Early prediction of diabetes by applying data mining techniques
A retrospective cohort study

Mohammed Zeyad Al Yousef, MBBSa,* Adel Fouad Yasky, MBBSa, Riyad Al Shammari, PhDbc, Mazen S. Ferwana, PhDd

Abstract
Background: Saudi Arabia ranks 7th globally in terms of diabetes prevalence, and its prevalence is expected to reach 45.36% by 2030. The cost of diabetes is expected to increase to 27 billion Saudi riyals in cases where undiagnosed individuals are also documented. Prevention and early detection can effectively address these challenges.

Objective: To improve healthcare services and assist in building predictive models to estimate the probability of diabetes in patients.

Methods: A chart review, which was a retrospective cohort study, was conducted at the National Guard Health Affairs in Riyadh, Saudi Arabia. Data were collected from 5 hospitals using National Guard Health Affairs databases. We used 38 attributes of 21431 patients between 2015 and 2019. The following phases were performed: (1) data collection, (2) data preparation, (3) data mining and model building, and (4) model evaluation and validation. Subsequently, 6 algorithms were compared with and without the synthetic minority oversampling technique.

Results: The highest performance was found in the Bayesian network, which had an area under the curve of 0.75 and 0.71.

Conclusion: Although the results were acceptable, they could be improved. In this context, missing data owing to technical issues played a major role in affecting the performance of our model. Nevertheless, the model could be used in prevention, health monitoring programs, and as an automated mass population screening tool without the need for extra costs compared to traditional methods.

Keywords: data mining, diabetes, diabetes prevention

1. Introduction

Diabetes is a major health problem in Saudi Arabia, with the second-highest rate of diabetes in the Middle East and the seventh highest in the world, with an estimated population of 7 million living with diabetes and more than 3 million with pre-diabetes.[1] The prevalence of type 2 diabetes in Saudi Arabia is 32.8%; however, it is predicted to reach 35.37% in 2020, 40.37% in 2025, and 45.36% in 2030.[2]

Based on data from the Institute for Health Metrics and Evaluation (IHME) in Saudi Arabia, the estimated cost of diabetes in 2014 was 17 billion riyals (US $4.5 billion), with the

This study was approved by the KAIMRC Institutional Review Board (RC 17/248/R) and funded by King Abdulaziz City for Science and Technology (KACST) (Grant number: 1-17-03-025-0003).

Mohammed Zeyad Al Yousef and Adel Fouad Yasky worked as physicians in King Abdullah Medical City, Riyadh, Saudi Arabia. A professor in the Department of Health Informatics, College of Public Health and Health Informatics, King Saud bin Abdul-Aziz University for Health Sciences. Mazen S Ferwana worked as a consultant, in family medicine, and the primary healthcare department in King Abdullah Medical City, Riyadh, Kingdom of Saudi Arabia.

The data that support the findings of this study are available from the King Abdullah International Medical Research Center (KAIMRC), but restrictions apply to the availability of these data, which were used under license for the current study, and are not publicly available.

* Family Medicine, King Abdulaziz Medical City, King Abdullah International Medical Research Center, Riyadh, Kingdom of Saudi Arabia. a Department of Health Informatics, College of Public Health and Health Informatics, King Saud bin Abdul-Aziz University for Health Sciences, Riyadh, Saudi Arabia. b Centre of Excellence in Health Informatics, Riyadh, Saudi Arabia. c Family Medicine and Primary Healthcare Department, King Abdulaziz Medical City, Riyadh, Kingdom of Saudi Arabia.

Copyright © 2022 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial License 4.0 (CCBY-NC), where it is permissible to download, share, remix, transform, and buildup the work provided it is properly cited. The work cannot be used commercially without permission from the journal.

How to cite this article: Al Yousef MZ, Yasky AF, Al Shammari R, Ferwana MS. Early prediction of diabetes by applying data mining techniques: a retrospective cohort study. Medicine 2022;101:29(e29588).

