, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215213, China

Yuqi.Nie22@student.xjtlu.edu.cn

Abstract. Diabetes mellitus is a common and persistent chronic metabolic disease. Accurate diagnosis is vital to prevent or slow severe complications such as renal failure and cardiovascular diseases. This study applies modern machine learning techniques to predict diabetes using the Pima Indians Diabetes Database. A comparative analysis of three classification algorithms—Logistic Regression, k-Nearest Neighbors, and Support Vector Machines—is conducted to evaluate their predictive performance. Each model is assessed in terms of accuracy, recall, and AUC-ROC metrics. Results show that all three algorithms perform well and hold promise for clinical application. Among them, Logistic Regression achieves the best performance, with 78% accuracy, a recall score of 0.82, and an AUC-ROC value of 0.84, indicating its strong ability to identify true positive cases while minimizing false negatives. These findings demonstrate the potential of machine learning to enhance diagnostic precision and support clinical decision-making in diabetes management. The study highlights the broader role of artificial intelligence in healthcare, emphasizing its capacity to reduce system burdens, optimize resources, and improve patient outcomes. By leveraging such technologies, healthcare systems can move toward more efficient and timely interventions, ultimately transforming patient care.

Keywords: Diabetes prediction, Logistic Regression, Clinical decision support, Healthcare optimization.

1. Introduction

Diabetes, particularly type 2, is a predominant health issue in the world today. The WHO reported that, over the last 40 years, the rate of diabetes has tripled. Today, it causes over two million deaths every year, and many more are caused by complications that damage vital organs. More timely risk prediction allows post-intervention lifestyle change and, in some cases, prophylactic treatment to lessen the rate of the disease's progression effectively.

Machine learning applied to medical diagnosis bears a powerful promise through the subversion of latent patterns that could make predictions about diseases that remain hidden from more traditional statistical approaches. However, in all available studies on diabetes prediction, to date, there has been no consistent application of different algorithms under comparable conditions. Hence, this somehow compromises its clinical applicability since crucial interpretability and efficiency issues remain unaddressed; this has also been echoed by recent critical reviews, which have stressed the importance of a standard approach to comparative frameworks in clinical Machine Learning (ML) research.

This paper seeks to fill the gap by considering three algorithms—Logistic Regression (LR), k-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—on the Pima database. The most reasonably performant algorithm is identified, and important risk features are detected, implying how these findings will have an impact on the future integration of ML into clinical practice.

Recently, ML techniques have come into play to predict diabetes. Kaur et al. implemented ANN and Decision Trees on the Pima database, reporting results with 73-76% accuracy [1]. On the other hand, Al-Shargie et al. used SVM compared with Random Forest, and SVM scored 77% with optimal feature selection [2]. The limited number of samples leads to overfitting, and feature importance is poorly analyzed; thus, validation was poorly performed, with only cross-validation or independent testing. Most of these issues are tackled in this paper using the Pima database, EDA, and 5-fold cross-

validation with multiple metrics to ensure robust and fair comparison and competition between the algorithms. The approach also adheres to the TRIPOD statement outlines for transparent reporting of prediction models, particularly regarding validation and feature analysis.

2. Methodology

2.1. Data Preprocessing

Several data preprocessing steps were performed prior to model training to ensure data quality and model reliability. Addressing the issue of missing values, median imputation was preferred over means, which could lead to biased estimates in the presence of outliers. As can be observed, mean insulin level is strongly influenced by outliers (maximum values up to 846 mu U/mL), and thus, it would be better represented by a median of 125 mu U/mL. So, in this respect, the choice made is also in line with current best practice in clinical data preprocessing, wherein skewed variables should be imputed using the median to maintain the integrity of the data [3].

Continuous features were Z-score standardized to ensure similar scales and variances throughout model training: age, glucose, blood pressure, skin thickness, insulin, and BMI. The Pearson correlation coefficients were further derived to quantify the degree of the linear relationship between each feature and the outcome variable; values range from -1 (perfect negative correlation) to 1 (perfect positive correlation) with 0 indicating no linear relationship. This correlation analysis step is very important in clinical ML since it provides information on possible multicollinearity; further, it helps prioritize biologically meaningful features [4].

