International Journal of Database Theory and Application
Vol. 11, No.3 (2018), pp. 39-48
hhtp://dx.doi.org/10.14257/ijdta.2018.11.3.04

Decision Support System for Prognosis of Diabetes Using Non-Clinical Parameters and Data Mining Techniques

Mirza Shuja[1*], Sonu Mittal[2] and Majid Zaman[3]

[1,2]*School of Computer and Systems Science, Jaipur National University, Jaipur, Rajasthan, India*
[3]*Directorate of IT&SS, University of Kashmir, Srinagar J&K, India*

Abstract

The main objective of this research was at designing a predictive model using data mining tools and technique which will not require lab parameters for prognosis of diabetes the most appalling disease mankind is currently facing. For this study we obtained a primary dataset from leading diagnostic centers in Kashmir valley (J&K). After preprocessing the data for removal of inconsistencies we applied data rebalancing algorithm SMOTE to remove class imbalance. The features that were used as input all were non-clinical attributes: age, waist, BMI, systolic, diastolic blood pressure, gender, family history and additional diagnostic variable class was added. J48 decision tree was using WEKA software to develop the model. After preprocessing the dataset with 734 records for data mining, we applied the required methods (viz J48 with SMOTE). The developed models accuracy was about 85.56% while as precision of the model is 82.86%. The ROC curve area was about 89.2% which was indicative of model being decent in prediction of disease. A predictive model using SMOTE and decision tree was developed for prognosis of diabetes. The developed model is special in an approach as it doesn't necessitate lab parameter tests for prognostic purposes and predicts the disease with decent accuracy rate.

Keywords: *Diabetes, Decision Tree, Prognosis, Class imbalance, SMOTE, WEKA*

1. Introduction

Data mining is an intense new innovation to find knowledge hidden inside the huge amount of the data. Data mining is considered as an imperative sub-field in knowledge management. Today, Data mining enables diverse association to center around the data rich in the information they have gathered about the conduct of their customer's. In recent years, research in data mining keeps developing in different fields of association, for example, Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition, business, training, restorative, logical and so on. Data mining is amongst the practical branches of Artificial Intelligence that discovers the covert patterns by looking for relations amid features in large databases. The discovered patterns should be significant and must be advantageous which include economic one [1].

Diabetes mellitus is an interminable ailment and most common amongst the most widely recognized endocrine disorders that include 90 to 95 percent of diabetic patients. A number of intelligent algorithms have been proposed by researchers in healthcare domain which are aimed at detection of disease. Diabetes, cancer detection, heart disease has remained the center of attention among research community. Diabetes Mellitus is a disorder portrayed by ceaseless hyperglycemia with unsettling influences of sugar, proteins and fat digestion coming about because of deformities in insulin emission, insulin

Received (April 10, 2018), Review Result (July 6, 2018), Accepted (July 19, 2018)
* Corresponding Author

ISSN: 2005-4270 IJDTA

activity, or both. Thirst, Polyurea, obscuring of vision, and weight loss are some of the main symptoms characterizing it. The stringent form of diabetes can give rise to ketoacidosis (also known as non-ketonic hyperosmolar) condition where acids known as ketones buildup in body. This condition is a very serious and can prompt semi-consciousness or unconsciousness and if left untreated can lead to death. The indelible effect of diabetes can lead to serious medical complications that include: retinopathy which has a potential to cause permanent blindness; nephropathy that causes kidney failure; neuropathy that can cause foot ulcers which results in amputation on foot joint; Charcot joint or neuropathic osteoarthopathy and sexual dysfunction leading to impotency. People suffering with diabetes are at high risk of cardio-vascular disease, peripheral vascular disease and cerebrovascular disease. The diabetes is a chronic disease which is characterized by increased blood sugar level. This increase can be caused by two reasons 1. The body does not produce sufficient insulin or no insulin at all, this condition is called as type I diabetes. 2. The cells in the body do not respond to insulin produced and the condition is called as type II diabetes, the insulin is an important hormone that regulates energy in our body.

