Pima Indian Diabetes Database and Machine Learning Models for Diabetes Prediction

Linshan Xie *

School of Life Science, Fudan university, Shanghai, China, 200433

* Corresponding Author Email: 20300220008@fudan.edu.cn

Abstract. As a matter of fact, diabetes mellitus is becoming a growing epidemic worldwide in recent years which attracts great attention of researchers. In reality, as a result of the illness and its complications, the strain on healthcare systems is rising. Therefore, it is crucial to identify diabetes early in order to shield individuals from major consequences. With this in mind, in this study, four ML models are used for the imputation of dataset (K-NN) and the prediction of diabetes (LR, SVM and RF). According to the analysis, the results show that the LR model is slightly better than the RF as well as SVM models, with a prediction accuracy of 0.7913 and a precision of 0.8571. Ultimately, it can be said that there is a lot of promise when employing ML models to diagnose diabetes early on based on the evaluations. Overall, these results shed light on guiding further exploration of diabetes prediction.

Keywords: Diabetes; random forest (RF); logistic regression (LR); support vector machine (SVM).

1. Introduction

Insulin insensitivity, insulin deficiency, and compromised cellular activity are hallmarks of diabetes mellitus, a metabolic disease brought on by a combination of environmental and genetic factors [1]. As to the 2021 report of the International Diabetes Federation, diabetes is diagnosed in 536.6 million persons globally between the ages of 20 and 79. Globally, it is anticipated that the prevalence of diabetes would rise from 10.5% in 2021 to 12.2% (or 783.2 million) in 2045 [2]. Type 1 and type 2 diabetes fall into two primary types. Five to ten percent of all cases of diabetes worldwide are caused by type 1 diabetes mellitus (T1D) [3]. Total insulin insufficiency is the end result of this diverse illness, which is typified by the death of pancreatic beta cells. Some cases are caused by idiopathic beta cell failure or destruction, while the majority are related to autoimmune-mediated beta cell death [4]. Two distinguishing features of type 2 diabetes mellitus (T2D) are the presence of insulin resistance and inadequate compensatory insulin secretion [5]. Diabetes mellitus is predominantly attributed to two underlying pathophysiological processes, namely, compromised insulin production and insulin resistance, mainly affecting the liver and skeletal muscle [6]. There are still a number of complex aetiologias of T2D that remain undiscovered by researchers, as the condition is complex and arises from the interaction of genes and environment. Over the past ten years, the use of genome-wide association studies (GWASs) has demonstrated that genetic variables have a significant influence on type 2 diabetes [7].

Due to dietary and lifestyle changes, Type 2 diabetes (T2D) has emerged as a prominent global contributor to heart disease, amputation, blindness, kidney failure, and premature death in recent years [8]. Consequently, it is critical to identify those who are very susceptible to acquiring type 2 diabetes because prompt treatment could halt or even reverse the disease's progression [9]. Artificial intelligence (AI) and machine learning (ML) supervised models in particular are useful instruments for diagnosing and treating diabetes as a chronic illness. At the same time, ML modeling has been shown to be an important tool for predicting the development of diabetes. ML techniques have been tried in numerous studies to predict diabetes. For example, Kahramani et al. predicted T2D instances more accurately by combining an ANN with a fuzzy neural network (FNN) in a hybrid approach [10]. Maniruzzaman et al. utilized various feature selection techniques, such as analysis of variance (ANOVA), principal component analysis (PCA), mutual information (MI), RF, and LR, in their study on PIDD research. These techniques were employed to investigate different subsets of features and

subsequently classify them using different classifiers [11]. This research aims to utilize the Pima Indian dataset, consisting of 768 patients, in order to employ Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) models for the prediction of diabetes.

2. Materials and Methods

2.1. Data

The information in this article comes from the Pima Indians Diabetes Database on Kaggle. The primary source of the Pima Indians Diabetes Database may be attributed to the National Institute of Diabetes and Digestive and Kidney Diseases. The respondents were all female Pima Indian indigenous subjects over 21 years of age, body mass index, living in a community close to Phoenix, Arizona, USA. Nine characteristics, including age, number of pregnancies in women, blood pressure, skin thickness, glucose level, insulin level, family history of diabetes, and the outcome (diabetes or not), are present in the dataset, which has 768 patient records in total.

