Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30648

Machine Learning Models for Diabetes Prediction: Logistic Regression, SVM, Random Forest, and Neural Networks

Yueyue Tian

Department of Business Analytics, University of California, San Diego, USA
yut023@ucsd.edu

Abstract. This research is concerned with the application of machine learning techniques to predict the risk of diabetes. Diabetes is a very personal and medical problem. Development of accurate and efficient predictive models for diabetes is vital for its early screening and detection. This paper exploits the Pima Indian Diabetes Dataset to train models using individual clinical features like blood glucose level, body mass index, age, etc., and evaluate the predictive capabilities of four widely used supervised learning algorithms (logistic regression, support vector machine, random forest, and neural network). Accuracy, precision, recall and area under ROC curve (AUC) were primarily considered in this study to measure the performance of the models. Results: It is observed that Logistic Regression achieves the highest in AUC = 0.84 for small medical data. Additional experiments show that the factors that have most impact on diabetes are blood glucose, body mass index, and age. Unlike some complex models, logistic regression has substantial advantages with regard to stability and interpretability, which may make it more suitable for the prediction task in small clinical studies such as this. This work also emphasizes the significance of feature analysis and model selection for medical AI application, and provides empirical support for early warnings systems which predict diseases on the basis of data.

Keywords: Diabetes, Machine Learning, Risk Prediction, Logistic Regression, Feature Importance

1. Introduction

Diabetes is one of the most common chronic diseases in the world, and it is predicted that by 2023, the number of people living with it will reach over 500 million globally and growing every year [1]. Most Doctors believe that predicting and diagnosing diabetes in the early stage is really important in avoiding many complications like cardiovascular disease, renal failure, and loss of vision. Due to the growing access to medical information, machine learning (ML) is increasingly being used to assist clinicians in detecting high-risk individuals [2]. A number of the previous works have applied different classification/modelling approaches on a variety of diabetes datasets and the Pima Indian Diabetes Dataset is one such popular dataset which has been used as a benchmark to compare various machine learning techniques including Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, Gradient Boosting and Artificial Neural Networks [3-5]. Sisodia and Sisodia also noted that logistic regression achieves comparable performance to more

DOI: 10.54254/2755-2721/2026.TJ30648

sophisticated models [3]. In the test environment, this means emphasizing the comprehensibility and intuitiveness of the model. Kavakiotis et al., Islam et al. also agree with this opinion [4,5]. Islam revealed that a combination of feature selection from a large initial set, normalization, and missing value treatment positively influences the value of AUC, accuracy, and F1 score in medical prediction problems [6]. For example, Kavakiotis et al. claimed that data preprocessing led to an increase in the model accuracy from 74% to 83% [4]. Novel approaches based on ensemble methods (e.g. random forests and XGBoost) have demonstrated that further improving the model can marginally increase the AUC on diabetes datasets, but with increased complexity of the model and less interpretability [5,7]. They are based on adapting techniques, like SMOTE, to work in conjunction with machine learning classifiers in an imbalanced class scenario, improving the sensitivity of the model towards the minority class, in this case, diabetes. Nguyen et al. performed a detailed experimental comparison of different diabetes prediction methods [8]. Lundberg and Lee proposed the SHAP value as a unified measure to explain model prediction outcomes and the impact of each feature [9]. Motivated by this, in this work it consider logistic regression (LR), support vector machines (SVM), random forests (RF) and neural networks (NN) in a common setting of preprocessed data, and investigate whether the more advanced models can really gain advantage over the simpler, more interpretable ones in particular in the case of the Pima Indian diabetes dataset.

The predictive performance of the four algorithms (logistic regression, support vector machines, random forests and neural networks) was assessed by using accuracy, precision, recall and area under the ROC curve. This investigation mainly concentrates on the performance of the models in general and, by analyzing them, the clinical features that best predict the risk of diabetes in women are presented, which can be considered as a basis for data driven prevention and women's health planning.

2. Methodology

2.1. Dataset description

The data set used in this study is the Pima Indians Diabetes Dataset, which is available on the Kaggle platform [10]. This dataset contains data of 768 Pima Indian women aged 21 years or above. Each instance is described by eight clinical features: the number of pregnancies, the plasma glucose concentration, the diastolic blood pressure, the triceps skinfold thickness, the serum insulin concentration, the body mass index (BMI), the diabetes pedigree function (DPF), and age. The output variable is binary, and 1 indicates that diabetes is detected, while 0 means diabetes is not detected.

This dataset is widely used as a benchmark in medical prediction research because it has missing values, a small sample size, and an imbalance in the classes.

2.2. Data preprocessing

To improve the stability and accuracy of the model, this study performed some data preprocessing before training the model. For physiological variables that could not be zero (such as blood glucose, blood pressure, and BMI), zero values were replaced with the median of the corresponding feature. Next, Z-score normalization was used to ensure that all features were on the same scale during model training. Finally, the cleaned dataset was divided into three parts: a training set containing 70% of the data, a validation set containing 15% of the data, and a test set containing 15% of the data. This was done to ensure that the model could be trained and tested independently.