The occurrence of diabetes mellitus has increased rapidly in recent years primarily because of changed lifestyle. According to annual statistics of WHO of year 2017, 422 million people are suffering from diabetes worldwide, among them 90 percent of them are suffering from type II diabetes mellitus with about 1.6 million deaths worldwide were directly attributed to diabetes [2]. Type II diabetes also called as adult onset diabetes is most common type of diabetes which is characterized by an asymptomatic stage between real beginning of diabetes and clinical diagnosis and stage is termed as pre-diabetes. As reported by [3] about 30-80 percent cases of people suffering from type II diabetes mellitus remain undiagnosed. On time diagnosis and prevention results in decreased mortality rate and prevents and decreases the complications caused by diabetes which results in improved quality of life [4]. The principle challenge in diabetes screening nonetheless, is the need of studying and collection of blood samples of many people which is a costly affair both financially and workforce related assets which are beyond capabilities of healthcare system, particularly in emergent nations.

By utilizing the capabilities of data mining and knowledge discovery which have the potential of identification of latent patterns that are associated with diagnostic decisions hidden within the data and can aid in prognosis and diagnosis of diabetes. A number of intelligent algorithms have been proposed by researcher in healthcare domain that are aimed at revealing of diseases, cardiac diseases, cancer detection and diabetes have remained center of attention among research community. For application of data mining and knowledge discovery the precondition of huge quantity of subject-matter associated data needs to be saved in databases. At present time the significance of data saving has gained much importance. The patients are provided with the electronic records by hospitals or healthcare centers and this is considered as an approach to enhance the level of health of general population in public arena. The argument about the setting up and creating the databases of electronic health records (EHR) would pave means for new examinations and studies in healthcare issue in the public arena, even in third world countries [5].

In this paper we presented the joint implementation of decision tree and SMOTE on dataset obtained from patients examined in Kashmir valley. The SMOTE was applied for reduction of class imbalance and for enhancement of prediction accuracy. The developed model uses only non-clinical attributes for prognosis of diabetes. The attributes used are void of requirement for patient to go through agonizing needle tests. During the clinical examination of patients we acquired a lot data. To develop a computer based model which will enable high probability recognition of diabetic condition would provide an efficient support for decision making in healthcare system.

2. Related Work

Medicinal data mining discovers concealed patterns from datasets proficiently and precisely. These patterns would then be able to be used for ailment finding and treatment. Following examination strategy centered on utilizing diverse data mining systems for restorative datasets. Following are some of the research methods that focus on applying data mining techniques on medical data.

In their research work [6] explored the possibility of people in certain age groups getting affected by diabetes in view of their life style. Not only this, they also explored the factors which are responsible for individuals to be diabetic.

To reveal the highlighting features that are commanding in taking the decision about presence of diabetes, data mining techniques like clustering and attribute oriented induction techniques were applied by [7], the study also tracked as to why maximum number of women experience diabetes and what are the key features that are associated with these women.

A hybrid predictive model for prognosis of diabetes was proposed by [8] by using classification algorithms viz SMO, Bagging, J48, Naïve Bayes, AdaBoost and random forests and combined them with K-means clustering to predict positive and negative cases of diabetes.

[9] Applied association rules and various classification techniques on diabetic data for supporting diabetic decision making.

[10] Applied two popular machine learning techniques Support Vector Machines (SVM) and Artificial Neural Network (ANN) for prediction of pre-diabetes in Korean population.

In [11, 12] researchers applied SVM (Support Vector Machine) on diabetic data to predict diabetes, however prediction accuracy was varying with formal having 94% accuracy and later-one having 78% accuracy.

Researcher in [13] applied K-means and KNN (K-nearest neighbor) algorithms on diabetic data and cataloged diabetic patients. The patients were classified using the results obtained by employing of the said algorithms. The proposed system had an accuracy of 82%.

[14] In their research work improved the prediction rate of diabetes by applying Fuzzy logic. The system predicts the diabetes based on knowledge about patients diagnosis and experience. By applying IF-THEN rule data was converted into fuzzy data and fuzzy input was improved to fuzzy output.

[15] Developed diabetic prediction system and awareness system which was based on ID3 classification algorithm. The system had an accuracy of 94%.

An effective machine learning algorithm for classification of diabetes mellitus is proposed by [16], the algorithm finds the “optimal hyperplane” that divides various classes.

3. Methods and Materials:

Most of the work that is available in literature is based on Pima Indian diabetic dataset and hence all have same attributes and similar conclusions. The highlights utilized as a part of this investigation for T2DM forecast were in fact the same as those utilized as a part of diabetes screening and suggested by most understood experts. Earlier investigations additionally have affirmed that these highlights are essential indicator factors.

