ISSN 2581 – 3781

AJCSE
ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

Available Online at www.ajcse.info
Asian Journal of Computer Science Engineering 2018;3(4):27-34

RESEARCH ARTICLE

Performance Evaluation of Data Mining Algorithm on Electronic Health Record of Diabetic Patients

Prakash Kuppuswamy[1], Rajan John[2], Shanmugasundaram Marappan[3]

[1]Department of Computer Networks and Engineering, College of Computer Science and Information System, Jazan University, Saudi Arabia, [2]Department of Computer Science, College of Computer Science and Information System, Jazan University, Saudi Arabia, [3]Department of Computer Science, College of Computer Science and Information System, Jazan University, Saudi Arabia

Received on: 10-09-2018; Revised on: 10-10-2018; Accepted on: 10-11-2018

ABSTRACT

Data mining is a process of finding interesting patterns from large databases. One of the important application areas of data mining is health-care sector. In health care, it is not only providing useful information to health-care professionals but it also provides for health insurance companies. Using these techniques, many diseases can be predicted at an earlier stage which gives betterment life for human being. In our research work, we have collected an electronic health record database for a disease of diabetic patients. Every day, the volume of health-care data is increasing. Using data mining techniques extract the knowledge from this enormous database efficiently. Many algorithms are available in data mining. We have used classification algorithms such as One R, Zero R, J48, random forest, and linear discriminate analysis. The performance evaluation of classifiers can be analyzed through confusion matrix and in terms of precision, recall, and error rate.

Key words: Classification algorithms, data mining, decision stump, diabetes, electronic health record, linear discriminate analysis, One R, J48, Zero R

INTRODUCTION

Health-care industry generates large amounts of complex data such as patient history, hospital resources, electronic records, and information about medical devices. These data serves as a key resource to process and analyze for knowledge extraction that enables the decision-making and to save cost. Research using data mining techniques have been applied in the diagnosis of various diseases such as cardiovascular diseases, AIDS, asthma, and diabetes.[1] Diabetes is one of the major health problems of all over the world.[2] Diabetes is a disease that occurs when the insulin production in the body is inadequate, or the body is unable to use the produced insulin in a proper manner; as a result, this leads to high blood glucose. The body cells break down the food into glucose, and this glucose needs to be transported to all the cells of the body. The insulin is the hormone that directs the glucose that is produced by breaking down the food into the body cells. Any change in the production of insulin leads to an increase in the blood sugar levels, and this can lead to damage to the tissues and failure of the organs[3] such as kidney, eye, heart, nerves, and foot.[5] In general, a person is considered to be suffering from diabetes, when blood sugar levels are above normal (4.4–6.1 mmol/L).[3]

Diabetes mellitus is classified into four broad categories: Type 1, type 2, gestational diabetes, and other specific types. All forms of diabetes increase the risk of long-term complications. These typically develop after many years but maybe the first symptom in those who have otherwise not received a diagnosis before that time.[2] Cause of diabetics are not yet entirely understood; scientist believes that both genetic factors and environmental triggers are involved therein.[4] Diabetes can be controlled using different measures such as insulin and diet. For this, it should be identified as early as possible and subsequently provide appropriate treatment. Most of the classifying, identifying and diagnosing