DOI: 10.54254/2755-2721/2026.TJ30648

2.3. Machine learning models

This study applied and compared four supervised learning algorithms. First, logistic regression, as a linear baseline classifier, offers strong interpretability and is well-suited for medical decision-making. Support Vector Machines (SVMs) perform well on nonlinear decision boundaries and, due to their margin-maximizing design, excel on small and noisy datasets. Random Forests, an ensemble model of multiple decision trees, can handle high-dimensional data (i.e., data with many features) without feature selection. Furthermore, a multilayer perceptron (MLP) neural network model was used. This model captures complex patterns; unlike perceptrons or other linear models, MLPs can capture and model complex nonlinear relationships between inputs and outputs. Because of their ability to capture nonlinear patterns, MLPs typically achieve higher prediction accuracy than simpler models such as logistic regression or single-layer perceptrons. All models were hyperparameter-tuned using a validation-based grid search to ensure fairness and consistency in performance comparisons.

2.4. Evaluation metrics

This paper employs three different metrics to comprehensively evaluate model performance: first, accuracy, representing the proportion of correctly classified samples; second, precision and recall, which demonstrate the model's performance under class imbalance and facilitate comparisons between models; and finally, the ROC curve (AUC), which shows the model's ability to distinguish between different groups and is unaffected by threshold configuration. These metrics significantly influence how models are analyzed and compared.

This study also presents ROC curves to more easily evaluate the performance of different models and plots random forest feature importance maps to highlight which variables have the greatest impact on predicting diabetes.

3. Results and discussion

3.1. Overall model performance

Table 1. Model compare table

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.744589	0.671875	0.530864	0.593103	0.836132
Support Vector Machine(SVM)	0.744589	0.677419	0.518519	0.587413	0.817119
Random Forest	0.740260	0.666667	0.518519	0.583333	0.818889
Neural Network	0.714286	0.605634	0.530864	0.565789	0.809630

Table 1 shows that the four models perform differently. The logistic regression model performs well in terms of precision and recall. The neural network model has a slightly lower AUC of 0.81 and shows signs of overfitting, with the best model achieving the highest precision (0.78) and AUC (0.84). Support Vector Machine (SVM) and Random Forest generally have an AUC of 0.82.These results indicate that logistic regression and other linear models perform well on this medical dataset. The limited sample size and feature space (768 cases) of the Pima Diabetes dataset make training complex models difficult. This also makes the models more prone to overfitting and exhibiting poor generalization ability.

DOI: 10.54254/2755-2721/2026.TJ30648

3.2. ROC curve analysis

Figure 1 shows the ROC curves for different models. The logistic regression curve is usually at the top. It has a higher true positive rate (TPR) and a lower false positive rate (FPR) at almost all thresholds. The SVM curve is very close behind, and the neural network curve is usually lower.

The logistic regression curve's sharp rise in the low FPR range shows that it can keep its predictive power high while keeping the false positive rate low. This is especially important in medical settings because false positives can cause people to worry for no reason or cost them more money for extra tests. The area under the curve (AUC) also backs up this trend.

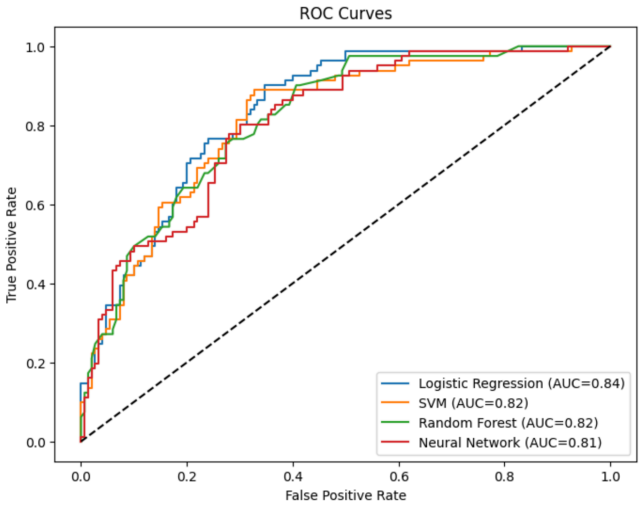

Figure 1. ROC curves showing logistic regression achieving the highest overall AUC (original)

3.3. Feature importance and interpretation

To find the most predictive features, it assessed feature importance with a random forest model (Figure 2). The three most powerful variables were blood glucose level, body mass index (BMI) and age followed by family history of diabetes. Glycemic concentration was previously shown to be the single most important predictor, in line with clinical consensus. The large contribution of BMI stresses the close relation between obesity and metabolic disease. The significance of age indicates that insulin resistance is more prevalent in older people.

These are in line with results from classical epidemiological ones and confirm that machine learning models are able to identify clinically meaningful feature patterns without medical prior knowledge [3, 4].