Received: 14 December 2021 / Received in final form: 7 April 2022 / Accepted: 28 April 2022
http://dx.doi.org/10.1097/MD.0000000000029588
expectation that it will increase to 27 billion riyals (US $7.2 billion) in 2030 if undiagnosed people are documented. Moreover, if pre-diabetics were to become diabetic, the cost would increase to 43 billion riyals (USD 11.43 billion). These costs include medications, visit, and laboratory tests, which vary based on the patient's stage and complications. Further, due to the high costs of treatment and the expected growth rate of diabetes, Saudi Arabia will encounter a health and financial dilemma in the near future. However, prevention and early detection can effectively address these challenges and decrease costs by preventing new cases and long-term complications. To this end, the current evolution of information technology and the large amount of data available that could be used by applying data mining should be used.

Data mining is a concept that emerged in the 1990s as a new method for data analysis and knowledge discovery. One definition of data mining is the analysis of large observational datasets to discover unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Another definition is the process of finding previously unidentified patterns and trends in databases and using this information to build predictive models. Data mining has evolved from its beginning to include pattern recognition, clustering, classification, machine learning, artificial intelligence, and visualization.

Unfortunately, the real application of and information on data mining applications in healthcare are usually not found in the scientific literature but on vendor websites as success stories that do not provide the complete technical details of which data mining algorithms were used and how. Examples of success stories include those on healthcare fraud prevention and the maximization of medical revenues by the detection of underdiagnosed patients. Additionally, we found that data mining is being used in healthcare research. In this context, there are 2 major types of predictive analytics: supervised and unsupervised. Supervised learning uses known data or information on a specific problem and produces a predictive model, whereas unsupervised learning does not require previously known data on a certain problem to train its model, but defines clusters or groups.

In clinical medicine, data mining can be used to extract knowledge from large complex datasets assessing disease risks, supporting clinical decisions, and predicting disease development. A systematic review of the application of data mining techniques in the field of diabetes research, which included 17 articles, concluded that data mining is a valuable asset for diabetes researchers because it can unearth hidden knowledge from a large amount of data. Thus, it can significantly help diabetes research and, ultimately, improve the quality of healthcare for diabetic patients.

Multiple studies have applied data-mining techniques to health data to construct predictive models for different diseases. Harper depicted the use of multiple healthcare datasets to compare classification algorithms and predict different health problems. He applied discriminant analysis (DA), regression models (multiple and logistic), Classification and Regression tree-based algorithms (CART), and artificial neural networks to build the model; CART (a decision tree algorithm) achieved the best overall accuracies. Another study was conducted on 395 colorectal cancer patients to predict the 5-year survival rate using 17 variables and comparing 18 algorithms. The results showed an area under the curve of more than 0.9; in other words, the accuracy of the model’s performance was greater than 90%.

Sayad and Halkarnikar built a model to predict heart disease in patients with a sample size of 170 records and 13 attributes. They applied a multilayer perceptron neural network as a training algorithm. Their model achieved 94% accuracy with a sensitivity of 92% and specificity of 92.5%. A study by Daghestani and Alshammari built a diabetes prediction model using 18 attributes from the NGHA database and found 66325 instances (diabetics 64.47%, and non-diabetics 35.53%); however, they excluded pre-diabetics.

In another chronic silent disease, such as hypertension, a study in Qatar found that it is possible to use noninvasive predictors through machine learning to achieve a predictive model that can achieve the targeted screening tool for such diseases. In this context, calculators for predicting and estimating the risk of developing diabetes are becoming more widely used. We found many such tools, but the most frequently mentioned in the literature were the American Diabetes Association (ADA) risk calculator, Cambridge Diabetes Risk Score, and Finnish Diabetes Risk Score (FINDRISC).