2.2. Variables

The dataset contains nine variables in total: age, the outcome, skin thickness, insulin, blood pressure, glucose, BMI, and pregnancies. One of the most important variables is outcome, which means whether a person has diabetes or not, with a value of 1 if they have diabetes and 0 if they do not. In this dataset, there are 268 people with diabetes and 500 without the disease. A woman's total number of pregnancies is indicated by the variable Pregnancies, and it can be observed that the maximum number of pregnancies is 17 and the minimum is 0, which is a large variation, and although the occurrence of 17 pregnancies is rare, it is reasonable considering the local culture and specificities. Among all the variables, Age describes the information about the age of the women in the dataset, where the mean age is 33.24 years, and the standard deviation is 11.76. The glucose variable, which has a normal range of 70-140 mg/dl, represents the blood glucose level at the 2-hour mark of the oral glucose tolerance test (OGTT). For the diagnosis of diabetes, OGTT is presently the gold standard [12]. The test is performed by dissolving 75 g of anhydrous dextrose orally in 300 ml of water in a fasted state (usually early in the morning) after 8 hours of fasting and water fasting, requiring the person to be tested to drink the glucose within 5 minutes, and determining the venous plasma glucose level 2 hours after the start of the glucose water intake [13]. The normal plasma glucose level after 2 hours of OGTT should be 70-140 mg/dl. As a variable correlated with the OGTT experiment, insulin measures the amount of insulin in plasma extracted at the two-hour OGTT time cutoff. Skin thickness refers to triceps skinfold thickness (TSF), which is an indicator of subcutaneous fat content and indirectly reflects caloric changes. The upper arm of the person to be tested is naturally lowered, and the dorsal side of the upper arm is taken from the shoulder peak to 2cm above the midpoint of the ulnar eagle's beak line, where the skin is pinched up together with the subcutaneous fat to form a fold, and then measured by a skinfold thickness gauge with a pressure of 10g/mm2. Body mass index, or BMI, is computed as follows: weight in kilograms divided by height in square meters [14]. The family history of diabetes is a function of the diabetic pedigree. This tool assigns a family history-based risk score for diabetes.

2.3. Models and Evaluations

A total of four ML models is used in this article, namely, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF). Among them, K-NN is primarily employed for post-deletion complementation of dataset outliers, with the remaining three models being utilized for diabetes prediction.

The K-Nearest Neighbor (K-NN) technique is a distance-based technique; the model determines the separation between a sample and every other sample in the data set as well as the samples that need to be categorized [15]. The premise is that an outlier's K-nearest neighbor distance is significantly larger than a regular point's K-nearest neighbor distance [16]. The K-nearest neighbors'

technique can really be applied to the imputation of missing data in addition to classification. So, in this study, the KNN approach was used to patch up the incomplete dataset after removing the anomalous 0s.

Logistic Regression and Linear Regression are both generalized linear models. One kind of machine learning model called an LR model is frequently used to forecast an event's likelihood based on a predictor variable, like true or false, success or failure, or yes/no. Logistic regression introduces a nonlinear element through a Sigmoid function to obtain a logit of the probability of one outcome category (Yˆ) relative to another (1-Yˆ) [17]:

$$\ln(Y^{\wedge}/1 - Y^{\wedge}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \quad (1)$$

An SVM model is a regression and classification technique. The model finds the hyperplane that minimizes the distance between samples and the two closest classes [18]. Its purpose is to build a judgement boundary between two classes so that one or more feature vectors may be utilised to forecast labels [19]. In the case of two classes or two dimensions, the hyperplane is represented by the following equation [20]:

$$\beta_0 + \beta_{1\times1} + \beta_{2\times2} = 0 \quad (2)$$

A random forest is an assembly or collection of Classification and Regression Trees (CART) trained on datasets the same size as the training set. These datasets are termed bootstraps, and they are produced by randomly resampling the training set [21]. The Gini index of a binary split node n is calculated as follows [22]:

$$\text{Gini(n)} = 1 - \sum_{j=1}^{2} (p_j)^2 \quad (3)$$

Where p_j is the node n's relative frequency of class j.