BMI and age were recognized as fundamental indicative factors by studies that distinguish and score the primary factors that influence the advancement of diabetes. Blood pressure, family history of Diabetes and sex were categorized as the high risk markers for prognosis undiscovered diabetes. For data mining on diabetes mellitus Pima Indians dataset has been used extensively. Because of the inclusion of plasma glucose and

serum insulin level in past experiments on Pima Indian dataset the developed models had high precision levels. According to [17] determination of endocrine disease isn't yet settled by just utilizing symptoms rather, and conclusion is made by estimating the secretion of hormone levels. The crucial examination of diabetes relies upon a couple of strategies for assessing the plasma glucose or serum insulin level, insisting that the assurance and usage of the pointer features requires more consideration. Classification technique of data mining have been extensively used in building of prediction models in diverse studies concerning healthcare [18], the competence of whose has been established in discovery of patterns and relationships amongst the features within the database and using these discovered patterns and associations for diagnostic and prediction purposes [19]. The capabilities of classification techniques for diagnosing diabetes mellitus have not thus far been wholly been demonstrated. The decision tree is one of the powerful classification methods, which has been exceedingly used in medical domain for studies allied to diagnosis and prediction; however, its application for prediction has extensively increased [20].

3.1. Dataset Description:

Clinical dataset containing the records of 734 patients that was collected from leading medical labs in Kashmir valley was used in this research. The dataset was collected under expert medical supervision. The privacy and anonymity of the patients was properly taken care of during dataset compilation. The attributes that are used in this work for prediction of type II diabetes mellitus are same as used in diabetic screening and are recommended by authentic authorities like American Diabetic Association [21] and National Diabetes Information Clearinghouse (NDIC) [22]. The attributes used in this research work are purely non-clinical parameters viz parameters which don't require patient to undergo painful needle tests that is why we removed the attributes plasma_glucose_fasting, Plasma_glucose_PP and HbA1c from the original dataset. After removal of these three attributes we were left with only 8 attributes that include body mass index (BMI), Systolic blood pressure, And Diastolic Blood Pressure, waist thickness and family history of diabetes and an additional classifying attribute to indicate diabetic or non-diabetic. The description of dataset is given in Table 1.

Table 1. Attributes used in our Dataset

Serial	Attributes	Description	Type
1	Age	Age of Patients in years	Numeric
2	Waist	Waist measurement	Numeric
3	BMI	Weight in kg's/height in m^2	Numeric
4	Systolic	Systolic blood pressure	Numeric
5	Diastolic	Diastolic blood pressure	Numeric
6	Gender	Gender of patient	Nominal
7	History	Family history	Numeric
8	Class	Diagnosis of disease	Nominal

3.2. Weka

To implement our method, we used WEKA toolkit. It is machine learning software which is developed at university of Waikato New Zealand. The tool is open source

software and is freely available under GNU (General Public Licenses) [23]. The Weka tool has a rich collection of standard machine learning methods that are applied on datasets large enough to be analyzed manually to obtain Knowledge [24].

3.3 SMOTE:

Class imbalance at present is one of the most sought-off topics in data mining research. The condition arises whilst number of instances of one class outnumbers another class. Majority of the data mining algorithms used in medicinal diagnosis tend to function fine when supplied with uniformly dispersed dataset, commonly the supplied dataset is not uniformly dispersed viz, one class tend to have numerous instances than another consequently leading to data imbalance [25]. This impasse of imbalanced data is awfully prevalent in medicinal dataset. This imbalancing datasets causes many hindrances and difficulties' for machine learning and data mining techniques in their performance capabilities. As data mining strategies achieve knowledge from accessible diagnostic data and after that this extricated knowledge is utilized for forecast of disease. But when data applied is imbalanced, this condition leads to poor prognostic accurateness for minority class while as it produces elevated accurateness for majority class. To augment predictive accuracy and to beat the bias of model towards the majority class due to data imbalance we applied SMOTE (Synthetic Minority Oversampling Technique), it is an oversampling technique that work at data level as proposed by Chawla *et.al.*, [26]. By applying this approach the instances in minority class within the original dataset are increased, by formation of new synthetic instances that are created with specification of two parameters 1st oversampling rate 2nd number of nearest neighbors (k).

