Diabetes Prediction by KNN, SVM, Random Forest and XGBoost

Shuqi Liu*

School of Computer Science, South-Central University for Nationalities, Wuhan, China

*Corresponding author: 2016120755@jou.edu.cn

Abstract. Diabetes is a chronic condition that is incurable, and it may even cause some serious complications, so early prevention is crucial. In order to benefit relevant researchers, this paper collects the Pima Indians dataset from the kaggle website. The dataset contains data of 768 patients, in the first eight columns of the dataset are the data of their body indicators and the last column represents the result of whether they have diabetes or not. In this paper, four ML algorithms KNN, SVM, RF and XGB are used to predict diabetes, and hyperparameter optimization is done using Gridsearch, Cross validation to get the best result of each model. After optimizing these four models and training them on the training set, the following results were obtained on the test set: the KNN algorithm got the highest precision score and accuracy score, while the XGBoost algorithm got the highest recall score and f1 score, and the remaining two algorithms got slightly lower results.

Keywords: Diabetes; machine learning; KNN; XGBoost.

1. Introduction

Diabetes is a chronic disease that is difficult to cure. There are two causes of the disease. Firstly, the destruction of insulin-producing beta cells is caused by the immune system, causing the pancreas to generate inadequate insulin levels, resulting in the development of type 1 diabetes. And secondly, the body has difficulty using the insulin produced by the pancreas properly, known as insulin resistance. This causes type 2 diabetes. If a person's body remains in a state of high blood glucose for an extended period and is not treated, it can be very harmful to the body and even lead to significant harm to several bodily systems, notably the nerves and blood vessels, over time [1].

In 2014, eight or nine out of every hundred adults worldwide were likely to have diabetes. By 2019, the situation will be even more critical: 1.5 million people will lose their lives to diabetes, and nearly half of them will be under the age of seventy [1]. Another 460,000 deaths from kidney disease are attributed to diabetes, and raised blood glucose is a factor in approximately 20% of deaths from cardiovascular disease [2]. Between 2000 and 2019, deaths related to diabetes grew by 3%, and in low- and middle-income countries by 13%.

Machine learning is widely implemented in the healthcare sector due to its significant role. Utilizing machine learning algorithms can aid in detecting diabetes [1]. Sunil Ghane et al [3] have suggested that Light Gradient Boosting Machine is the most effective technique for predicting diabetes. The PIMA India dataset was employed for this purpose. Comparative evaluations were conducted against other methods such as DT, SVM, KNN, RF, and Adaboost algorithms to assess performance. To forecast diabetes, Sadhana Tiwari and other two researchers suggested the stacking and stacking with hyperparameter tweaking approach [4].However, further investigations are still necessary as relevant research in this field is limited.

In order to improve ML in the field of diabetes research, this study investigates four ML algorithms with and without hyperparameter adjustment.

2. Evaluation metrics

Confusion matrix, commonly, the matrix is shown as Table 1.

Table 1. Confusion matrix

	0	1
0	TN	FN
1	FP	TP

In this context, TP stands for True Positive, while TN is the opposite in this aspect of Masculine Negative. FP is the opposite in this aspect of True False, and FN differs greatly in both aspects. It is crucial to make sure that these technical phrases are used correctly throughout the whole content. It is categorized as a true positive if the sample label in the dataset is positive and the classifier also forecasts a positive label. A real negative is one in which both the sample label and the prediction are negative. A false negative arises when the sample label is positive, but the classifier incorrectly predicts a negative outcome. In other words, the classifier fails to identify actual positives. A false positive arises when the sample label is negative, but the classifier foresees a positive outcome. Put simply, the classifier makes incorrect forecasts of positive results.