Diabetes is one of the main topics of medical research because of its longevity and high cost in the healthcare system. We aimed to build a model to predict diabetic and pre-diabetic patients using demographic information and laboratory tests without the use of diagnostic tests for diabetes, hemoglobin A1c (HbA1c), random blood sugar (RBS), and fasting blood sugar (FBS), while considering that the attributes and variables are related to diabetes and are available in the database. We used a large sample size to include as much information as possible, along with multiple measures to assess the performance of our model.

2. Materials and Method

This study was conducted at the National Guard Hospital in Riyadh, Saudi Arabia. Data sets were collected from 5 hospitals that store information in the NGHA databases in the 3 highest-populated regions: the central region (Riyadh city), the western region (Jeddah and Al Madinah cities), and the eastern region (Al Ahsa and Dammam cities). This was a chart review and retrospective cohort study. All National Guard employees, dependents, and other eligible patients with available data in the NGHA database from January 2015 to January 2019 were included. Several phases were required to achieve the study objective: data collection, data preparation, data mining, model building, and model evaluation and validation.

2.1. Data collection and attribute selection

Initially, we identified the attributes that needed to be extracted from the database by reviewing the literature on diabetes (Fig. 1). The requested attributes are listed in Table 1. Complete blood count (CBC) and basic screening tests were added because almost all patients had available test results in their records. HgA1c, FBS, and RBS were requested only to classify patients as diabetic, pre-diabetic, or non-diabetic. Subsequently, they were removed from the dataset before building the model. However, other important attributes related to diabetes were unavailable (family history of diabetes, hypertension, smoking and alcohol history, and an active or sedentary lifestyle). Initially, we retrieved 1,256,898 records for demographic data (sex, age, and region), 972,239 for vital signs (systolic blood pressure [SBP], diastolic blood pressure [DBP], body mass index [BMI]), and 2,598,103 records for laboratory tests, all of which were in separate files and sheets. Table 1 summarizes the attributes used in this study.

2.2. Data preparation, cleaning, and preprocessing

Data pre-processing is an essential step that affects the prediction quality. It includes missing values, smooth noisy data, identifying or removing outliers, normalization, and transformation. Initially, we combined all data based on the medical record number (MRN) and dates when the demographic and laboratory data were entered, which resulted in the gathering
of 120,758 records. Healthcare data are usually not well organized and have a lot of missing data and noise; therefore, they must be prepared first. To this end, we identified the number of unique MRNs by removing duplicates from the dataset; we had 45,365 unique MRNs. We then extracted a single record with the most complete attributes for each unique MRN from the 120,758 records. The missing data included either demographic information or laboratory tests that were not retrieved properly from the database due to technical issues or were not performed or requested by the medical team for the patient. Demographic information accounted for 15% (6 out of 41) of the attributes (region, sex, age, BMI, SBP, and DBP) in this study. Patients who had 0 of 6 demographic attributes were excluded.

### 2.3. Labeling the patient with the diagnosis

We used the ADA guidelines to diagnose diabetes and pre-diabetes status (Fig. 2)\(^\text{[21]}\) and HgA1c as the main diagnostic test to label patients with diabetes, pre-diabetic, or non-diabetic. If HgA1c was missing, we used it as a diagnostic test. If FBS was missing, we used the RBS, and if RBS was missing, we labeled the patient as “non-diabetic.” The flow charts in Figures 1 and 2 demonstrate how the data were retrieved, cleaned, and prepared for the model, and Tables 2 and 3 show the distribution of the demographic variables and statistical values. Additionally, manual inspection of the data was performed to ensure that the data were consistent and accurate. The MRNs were removed after preparing the data so that no patients could be identified. After preparing the data, we obtained a sample size of 21431 unique records that included 41 attributes. Table 4 presents the number of missing values for each attribute.

