Indian journal of
Engineering

Prediction of diabetes in adults using supervised machine learning model

Sunil Upadhyay[1], **Yogesh Kumar Gupta**[2]

To Cite:
Upadhyay S, Gupta YK. Prediction of diabetes in adults using supervised machine learning model. *Indian Journal of Engineering*, 2023, 20, e26ije1657 doi: https://doi.org/10.54905/disssi/v20i53/e26ije1657

Author Affiliation:
[1]Department of Computer Science, Banasthali Vidyapith-304022, Rajasthan, India
Email: sunilhit120@yahoo.com
ORCID: 0000-0003-0814-8408
[2]Department of Computer Science, Banasthali Vidyapith-304022, Rajasthan, India
Email: gyogesh@banasthali.in
ORCID: 0000-0002-4572-178X

Peer-Review History
Received: 08 May 2023
Reviewed & Revised: 11/May/2023 to 10/June/2023
Accepted: 14 June 2023
Published: 18 June 2023

Peer-Review Model
External peer-review was done through double-blind method.

Indian Journal of Engineering
pISSN 2319–7757; eISSN 2319–7765

ABSTRACT

Diabetes disease is caused by increase in blood sugar in human body. Pancreas secretes insulin to regulate glucose in blood. When pancreas not able to secrete or enough insulin or not able to use insulin led to diabetes. Biomarkers related to diabetes are Diet, life style, age, pregnancies, physical activity, tension, blood pressure etc. Diabetes is mainly responsible for diseases like kidney failure, eyes issues, nerves damage, heart attack etc. In current scenario tests are the only methods to detect diabetes, but this is a time-consuming process. Machine learning helps in early detection of diseases through identification of hidden pattern and analyses of various biomarkers. The purpose of this research is to propose a model for early detection of diabetes on PIMA dataset, which is sourced from Kaggle repository and to identify highly related biomarkers related to diabetes. Data set includes 768 records and 8 attributes. In this paper nine supervised classification algorithms are used like Logistic regression, KNeighbours Classifier, Support vector classifier, Extra tree, Bayes classifier, Gradient boosting classifier, Random and Decision classifier. Logistic regression performed best with accuracy of 82% when compared with others classification algorithms. Two biomarkers identified as Glucose and BMI which are directly linked with increase in diabetes. Higher values of glucose and BMI higher the risk of diabetes.

Keywords: Machine learning, Diabetes, Ensemble, Supervised learning, Artificial intelligence, Health analytics, Adaboost.

1. INTRODUCTION

Diabetes originates when glucose increase in the human body (Aljumah et al., 2013). Pancreas secretes insulin which regulates glucose in human body (Gujral, 2017). When pancreas not able to secrete insulin or not produce enough insulin that may lead to diabetes (Hathaway et al., 2019). Diabetes may lead to various diseases like heart attack, nerves damage, eyes issues, kidney related issues etc. (Kumari and Chitra, 2013). Diabetes number increasing day by day and is major cause's death, heart attack, blindness all over the world (Alkhatib et al., 2020). As per report, 8.5% of adults had diabetes in year 2014 and 1.5 million deaths due to diabetes in 2019 (Breault et al., 2002).

DISCOVERY
SCIENTIFIC SOCIETY

Diabetes can be managed through change in life style, exercise, food hobbit and manage weight. Class of diabetes are- Type 1, Type 2 and gestational (Yu et al., 2010). Type 1 when pancreas not able to secrete insulin, Type 2 when body not able to use insulin and gestational diabetes occurs at the time of pregnancies (Yu et al., 2010). Age, gender, BMI, Blood pressure, pedigree function, skin thickness, glucose etc. are few biomarkers related to diabetes (Plis et al., 2014). Machine learning and Artificial intelligence play an important role in early detection of diabetes (Hasan et al., 2020). ML find hidden pattern and correlation in data and predict outcome (Dagliati et al., 2018).

After collection of datasets, data divided into two stages after pre-processing i.e., training and testing. On training data set various classification models run and predict outcome once training complete then in second stage testing to be done to check the accuracy of model. Machine learning types are Supervised, unsupervised and reinforcement. In Supervised learning input and output are mapped, this learning classified into two types- classification and regression. Unsupervised learning means data not labelled based on available patterns output generated. Unsupervised learning classified into two types: Association and clustering. Reinforcement learning in which, agent perform action in an environment.