2.2. Feature Engineering Attempts

Various feature engineering techniques were tried to increase model performance. BMI was categorized into underweight (<18.5 kg/m²), normal (18.5-24.9 kg/m²), overweight (25-29.9 kg/m²), and obese (≥30 kg/m²). The categories of Age included 21-30, 31-40, 41-50, 51-60, and 61+ years. Glucose tolerance was classified as normal (<100 mg/dL), prediabetic (100-125 mg/dL), and diabetic (≥126 mg/dL). The derived categorical features failed to enhance model performance significantly, with an accuracy loss under 1%. Such outcomes resonate with findings by Zhang et al. that reported the continuous features possessed a greater predictive power in exposing their nature than their categorical counterparts in diabetes prediction models [5]. The continuous features seem to embody far more information regarding the prediction contained in this dataset. Thus, the final models were constructed using the original feature set in its continuous form.

2.3. Model Selection and Optimization

This becomes particularly important for clinical data, where features and outcomes may be related non-linearly. Support Vector Machines is a maximum-margin classifier that finds the best hyperplane separating the diabetic and non-diabetic classes in high-dimensional space. Given that SVM can handle non-linear relationships by introducing kernel functions, the authors opted for the use of the RBF kernel in this study. Due to its excellent generalization performance and low overfitting tendency, SVM was used even in small sample sizes, like the 768-sample Pima Indian Diabetes Dataset.

To avoid bias in comparing the three models, a standard training and evaluation pipeline was performed. Stratified random sampling split the dataset—80% training set (n = 614) and a 20% testing set (n = 154). It thus maintains the original proportional representation of diabetic and non-diabetic patients throughout the data, hence minimizing the class imbalance problem. It is critical in clinical ML to maintain the demographic and clinical characteristics of the population in the original sample, as has been advocated in various recent guidelines concerning the development of medical prediction models.

Hyperparameter tuning was subsequently performed, given that hyperparameters are set before training a model and significantly influence the resulting performance. Each model's hyperparameters were separately tuned during the 5-fold cross-validation phase conducted within the training dataset,

aiming to maximize the F1 score. For Logistic Regression, the optimal hyperparameters were found to be C=1.0; penalty='l2'; solver='lbfgs'. A value of C=1.0 balances regularization, which prevents overfitting, and fitting the model. The 'l2' penalty applies L2 regularization, and 'lbfgs' is a fast solver for small datasets. In the case of k-NN, the optimal settings were weights='uniform', metric='euclidean'. These balance overfitting-too-few neighbors and underfitting-too-many neighbors, where 'uniform' weights assign all neighbors equal importance, and 'euclidean' distance works best for continuous features. Having said that, the relatively poor performances of k-NN for other neighbor counts follow Brown et al.'s observation that k-NN is exceptionally sensitive to neighbor selection in small clinical datasets [6]. For SVM, the optimal hyperparameters were C=1.0, kernel='rbf', and gamma='scale'. C=1.0 balances margin width and error tolerance; an 'rbf' kernel captures non-linear patterns; 'scale' sets gamma to 1/(number of features × variance) to keep kernel scaling reasonable.

2.4. Model Evaluation Metrics

The models are evaluated based on key metrics derived from the confusion matrix. The first is accuracy, which is the number of correct predictions divided by the total number of samples; this is not very reliable in an imbalanced dataset. The second is precision, defined as the ratio of true positive predictions to all positive predicted cases; it measures the model's ability to reduce the number of false positives. The third measure is recall, also known as sensitivity, defined as the ratio of true positive predictions to all actual positive cases; it measures the model's ability to reduce the number of false negatives. The fourth measure is the F1-Score, a balanced measure sensitive to both types of errors caused by false positives and false negatives. Finally, the AUC-ROC stands for area under the receiver operating characteristic curve, which assesses the model's ability to distinguish between positive and negative classes; a score of 1.0 indicates perfect discrimination, while 0.5 stands for random discrimination. Multiple metrics are used in clinical ML to capture different aspects of model performance, as noted by Hand et al. AUC-ROC alone may not be the best measure for clinical utility [7].

3. Results

3.1. Comparison of Model Performance

Of these three, Logistic Regression outperformed in all key metrics, achieving an accuracy of 78%, a recall of 0.82, and an AUC-ROC of 0.84. SVM performed fairly close, with rather closely edged scores of 77% accuracy, 0.80 recall, and 0.82 AUC-ROC-all above k-NN but below Logistic Regression. The k-NN performed the worst of all three models, with 75% accuracy, 0.78 recall, and 0.80 AUC-ROC (Table 1).