3.4. Decision Tree

Decision tree is one of the most popular classification techniques and is most useful for classification problems. The technique constructs a binary tree for modeling classification processes. The tree has three types of nodes: root node, child node, leaf node. The rules generated by decision tree are easy to interpret and almost effortless in understanding. Due to its simplistic nature the algorithm helps in valuing the elementary methods that separates the samples in distinct classes. C4.5 decision tree algorithm is well established and widely used for classification problems. To craft a decision about the most biased element at each progression, the algorithm uses information gain ratio principle for its decision tree induction process, at every round of determination, the information gain ration chooses the characteristic with maximum ration of its gain divide by entropy amongst features that had an average or better information gain. C4.5 stops building sub-tree when:

1. An obtained data subset contains samples of just distinct class.
2. No more features are available (in that case leaf node is labeled as majority class).
3. When number of samples contained in obtained subset is less than a specific threshold (in that case leaf node gets labeled by the majority class) [27].

4. Experimentation

We carried out experiment on diabetic dataset by applying decision tree classifier J48 combined with SMOTE for enhancing the classification accuracy by reducing class imbalance of our predictive model. Following attributes were used as input for decision tree classifier: Age, Sex, Systolic blood pressure, Diastolic blood pressure, Diabetic history within family, and Body mass index (BMI), attribute class was used as diagnosis feature. Description of attributes is given in Table 1. Clinical parameters fasting, post-Prandial and HbA1c levels were omitted. We calculated BMI attribute using height and

weight of patient. At the end of experiment, the performance evaluation of our model was carried out.

4.1. Performance Evaluation of Proposed Model

4.1.1. K-fold Cross Validation

To evaluate the performance of developed model, K-fold cross validation method was used to having a good measure of performance of the model [28]. The technique divides the dataset into 'K' subsets and repeats the holdout method 'K' times at each iteration one among 'K' subsets is used as test set while other K-1 subsets are grouped forming a training set. At the end an average error of all 'K' iterations is calculated thus giving the test accuracy of our model. One of the most significant advantages of this technique is that, how data is divided is of least significance.

4.1.2. Accuracy, Specificity, Sensitivity

Every single prediction has four attainable outcomes viz True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). True Positive (TP) and True Negative (TN) are correct classifications. True Positive occurs when the sample is predicted positive and is actually positive. False Positive occurs when the sample is predicted as positive but is actually negative. True Negative occurs when sample is predicted as negative and actually is negative. False Negative occurs when sample is predicted as negative but actually is positive. For this study we used: 1. Accuracy. 2. Specificity. 3. Sensitivity. 4. Precision; equations for evaluation and analysis [28].

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

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$Precision = \frac{TP}{TP + FP}$$

5. Discussion

After proper data preprocessing and preparation, clinical attributes viz the attributes that need blood samples were removed from dataset with the aim to develop a model devoid of requirement of any blood test. The combination of Decision tree and SMOTE was classified, trained and tested on diabetic dataset. The obtained test classification accuracy of method was 85.5491% by using 10-fold cross validation. The obtained accuracy, sensitivity and specificity of the proposed model are 85.5491%, 82.09% and 87.94% respectively and are shown in Table 3. The detailed accuracy of our model is given in Table 5 while as Table 4 depicts the overall error report. Our model had the ROC value of 0.892 see Figure 1.

Table 3. Performance Measure

Metrics	Value
Accuracy	85.5491%
Sensitivity	82.09%
Specificity	87.94%

Table 4. Error Report

Statistic	Value
Kappa	0.7018
Mean absolute error	0.169
Root mean squared error	0.3492
Relative absolute error	34.8263%
Root relative absolute error	70.8984%

Table 5. Detailed Accuracy

Classifier	TP rate	FP rate	Precision	Recall	ROC
Values	0.855	0.155	0.855	0.855	0.892

