Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms

Masoud M. Hassan
University of Zakho
6 PUBLICATIONS **7** CITATIONS

SEE PROFILE

Najmeh Njmh Amiri
Shahid Beheshti University
1 PUBLICATION **2** CITATIONS

SEE PROFILE

IV.International Conference on
Theoretical and Applied Computer Science and Engineering
(ICTACSE, 2019)
İstanbul, Turkey, October 11-12, 2019

Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms

Masoud Muhammed Hassan
Computer Science Department
University of Zakho
Zakho, Iraq
Masoud.Hassan@uoz.edu.krd

Najmeh Amiri
Computer Science Department
Shahid Beheshti University of Tehran
Tehran, Iran
amiri.najmeh@gmail.com

Abstract—Diabetes is a widely distributed disease in the world. It is a chronic disease characterized by hyperglycemia. This disease causes various complications and high mortality and morbidity rate. Early diagnose of diabetes is very crucial for timely treatment. There are different ways to detect the diabetes; one of which is the use of machine learning algorithms. Machine learning techniques have been widely used in medical health field to diagnose and predict the occurrences of diseases. In this study, six algorithms, such as Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (K-NN), Naïve Bayes (NB), Support Vector Machine (SVM) and Artificial Neural Networks (ANN) were used to diagnose and classify type II diabetes of Pima dataset. However, classes are not always balanced, and imbalanced data occurs when one of the classes is minority and the other one is majority. The main issue of imbalanced data is usually resulted in misclassification where the minority class tends to be misclassified. To overcome this problem, different methods can be used. In this paper, we investigated the use of SMOTE resampling method to balance the data, and the results of the six algorithms were compared for balanced and imbalanced data. Experimental results showed that classification with resampling/ balancing has significant improvement up to 20% of accuracy for some models.

Keywords— Diabetes; Machine Learning; Classification; Imbalanced Class; Support Vector Machine; Naïve Bayes.

I. INTRODUCTION (HEADING 1)

Diabetes is considered as one of the most common diseases worldwide. Diabetic patients usually suffer from high blood sugar. The main distinction merit of this disease are over-eating and over-drinking. The major cause of diabetes is the reduction of insulin production in the body as well as losing the efficiency of insulin in the metabolism of sugar [1]. Diabetic patients might have serious complications such as blindness, nerve damage, heart attack, kidney failure and strokes. There are two types of diabetes, type-1 and type-2. Type-1 diabetes commonly occurs before age of thirty, while type-2 usually appears in patients aged forty and over [2]. The main symptoms of diabetic patients include intensify thirst, intensify hunger and frequent urination.

Data mining techniques have been widely used in the medical field. The main purposes of applying these methods are pattern recognition, classification, clustering and prediction. Data mining and machine learning techniques are used to extract and discover interesting, unknown, hidden features from large amount of data. A variety of machine learning algorithms have been proposed for diagnosing the diseases using different classification and clustering methods of data mining. Hence, the automated detection of diabetes is possible using classification algorithms [3]. In this paper we used various classification algorithms, such as LR, DT, K-NN, NB, SVM and ANN, to discover patterns of the diseases, and classify and predict the diabetic patients. We applied this algorithms using public dataset of Indian pregnant women, known as PIMA dataset. Experimental performances of all the six classifiers are compared based on different normalization and resampling methods applied for preprocessing data, and achieved good accuracy.

The rest of the paper is organized as follows. Section 2 briefs related work of data mining in the group of diabetic patients. Section 3 describes the methodology, dataset, and classifiers used. Section 4 discusses obtained results from the experiments. Section 5 determines the conclusion of the research with some directions for future work.

II. RELATED WORK

In recent decades, data mining and machine learning algorithms have been studied by different researchers and have been used to predict the possibility of diseases. In this section, we briefly present several related studies that are related to our proposed method.