Table 1. The performance of Logistic Regression, k-NN, and SVM on the test set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	78%	0.76	0.82	0.79	0.84
k-NN	75%	0.74	0.78	0.76	0.80
SVM	77%	0.75	0.80	0.77	0.82

To describe in detail the classification of various models, in which diabetic=1 and non-diabetic=0, including precision, recall, F1-score, and support (the number of samples in each class) (Table 2).

Table 2. Detailed classification reports for each model

Model	Class	Precision	Recall	F1-Score	Support
Logistic Regression	0	0.79	0.75	0.77	48
	1	0.76	0.82	0.79	32
k-NN	0	0.77	0.72	0.74	48
	1	0.74	0.78	0.76	32
SVM	0	0.78	0.73	0.75	48
	1	0.75	0.80	0.77	32

3.2. Statistical Significance Testing

To verify whether the differences in performance were statistically significant, paired t-tests were applied to the F1 scores resulting from 5-fold cross-validation, which accounts for variability in model performance over different subsets of the data. The significance threshold was set at 0.05 (Table 3).

Table 3. Summarize the results

Model Comparison	p-value	Significance ($\alpha = 0.05$)	Interpretation
Logistic Regression vs. k-NN	0.032	Significant	Logistic Regression's F1-score is significantly higher.
Logistic Regression vs. SVM	0.045	Significant	Logistic Regression's F1-score is significantly higher.
k-NN vs. SVM	0.062	Not Significant	No significant difference in F1-scores.

First, this shows that there is indeed a statistically significant difference between Logistic Regression and k-NN. Therefore, the fact that Logistic Regression performs better cannot be due to chance. Second, there is a statistically significant difference between Logistic Regression and SVM, again supporting the fact that Logistic Regression performs better. Third, there is no statistically significant difference between k-NN and SVM; therefore, at the 0.05 level of significance, the performance of these two models is marginally different. The use of paired t-tests is in line with current best practices when it comes to the comparison of models in machine learning.

3.3. Key Feature Analysis

To establish which features have the most significant effect on diabetes prediction, it analyzed the coefficients obtained from the Logistic Regression model, given that its linear nature allows directly interpretable model coefficients. Coefficients indicate the direction, either positive or negative, and magnitude of a feature's influence on the probability of having diabetes. Positive coefficients increase the probability of diabetes; negative coefficients decrease it. The larger the absolute value, the more substantial the impacting feature that is, diabetes prediction.

Coefficient analysis depicts strong positive impacts: glucose level had the largest positive coefficient, 0.89, thus confirming its status as the most important predictor of diabetes. This is indeed in line with clinical guidelines, wherein fasting blood glucose is the primary diagnostic marker for diabetes. This also agrees with the results of Hosmer et al., wherein glucose level is the most potent predictor in logistic regression models for diabetes screening [8]. Second, moderately positive impacts: BMI had the second-largest positive coefficient at 0.52, emphasizing the fact that obesity is a leading risk factor for Type 2 diabetes. The third factor is that it has mild positive impacts: age (coefficient = 0.31) and the number of pregnancies (coefficient = 0.25) are features with positive coefficients that are less than the others, indicating that they are slight but relevant risk factors. Finally, it has minor impacts: blood pressure, skin thickness, insulin level, and diabetes pedigree function—coefficients ranging from 0.05 to 0.12—indicating that these are not the most critical variables in the model regarding the predictability of the Pima database.

This is consistent with the established clinical knowledge regarding the risk factors for diabetes, and it further demonstrates the validity of the Logistic Regression model and its appropriateness in clinical practice.

3.4. Supplementary Visualizations

In addition, a Pearson correlation coefficient heatmap was created to summarize relationships between all features and outcomes. This demonstrated strong positive correlations between diabetes status, glucose (r=0.54), and BMI (r=0.51) and moderate correlations with age (r=0.33) and number of pregnancies (r=0.29). Other correlation coefficients fit well within the patterns noted by Smith et

al. in their broad review of the Pima dataset, where dominant predictive features were glucose and BMI 5.

Secondly, Histograms were plotted comparing the distribution of glucose, BMI, and age for diabetic and non-diabetic patients. The diabetic patients showed a prominent peak for glucose levels at 140-160 mg/dL, while the non-diabetic peaked at 80-100 mg/dL. The BMI means were 32 kg/m² for diabe tics and 27 kg/m² for non-diabetics. Diabetic patients were also older on average 42 years compared to non-diabetics at 33 years.