2. METHODOLOGY

Data set

Dataset sourced from National Institute of Diabetes & Digestive and kidney diseases of 21 years females. Data have 768 records and 9 biomarkers.

Algorithms

The following classification algorithms compared on PIMA dataset.

Logistic regression

Algorithm used for predicting categorical dependent variable. This algorithm is based on probability concepts and predicted value fall between 1 and 0.

Decision tree

Algorithm used for both classifications as well as for regression. This algorithms design tree in which node act as features, branches as decision rule and leaf as output. It uses CART algorithm. Pruning helps in decision tree to remove unwanted branches and Attribute selection method is used to select best attributes for root node and sub node.

KNeighbours classifier

KNN support both classifications as well as for regression. KNN is a type of lazy learner means at taring phase it stores data and wait for new data, once received new data then it classifies the new data on similarity based on store data. K can be any value but the most preferred value is 5. To improve accuracy value of k will be larger.

Support vector machine

SVM create hyper plane to divided space into classes, so data when arrived can easily be segregated into the correct class. SVM select extreme data points for creating hyperplane. SVM classified into two types- linear and non- linear. Linear SVM segregate two classes through a straight line while non-linear SVM not able to segregate data through a straight line.

Naïve Bayes

NB is used for categorical data problem and this use probability concept for solving problem. This is used for imbalance or missing value dataset.

Random Forest

RF support both classification and regression problem. RF is based on hybrid technique in which multiple classifiers combined together to improve the model performance. This algorithm combines multiple decision trees together and generates output using voting technique on predication of each tree. Higher number of trees in classifier will improve accuracy of model and avoid issue of overfitting.

Adaboost

Adaboost is also known as adaptive boosting and used for both classifications as well as regression problem. Adaboost is based on ensemble technique in which multiple decision tree with one split combined together to improve the model performance. This algorithm combines the output of each weak classifier by assigning weight sum as final result.

Gradient boosting

Gradient boosting supports both classifications as well as regression problems. The purpose of this algorithm is to minimize the loss function by adding weak classifier in gradient descent manner. Log loss is used in sification model as a cost function and mean square error (MSE) is used in regression model as a cost function.

Extra tree classifier

ETC is based on ensemble technique in which multiple decision trees combine combined together to improve the performance of model. ETC reduces both variance and bias. In this technique whole original sample data selected randomly.