As most of the data mining investigations of diabetes in the literature are using Pima Indians Diabetes Dataset [14], we only review those works using this dataset. Nilasi et al. [1] developed an intelligent system based on machine learning algorithms for automatic classification of diabetes patients. In their proposed system, they used clustering, removing noise and then classifying patients. They Self Organizing Map (SOM) method for clustering, Principle Component Analysis method (PCA) for removing the noise, and Neural Networks (NA) for classification. They use PIMA dataset for evaluating their approach. They claimed that their new developed method has significantly improved the accuracy of diabetes prediction. Sisidia et al. [2] designed a model that prognosticates the probability of diabetes in patients with maximum accuracy. They utilized Pima dataset using three different machine learning algorithms (Decision Tree, SVM and Naive Bayes) to evaluate their approach. Swapna et al. [3] presented a new approach for classifying diabetic patients based on deep learning algorithms. They employed long short-term memory (LSTM), convolutional neural network (CNN) and its combinations for extracting features, hence passing these features into SVM algorithm for classification. Sujni et al. [4] proposed a method to predict diabetes of Pima dataset using recursive partitioning algorithm. They conclude that the performance of the classification algorithms can be improved by selecting appropriate features. Hina et al. [5] compared different algorithms with regression model to predict diabetes patients using Pima. Prema et al. [6] used ensemble voting classifications for Pima dataset, and compared their method with different algorithms in terms of accuracy performances. Kadhm et al. [7] proposed a system for diabetes prediction besed on K-nearest neighbor algorithm for eliminating the undesired data. Wu et al. [8] used a double-level algorithm, based K-means algorithm with logistic regression model, to improve the accuracy of the Pima diabetes dataset. References [9 - 13] made a comparison of six main classification algorithms to predict diabetic patients using Pima dataset.

Our literature survey revealed that the research on analysing and predicting diabetes of Pima dataset was focused only on using data mining techniques and comparing the accuracy of each algorithm used. However, the data is not always balanced. Imbalanced class occurs when one of the classes has smaller amount, named minority class, than the other class (majority class). The issue of imbalanced data usually leads to misclassification problem where the minority class tends to be misclassified as compared to the majority class. Therefore, it is rather challenging to use the raw data without normalizing and balancing to get higher accuracy. Unlike those works, we aim to resample the data with a view to get balanced classes using SMOTE method. We also used two normalization methods to make the features identical in their scales.

III. Methods and Materials

A. Data

We used Pima Indians Diabetes Dataset [14] which consists of 768 patients, nine variables for each patient. Type 2 diabetes can be diagnosed from those features using data mining techniques. Table 1 shows the description of the data, and Fig.1 shows the distributions of these variables. The first eight features are inputs, and feature class is the output, which takes 0 as normal, and 1 as sick.

TABLE I. STATISTICAL INFORMATION OF INPUT FEATURES

Feature	Min	Max	Mean	Standard Deviation
Number of times pregnant	0	17	3.84	3.4
Glucose Concentration	0	199	120.89	32.0
Blood Pressure	0	122	69.11	19.4
Skin fold thickness	0	99	20.54	16.0
Serum insulin	0	846	79.79	115.2
Body mass index	0	67	31.99	7.9
Diabetes pedigree	0.8	2.4	0.47	0.3
Age (years)	21	81	33.24	11.8
Outcome	Normal = 0, Sick = 1			

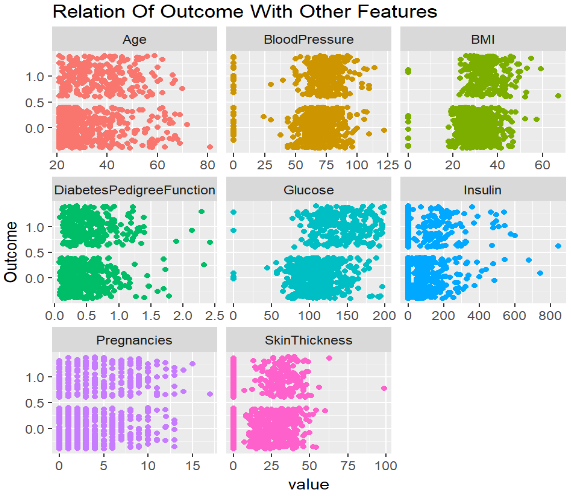

Fig. 1. Distribution of input features

The distribution of the outcome class is as follows: from total instance of 768, 65.10 % are normal and 34.90 are sick. This shows that the data is imbalanced.

