Comparisons of Machine Learning Models for Prediction of Susceptibility to Diabetes

Yutian Jiao

Department of Mathematical Sciences, Carnegie Mellon University, 15213, Pittsburgh, United States

Abstract. Diabetes is a chronic disorder causing millions of people to suffer from severe complications such as heart attacks, kidney failures, and permanent vision loss. This study aims to find an optimal choice among the five selected models that perform the best on diabetes prediction, and thus provide valuable insights in early detection of diabetes. This study compares the predictive performance of machine learning models such as Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). The study preprocessed the Pima Indians Diabetes (PID) dataset, and the models were trained on it before being assessed using four assessment criteria. According to the results, LR had the best accuracy of 0.76, with RF and SVM coming in second and third, respectively. Results showed that LR achieved the highest accuracy of 0.76, closely followed by RF and SVM. While SVM has the highest precision, it performs poorly on recall, limiting its overall performance on diabetes prediction. On the contrary, LR and RF achieved good results in the F-score, making them outperform the other models in terms of overall performance score in predicting diabetes.

1 Introduction

Diabetes is a chronic metabolic disease that causes high glucose levels in the blood (hyperglycemia) either in a fasting state or after meals. Complications from diabetes might include heart attacks, kidney failure, and blindness. The International Diabetes Federation (IDF) estimates that 537 million people aged 20 to 79 have diabetes, and that number is projected to increase to 643 million by 2030 and 783 million by 2045 [1]. According to the World Health Organization (WHO), the prevalence of diabetes is still rising, and 2 million people died from diabetes-related kidney disease in 2019 [2]. Hence, early detection of diabetes is a crucial area of research. Predictive machine learning models can evaluate one’s risk of suffering from diabetes and thus enable early interventions, such as changing lifestyle and converting to a healthier diet.

One important technique for enhancing diabetes prediction is machine learning. Using the Pima Indians Diabetes (PID) dataset, Jobeda Jamal Khanam and Simon Y. Foo compared seven machine learning techniques for diabetes prediction, such as DT, SVM, and LR. In their study, LR and SVM both perform well in diabetes prediction. They also built neural network (NN) models with hidden layers and found out that NN models with two hidden

Corresponding author: yutianj@andrew.cmu.edu

layers achieved a high accuracy of 88.6%. In addition, Aishwarya.M and Dr. Vaidehi V. proposed a pipeline machine-learning model for diabetes prediction. Their model outperforms other machine learning models like LR, KNN, and Perceptron; the AdaBoost Classifier, for example, performs better when it comes to diabetes prediction, with an accuracy of 98.8%.

In another study, Dagliati used electronic health records (EHR) data and machine learning methods to predict the risk of getting diabetes [3]. The researcher applied a data mining pipeline to train models with the target of nephropathy, neuropathy and retinopathy over time horizons of 3, 5, and 7 years. The study finally achieved an 83.8% accuracy of prediction. The results showed that logistic regression is the most interpretable model due to its ability of handling missing data. Additionally, Hang Lai developed Gradient Boosting Machine (GBM) and other machine learning techniques in his study and at last achieved an area of 84.7% under the receiver operating characteristic curve (ROC) [4].

This study's objective is to assess and contrast how well machine learning algorithms predict a person's risk of developing diabetes, including Random Forest (RF), KNN, Logistic Regression (LR), SVM, and NN. The PID dataset is used as both the training and test sets, containing biological records of 8 attributes such as glucose level, BMI, and skin thickness. This study detects and handles the missing values in the dataset and uses min-max normalization to scale the data. After feature selection based on the correlation matrix, the study splits the data 85-15 into a train set and a test set and trains the models based on these sets. The models are evaluated and compared based on metrics including accuracy, precision, recall, and f-score.

2 Data & Methodologies

2.1 Dataset Selection

The study uses the PID Database, which originates from the National Institute of Diabetes and Digestive and Kidney Diseases [5]. The database contains 768 records with 9 attributes. Eight of these attributes are key features for diabetes prediction: number of pregnancies, glucose, blood pressure (BP), and so on. The outcomes are labeled binarily, where 1 represents the subject diagnosed with diabetes and 0 otherwise. Descriptions and types of the attributes are shown in Table 1.