Proposed Model

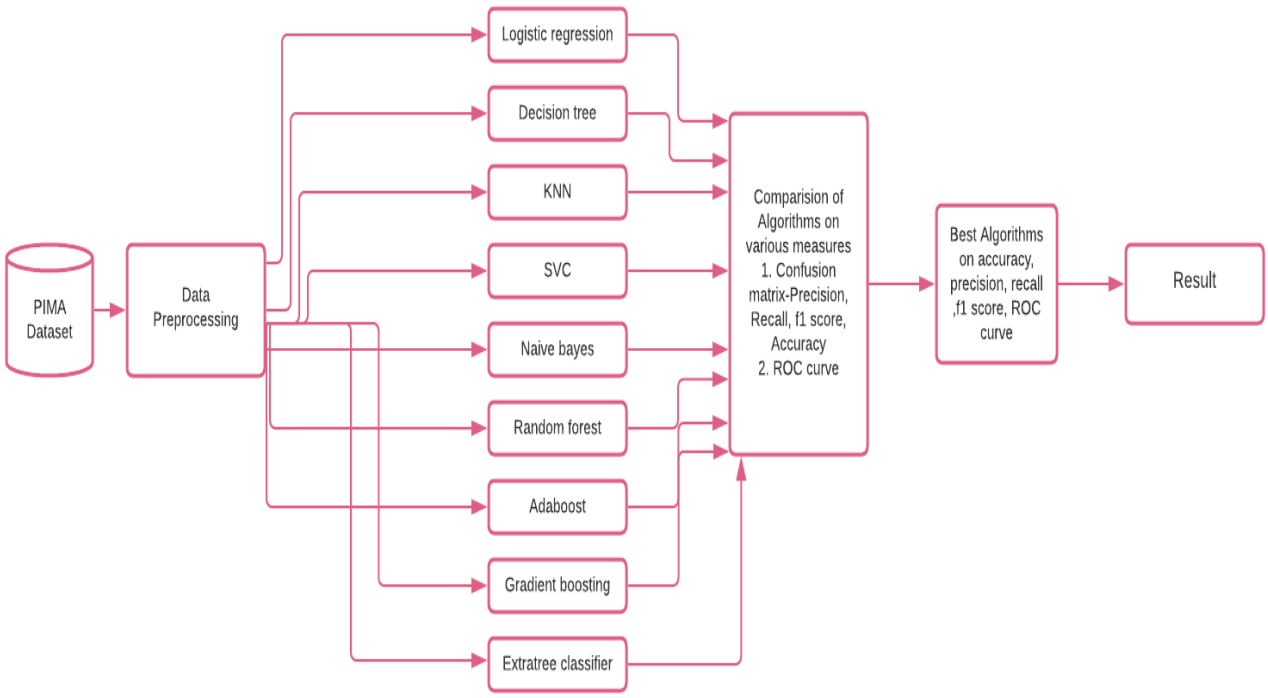

Figure 1 Proposed Model

Accuracy measures

Classification algorithms used in this paper are Logistic regression, Support vector machine, Random Forest, KNeighbours Classifier, Extra tree classifier, Naïve bayes, Adaboost, Gradient Boosting and Decision tree. To measure performance of models Precision, recall, accuracy and f1-score and ROC curve used.

Confusion matrix- Matrix used to measures the performance of models in machine learning.

True Positive (TP) - The predicted values match with actual values.

True Negative (TN) - The predicted values match with actual values.

False Positive (FP) - Actual value negative while model predicted value positive.

False Negative (FN) - Actual value positive while model predicted value negative.

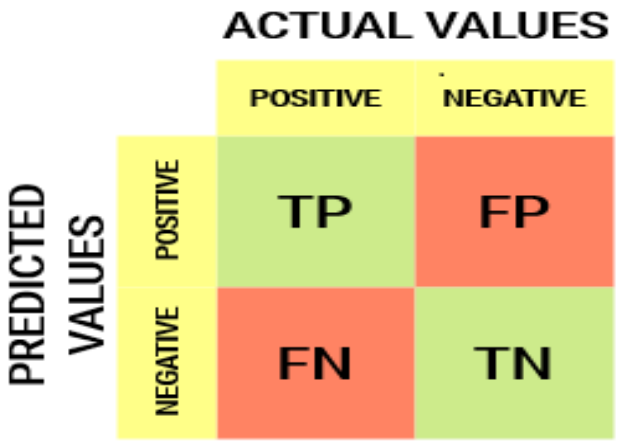

Figure 2 Confusion matrix

Precision- Precision measure quality. High precision means low false positive. The best precision score is 1 and poor is 0. Precision= TP/(TP+FP).

Recall- Recall measures of quantity mean number of true positives was found. The best recall score is 1 and poor is 0. Recall= TP/(TP+FN)

F1-score- F1 measures weighted average of recall and precision. F1 measure is high when both precision and recall measures is high. F= 2*(Precision * Recall) / Precision +Recall

Accuracy- Measures correct prediction in term of percentage on the test data. Accuracy= TP+TN / Total samples.

ROC curve- ROC curve compare the performance of all classification models in usefulness of test.