B. Data Preprocessing

Raw data are usually incomplete and have some noise, and the quality of data has impact on the model performance used for classification and prediction [4]. Data preprocessing is the process of converting raw data into logical or comprehensible format so that the data will have the same domain values and the features will be comparable. Data preprocessing comprises of data cleaning, data normalization, data transformation and data reduction [5]. It is crucial to preprocess the dataset before training it in the model with a view to better learning the hidden patterns in the dataset. In our dataset used in this paper, input features are not having the same scale, and the number of classes are unbalanced. We used two different normalization methods (Min-Max and Z-score) for rescaling the data, and Synthetic Minority Over Sampling Technique (SMOTE) method for resampling the data to be balanced.

SMOTE Balancing Method

Imbalanced class usually exists in any real dataset. Imbalanced class occurs when one of the classes has smaller number, called minority class, than other class (majority class). We used SMOTE method to balance the classes. The formula to generate synthetic data by SMOTE is defined as follows [15]:

$$C_{new} = C_i + \left(\widehat{C_i} - C_i\right) \times \delta \qquad (1)$$

where C_{new} represents synthetic data, C_i is an example from minority class, $\widehat{C_i}$ is one of k-nearest neighbor from C_i and δ is random number between 0 and 1.

Min-Max Normalization

This method is used to rescale the feature to have a new range of values between 0 and 1. The formulation of this method is as follows [16]:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \qquad (2)$$

Z-score Normalization

In this technique the mean and standard deviation are used to rescale the input features. The formula of Z-score is given as follows [16],

$$z_i = \frac{x_i - \mu}{\sigma}, \qquad (3)$$

where μ is the mean of the feature x, and σ is the standard deviation. Values of Z will be between -1 to 1 with mean = 0 and standard deviation = 1.

C. Classification Algorithms Used

Six different classification algorithms are used for comparing analysis for prediction of diabetes.

1- Support Vector Machine (SVM)

SVM is a type supervised machine learning algorithm used for classification [7]. If we have a two-class training sample, SVM tries to identify the best highest margin separating hyper-plane between the two classes [9]. The good hyper-plane should not be close to the data points. Hyper-plane is selected in such a way that it is far from the data points from each class. Any point lies neare to the margin of the classifier is called support vector [11]. The SVM choose the best hyper-plane which maximize the distance between the two decisions boundaries. Mathematically, we are maximizing the distance between the hyper-plane that is defined as $w^T x + b = -1$ and the hyper-plane defined as $w^T x + b = 1$. This equals to $\frac{2}{||w||}$, and the target is to maximize $\frac{2}{||w||}$, which is implicitly equivalent to minimize $\frac{||w||}{2}$.

2- Naive Bayes Classifier

Naive Bayes is a probabilistic classification algorithm which based on Byes Theorem [4]. In this algorithm, all features are assumed to be independent and unrelated to each other, which means that any feature in a class does not influence the behavior of other features [6]. As it relies on the conditional probabilities of each class, this algorithm is one of the powerful algorithms used for classification, especially when the features are not related. This algorithm is working well even if the data are imbalanced, because it takes into consideration all possible probabilities [8].

Assume we have $X = (x_1, \dots, x_n)$ features, and $C = (C_1, \dots, C_k)$ is the vector of all classes. Using Bayes theorem, we can find the posterior probability $P(C_i|X)$ of any class as follows [10]:

$$P(C_k|X) = \frac{P(X|C_k)\,P(C_k)}{P(X)}, \qquad (4)$$

where $P(C_k|X)$ represent the posterior probability of the target class, $P(X|C_k)$ is the predictor class's probability, $P(C_k)$ is class C_k probability being true and $P(X)$ is the predictor's prior probability.

3- Decision Tree Classifier

Decision Tree is another useful supervised machine learning algorithm utilized for classifying data [12]. The main idea behind this algorithm is to predict the target class based on identifying decision rules which possessed from prior knowledge [8]. Decision Tree works as follows: it creates nodes and internodes for the prediction and classification, then the root node classifies the instance with different features. Every root node has at least two branches, and the leaf nodes are used for representing the target class. In each level, we choose the best node with the highest information gain among all the features used [13].