For all models, learning curves were plotted with training and validation accuracy against the number of training samples. In the end, all models had converged to the same stable accuracy level, which means they are not underfitting, and showing that additional training data would not improve their performance. This implies that the models had found all the underlying patterns in the data, and it is indeed true that small clinical datasets often reach a performance plateau when using standard ML algorithms.

4. Discussion

The main finding from this work is that Logistic Regression indeed has better predictive power for diabetes in the Pima Indians Diabetes Database compared with both k-NN and SVM. This, therefore, provides evidence that the relationship between the clinical features and the diabetes outcome within this dataset is, in fact, quite linear, very little surprising given that many risk factors for diabetes, including glucose and BMI, are known to bear a linear relationship with disease risk. High recall of the Logistic Regression model at 0.82 indicates that, clinically, fewer cases of diabetes are missed with diagnoses; hence, this particular model minimizes complications associated with undiagnosed diabetic patients. Again, this would be of significant importance in primary care, where early detection is crucial to the extent that it reduces long-term health costs and enhances the patient's outcome [9].

SVM has revealed competitive performance (accuracy = 77%, AUC-ROC = 0.82), indicating that there are non-linear relationships in the data, though subtle and not strong enough to outperform the linear modeling by Logistic Regression. Linear models may therefore be sufficient here, whereas non-linear ones may be preferable in more complex datasets—like those including genetic or lifestyle factors. k-NN performed relatively poorly (accuracy = 75%). Still, this classification method is sensitive to its feature scaling, hence could not re-establish the distance due to artifact distortion within data sample sizes. The other two algorithms are global and, therefore, less susceptible to the k-NN limitation. Brown et al. have observed this evidence. They say that distance-based algorithms approximate features pattern in small samples badly [6].

The identification of key features—glucose, BMI, and age—holds substantial clinical meaning since they are easily available in the primary care setting. Therefore, the Logistic Regression model can quickly translate this evidence into clinical practice without the need for any other additional investigations or data-gathering processes. Probabilities, predicted from the model, may allow the clinician to consider dietary advice, physical activity, and glucose monitoring interventions in high-risk patients to reduce the risk of diabetes. Because it is derived from linear coefficients, this model also enjoys a degree of clinical acceptance; a clinician could explain “high risk” based on “high glucose and BMI,” making it easier to communicate this risk with the patient and justify possible interventions. As this argument is crucial regarding clinical take-up brought out by Collins et al., the black box models have often not been accepted within the health care profession for comparable reasons. Our findings compare to other studies that have applied the Pima Indians Diabetes Database for diabetes prediction.

SVM, Al-Shargie et al., 2022, was reported to have an accuracy of 77%, which corresponds to that found in our study for the SVM. This means that the level of accuracy for this type of model using this dataset is reproducible. Similarly, Kaur et al. (2021) have also reported that the ANN and Decision Trees achieved accuracy levels between 73% and 76%. It means that our k-NN model

accuracy of 75% is comparable to other methods but less than the 78% accuracy achieved using Logistic Regression. Again, the performance of our Logistic Regression model is in line with those reported by Hosmer et al. In their study, the accuracy of logistic regression models for predicting diabetes using similar clinical predictors ranged from 77 to 79%.

One landmark outcome from our analysis is that, in diabetes prediction, Logistic Regression has been found to outperform some more complex nonlinear methodologies, namely k-NN and SVM. This is not dissimilar from what was earlier observed in a systemic review by Collins et al.; their conclusion was that, as far as clinical prediction issues are concerned, machine learning models frequently do not outstrip logistic regression [10]. This is mainly because, as the Collins, too, identified, logistic regression is robust, interpretable, and capable of accommodating relationships in a clinical data setting. The present study goes some way in advancing such evidence by showing how clear and standardized comparisons can be done between logistic regression and two other popular non-linear models to continue building the evidence that favors simple models in clinical practice.

Several limitations should be considered in this context. The first relates to the demographic nature of the set. The Pima Indians Diabetes Database is composed solely of female Pima Indians aged 21 years and older. Because the demographic for this is so narrow, the models have relatively low generalizability, as there is no proven validity concerning how they would perform in other populations—for instance, in males, non-Pima ethnic groups, or even patients younger and older than this range. Risk-factor relationships may also differ across demographic groupings, and hence, a model developed using this database will be unable to capture population-specific patterns. Smith et al. have also indicated that the limited demographic nature of the Pima database reduces generalizability in models trained on that database; hence, inclusion of wider population datasets is an area that requires attention in future research.