Table 1 Classification algorithms performance on various measures

Algorithms		Precision	Recall	f1-score	Accuracy	Training score	ROC
Logistic Regression		0.76	0.62	0.68	0.82	76.22	0.87
SVC		0.73	0.51	0.6	0.79	75.89	0.85
Random Forest		0.71	0.64	0.67	0.81	100	0.86
KNN	K=1	0.39	0.45	0.42	0.62	100	
	K=3	0.54	0.60	0.57	0.72	85.17	
	K=5	0.59	0.62	0.60	0.75	78.50	
	K=7	0.61	0.57	0.59	0.76	78.15	
	K=9	0.63	0.62	0.62	0.77	78.33	
	K=11	0.63	0.57	0.60	0.77	77.52	
	K=13	0.67	0.62	0.64	0.79	78.33	
	K=17	0.68	0.55	0.61	0.79	76.71	0.82
Extra tree Classifier		0.69	0.66	0.67	0.81	100	0.86
Naïve Bayes		0.67	0.62	0.64	0.79	75.33	0.84
Adaboost		0.63	0.66	0.65	0.78	81.75	0.85
Gradient Boosting		0.63	0.66	0.65	0.78	91.85	0.84
Decision Tree		0.59	0.43	0.45	0.73	82.08	0.75

3. RESULT

In this paper, nine classifier models are proposed to find the best classifier model in term of accuracy on PIMA dataset. All models trained and tested on PIMA dataset after pre-processing. As a result, Logistic regression perform best in term of accuracy i.e., 82%, Precision – 76%, f1-score-68% and ROC curve is 87% while recall is 62%. Higher glucose and BMI increase the risk of diabetes. Glucose below 125 most likely not to be diabetic while glucose level above 126 will be diabetic. BMI below 29 most likely not to be diabetic while BMI level above 30 will be diabetic.