### 2.4. Missing values

Missing values were replaced with the mean values of the associated attributes. This step aims to reduce the number of values for continuous attributes. This is achieved by splitting the range of the continuous attribute into intervals. Furthermore, discretization reduces the time needed to build the prediction model and improve the prediction results.\(^\text{[22]}\) The age was thus discretized into 4 groups: group 1, 0–36; group 2, 36–54; group 3, 54–72; and group 4, 72–92.

### 2.5. Sampling

The dataset used in this study consisted of 21,431 unique records with 12,791 who experienced diabetes; however, 4567 did not have diabetes, and 4073 had pre-diabetics. The most common metric used to evaluate machine-learning techniques is accuracy; however, this measure will not work here because of the imbalanced nature of the 3 classes. In general, there are 2 ways to address this issue: oversampling the minority class or sampling the majority class. In this study, we used both undersampling and oversampling to solve the imbalanced data problem and compared the performance of both techniques. To this end, we used the Synthetic Minority Oversampling Technique (SMOTE).\(^\text{[23]}\) The percentage of synthetic examples generated by SMOTE from class “None Diabetes Miletus (DM)” and “Pre-Diabetic” was 100% for both classes, meaning that the number of instances of “None DM” and “Pre-Diabetic” were 9134 and 8146, respectively.

### 2.6. Data mining and building the model

#### 2.6.1. Feature selection

Feature selection is an important aspect of building a high-performance machine learning model, and one of its main benefits is reducing the data dimensionality,
which in turn reduces the number of features used in building the model, and subsequently, the time needed to build the model.\textsuperscript{[24]} This can be achieved by choosing the most important attributes that improve the prediction accuracy. We used an automated R-based machine learning feature selection algorithm that ranks the attributes based on their information gain.\textsuperscript{[25]} It evaluates the importance of an attribute by measuring the entropy gain with respect to the outcome and then ranks the attributes based on their individual evaluations. Figure 3 shows the attributes after feature selection.

In this study, we compared the following 6 different machine learning algorithms:

1. K-nearest neighbors (KNN) are identified from the neighbors with K similar points in the training data that are closest to the test observation. These are then classified by estimating the conditional probability of belonging to each class and choosing the class with the highest probability.\textsuperscript{[26]}

2. The random tree forest (RTF) is a classification algorithm that works by forming multiple decision trees during training and outputting the class that is the mode of the classes (classification) at testing.\textsuperscript{[27]} Decision trees work by learning simple decision rules extracted from data features. The deeper the tree, the more complex are the decision rules and fit of the model. However, random decision forests overcome the problem of overfitting decision trees.

3. Support vector machine (SVM) represents the instances as a set of points of 2 types in an N-dimensional place and generates an (N – 1)-dimensional hyperplane to separate those points into 2 groups.\textsuperscript{[28]} SVM attempts to find a straight line that separates those points into 2 types and is situated as far as possible from all those points.

4. The naïve Bayesian classifier (BC) is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high.\textsuperscript{[29]} Despite its simplicity, naïve Bayes often outperforms more complex machine learning techniques such as SVM.

5. Bayesian network (BN) is a simple probabilistic classifier that is considered a generalization of the naïve BC that removes the dependencies between variables.\textsuperscript{[30]} BN is designed for modeling under uncertainty where the nodes represent variables, and arcs represent direct connections between them. The BN model allows probabilistic beliefs about variables to be updated automatically as new information becomes available.

6. Logistic regression (LR) is a type of regression that is used to predict the outcome of the categorical dependent variable. (i.e., categorical variables have a limited number of categorical values) based on 1 or more independent variables.\textsuperscript{[31]}

All machine learning algorithms were conducted using WEKA software (version 3.8) and R-based machine learning packages (version 3.3.1).\textsuperscript{[32,33]}

2.7. Model evaluation and validation

The model was evaluated using the hold-out method,\textsuperscript{[34]} in which the dataset was partitioned into 2 separate datasets: one for training the machine learning model and another for testing it. For the holdout method, 2 data splits were used: training with 70% of the dataset and testing with 30% of the dataset.