Address for correspondence:
Prakash Kuppuswamy
E-mail: prakashcnet@gmail.com

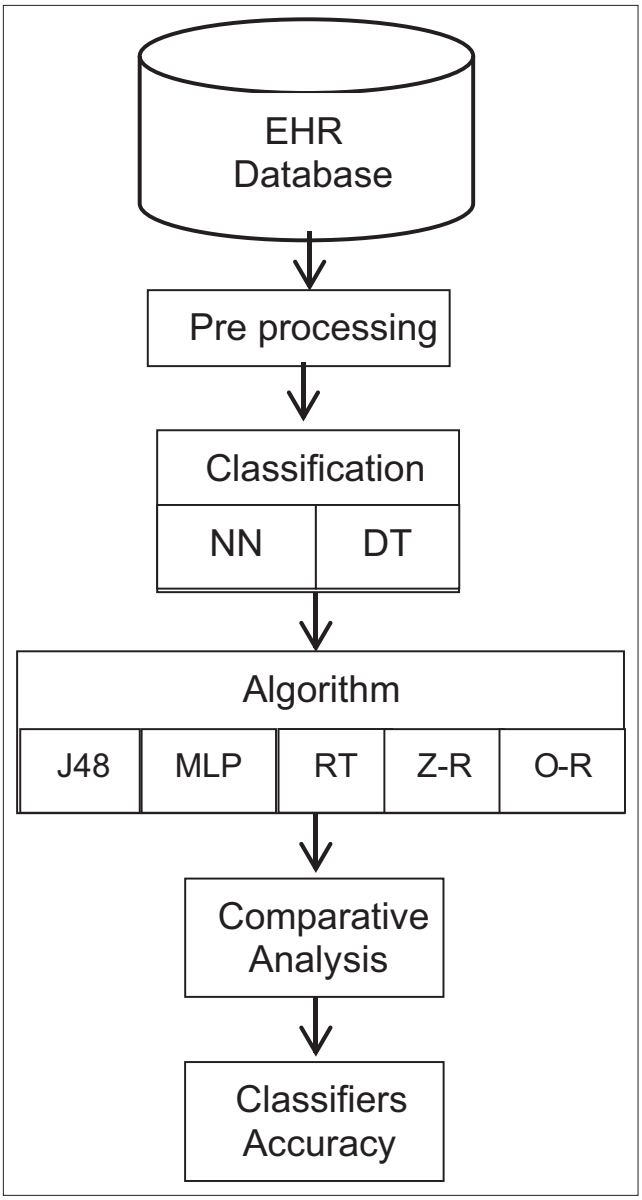

Figure 1: Data processing architecture model

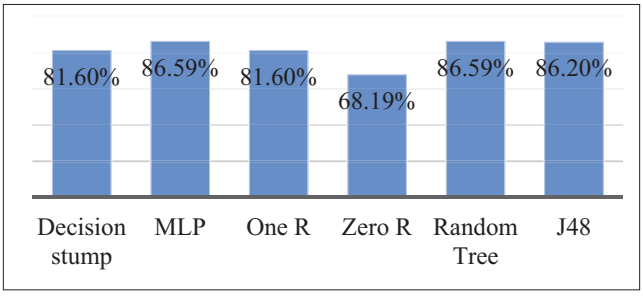

Figure 2: Accuracy of classifiers

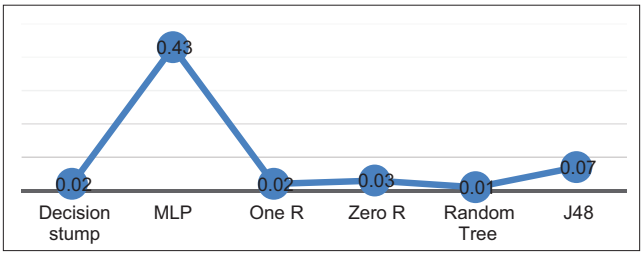

Figure 3: Time taken by the classifiers

treatments are based on chemical and physical tests.[6]

Data mining is the process of selecting, exploring, and modeling large amounts of data. This process has become an increasingly pervasive activity in all areas of medical science research. Data mining has resulted in the discovery of useful hidden patterns from massive databases.[7] It is a multidisciplinary field of computer science which involves a computational process.[5] Data mining is one of the "Knowledge Discovery in Databases" processes. The overall goal of the data mining process is to extract information from a dataset and transform it into an understandable structure for further use. This process has become an increasingly pervasive activity in all areas of medical science research. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization, and statistics, including hypothesis testing, clustering, classification, and regression techniques [Figure 1]. In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, and education.[8]

There are several major data mining techniques have been developed and used in health-care management for, diagnosis and treatment, health care resource management, customer relationship management, and fraud and anomaly detection. Data mining techniques can help physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services [Figure 2]. There are some famous data mining methods are broadly classified as on-Line Analytical Processing, classification, clustering, association rule mining, temporal data mining, time series analysis, spatial mining, and web mining.[9]

The purpose of this study was to compare multiple prediction models for diabetes incidence based on common risk factors. This study developed three widely used data mining classification models, logistic regression, artificial neural networks (ANNs) and decision tree, and along with a 10-fold cross-validation technique. Accuracy, sensitivity, and specificity were used to evaluate them [Figure 3].