Three models' prediction outcomes in this study are evaluated using the confusion matrix and its associated evaluation metrics F1-score, recall, accuracy and precision. A table containing the dimensions "actual" and "predicted" along with the columns "true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN)" makes up a confusion matrix. Moreover, the following formulas are used to compute the accuracy, precision, recall, and F1-score:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}, \text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN}, \text{F1} - \text{score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

Table 1. Demographic information of Pima Indian diabetes dataset

	Average	Standard Deviation	Median	Min	Max	Range	Skewness	Kurtosis
Pregnancies	3.845052083	3.369578063	3	0	17	17	0.898154873	0.142183957
Glucose	121.6867628	30.53564107	117	44	199	155	0.528902593	-0.290197983
BloodPressure	72.40518417	12.38215821	72	24	122	98	0.133604175	0.886154943
SkinThickness	29.15341959	10.47698237	29	7	99	92	0.686794053	2.875579478
Insulin	155.5482234	118.7758552	125	14	846	832	2.149995838	6.227754173
BMI	32.45746367	6.924988332	32.3	18.2	67.1	48.9	0.591617914	0.839606953
Family history of diabetes	0.471876302	0.331328595	0.3725	0.078	2.42	2.342	1.912417924	5.528538857
Age	33.24088542	11.76023154	29	21	81	60	1.125188044	0.621726908
Outcome	0.348958333	0.476951377	0	0	1	1	0.632538263	-1.601976168

2.4. Statistical Analysis

For all the statistical analyses, the study used the R software version 4.2.2. First of all, there are many abnormal 0 values in the original dataset, such as BMI is 0 (since the calculation of BMI is weight divided by the square of height, it is impossible for its BMI value to be 0). The abnormal 0 values were counted as: blood presure 35 times, skin thickness 227 times, insulin level 374 times, and BMI 11 times. After labeling all these abnormal values as NA, the overall statistics and description of the data information were carried out and the results are shown in Table 1. Then, the missing data just labeled as NA were supplemented using the K-NN method. The newly added dataset was split

into training and test sets using the following ratio for the training sets: test group: 8:2, and the pertinent parameters were counted for the training set and test set, respectively, along with their levels. The findings demonstrate that the training set and test set's data compositions are essentially the same, with no discernible variations, and they can be utilized for further prediction. As a result, three models—LR, RF, and SVM—were utilized to forecast the onset of diabetes mellitus. The models were built using the training set, and the test set was used to evaluate the models' prediction performance and compare the three models in respect to relevant metrics (F1-score, accuracy, precision and recall).

Table 2. Demographic information of the test and train dataset

Test - Data Set								
	Average	**Standard Deviation**	**Median**	**Min**	**Max**	**Range**	**Skewness**	**Kurtosis**
Pregnancies	3.69318181	3.047651853	3	0	12	12	0.814366269	-0.34280843
Glucose	118.539772	29.72845886	114	68	198	130	0.701548643	-0.07408688
BloodPressure	70.3443181	9.512089846	70	48	94	46	-0.066961934	-0.26000900
SkinThickness	26.7704545	8.466055369	27.1	10	50	40	0.207320038	-0.43480879
Insulin	136.622727	62.88631871	129.55	18	328	310	0.503702611	-0.34685935
BMI	31.0682386	5.321131971	32	18.2	45.8	27.6	0.019524774	-0.29920540
Family history of diabetes	0.4120625	0.229511809	0.353	0.1	1.144	1.044	0.957742557	0.239537505
Age	30.8579545	9.387970681	28	21	58	37	1.078282742	0.393814971
Outcome	0.34659090	0.477241805	0	0	1	1	0.639245962	-1.60036089
Train - Data Set								
	Average	**Standard Deviation**	**Median**	**Min**	**Max**	**Range**	**Skewness**	**Kurtosis**
Pregnancies	3.67228915	3.175019999	3	0	13	13	0.790322334	-0.17456888
Glucose	117.269638	27.3847668	113	44	196	152	0.56589904	0.06539153
BloodPressure	71.4315662	10.7197385	71.6	38	106	68	-0.00615881	-0.10085363
SkinThickness	28.4896385	8.249691174	29	10	50	40	-0.025871514	-0.59445182
Insulin	130.133734	58.72889255	121.9	15	318.2	303.2	0.775823267	0.515979738
BMI	32.1994216	6.456910925	31.9	18.2	50	31.8	0.252120771	-0.30466512
Family history of diabetes	0.40232048	0.219745958	0.342	0.078	1.182	1.104	0.830803849	0.095541911
Age	31.0530120	9.230650703	28	21	60	39	0.922247245	-0.10730975
Outcome	0.30602409	0.461396064	0	0	1	1	0.83879559	-1.29953596