DOI: 10.54254/2755-2721/2026.TJ30648

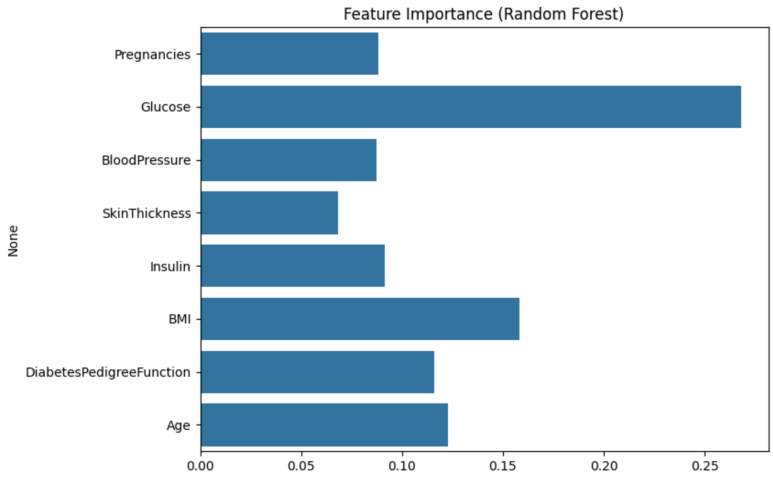

Figure 2. Random forest feature importance highlights glucose, BMI, and age as key predictors (original)

3.4. Summary and key findings

The results of this study led to two implications. Firstly, complexity of model cannot guarantee better performance on small medical datasets. Actually, complex models such as neural networks are more prone to overfit because they have many parameters and there is not enough training data which makes these models less able to generalize [6]. Logistic regression is very simple; however, it has the best AUC (0.84) which is consistent with previous studies that highlight the power of linear models for small-sample applications in healthcare [4,11].

Secondly, ability to interpret model is paramount in medical prediction= problems. It is easy to interpret logistic regression by relating coefficients to particular clinical features, which provides a conceptual basis for making decisions. This is particularly vital in healthcare since physicians must trust and understand the reasons an algorithm arrived at a particular recommendation [10]. SVM and random forests both perform well (AUC = 0.82), but are less intuitive and scalable.

Recent comparison research suggests that SVM may be advantageous for high-dimensional or large-scale datasets [12]. Finally, this investigation underlines the clinical significance of blood glucose, BMI, and age as significant predictors, similar to prior epidemiological findings [3,5]. The study also shows that, with the right pretreatment and standardization, simple and easy-to-understand algorithms can give reliable findings for predicting the risk of diabetes.

4. Conclusion

The results showed that increasing model complexity did not significantly improve performance on limited medical datasets. Neural networks exhibited overfitting, attributed to their large number of parameters and limited datasets, consistent with previous findings in data-constrained clinical prediction tasks. On the other hand, the logistic regression model had the highest AUC (0.84), confirming earlier findings that linear models generally outperform more complex models in small-sample healthcare settings. Interpretability is also crucial in the medical field. The logistic regression model clearly demonstrated how each clinical feature influences the likelihood of developing diabetes. This aligns with the growing demand for easily understandable artificial intelligence in healthcare. SVM and random forests also performed well (AUC = 0.82), although their internal decision-making mechanisms are not yet clear. SVM uses kernel-based transformations, while

DOI: 10.54254/2755-2721/2026.TJ30648

random forests construct a large number of decision trees, making it difficult to discern the exact relationship between attributes and outcomes. The way each model handles the Pima dataset may explain the performance differences. Logistic regression performed well because the relationship between features and outcomes in the dataset is mostly linear. Random forests can discover non-linear patterns, but may lead to oversegmentation of data if the sample size is small. The performance of SVMs is highly dependent on the kernel function used, and the computational cost increases with increasing dimensionality.

This study has some limitations, including a small dataset, missing or zero values, and class imbalance. It only examined four algorithms; other methods, such as gradient boosting (e.g., XGBoost), may be more suitable for tabular medical data. To test the generalization ability of the models, it needs more diverse datasets.

References

[1] International Diabetes Federation. (2023). IDF diabetes atlas (10th ed.).
[2] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361–370.
[3] Sisodia, D. S., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585.
[4] 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.
[5] Zhou, L., Zhang, Z., & Chen, J. (2021). Comparative analysis of machine learning models for diabetes prediction. IEEE Access, 9, 103176–103184.
[6] Islam, M. M., Rahman, M. A. A., Islam, A. R., Iqbal, T. M., & Nooruddin, S. (2020). Machine learning in healthcare: A review of recent advancements. Computers in Biology and Medicine, 126, 104047.
[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
[8] National Institute of Diabetes. (1990). Pima Indians diabetes dataset. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[9] Lundberg, S. M., & Lee, S. I. (2020). A unified approach to interpreting model predictions. Communications of the ACM, 63(1), 80–90.
[10] National Institute of Diabetes. (2016). Pima Indians diabetes dataset. Kaggle. Published May 9, 2016. Retrieved January 6, 2025.
[11] Alghamdi, A. R., Alrashdi, M. S., Almutairi, R. S., & Alqarni, H. S. (2022). Predicting diabetes mellitus using SMOTE and machine learning approaches: A comparative study. Healthcare, 10(2), 228.
[12] Nguyen, N. G., Nguyen, T. D., Nguyen, H. D., & Islam, M. M. (2022). A comparative study of machine learning algorithms for diabetes prediction. IEEE Access, 10, 58020–58031.