4- K-Nearest Neighbors (K-NN)

K-NN is a very simple machine learning algorithm which can be used to solve both classification and regression problems [3]. K-NN is a type of instance-based learning, which classifies objects via taking the closest distance of the point from the training data as much as k data. It stores all available points and classifies new points using similarity measures (distance functions) [11]. A point is classifying by a majority vote of its surrounding local neighbors. The Euclidean distance between two points x_1 and x_2 is defined as follows [5]:

$$Euclidean = \sqrt{\sum_{i=1}^{k} (x_1 - x_2)^2}, \qquad (5)$$

where k is the number of related neighbors. To choose the optimal values of k, we usually run the algorithm many times with different values of k, and the optimal value of k is the one that minimize the error [12]. In most cases the best values of k lies between 3 and 10.

5- Artificial Neural Network

Artificial Neural networks (ANN) are data processing algorithm which mimic the biological neural networks [6]. Multilayer Perceptron (MLP) is one of the forward Neural Networks which can be used for classification. MLP consists of three main layers (see Fig.2): input layer, the hidden layer, and the output layer. ANN was formed as a generalization of the mathematical model of neural network biology [7]. Data processing happens in neurons, and the signals are sent between the neurons through the connectors. Each connection between the neurons has a single weight wich indicate whether the signal is strong or weak [10]. To identify the output, each neuron is using an activation function which imposes on the sum of inputs received. Then, the amount of the output will compare to a specified threshold to classify the signal [12]. In our paper, we have 8 nodes as input layer, the hidden layer has 16 nodes, and the output has 2 nodes (normal or sick), and we used the sigmoid activation function.

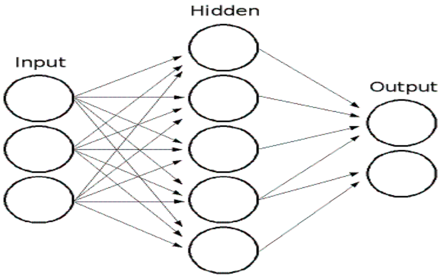

Fig. 2. Simple Model of Artificial Neural Network with Multilayers

6- Logistic Regression

Logistic Regression (LR) is another useful and simple machine learning classification algorithm. It based on predicting the probability of a categorical output feature [3]. This algorithm measures the relationship between the categorical output feature and one or more input features by estimating probabilities using a logistic function [9]. In most cases, the output feature of the logistic regression algorithm is binary. In our classification problem we aim to predict whether the patient is diabetic or not. The logistic regression algorithm relies on the linear regression model expressed as.

$$y_i = \beta_1 x_1 + \beta_2 x_2 \ldots . \beta_m x_m, \qquad (6)$$

where y_i is the output, x_i are input features, and β_i are regression coefficients [11]. The predictive value is either 0 or 1, it is 1 if the value is bigger than the 0.5, otherwise the output is 0. Based on linear regression, the logistic regression adds a layer of sigmoid function (non-linearity). The formula of the logistic regression algorithm is as follows [12],

$$sigmoid\ (y) = \frac{1}{1+e^{y}}. \qquad (7)$$

D. Accuracy Measuremnets

Six different classification algorithms (SVM, NB, DT, K-NN, ANN, LR) have been used in this paper. To check the performance of each classifier, we used 10-fold cross-validation method for splitting the data into training and test. Samples were divided into 10 sub samples, one sample was used for testing/validation, and the rest 9 samples were used for training the model. Each of the six models were trained and tested 10 times, and the average of the 10 were calculated and used to examine the performance of the models. Confusion matrix is commonly used in machine learning applications to evaluate the performance of the classification algorithms. The metrics that are utilized for comparing the effecincy of each classifier are the Accuracy (Ac), Precision (Pr), F-measure (F) and ROC (Receiver Operating Curve) AUC (Area Under Curve). These metrics are calculated from the confusion matrix. The formulas for these metrics are as follows [3],

$$Ac = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (8)$$

$$Pr = \frac{TP}{(TP+FP)} \qquad (9)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative, all taken from the confusion matrix

IV. Results and Discussion

In this section, we will summaries the main results obtained from the models used. Six different classification algorithms are applied to automatically classify the patient into diabetic or normal using the pima Indian diabetes. Results are shown in Table II. The 10-fold cross-validation was used for splitting data into training/testing, and the performance of the six classifiers were compared.