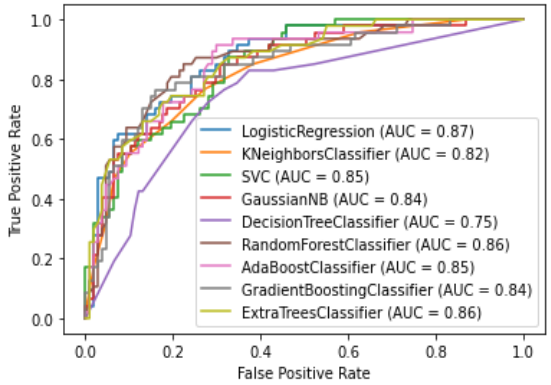

Figure 3 ROC curve

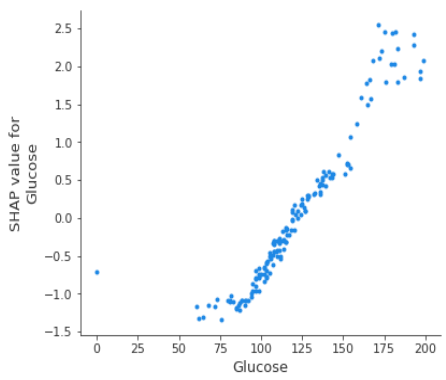

Figure 4 SHAP dependent plot

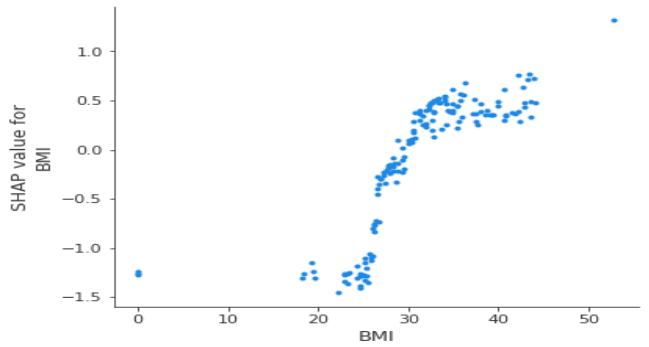

Figure 5 SHAP dependent plot

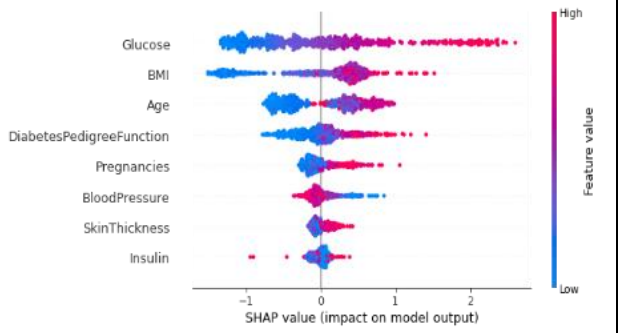

Figure 6 SHAP summary plot

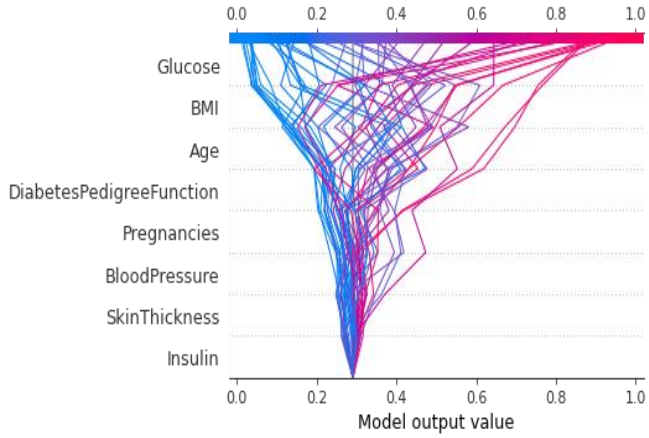

Figure 7 SHAP Output plot

```
Classification Report is:
              precision    recall  f1-score   support

         0.0       0.84      0.92      0.88       107
         1.0       0.76      0.62      0.68        47

    accuracy                           0.82       154
   macro avg       0.80      0.77      0.78       154
weighted avg       0.82      0.82      0.82       154

Confusion Matrix:
 [[98  9]
 [18 29]]
Training Score:
 76.2214983713355
```

4. DISCUSSION

Abdillah and Suwarno, (2016), discussed about diabetes diagnosis using support vector classifier. Algorithms applied in this paper are radial basis function with support vector classifier. As a result, accuracy=80.22%, sensitivity=82.56%, specificity=79.12%, auroc=0.8084 on while using 500 training data. This technique if combined with other kernel function or used pca for feature selection will increase accuracy in prediction.

Alkhatib et al., (2020) discussed about skin thickness may lead to diabetes. Anova test used in this paper. Skin thickness is higher in normal patients, while in case of prediabetic skin thickness decrease and in case of diabetic skin thickness is close to prediabetic. Skin thickness influenced by insulin not by glucose. In case of insulin is high in normal patient while decrease in prediabetic patient and insulin goes high in diabetic patients. To predict diabetes-2 skin thickness is inexpensive and simple method. Improving the accuracy by increasing the size of data.

Aljumah et al., (2013) and Kumari and Chitra, (2013) discussed-on diabetes classification through support vector classifier. 10-fold cross validation method yield accuracy=78%, sensitivity=80%, specificity=76.5%. This technique can be improved by feature subset selection. Ban et al., (2010) and Dagliati et al., (2018) reviewed literature on machine learning & data mining applications. As

result, 85% used supervised learning and 15% used unsupervised learning and support vector machine is the successful machine learning algorithm.

Breault et al., (2002) and Gujral, (2017) discussed role of machine learning in diabetes detection. In this paper various articles on diabetes reviewed using artificial intelligence. After reviewing the literature on diabetes, it has been observed that hybrid approach with svm, pca along with ann and genetic algorithms provide good result while comparing with single approach. Random forest is better than decision tree. Hybrid approach when combined with iot will develop real time applications for health care systems. Dagliati et al., (2018) and Breault et al., (2002) use cart algorithm to predict variables hgba1c and comorbidity index. Result shows that bad glycemic is the directly associated with diabetes.

Dagliati et al., (2018) and Repalli, (2011) discussed how peoples of different age are affected by diabetes. The dataset has 50784 records and 37 attributes. This research also found other factors responsible for diabetes. Garcia-Carretero et al., (2021) and Zou et al., (2018) discussed on neural network, decision tree & random forest to predict diabetes. Random forest classifier is at 0.8084 while compared with other algorithms. Gujral, (2017) and Yu et al., (2010) discussed-on prediction of disease using support vector classifier for two classification schemes. Scheme i accuracy is 83.47% and ii is 73.18% using two parameter gamma and c.