For all classifiers, the following evaluation metrics were calculated: precision, recall, F-score root, mean squared error, and receiver operating characteristic (ROC) curve.

3. Results

The final sample size was 21,431 patients. Finally, 15 out of 41 attributes were used to build the final model, which was chosen based on the information gained by the initial model (Fig. 3). Moreover, in our sample, 46.9% were female, 34.74% were male, and 18.36% were of unknown sex (Table 2).

Table 3 shows pre-diabetic, diabetic, and non-diabetic percentages for all genders, as well as the relevant trends. However, some of the records retrieved had missing attribute values. The 3 highest missing values out of the 15 attributes used in building the model were RBC, triglyceride, and eGFR (Table 4).
Table 5 presents a comparison of the performances of the different classification models with and without using the SMOTE sampling method. The results of this experiment show that BN outperforms all the other classifiers in terms of precision, recall, area under the curve, F-score, root mean square error (RMSE), and accuracy.

Table 6 shows the same comparison but with the SMOTE sampling method. The results of this experiment show better results and that BN outperforms all other classifiers in terms of precision, recall, AUC, F-score, RMSE, and accuracy.

4. Discussion
One of the goals of the Saudi Vision 2030 strategic framework in the healthcare sector is to focus on primary care, preventive medicine, and tackling chronic diseases. However, when considering mass screening for diabetes to identify undiagnosed or at-risk individuals, the costs are always considered. In this context, a study was conducted in Brazil on a population screening program for type 2 diabetes, in which 22 million capillary glucose tests were performed in individuals aged 40 years and older. They concluded that the screening program will yield a large health benefit, but higher costs compared to no screening resulted in US$ 31,147 per quality-adjusted life-year (QALY) gained. Using prediction models on available data from electronic health records to aid in diagnosing and identifying people at risk could be useful for early intervention and cost-effectiveness, as data are already available, and retrieving it will cost much less than conventional mass screening.
We used 5 measures to assess the performance of all the classifiers. We then compared the 6 classifiers with and without SMOTE. As shown in Tables 5 and 6, BN was the best classifier with an AUC of 0.71 and 0.75, RMSE of 0.41 and 0.42, the precision of 59% and 62%, recall of 63% and 66%, the accuracy of 63% and 66%, and F-score of 60% and 61%, respectively. Thus, the results show a fair model despite the missing data in our dataset.

Most prior studies on the prediction of diabetes have focused on discussing the technical aspects of machine learning, data mining, and prediction models. However, clinicians and patients are more concerned about the clinical aspects of the models, such as why the authors chose the attributes, the relevance of attributes to diabetes, and the availability of the attributes in all electronic health records, so it can be generalized to a different population.[4]

In comparison, our study included attributes that are relevant to diabetes or usually available in almost every electronic medical record. We excluded any diagnostic laboratory test that could diagnose diabetes, which would aid the model in identifying patients with the disease and eventually lead to
high-performance bias. This was not the case in the other studies examined. Nevertheless, according to ADA guidelines, there are cut-off points for FBS and RBS levels to diagnose patients as diabetic, pre-diabetic, or non-diabetic, and a model will be able to find patterns similar to the cut-offs, which would increase its accuracy. [33]

Furthermore, most studies on prediction models do not include an important class: pre-diabetics. In our model, we included all 3 groups—diabetic, pre-diabetic, and non-diabetic—which ensures that no patient at risk is left out. Moreover, the data used in most models were from a specific population, whereas our data were collected from multiple hospitals in different cities. Many studies have confirmed that prediction models are a promising method for use in healthcare because of their good performance in prediction. However, not all models use a measure that can assure the significance of their model performance, as well as no bias, under, or overfitting. Table 7 summarizes a comparison of the diabetes prediction models mentioned above.