The remainder of the paper is organized as follows: In Section 2, the literature review

discussed. In Section 3, the proposed structure and methodology used in this paper have been given. Section 4 deals with result analysis of data mining techniques used along with its standard tasks are presented. Further, we discussed about the results and analysis of the model and conclusion.

Literature review

Aljumah *et al*. discussed predictive analysis of diabetic treatment using a regression-based data mining technique. The Oracle Data Miner was employed as a software mining tool for predicting modes of treating diabetes. Vector machine algorithm was used for experimental analysis. Datasets of non-communicable diseases risk factors in Saudi Arabia were obtained from the World Health Organization and used for analysis. Authors conclude that drug treatment for patients in the young age group can be delayed to avoid side effects. Furthermore, they conclude that elderly diabetes patients should be given an assessment and a treatment plan that is suited to their needs and lifestyles. In this study, predictions on the effectiveness of different treatment methods for young and old age groups were elucidated.[7]

Meng *et al*. in this study authors compare the performance of logistic regression, ANNs, and decision tree models for predicting diabetes or prediabetes using common risk factors. A standard questionnaire was administered to obtain information on demographic characteristics, family diabetes history, anthropometric measurements, and lifestyle risk factors. They used three predictive models using 12 input variables and one output variable from the questionnaire information. This study may assist future researchers in choosing the optimal predictive models for implementing community lifestyle interventions to decrease the incidence of diabetes.[20]

Kumari *et al*., this paper predicting diabetes by applying data mining technique. The discovery of knowledge from medical datasets is important to make an effective medical diagnosis. In this paper, Bayesian Network classifier was proposed to predict the persons whether diabetic or not. The dataset used is collected from a hospital, which collects the information of persons with and without diabetes. Pre-processing is used to improve the quality of data. The techniques of pre-processing applied are attributes identification and selection, data normalization, and numerical discretization. Next, the classifier is applied to the modified dataset to construct the Bayesian model. Weka tool used to do simulation, and the accuracy of the model is calculated and compared with other algorithms efficiency.[21]

Kandhasamy and Balamurali this research study compare the performance of algorithms that are used to predict diabetes using data mining techniques. This article compares machine learning classifiers such as J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines to classify patients with diabetes mellitus. Authors compared four prediction models for predicting diabetes mellitus using eight important attributes under two different situations. One is before pre-processing the dataset. Here, the studies conclude that the decision tree J48 classifier achieves higher accuracy of 73.82% than other three classifiers. After pre-processing, the dataset is given more accurate result when compared to the previous studies. In this case, both KNN (k=1) and random forest performance much better than the other three classifiers and they provide 100% accuracy. From this, we can come to know that after removing the noisy data from our dataset it will provide good result for our problems.[2]

Perveena *et al*. discussed new conventional systems, which are typically, based either just on a single classifier or a plain combination thereof. Recently extensive endeavors are being made for improving the accuracy of such systems using ensemble classifiers. This study follows the adaboost and bagging ensemble techniques using J48 decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors. Decision tree is one of the most powerful and widely applied techniques for classification and prediction which constructed reasonably good models with higher performance to classify diabetic patients, across three age groups in the Canadian population, using bagging Adaboost as well as J48 decision tree. The author informed that, in future, this approach can be applied on other disease datasets such as hypertension, coronary heart disease, and dementia. Furthermore, diverse individual techniques such

as Naïve Bayes, SVM, and neural networks can be incorporated as base learners in an ensemble framework.[4]

Proposed architecture and methodology

Preprocessing

Data preprocessing is a data mining technique. It is used to reduce the data. There are many data reduction techniques which are available such as data compression, numerosity reduction, dimensionality reduction, and discretization. In our proposed method, we have used selection attributes filters for dimension reduction to select the subset of attributes from original data to improve the classifiers accuracy.