Table 1. Description of PID Dataset

Attributes	Description	Type
Pregnancies (Preg)	Number of times pregnant	Numeric
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test	Numeric
Blood Pressure (BP)	Diastolic blood pressure (mm Hg)	Numeric
Skin Thickness (ST)	Triceps skin fold thickness (mm)	Numeric
Insulin	2-Hour serum insulin (muU/ml)	Numeric
BMI	Body mass index (weight in kg/(height in m)^2)	Numeric
Diabetes Pedigree Function (DPF)	Diabetes pedigree function	Numeric

Age	Age (years)	Numeric
Outcome	Class variable (0 or 1), 268 of 768 are 1, the others are 0	Binary

2.2 Data Preprocessing

2.2.1 Missing Values

The dataset identifies missing values in the following columns: BMI, insulin, skin thickness, blood pressure, and glucose. A value of 0—which is biologically impossible—represents the missing data. Therefore, median imputation is used in this work to manage the missing values [6]. A technique called median imputation substitutes the median value of all the attribute's non-missing values for the missing ones. For example, since the mean value of glucose level before imputation is 120.89, all the missing values in the Glucose column will be replaced with 120.89. Table 2 reveals the number of missing values and compares the mean values before and after imputation.

Table 2. Number of missing values and mean values before and after imputation

Attributes	**Number of missing values**	**Mean before imputation**	**Mean after imputation**
Preg	0	3.85	3.84
Glucose	5	120.89	121.65
BP	35	69.11	72.39
ST	227	20.54	29.11
Insulin	374	79.80	140.67
BMI	11	31.99	32.46
DPF	0	0.47	0.47
Age	0	33.24	33.24
Outcome	0	-	-

2.2.2 Normalization

After handling missing values, this study first detects outliers and then scales the dataset by normalization. Normalization is crucial in this case because the mean values of the attributes vary significantly (140.67 for Insulin compared to 0.47 for DPF), and thus the unscaled data will cause the models to disproportionately weight attributes. Min-max normalization is used in the scaling process, and Table 3 shows the mean values of the attributes before and after normalization [7].

Table 3. Mean values before and after normalization

Attributes	Mean before normalization	Mean after normalization
Preg	3.84	0.22
Glucose	121.65	0.50
BP	72.39	0.49
ST	29.11	0.24
Insulin	140.67	0.15

BMI	32.46	0.29
DPF	0.47	0.17
Age	33.24	0.20
Outcome	-	-

2.2.3 Feature Selection & Data Splitting

To enhance the efficiency of models, this study selects features based on the correlation matrix (Fig. 1) [8]. The coefficients range from –1 to 1, where 1 and –1 respectively indicate strong positive and negative correlation, while 0 indicates no correlation. A threshold of 0.2 is set, so the attributes that have a correlation that is less than 0.2 with outcomes will no longer be considered in this study. The figure reveals the correlation values: pregnancies (0.22), glucose (0.49), BP (0.17), skin thickness (0.21), insulin (0.2), BMI (0.31), DPF (0.17), and age (0.24). Thus, this study will build models based only on 6 attributes: pregnancies, glucose, skin thickness, insulin, BMI, and age.

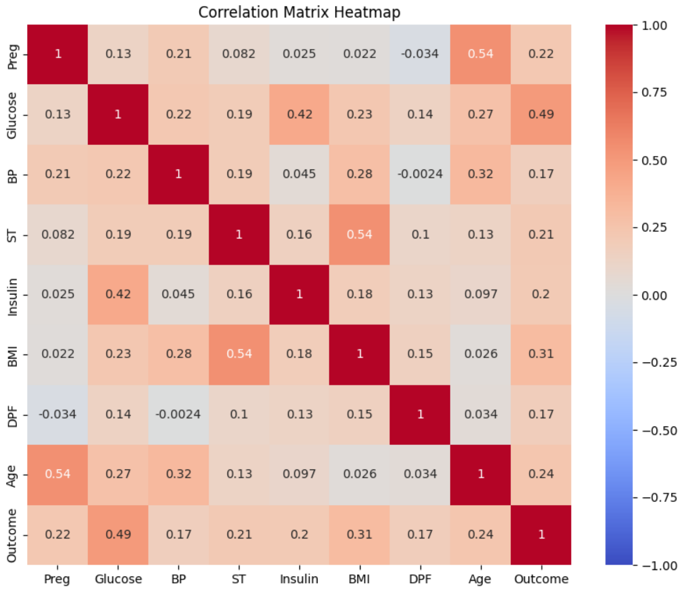

Fig. 1. Correlation Matrix Heatmap (Photo/Picture credit: Original)