3. Results and Discussion

3.1. Data Processing

One of the most significant responsibilities in the identification and classification of diabetes using ML models is analyzing how the variables in the dataset are connected to one another, which necessitates the use of data analysis techniques and software tools. For this assignment, the correlation matrix is computed and visualized. As shown in Fig. 1, as one previously recognized, diabetes was most strongly associated with blood glucose levels 2h after performing the OGTT, with a correlation of 0.52, while the correlation between the other variables and diabetes was low, with values below 0.5. Due to the small dataset itself and the large number of anomalous values, directly discarding the NA data would have a large impact on the capacity of the dataset and the accuracy of the prediction results, so the K-NN method was used for the supplementation of the NA data.In this case, the number of 10 closest neighbors was used. Following imputation, the dataset was split into training and test sets with a training set to test set ratio of test set = 8:2. Following imputation, the dataset was split into training and test sets with a training set to test set ratio of test set = 8:2. Table 2 displays the statistics pertaining to the training and test sets following the division and supplementation of the dataset.

Fig. 1 Correlation matrix

3.2. Model Analysis & Comparison

In order for the algorithms to learn, identify patterns, and generate precise predictions, data had to be fed to them. Following the training and fine-tuning of the three models, the trained machine learning model was applied to a test set to evaluate its ability to predict diabetes. Further, the prediction results (confusion matrix) of the three models were visualized using the fourfoldplot package in R. Fig. 2 depicts the predicted results—confusion matrix of the three models. The essential assessment metrics (Recall, Accuracy, F1-score, and Precision) were calculated in R using the Confusion Matrix package in order to more accurately examine the prediction impacts of the three models (Table 3).

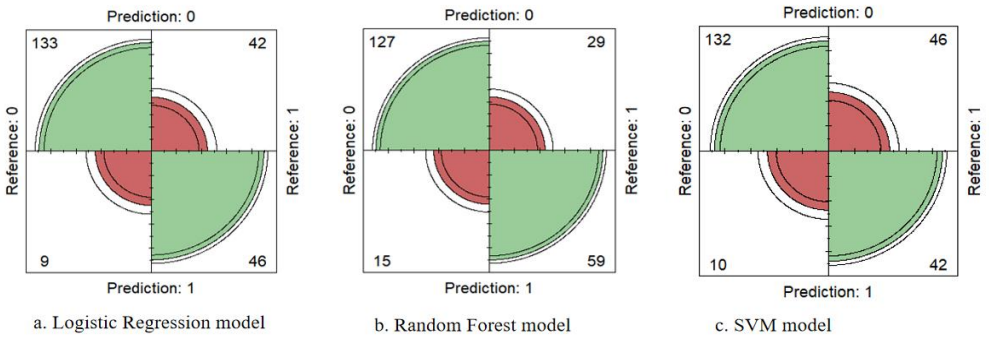

Fig. 2 Confusion matrix of the 3 models

Table 3. Comparsion between models

	Accuracy	**Precision**	**Recall**	**F1-Score**
LR	0.7913	0.8571	0.5455	0.6667
RF	0.7739	0.8000	0.5455	0.6486
SVM	0.7565	0.8077	0.4773	0.6000

With the highest recall of 0.5455, accuracy of 0.7913, F1-score of 0.6667 and precision of 0.8571 among the three models, as indicated in Table 3, logistic regression produces the best prediction results. At 0.5455, 0.7739, 0.6486 and 0.8000 for recall, accuracy, F1-score and precision respectively, Random Forest's prediction performances are the second best. SVM has the lowest prediction result

of the three models, with values for recall, accuracy, prediction and F1-score of 0.6000, 0.7565, 0.8077 and 0.4773. Overall, all the three models had better prediction results, with accuracy values greater than 0.75.