Classification

Classification is one of the data mining Techniques. It is used to group the instances which belong to the same class. It is a supervised learning, in which predefined training data are available. Most popular data mining classification techniques are decision trees and neural networks. They are given below.

Neural networks

A multilayer perceptron (MLP) is a feed-forward ANN model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network.[10,11,14]

Decision tree

Decision tree is one of the classification techniques in data mining. It is tree-like graph.[14] The internal node denotes a test on an attribute, each branch represents an outcome of the test, and the leaf node represents classes. It is a graphical representation of possible solutions based on condition from these solutions optimum course of action is carried out. In our work, we have used two decision tree classifier such as decision stumps and J48 to classify the hypothyroid data set. The algorithm of J48 and decision stump is given below.

Algorithms

A. J48 Algorithm

J48 is a tree-based learning approach. It is developed by Ross Quinlan which is based on Iterative Dichotomiser 3 (ID3) algorithm.[14] J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions until leaf node (target node) occurs in tree. Given a set T of total instances the following steps are used to construct the tree structure.[12]

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most frequent class in T.

Step 2: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of each outcome of the test, partition T into corresponding T1, T2, T3., according to the result for each respective cases, and the same may be applied in recursive way to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

Decision stump

A decision stump is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes. Decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rule.[14]

Random tree

Random tree is a supervised classifier; it is an ensemble learning algorithm that generates lots of individual learners. It employs a bagging idea to construct a random set of data for constructing a decision tree. In standard tree every node is split using the best split among all variables. In a random forest, every node is split using the best among the subset of predicators randomly chosen at that node. Random trees have been introduced by Breiman and Adele Cutler. The algorithm can deal with both classification and regression problems. Random trees are a group (ensemble) of tree predictors that is called forest. The classification mechanisms as

follows: The random trees classifier gets the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes." In case of a regression, the classifier reply is the average of the responses over all the trees in the forest. Random trees are essentially the combination of two existing algorithms in Machine Learning: Single model trees are merged with random forest ideas. Model trees are decision trees where every single leaf holds a linear model which is optimized for the local subspace explained by this leaf. Random forests have shown to improve the performance of single decision trees considerably: Tree diversity is created by two ways of randomization.[13,15,20] First, the training data are sampled with the replacement for each single tree-like in Bagging. Second, when growing a tree, instead of always computing, the best possible split for each node only a random subset of all attributes is considered at every node, and the best split for that subset is computed. Such trees have been for classification Random model trees for the first time combine model trees and random forests. Random trees use this product for split selection and thus induce reasonably balanced trees where one global setting for the ridge value works across all leaves, thus simplifying the optimization procedure.[17-19,21]

Zero R

Zero R is the simplest of the rule-based classifiers which rely on the target and ignores all predictors.[14] It simply predicts the majority class. It is based on frequency table. The Zero R classifier takes a look at the target attribute and its possible values. It constructs the frequency table and selects its most frequent value. It will ever output the value that is most frequently found for the target attribute in the given dataset. Zero R as its names suggests; it does not include any rule that works on the non-target attributes. Hence, more specifically it predicts the mean (for a numeric type target attribute) or the mode (for a nominal type attribute).[14]

One R

One R is a short for "One Rule." It is a simple and accurate, classification algorithm that generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule." To create a rule for a predictor a frequency table for each predictor against the target has been constructed. It has been shown that One R produces rules only slightly less accurate than state-of-the-art classification algorithms while producing rules that are simple for humans to interpret.[14]

Experiment and result analysis

The open source software Waikato environment for knowledge Analysis 3.7 (WEKA) is used for the experiment. It is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, feature selection, and visualization. Weka can download from the website.[16]

Performance measure of classifiers

In our experiment, data are supplied to the classifier of One R, Zero R, Random tree, Multi-Layer Perceptron, J48 Algorithm, and decision stump to classify the data. The classifiers performance is evaluated through the confusion matrix.

Confusion matrix

It is used for measuring the performance of classifiers. In the confusion matrix, correctly classified instances are calculated by the sum of diagonal elements true positive (TP) and true negative (TN) and others as well as false positive and false negative are called incorrectly classified instances.