After feature selection, the study splits the dataset using the 85-15 method, where a random 85% of the dataset becomes the training dataset and the remaining 15% becomes the test set [7,8]. By training on 85% of the dataset, the models are ensured to learn based on enough data, and the remaining 15% also allows a solid evaluation of the models' performance. This study evaluates the models by a 10-fold cross-validation method. The dataset will be divided into 10 folds with 9 folds used for model training and the remaining used for evaluation. After repeating the process 10 times, the evaluation finally takes the average to give a more thorough assessment of the model’s generalization ability by reducing data variability [9,10].

2.3 Evaluation Metrics

This study evaluates the models by 5 key metrics: Accuracy, Precision, Recall, F score, and Confusion Matrix. This section explains how these metrics work.

The confusion matrix for a binary model divides the predictions into 4 parts: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Table 4 shows how the matrix divides the predictions. All the 4 other metrics are calculated based on the matrix.

Table 4. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Accuracy is determined by the sum of TP and TN divided by the total number of predictions. A higher value on the accuracy scale, which goes from 0 to 1, denotes better performance.

$$\text{accuracy} = \frac{\text{TP}}{\text{TP+FP+TN+FN}} \tag{1}$$

The ratio of TP to the total number of positives is known as precision. The precision scale goes from 0 to 1, with 1 being the highest level and 0 being the lowest.

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

Recall is the ratio of TP to the actual number of positives. Better performance is indicated by a greater recall value, which runs from 0 to 1.

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

Precision and recall are used to generate the F-Score, often known as the F measure. F-score = 0 if either precision or recall is zero, and F-score = 1 if both are one.

$$\text{F} - \text{score} = \frac{2\times\text{recall}\times\text{precision}}{\text{precision} + \text{recall}} \tag{4}$$

2.4 Model Training

The study creates an algorithm to train the intended machine learning models and assesses them using the evaluation metrics after the data has been preprocessed. The algorithm stores all the target models in a list and loops through the list by calling the train and predicting functions on each model. Then, the algorithm calls evaluation metrics on each model and eventually prints them out.

3 Result & Discussion

3.1 Result Analysis

This section discusses the results after implementing the algorithms and compares the performance of the models. A bar chart (Fig. 2) is generated to visualize the comparison of these models.

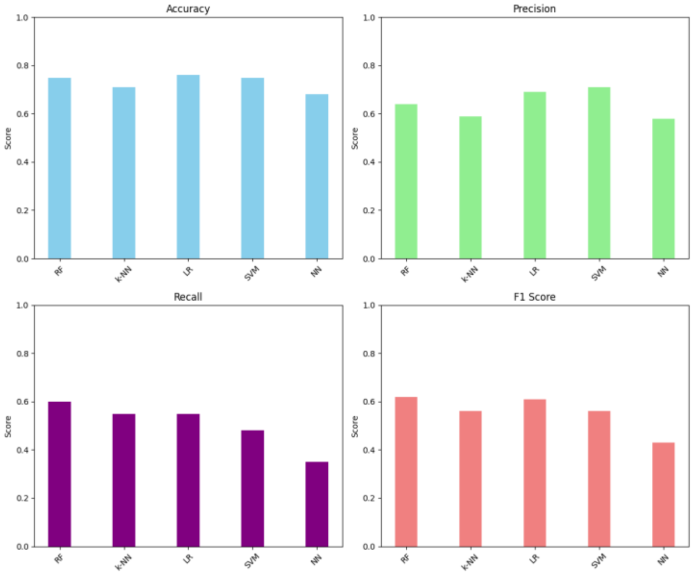

Fig. 2. Result Comparisons (Photo/Picture credit: Original)