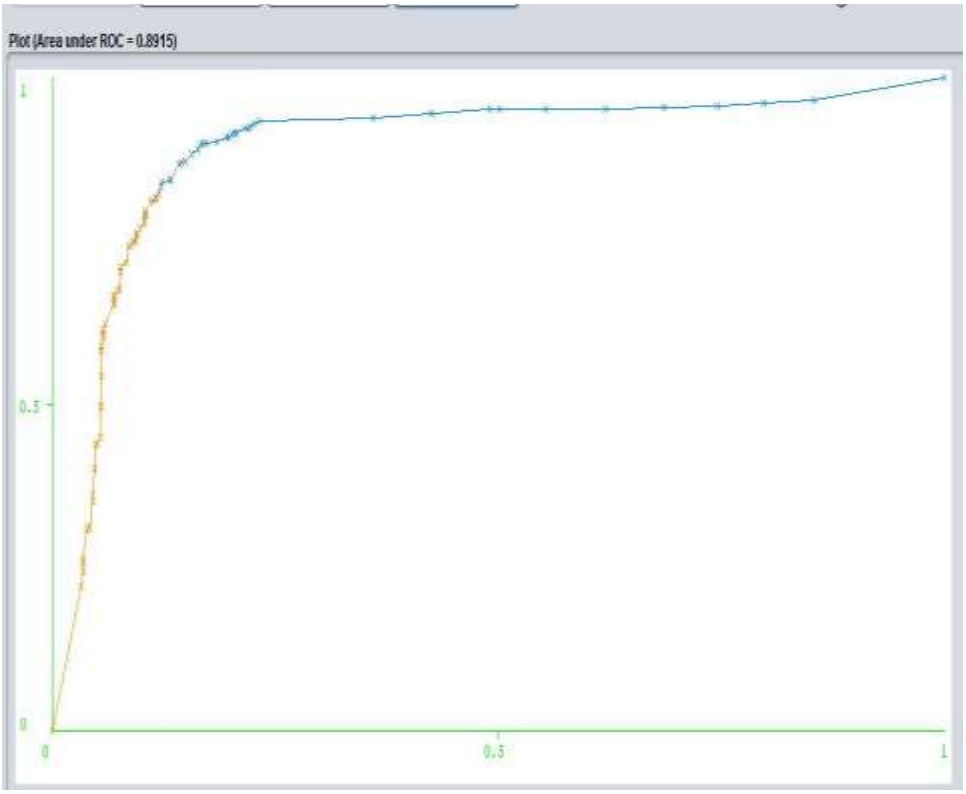

Figure 1. ROC Curve Performance

Although prior studies have reported higher accuracies and precision in some cases perhaps because we omitted the clinical features. In current study the decision tree was trained and tested using real world dataset and it was established that decision tree is effective in prediction of diabetes and could aid in automated screening of patients by automating screening with electronic system.

6. Conclusion

A predictive model for the prognosis of diabetes mellitus was developed in this study. The model used J48 decision tree and SMOTE. The developed model doesn't necessitate any blood or lab parameter tests to diagnose diabetes. The proposed model can be differentiated from prior ones for the following reasons: 1. a real life dataset was used for

study; 2. Primary screening features were applied without requirement for patient to undergo painful blood tests. 3. Data imbalance was reduced by applying SMOTE rebalancing algorithm to enhance prediction accuracy of decision trees. Although the precision and sensitivity of our model is on lower side as lab parameters have been removed, but we believe the model could be step forward in early prognosis of diabetes without the requirement for patient for undergoing painful blood tests.