Accuracy

It is defined as the ratio of correctly classified instances to the total number of instances in the dataset.

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

RESULT ANALYSIS

There are 768 records in diabetes dataset. In that 66% of records are taken as training dataset and the remaining 261 records in the diabetes test dataset. All the records are classified as TP and

Table 1: Confusion matrix for decision stump

Target class	Test positive	Test negative
Test positive	161	17
Test negative	31	52

Table 2: Confusion matrix for J48 algorithm

Target class	Test positive	Test negative
Test positive	168	10
Test negative	26	57

Table 3: Confusion matrix for MLP algorithm

Target class	Test positive	Test negative
Test positive	169	9
Test negative	26	57

MLP: Multilayer perceptron

Table 4: Confusion matrix for One R algorithm

Target class	Test positive	Test negative
Test positive	161	17
Test negative	31	52

TN. The following Table 1 represents confusion matrix for decision stump algorithm.

In decision stump classifier, the correctly identified instances are 213 and incorrectly identified instances are 48.

The following Table 2 represents a confusion matrix for J48 algorithm.

In J48 classifier, the correctly identified instances are 225 and incorrectly identified instances are 36.

The following Table 3 represents a confusion matrix for MLP algorithm.

In MLP classifier, the correctly identified instances are 226 and incorrectly identified instances are 35.

The following Table 4 represents a confusion matrix for One R algorithm.

In One R classifier, the correctly identified instances are 213 and incorrectly identified instances are 48.

The following Table 5 represents a confusion matrix for Zero R algorithm.

In Zero R classifier, the correctly identified instances are 178 and incorrectly identified instances are 83.

The following Table 6 represents a confusion matrix for random tree algorithm.

In random tree classifier, the correctly identified instances are 169 and incorrectly identified instances are 57.

The following Table 7 shows the accuracy and time taken to build the model of classifiers before preprocessing.

Table 5: Confusion matrix for zero R algorithm

Target class	Test positive	Test negative
Test positive	178	0
Test negative	83	0

Table 6: Confusion matrix for random tree algorithm

Target class	Test positive	Test negative
Test positive	169	9
Test negative	26	57

Table 7: Accuracy and time taken to build a model of classifiers

Classifier	Accuracy (%)	Time taken to build model (in Sec)
Decision stump	71.87	0.02
MLP	75.39	0.66
One R	71.48	0.03
Zero R	65.10	0
Random tree	68.09	0.03
J48	74.81	0.08

MLP: Multilayer perceptron

Table 8: Accuracy and time taken to build a model of Algorithm

Classifier	Accuracy (%)	Time taken to build model (in Sec)
Decision stump	81.60	0.02
MLP	86.59	0.43
One R	81.60	0.02
Zero R	68.19	0.03
Random tree	86.59	0.1
J48	86.20	0.07

MLP: Multilayer perceptron

The following Table 8 depicts detailed accuracy and time taken to build the model of classifiers after preprocessing.

In this chart, X-axis represents the algorithm and Y-axis represents the accuracy. It shows that the accuracy of random tree is 86.59% which is more than other classifiers.

The following chart 2 shows the time taken by the classifiers.

In this chart, X-axis represents the algorithm and Y-axis represents the time to build the model. It shows that the time to build the model of the random tree is 0.01 s which is less than other classifiers.

DISCUSSION

In WEKA, there are many classification techniques available. These classification techniques are

used to diagnose diabetes and some other clinical diagnosis issues. Studies showed that many researchers used different methods to diagnose the diabetes and achieved the high accuracy of classifiers for the dataset is taken from UCI machine learning repository. (Murat Koklu and Yavuz Unal (2013), proposed MLP, J48, Naive Bayes classification algorithm on diagnosing diabetes they showed that the prediction accuracy of Naïve Bayes classifier is 76.3%. Kumar. and Anandakumar (proposed C4.5 and incremental classification algorithm in which C4.5 gives prediction accuracy of 68%. In our study, we have used the selection attribute filter for dimensionality reduction to select the subset of attributes from original data to improve the performance of the classifier. We proposed five classification algorithms namely J48, One R, Zero R, MLP, and random tree to diagnose diabetes. The performances of classifiers are evaluated through the confusion matrix. In this experiment, the classifier random tree is giving high accuracy of 86.59% and minimum time to build the model of 0.01 s which is better than other algorithms.