4. Conclusion

To sum up, this study used some statistical, machine learning, and visualization methods to study the Pima Indians Diabetes Database, the K-NN method was used for the addition of anomalous missing values and Logistic, Random forest and SVM models were used for the prediction of diabetes. The outcomes demonstrated that all three models' prediction accuracy was more than 0.75, but the model of logistic regression had a better prediction effect in comparison, which might be related to the characteristics of logistic regression model specializing in dichotomous type of data. However, there are many shortcomings and future improvements in this study, such as the fact that the women in this dataset are of Pima Indian indigenous origin, which is not representative enough. In addition to this, the dataset is relatively small in size and contains more anomalous 0-values, so in this paper the K-NN method was used for the addition of anomalous missing values. Another noteworthy issue is that in this dataset, the percentage of diabetic patients is high, which is much higher than the results of epidemiologic statistics. Therefore, perhaps further supplementation and simulation of the dataset can further improve the accuracy of prediction. In conclusion, predicting diabetes is one of the most challenging obstacles in the field of medical engineering, and will require continuous investigation by data analysts in the future.

References

[1] Li Z, Han D, Qi T, Deng J, Li L, Gao C, Gao W, Chen H, Zhang L, Chen W. Hemoglobin A1c in type 2 diabetes mellitus patients with preserved ejection fraction is an independent predictor of left ventricular myocardial deformation and tissue abnormalities. BMC Cardiovasc Disord, 2023, 23(1): 49.

[2] Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract, 2022, 183: 109-119.

[3] Maahs D M, West N A, Lawrence J M, Mayer-Davis E J. Epidemiology of type 1 diabetes. Endocrinol Metab Clin North Am, 2010, 39(3): 481-97.

[4] Gillespie K M, Bain S C, Barnett A H, Bingley P J, Christie M R, Gill G V, Gale E A. The rising incidence of childhood type 1 diabetes and reduced contribution of high-risk HLA haplotypes. Lancet, 2004, 364(9446): 1699-700.

[5] Vehik K, Hamman R F, Lezotte D, et al. Trends in high-risk HLA susceptibility genes among Colorado youth with type 1 diabetes. Diabetes care, 2008, 31(7): 1392-1396.

[6] Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. Lancet, 2017, 389(10085): 2239-2251.

[7] Langenberg C, Lotta LA. Genomic insights into the causes of type 2 diabetes. Lancet, 2018, 391(10138): 2463-2474.

[8] Wu B, Niu Z, Hu F. Study on Risk Factors of Peripheral Neuropathy in Type 2 Diabetes Mellitus and Establishment of Prediction Model. Diabetes Metab J, 2021, 45(4): 526-538.

[9] Laakso M. Biomarkers for type 2 diabetes. Mol Metab, 2019 Sep, 27S(Suppl): S139-S146.

[10] Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. Expert Syst Appl, 2008, 35(1): 82-89.

[11] Maniruzzaman M, Rahman M J, Al-MehediHasan M, Suri H S, Abedin M M, El-Baz A, Suri J S. Accurate diabetes risk stratifcation using machine learning: role of missing value and outliers. J Med Syst, 2018, 42(5): 92.

[12] Phillips P J. Oral glucose tolerance testing. Aust Fam Physician, 2012, 41(6): 391-3.

[13] Greenspoon J S. Oral glucose tolerance test. Mayo Clin Proc, 1988, 63(8): 838.

[14] Uloma I U, Christopher C K. Age, gender, and racial/ethnic differences in the association of triclocarban with adulthood obesity using NHANES 2013–2016. Arch Environ Occup Health, 2022, 77(1): 68-75.

[15] Fathabadi A, Seyedian S M, Malekian A. Comparison of Bayesian, k-Nearest Neighbor and Gaussian process regression methods for quantifying uncertainty of suspended sediment concentration prediction. Sci Total Environ, 2022, 818: 151760.

[16] Hu Y H, Lin W C, Tsai C F, Ke S W, Chen C W. An efficient data preprocessing approach for large scale medical data mining. Technol Health Care, 2015, 23(2): 153-60.

[17] Stoltzfus J C. Logistic regression: a brief primer. Academic Emergency Medical, 2011, 18(10): 1099-104.

[18] Tanveer M, Rajani T, Rastogi R, et al. Comprehensive review on twin support vector machines. Ann Oper Res, 2022, 3: 1–46.

[19] Huang S, Cai N, Pacheco P P, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics, 2018, 5(1): 41-51.

[20] Iparraguirre-Villanueva O, Espinola-Linares K, Flores Castañeda R O, Cabanillas-Carbonell M. Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes. Diagnostics (Basel), 2023, 13(14): 2383.

[21] Breiman L. Bagging predictors. Mach Learn, 1996, 24: 123–140.

[22] Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front Aging Neurosci, 2017, 9: 329.