References

[1] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, **(2005)**.
[2] http://www.who.int/mediacentre/factsheets/fs312/en/.
[3] N. Brown, J. Critchley, P. Bogowicz, M. Mayige and N. Unwin, "Risk scores based on self-reported or Available clinical data to detect undiagnosed type 2 diabetes: a systematic review", Diabetes research and clinical practice, vol. 98, no. 3, **(2012)**, pp. 369-385.
[4] E. W. Gregg, L. S. Geiss, J. Saaddine, A. Fagot-Campagna, G. Beckles, C. Parker and W. Visscher, "Use of diabetes preventive care and complications risk in two African-American communities", American journal of preventive medicine, vol. 21, no. 3, **(2001)**, pp. 197-202.
[5] J. E. DeVoe, R. Gold, P. McIntire, J. Puro, S. Chauvie and C. A. Gallia, "Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers", The Annals of Family Medicine, vol. 9, no. 4, **(2011)**, pp. 351-358.
[6] P. Repalli, "Prediction on diabetes using data mining approach", Oklahoma State University, **(2011)**.
[7] S. W. Purnami, S. P. Rahayu and A. Embong, "Feature selection and classification of breast cancer diagnosis based on support vector machines", In Information Technology, 2008. ITSim 2008. International Symposium on IEEE, vol. 1, **(2008)** August, pp. 1-6.
[8] P. Hemant and T. Pushpavathi, "A novel approach to predict diabetes by Cascading Clustering and Classification", In Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on IEEE, **(2012)** July, pp. 1-7.
[9] S. M. Nuwangi, C. R. Oruthotaarachchi, J. M. P. P. Tilakaratna and H. A. Caldera, "Utilization of data mining techniques in knowledge extraction for diminution of diabetes", In Information Technology for Real World Problems (VCON), 2010 Second Vaagdevi International Conference on IEEE, **(2010)** December pp. 3-8.
[10] S. B. Choi, W. J. Kim, T. K. Yoo, J. S. Park, J. W. Chung, Y. H. Lee and D. W. Kim, "Screening for prediabetes using machine learning models", Computational and mathematical methods in medicine, **(2014)**.
[11] R. Aishwarya and P. Gayathri, "A Method for Classification Using Machine Learning Technique for Diabetes", International Journal of Engineering and Technology, vol. 5, no. 3, **(2013)**, pp. 2903-2908.
[12] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine", International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1797-1801.
[13] A. G. Karegowda, M. A. Jayaram and A. S. Manjunath, "Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients", International Journal of Engineering and Advanced Technology, vol. 1, no. 3, **(2012)**, pp. 147-151.
[14] V. Jain and S. Raheja, "Improving the prediction rate of diabetes using fuzzy expert system", IJ Information Technology and Computer Science, vol. 10, **(2015)**, pp. 84-91.
[15] S. P. Shetty and S. Joshi, "A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique", **(2016)**.
[16] C. Thiyagarajan, D. K. A. Kumar and D. A. Bharathi, "A survey on diabetes mellitus prediction using machine learning techniques", vol. 11, **(2016)**, pp. 1810-1814.
[17] J. Larry Jameson, A. S. Fauci, D. L. Kasper, S. L. Hauser, D. L. Longo and J. Loscalzo, "Harrison's principles of internal medicine", New York: McGraw-Hill Medical, The McGraw-Hill Companies, pp. 2393-2404.
[18] R. Bellazzi, F. Ferrazzi and L. Sacchi, "Predictive data mining in clinical medicine: a focus on selected methods and applications", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no. 5, **(2011)**, pp. 416-430.
[19] S. Upadhyaya, K. Farahmand and T. Baker-Demaray, "Comparison of NN and LR classifiers in the context of screening native American elders with diabetes", Expert Systems with Applications, vol. 40, no. 15, **(2013)**, pp. 5830-5838.
[20] S. H. Liao, P. H. Chu and P. Y. Hsiao, "Data mining techniques and applications–A decade review from 2000 to 2011", Expert systems with applications, vol. 39, no. 12, **(2012)**, pp. 11303-11311.
[21] American Diabetes Association, Screening for type 2 diabetes. Diabetes care, 27, S11, **(2004)**.
[22] National Diabetes Information Clearinghouse (NDIC) (Producer). (2014) Am I at risk for type 2 diabetes? Retrieved from http://www.diabetes.niddk.nih.gov/dm/pubs/riskfortype2/index.aspx.

[23] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald and D. Scuse, "WEKA Manual for Version 3-7-8", **(2013)**.
[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update", ACM SIGKDD explorations newsletter, vol. 11, no. 1, **(2009)**, pp. 10-18.
[25] N. V. Chawla, A. Lazarevic, L. O. Hall and K. W. Bowyer, "SMOTE Boost: Improving prediction of the minority class in boosting", In European Conference on Principles of Data Mining and Knowledge Discovery, Springer, Berlin, Heidelberg, **(2003)** September, pp. 107-119.
[26] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, vol. 16, **(2002)**, pp. 321-357.
[27] J. R. Quinlan, "C4.5 programs for machine learning", San Mateo, CA: Morgan Kaufmann Publishers, **(1993)**.
[28] D. Delen, G. Walker and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial intelligence in medicine, vol. 34, no. 2, **(2005)**, pp. 113-127.