Precision score: The percentage of real positive cases in the positive sample is the forecast outcome. The equation reads as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

Recall score: The ratio of actual positive samples to those that were anticipated to be positive. The equation is displayed below:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

F1 score: The harmonic mean of the recall rate and accuracy value is represented by the F1 score. Listed below is the formula:

$$\text{F1 score} = \frac{2*Precision*Recall}{Precision+Recall} \tag{3}$$

Accuracy: The proportion of samples for which the classifier makes accurate predictions to all samples. Below is the calculation for this:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

3. Methods and Materials

3.1. Dataset

This study makes use of the PIMA criteria Indians data set [5-7]. The dataset comprises solely of female patients aged 21 and above, being a resident of Phoenix, Arizona, the United States. Furthermore, the dataset includes medical predictor factors. Similarly, it addresses the single target variable "Outcome". Pregnancy can have a significant impact on various physiological parameters such as glucose levels, skin thickness, and insulin sensitivity. In addition, BMI may also influence these parameters during pregnancy. It is important to monitor these changes closely to identify and manage any potential complications that may arise.

In conclusion, the dataset consists of eight predictors, and an outcome parameter that distinguishes a total of 768 observations, 268 of which represent positive cases of diabetes and 500 of which represent negative instances. Class 1 represents individuals who have diabetes, and class 0 represents those who do not. These two classes make up the outcome variable.

3.2. Data Preprocessing

Data preprocessing is essential for developing high accuracy machine learning models. Outliers are removed, missing values are replaced, data are normalised, and features are selected to improve data quality [7].

3.2.1 0 Value

By using the describe() method and transposing it, minimum values for pregnancy, glucose, bloodpressure, skinfold, insulin, BMI and outcome can be found at 0. This is shown in the figure below. In this, for the attribute of Pregnancies, the value of 0 is also interesting to study. Meanwhile, for Outcome, the output value, originally there are only two possibilities, 0 and 1, so it can also be 0. However, for the remaining five attributes, if there is a 0 in them means that they are meaningless data, so they have to be processed.

First, use the replace() method to replace the 0 value with a NaN value, which makes it easy to count the 0 values. Then, the selection of the value to replace the 0 value is made based on the distribution graph of each variable. Some columns exhibit a skewed distribution, causing the mean to be impacted by outliers more than the median. Glucose and blood pressure follow a normal distribution, meaning the mean replaces 0 values in these columns. For SkinThickness, Insulin, and BMI, a median value is preferable, as these columns have skewed distributions and are less influenced by outliers. Eventually the 0 value is eliminated (Details are shown in Table 2).

Table 2. Characteristics of the dataset

	COUNT	MEAN	STANDARD DEVIATION	MINIMUM	1/4	1/2	3/4	MAXIMUM
P	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
G	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
B	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
S	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
I	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
D	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
A	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
O	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

3.2.2 Outliers Removal and Standardization

In this study, the quantile transformation method was used to reduce the effect of outliers on the prediction results and it also normalises the dataset. The Quantile Transformer method converts features to adhere to a uniform or normal distribution, resulting in the spreading out of the most common values for a given feature. It also lessens the influence of (marginal) outliers, making it a durable preprocessing approach [6].

3.2.3 Feature Selection

The correlation coefficients between the attributes were calculated and a heat map was drawn. The three attributes BloodPressure, Insulin, DiabetesPedigreeFunction had the smallest correlation coefficients with the final output. These three attributes were directly deleted and the remaining five attributes were retained for predictive analysis.

3.3. Algorithms Used

3.3.1 KNN

A similarity metric is used by KNN classification to determine the class of input data. This is accomplished by locating similar samples from the training data [8-11]. The class of the neighboring samples affects the anticipated class of the incoming data. Depending on the nature of input, different distance methods are employed to determine the neighbors of input data [10].

In this study, the first algorithm that was used was the KNN algorithm. Hyperparameter optimization is performed for this algorithm, i.e., it is time to determine k value. Combining the grid search and cross-validation methods, the final optimal k value is 27. The optimal k value is verified by iterating through dozens of sets of hyper-parameter values and plotting a line graph of the F1 scores as a function of the k value. The line graph is shown below in Fig. 1.