Logistic Regression achieves the highest accuracy of 0.76, which indicates that LR performs the best in classifying both positives and negatives. Also, the accuracy of SVM and RF closely follows LR with a difference of only 0.01. For precision, SVM performs the best, and LR and RF also perform well. However, SVM performs poorly on recall at only 0.35, whereas RF keeps outperforming the other models on recall value. Finally, RF and LR achieved the best two F-scores. From these results, logistic regression (LR) and random forest (RF) reveal their outstanding capacities for diabetes prediction. On the contrary, k-NN and NN models perform the worst among the 5 models. The SVM model performs well in accuracy, precision, and F-Score, but the recall of the SVM model being too low makes it not a good choice for diabetes prediction. One possible reason why LR and RF outperformed the other models may be that they are better at analyzing unbalanced class distributions.

3.2 Future Improvements

Of course, this study has several limitations. First, the current dataset only covers records for Indian women subjects with only 6 key attributes taken into consideration. Some other factors causing or correlated to diabetes, such as diet, income, and environmental factors can also be explored. Besides, the susceptibility to diabetes may also vary according to different genders and ethnicities. Therefore, adding records from a larger variety of populations to the dataset will increase the results' generalizability. Furthermore, only a small number of machine learning models are currently evaluated in the study. In addition, the study can evaluate the effectiveness of additional machine learning models, such as AdaBoost, Perceptron, and Gradient Boost Classifier.

4 Conclusion

This study uses the PID dataset to examine how well five machine learning algorithms predict diabetes. The data-preprocessing procedure uses techniques like correlation matrix for feature selection, min-max normalization for data scaling, and median imputation for handling missing values. After training and validating the data using 10-fold cross-validation, the study compares and assesses the models' performance using four important metrics: accuracy, precision, recall, and f-score. The results indicate that LR and FR achieved an overall optimal performance among all five chosen models. LR and RF performed relatively well, reaching 0.76 and 0.75 accuracy, 0.69 and 0.64 precision, 0.55 and 0.6 recall, and 0.61 and 0.62 F-score, respectively. It is worth noting that SVM also performed well in terms of accuracy (0.75) and precision (0.71), but its recall was only 0.48, showing some shortcomings. In addition, KNN and NN were less effective in diabetes prediction.

The study demonstrates the capacity of machine learning models for aiding the early detection of diabetes, which enabled early intervention and thus effectively restrained the growth of several patients suffering from diabetes. The results of the study align with some key findings in the previous studies. For example, the accuracy that LR achieves in this study is comparable to the findings in Lai's stud and align with the conclusion from Dagliati's study that LR outperforms the other machine learning models.

Furthermore, improvements can be made: the study can continue to estimate other machine learning models such as Gradient Boost Classifier, Perceptron, and AdaBoost; the dataset can also be expanded to include more diverse populations in addition to Indians, which may allow a more generalizable prediction with races taken into considerations.

References

1. World Health Organization (WHO), Diabetes, April 5 (2024), https://www.who.int/news-room/fact-sheets/detail/diabetes.
2. International Diabetes Federation (IDF), Facts & Figures (2021) https://idf.org/about-diabetes/diabetes-facts-figures.
3. A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, P. De Cata, L. Chiovato, and R. Bellazzi, Machine learning methods to predict diabetes complications, Journal of Diabetes Science and Technology. **12**, 2, 295–302 (2017)
4. H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, Predictive models for diabetes mellitus using machine learning techniques, BMC Endocrine Disorders. **19**, 1, (2019)
5. UCI Machine Learning Kaggle Team, Pima Indian Diabetes Database (2016) https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.
6. J. Sessa and D. Syed, Techniques to deal with missing data, in 2016 5th International Conference on Electronic Devices, Systems, and Applications (ICEDSA), Ras Al Khaimah, United Arab Emirates. 1-4 (2016)
7. S. Gopal Krishna Patro and K. Kumar Sahu, Normalization: A Preprocessing Stage. (2015)
8. Z. Vujovic, Classification Model Evaluation Metrics, International Journal of Advanced Computer Science and Applications.**12**,599-606 (2021)
9. J. Jamal Khanam and S. Y. Foo, A comparison of machine learning algorithms for diabetes prediction, ICT Express. **7**, 4, 432-439 (2021)
10. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, Machine learning and data mining methods in diabetes research, Computational and Structural Biotechnology Journal. **15**,104–116 (2017)