CONCLUSION AND FUTURE SCOPE

Identification and diagnosis of disease are a very challenging task in the field of health care. In general, various data mining techniques are used in decision-making process. In our proposed method, we have used attribute selection filter to select the subset of attributes from original data, and then we have applied J48 and decision stump data mining classification techniques which are used to predict diabetes. The performances of classifiers are evaluated through the confusion matrix in terms of accuracy and execution time. The random tree algorithm gives 86.59% which is providing better accuracy than other classifier's accuracy and also random tree algorithm takes very minimum time to classify data sets than other classifiers. The same procedure is used to apply for other disease datasets such as kidney disease, cervical cancer, heart disease, breast cancer, lung cancer, and so on.

REFERENCES

1. Thirumal PC, Nagarajan N. Utilization of data minining techniques for diagnosis of diabetes mellitus a case study. ARPN J Eng Appl Sci 2015;10:12-15.
2. Kandhasamy JP, Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science 2015;47:45-51. Available from: http://www.sciencedirect.com. [Last accessed on 2001 Oct 24].
3. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabeters using classification mining techniques. Int J Data Min Knowl Manag Process 2015;5:1-14.
4. Perveena S, Shahbaza M, Guergachib A, Keshavjeec K. Symposium on Data Mining Applications. SDMA 2016. Riyadh, Saudi Arabia: Performance Analysis of Data Mining Classification Techniques to Predict Diabetes; 2016.
5. Devi MR, Shyla JM. Analysis of various data mining techniques to predict diabetes mellitus. Int J Appl Eng Res 2016;11:727-30.
6. Vijayan VV, Ravikumar A. Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. Int J Comput Appl 2014;95:975-8887.
7. 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;25:127-36.
8. Kumar DA, Govindasamy R. Performance and evaluation of classification data mining techniques in diabetes. Int J Comput Sci Inf Technol 2015;6:1312-9.
9. Frank XR. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington, DC: Spartan Books; 1961.
10. David ER, Hinton GE, Williams RJ. Learning internal representations by error propagation. Rumelhart DE, McClelland JL, PDP Research Group, editors. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1. United States: MIT Press; 1986.
11. Han J, Micheline K. Datamining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publisher; 2009.
12. Wikipedia Contributors Random Tree Algorithm Wikipedia. The Free Encyclopedia. Wikimedia Foundation; 2015. Available from: http://www.en.wikipedia.org. [Last accessed on 2015 Nov 02].
13. Breiman L, Random Forests. Mach Learn 2001;45:5-32. Available from: http://www.cs.waikato.ac.nz/ml/weka. [Last accessed on 2001 Oct 24].
14. Breiman L, Random Forests. Mach Learn 2001;45:5-32. Available from: http://www.cs.waikato.ac.nz/ml/weka. [Last accessed on 2001 Oct 24].
15. Andy L. Documentation for R Package Random Forest; 2006. U.S. Trademark Registration No. 3185828. Available from: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. [Last accessed on 2013 Mar 15].
16. Yali A, Donald G. Shape quantization and recognition with randomized trees (PDF). Neural Comput 1997;9:1545-88.
17. Landwehr N, Hall M, Frank E. Logistic model trees. Mach Learn 2005;59:161-205.

18. Koklu M, Unal Y. A New Approach to Classification Rule Extraction Problem by the Real Value Coding. Anal Int J Appl Eng Res 2016;11:727-30.
19. Kumar UM, Anandakumar KR. Predicting early detection of cardiac and diabetes symptoms using data mining techniques. Int Conf Comput Des Eng 2012;49:22-6.
20. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 2013;29:93-9.
21. Kumari M, Vohra R, Arora A. Prediction of diabetes using bayesian network. Int J Comput Sci Inf Technol 2014;5:5174-8.