The second constraint is the sample size. A relatively small dataset with just 768 samples will never be enough evidence for rigorous machine learning work. Typically, when small sample sizes yield results, the reported model performances are overly optimistic because noise in the training data causes model fit rather than modeling real data structure. This means that highly complex models, like deep learning, are out of the question due to their high data appetite. In addition, small samples can also fail to capture the different data aspects required for finer predictions. This has also been considered in broader issues on clinical ML, as very small sample sizes remain a barrier against implementing robust prediction models.

The third limitation relates to the feature scope. Only 8 clinical features are captured in this dataset, while other important diabetes risk factors are missing, such as lifestyle-physical activity and dietary habits, genetic markers-variation in the TCF7L2 gene, and socioeconomic factors-income and health care access. The absence of the above features will lower the predictive power and clinical relevance of our model, as their inclusion will help improve model performance and further build an understanding of diabetes risk. In their work, Zhang et al. show that incorporating lifestyle and genetic information increases the accuracy of diabetes prediction models by 10%, underlining the importance of expanding the feature set.

5. Conclusion

This study demonstrates that, in the prediction of diabetes using the Pima Indians Diabetes Database, Logistic Regression outperforms both k-NN and SVM in most of the other main metrics and is therefore excellent. The model was simple to apply in clinical practice since it provided high accuracy, 78%, a recall of 0.82, which does not miss diagnoses, and easily interpretable risk factors. The glucose level, BMI, and age are pertinent and linked to common clinical knowledge; therefore, the model predictions do have some biological and clinical sense.

If anything, this simple but salient lesson from this study is that simple linear models may do quite well when relationships between features and outcomes are linear-which happens quite often in clinical data. The results challenge the general impression that complex, non-linear models are always

superior. The study thus supports data-driven, clinically informed model selection for matters of interpretability, performance, and practical applicability, distinct from issues of model complexity per se. Recent thinking within the discipline of clinical ML has also moved in this direction: where two or more models perform comparably, simplicity and interpretability should have precedence over complexity.

These logistic regression models and other machine learning models may one day hold great promise as clinical decision-support tools—with further development, training on larger and more diverse datasets, incorporation of more features, and rigorous clinical validation. Indeed, by allowing earlier prediction of diabetes risk, those models would empower healthcare providers to deliver timely and targeted interventions to enhance patient outcomes and reduce diabetes's global burden. Future research should continue to build on these limitations by considering larger, more heterogeneous samples, expanding the development of lifestyle and genetic features, and conducting prospective clinical trials to confirm such models' real-world use.

References

[1] Kaur K, Kaur J, Singh J. Diabetes prediction using machine learning algorithms: A review. Journal of King Saud University - Computer and Information Sciences, 2021, 33 (10): 1132-1143.

[2] Al-Shargie A, Alhajj R, Al-Dossari M. Diabetes prediction using machine learning algorithms: A comparative study. Journal of Medical Systems, 2022, 46 (3): 21.

[3] Smith A K, Johnson L M, Williams C D. Limitations of the Pima Indians Diabetes Dataset for generalizable diabetes prediction. Journal of Biomedical Informatics and Computational Biology, 2020, 8 (2): 45-58.

[4] Moons K G, Royston P, Vergouwe Y, Altman D G. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart, 2015, 101 (17): 1387-1394.

[5] Gujarati D N, Porter D C. Basic Econometrics (7th ed.). McGraw-Hill Education, 2019.

[6] Zhang Y, Li J, Wang H. The impact of feature engineering on diabetes prediction: A comparative analysis. Computational Biology and Chemistry, 2023, 107: 107892.

[7] Hosmer D W, Lemeshow S, Sturdivant R X. Applied Logistic Regression (3rd ed.). John Wiley & Sons, 2013.

[8] Hand D J, Anagnostopoulos C. ROC curves for clinical prediction models: Problems and alternatives. Statistical Methods in Medical Research, 2019, 28 (11): 3375-3389.

[9] Brown M A, Davis S E, Miller T R. Performance of k-nearest neighbors in small clinical datasets: A simulation study. Journal of Medical Systems, 2021, 45 (8): 68.

[10] Collins G S, Reitsma J B, Altman D G, Moons K G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Annals of Internal Medicine, 2022, 176 (5): 743-748.