5. CONCLUSION

In this paper, a systematic approach was made to design a machine learning model for the early prediction of diabetes using PIMA dataset sourced from Kaggle repository. This model includes nine classification algorithms. Models evaluated on various measures to check accuracy, precision, f1 score, recall and ROC curve. In this evaluation logistic regression preformed best with accuracy of 82 % and ROC curve 87% while compared with other classification models. Two biomarkers identified as Glucose and BMI which are directly linked with increase in diabetes. Higher the values of glucose and BMI higher the risk of diabetes. Further this work can be extended to detect other life style related diseases using other machine learning techniques with more accuracy and reduction in processing time.

Authors Details

Sunil Upadhyay: PhD student at Department of Computer Science, Banasthali Vidyapith, Newai (Rajasthan), has experience of more than Thirteen years in academics and research. He completed his Master in Business Analytics from Birla Institute of Technology & Science (BITS), Pilani. He is an Oracle certified Professional (OCP) from Oracle corporation an also certified in Big Data & Data Analytics from University of California, Indian Institute of Technology- Roorkee, Indian Institute of Technology-Ropar, Motilal Nehru National Institute of technology (MNIT) Allahabad and National Institute of Technology (NIT) Srinagar. He has over 13 years of experience in academics and research. He has published research papers in reputed national, international conferences and journals. He has conducted hands-on session in Data Science using Python, Tableau and Programming in R for students and staff. He has experience in designing Oracle databases. His areas of interest are Business Analytics, Data Visualization, Database Management System, Decision making using Excel and ERP. Email: sunilhit120@yahoo.com

Yogesh Kumar Gupta: Assistant Professor, Department of Computer Science, Banasthali Vidyapith, Newai (Rajasthan), has experience of more than Thirteen years in academics and research. He obtained his PHD degree in computer science from Banasthali Vidyapith (Rajasthan). His areas of research interest in Big Data Analytics, Medical Image Processing, Cloud Computing, Web Based Implementation and Databases. He has published various research papers in peer reviewed Scopus/SCI Indexed journals. Email: gyogesh@banasthali.in

Ethical issues

Not applicable.

Informed consent

Not applicable.

Funding

This study has not received any external funding.

Conflict of Interest

The author declares that there are no conflicts of interests.

Data and materials availability

All data associated with this study are present in the paper.

REFERENCES AND NOTES

1. Abdillah AA, Suwarno S. Diagnosis of diabetes using support vector machines with radial basis function kernels. Int J Technol 2016; 7(5):849-858.
2. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. J King Saud Univ Comput Inf Sci 2013; 25(2):127-136.
3. Alkhatib AJ, Sindiani AM, Funjan KI, Alshdaifat E, Alkhatib A. Skin thickness can predict the progress of Diabetes Type 2: A New Medical Hypothesis 2020; 4:8.
4. Ban HJ, Heo JY, Oh KS, Park KJ. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. BMC Genet 2010; 11:26.
5. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artif Intell Med 2002; 26(1-2):37-54.
6. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, Cata P, Chiovata L, Bellazzi R. Machine learning methods to predict diabetes complications. J Diabetes Sci Technol 2018; 12(2):295-302.
7. Garcia-Carretero R, Vigil-Medina L, Barquero-Perez O. The Use of Machine Learning Techniques to Determine the Predictive Value of Inflammatory Biomarkers in the Development of Type 2 Diabetes Mellitus. Metab Syndr Relat Disord 2021; 19(4):240-248.
8. Gujral S. Early diabetes detection using machine learning: A review. Int J Innov Res Sci Technol 2017; 3(10):57-62.
9. Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020; 8:76516-76531.
10. Hathaway QA, Roth SM, Pinti MV, Sprando DC, Kunovac A, Durr AJ, Cook C, Fing G, Cheuvront T, Grossman J, Aljahli G, Taylor A, Giromini A, Allen J, Hollander JM. Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics. Cardiovasc Diabetol 2019; 18(1):1-16.
11. Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. Int J Eng Res Appl 2013; 3(2): 1797-1801.
12. Plis K, Bunescu R, Marling C, Shubrook J, Schwartz F. A machine learning approach to predicting blood glucose levels for diabetes management. In Workshops at the Twenty-Eighth AAAI conference on artificial intelligence 2014.
13. Repalli P. Prediction on diabetes using data mining approach. Oklahoma State University 2011.
14. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 2010; 10(1):16.
15. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet 2018; 9:515.