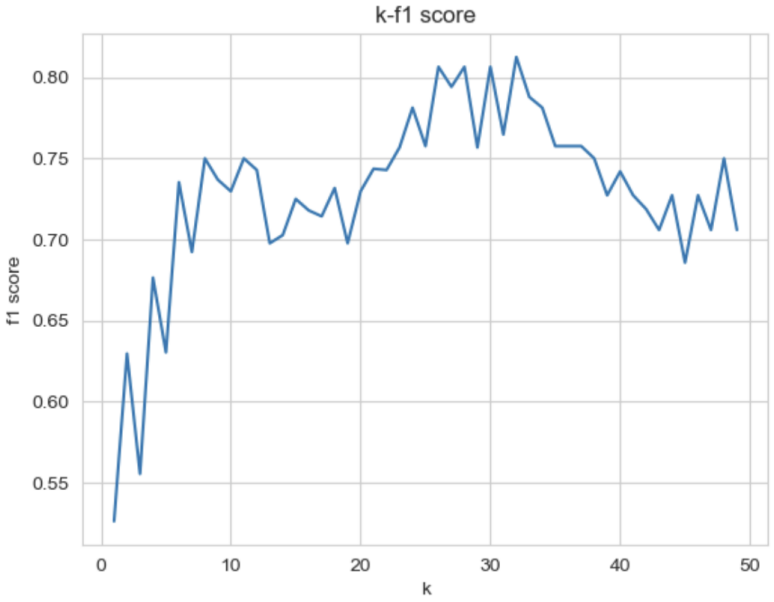


Fig. 1 The line graph of k and f1 score

The prediction results obtained after training the model are shown below in Table 3 and Fig. 2:

Table 3. The scores of KNN

F1 score	0.70
P score	0.77
R score	0.64
A score	0.83

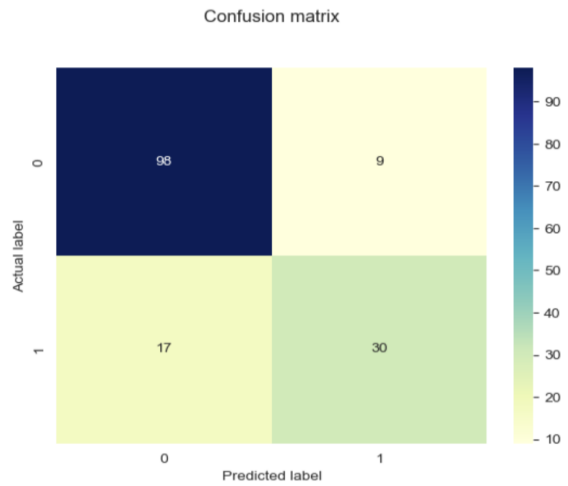


Fig. 2 The confusion matrix of KNN

3.3.2 SVM

An SVM algorithm is utilised to assess data and establish categorisations [8]. During training, the SVM algorithm employs marked training data to identify membership of distinct categories. Using Support Vector Machines (SVM), it is possible to construct a model that classifies new data into predefined categories. This shows that SVMs can be built and applied to specific examples from other categories than binary linear classification [9]. Examples from each category that are mapped to one another so that they all make sense form the basis of the model [7].

The second algorithm used was the SVM algorithm with the following results in Table 4 and Fig. 3:

Table 4. The scores of SVM

F1 score	0.60
P score	0.73
R score	0.51
A score	0.83

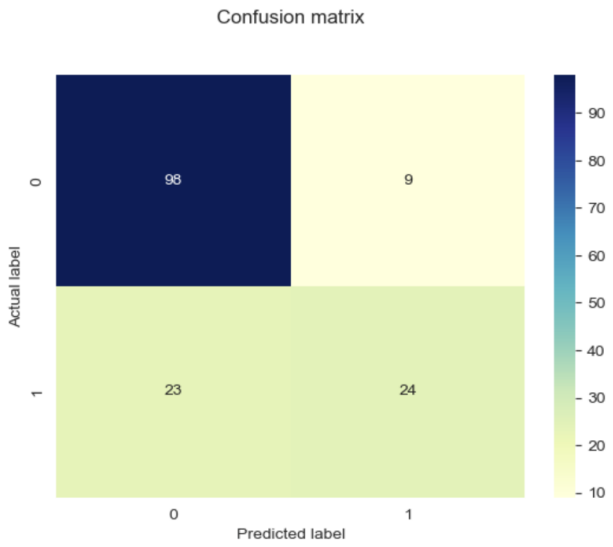


Fig. 3 The confusion matrix of SVM

3.3.3 Random Forest

The RandomForest method is an upgraded version of the bagging technique that combines a number of decision tree algorithms. This results in a robust data analysis tool. The RandomForest algorithm differs from the bagging technique in that the input number for each DT is chosen at random in the former. It indicates that the characteristics are chosen at random the moment the DT is produced. This minimizes decision trees' relationship with one another, reduces model variance, and improves model stability [12].

The third algorithm is the Random Forest algorithm with the following results in Table 5 and Fig. 4:

Table 5. The scores of Random Forest

F1 score	0.65
P score	0.65
R score	0.66
A score	0.79

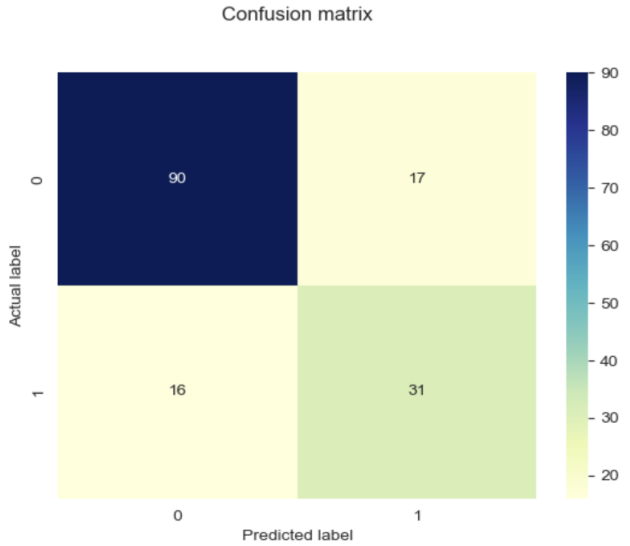


Fig. 4 The confusion matrix of Random Forest

3.3.4 XGBoost

It excels in a variety of tasks, including classification and regression [13]. In order to change the gradient boosting (GBM) anatomy, Extreme Gradient Boosting (XGBoost) joined boosting techniques. It is a networked, efficient gradient boosting library. The sequential technique known as XGBoost and its function are based on the idea of an ensemble approach. To improve prediction accuracy, a group of inexperienced learners was used. The outcomes that were incorrectly categorised are given more weight than the results that were predicted [14].

The last algorithm is the XGBoost algorithm, and the results are as follows in Table 6 in Fig. 5:

Table 6. The scores of XGBoost

F1 score	0.71
P score	0.72
R score	0.70
A score	0.79

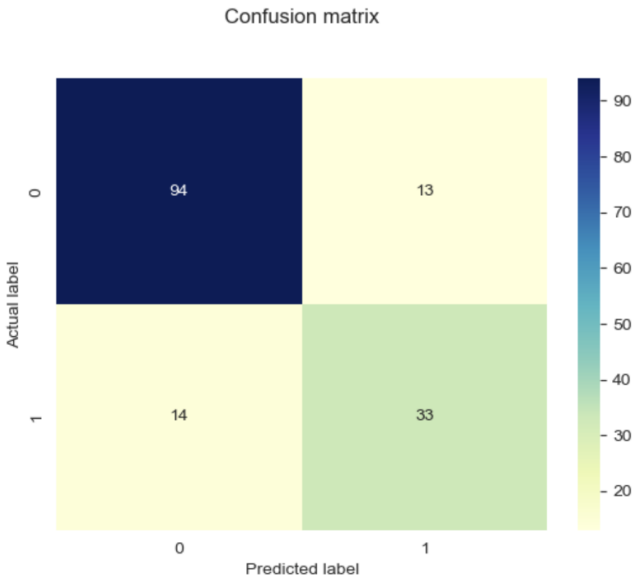


Fig. 5 The confusion matrix of XGBoost

4. Comparision of models

This study gets the results of each model and compare them, it is not difficult to find that the KNN algorithm gets the highest P and A, and the XGB algorithm gets the highest R and F1, and these two algorithms are used to predict this dataset excellently, as shown in Table 7:

Table 7. Comparision of models

	P SCORE	R SCORE	F1	ACCURACY
KNN	**0.77**	**0.64**	**0.7**	**0.83**
SVM	**0.73**	**0.51**	**0.6**	**0.79**
RF	**0.65**	**0.66**	**0.65**	**0.79**
XGB	**0.72**	**0.70**	**0.71**	**0.82**

5. Limitations

The dataset selected for this paper is the Pima Indian dataset with 500 diseased and 268 non-diseased individuals, and the dataset is relatively unbalanced. The dataset was not balanced in this experiment, which is the limitation of this experiment. The plan is to balance the dataset by sampling the dataset first in the next step, and then train the model separately and evaluate how things turn out.

6. Conclusion

Diabetes mellitus is a widespread issue affecting all humans. This paper is concerned with the problem of diabetes prediction, Pima Indians datase is selected as a dataset, four different models are trained on this dataset and their different results are compared. The best results were achieved by using cross-validation for hyper-parameter selection to optimize the models. But this dataset is small in number and the results obtained may be difficult to apply to larger datasets. And the dataset is also unbalanced. To address these two issues, the next step will be to select a larger dataset for model training and to sample it to get more accurate prediction results. One thing that must be firmly believed is that deep learning algorithms are likely to be increasingly prevalent within the medical field, enabling more effective disease treatment and a reduction in the current burden on healthcare systems.

References

[1] World Health Organization. Diabetes, 2023.

[2] Institute for Health Metrics and Evaluation. Global Burden of Disease Study 2019: Results. Seattle, WA: IHME, 2020.

[3] Olisah, C., Smith, L., & Smith, M. Diabetes mellitus prediction and diagnosis from a data pre-processing and machine learning perspective. Computer Methods, 2022.

[4] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access, 2020, 8: 76516-76531.

[5] UCI Machine learning. PIMA Indians Diabetes Dataset. Retrieved from https://kaggle.com

[6] Khanam, J. J., & Foo, S. Y. A comparison of machine learning algorithms for diabetes prediction. ICT Express, 2021, 7(4): 432-439.

[7] K. V. K. G., Shanmugasundaram, H., E, A., M, R., C, N., & SJ, B. Analysis of Pima Indian Diabetes Using KNN Classifier and Support Vector Machine Technique. In 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies, 2022: 1376-1380.

[8] Bi, Y., Wang, P., Pan, Y., et al. Comparison of risk prediction models for type 2 diabetes. Computer Knowledge and Technology, 2020, 16(28): 8-10+17.

[9] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 20187, 9: 515.

[10] Fan, B. Research on diabetes risk prediction model based on convolutional neural network. Thesis for doctoral dissertation, Nanjing University of Posts and Telecommunications, 2023.

[11] Luan, X. Research on diabetes prediction based on clustering undersampling hybrid integration algorithm. Thesis for doctoral dissertation, Jilin University, 2023.

[12] FU, H., DING, X., CHEN, D., et al. Establishment of a predictive model for hypotension in haemodialysis patients with diabetic nephropathy based on the random forest algorithm. Chinese Journal of Integrative Nephrology, 2023, 24(06): 493-496.

[13] Fan, J. Research on multiple algorithms for prediction of early diabetes. Thesis for doctoral dissertation, Chongqing University, 2022.

[14] Liu, Q., Ma, Y., & Cai, Y. Diabetes classification prediction model and application based on XGBoost algorithm. Modern Instrumentation and Medical Treatment, 2023, 29(04): 1-6+11.