Comparing our results with the risk calculator tools, we found that the ADA risk calculator had a sensitivity of 79%, specificity of 67%, and positive predictive value of 10%, and was validated in a Chinese population with an AUC of 0.725.[18,36,37] The Cambridge Diabetes Risk score had a specificity of 72%, a sensitivity of 77%, and a likelihood ratio of 2.76. The area under the receiver-operating characteristic curve was 80% and was validated in different populations, but mostly in European countries; the results were lower with an AUC ranging from 63% to 74.5%. [19-31] FINDRISC had a sensitivity of 81%, a specificity of 76%, and a positive predictive value of 0.05. The tool was validated in other populations, and the results showed an AUC of 0.724 and 0.75.[20,42,43] All risk calculator tools were validated and showed good results in identifying at-risk patients. Some of the questions were found to be similar in all risk calculator tools, indicating their importance in identifying at-risk patients. Unfortunately, although we know the importance of these questions and how they will increase the performance of our model, we could not retrieve some of these data because they were not available in the system. On the other hand, comparing the results of identifying patients in our model had an AUC of 0.71 and 0.75, which is promising and similar to the validated risk tool measures; thus, our model could be improved if we added important missing attributes.

Moreover, all risk calculator tools need to be filled manually by a physician, nurse, or person who speaks English and would require either an Internet connection to access it or have it printed. This would lead to increased costs due to time loss, papers used, availability of manpower, and internet connectivity to run these tools. This increase would affect the number of patients assessed using the risk tools. Nevertheless, it is noteworthy that our model was capable of screening 21,431 patients in a couple of minutes at no cost.

Integrating an automated prediction model, such as our model, into the electronic health system would give the healthcare provider a live identification of those at risk of developing diabetes. It can be sent automatically or after the patient’s health information has been reviewed by a healthcare provider, so that proper diagnostic screening tests can be performed. This eventually leads to early identification, intervention, and management for better health outcomes and reduction of disease complications and costs.

### 4.1. Limitations

Because of the unavailability and missing data, as shown in Table 4, our accuracy was affected, which consequently affected the feature selection for the attributes. Nevertheless, more attributes that are clinically related to the disease can be added to increase the possibility of prediction. However, because these data were unavailable in the database, they would either have to be collected manually or the stakeholders in our hospital and the health information system department would have to be contacted to implement some changes to the system that would ensure that these data are available in the future.

Furthermore, we recommend identifying technical problems with the information technology department regarding missing data, adding more attributes that are related and known to increase the risk of developing diabetes from the literature to the prediction model, collaborating with other hospitals to cover more population, increasing the sample size, excluding pediatrics, and identifying better methods for imputing missing data.

### 5. Conclusion

The results from our model are acceptable but can be improved. Missing and unavailable data owing to technical issues play a major role in affecting the performance of our model. Nevertheless, it could still be used in preventive and health monitoring programs and as an automated mass population screening tool to identify diabetics and individuals at high risk of developing diabetes with the available data in the database and without the need for extra costs as compared to traditional methods. Moreover, with the use of better data entry and sorting methods in our health attribution datasets, the implementation of such prediction models would be more effective in the long run, although we have achieved similar outcomes to those of similar models.

---

**Table 7**

A comparison summary of the models.

| Results | Performance measures | Validation and training | # of attributes | class | # of records | Data set | Author |
|---------|----------------------|------------------------|----------------|-------|-------------|---------|--------|
| RF: Recall (90%), precision (68%) | Sensitivity (recall) and PPV (precision) | training/testing, percentage not mentioned | 18 | Dm, Non-Dm | 66,325 | NGHA | Daghistani T[4] |
| PNN: accuracy 81.49% clustering + SVM: Accuracy 98.93%, sensitivity 99.33, specificity 98.73% a and AUC of 0.97 clustering + C4.5: accuracy 92.38 %, sensitivity (90.38), and specificity (93.29). | Accuracy, sensitivity, specificity, AUC | 76% training/25% testing Cross validation | 9 | Dm, Non-Dm | 768 | PIMA | Soltoni Z[18] |
| BN: Precision (62%), Recall (66%), AUC (0.75), F-Score (61%), RMSE (0.42), and Accuracy (66%) | Sensitivity (recall), PPV (precision), AUC, F score, RMSE, accuracy | 70% training/30% testing | 41 | Dm, Pre Dm, Non-Dm | 18,181 | NGHA | Current work |

BN = Bayesian Network, Dm = Diabetic, Non-Dm = Non-Diabetic, PNN = probabilistic neural network, Pre Dm = Pre-diabetic, RF = Random Forest, SVM = support vector machine.
Acknowledgments

The authors would like to thank KAIRMC and KACST for their cooperation and support shown by them during the study. The authors would also like to acknowledge and value the input of Prof. Sherif Sakr as he passed away recently and could not witness the final result.

References

[1] Robert AA, Al Dawish MA, Braham R, et al. Type 2 diabetes mellitus in Saudi Arabia: major challenges and possible solutions. Curr Diabetes Rev. 2017;13:59–64.
[2] Moe SA. Prevalence and future prediction of type 2 diabetes mellitus in the Kingdom of Saudi Arabia: a systematic review of published studies. J Pak Med Assoc. 2016;66:722–5.
[3] El Bcheraoui C, Basulaiman M, Tuffaha M, et al. Status of the diabetes epidemic in the Kingdom of Saudi Arabia, 2013. Int J Public Health. 2014;59:1011–1021.
[4] Daghistani T, Alshammari R. Diagnosis of diabetes by applying data mining classification techniques. Int J Adv Comput Sci Appl. 2016;7.
[5] Yoo I, Alafaireet P, Marinov M, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst. 2012;36:2431–48.
[6] Kincade K. Data mining: digging for healthcare gold. Inf Technol. 1998;23:IM2–IM7.
[7] Eckerson W. Predictive analytics. Transforming data with intelligence. USA: John Wiley & Sons; 2010.
[8] Sato F, Shimada Y, Selaru FM, et al. Prediction of survival in patients with colorectal cancer patients. Acta Inform Med. 2017;25:254–8.
[9] Dong W, Huang Z, Ji L, et al. A genetic fuzzy system for unstable angina risk assessment. BMC Med Inform Decis Mak. 2014;14.
[10] Rahman M, Simmons RK, Harding AH, et al. A simple risk score identifies individuals at high risk of developing type 2 diabetes: the Oxford identifying diabetes risk using clinical features (ODIN) study. Diabetol Metab Syndr. 2015;7.
[11] Poltavskiy E, Kim DJ, Bang H. Comparison of screening scores for diabetes and prediabetes. Diabetes Res Clin Pract. 2016;118:146–53.
[12] Woo YC, Lee CH, Fong CHY, et al. Validation of the diabetes screening tools proposed by the American diabetes association in an aging Chinese population. PLoS One. 2017;12:e0184840.
[13] Kengne AP, Beulens JW, Peelen LM, et al. Noninvasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. Lancer Diabetes Endocrinol. 2014;2:19–29.
[14] Hearst MA, Dumais ST, Osuna E, et al. Support vector machines. IEEE Comput. 1998;31:13–26.
[15] Ho TK. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition IEEE. Wiley-IEEE Press; 2002.
[16] Hearst MA, Dumais ST, Osuna E, et al. Support vector machines. IEEE Intell Syst. 1998;13:18–28.
[17] Webb GI, Boughton JR, Wang Z. Not so naive Bayes: Aggregating one-dependence estimators. Mach Learn. 2005;58:5–24.
[18] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn. 1997;29:131–63.
[19] Balakrishnan N. Handbook of the logistic distribution. New York: Marcel Dekker; 1991.
[20] The R Foundation. The R project for statistical computing. 2020 Oct 29. Available at: https://www.r-project.org/