TABLE II. PERFORMANCE MEASURES OF THE SIX CLASSIFIERS

Algorithm	Accuracy	Precision	F-measure	AUC
KNN	0.74	0.73	0.73	0.78
DT	0.72	0.72	0.72	0.65
SVM	0.60	0.69	0.61	0.69
ANN	0.77	0.77	0.77	0.83
NB	0.74	0.75	0.74	0.82
LR	0.77	0.77	0.76	0.83

Corresponding algorithms performance over Accuracy, Precision, F-measure, are presented in Table II. Results in Table II show the performance of the six algorithms used when raw data (imbalanced) were used. The performance of Artificial Neural Networks and Logistic Regression models were similar and showing the maximum accuracy of 77%, which is higher than the other classifiers, while the lowest performance was obtained from the Support Vector Machine algorithm, with accuracy of 60%.

Fig.3 shows the ROC-AUC (Receiver Operating Curve with Area Under Curve) for imbalanced data. The ROC curve in any classifier portrays the performance of the algorithm when its discrimination threshold is diverse. ROC curve was applied to make comparison between the six algorithms used. The model with the highest area will be identified as the best model. The area under curve for the Support Vector Machine model (purple line) was only 50%, while the highest area was obtained the Naïve Bayes model with 73%. The results show that the Naïve Bayes is the best model.

Results of Table II and Fig. 3 indicate that there is a need for resampling the data as a preprocessing method to create a balanced dataset. Therefore, we used SMOTE method which aims to makes the data balanced, so that there is no minority or majority class in the data. Fig. 4 shows the ROC curve and the area under curve results of the same dataset after balancing.

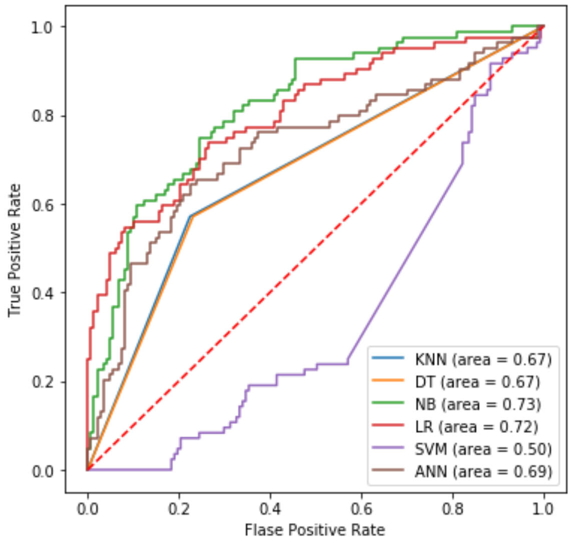

Fig. 3. ROC Curve For KNN, DT, NB, LR, SVM, ANN for (Im-Balanced)

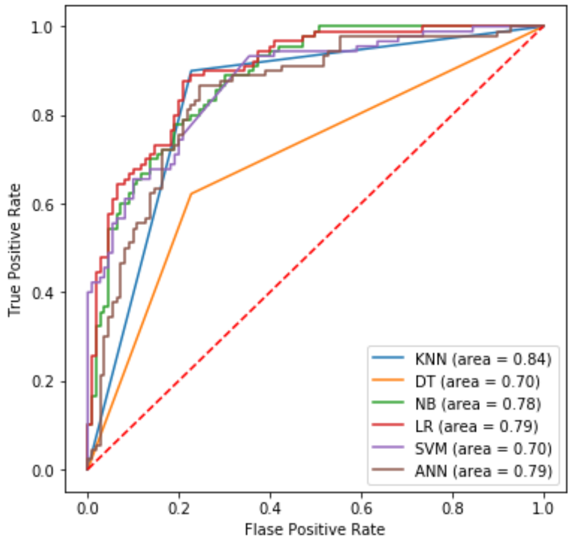

Fig. 4. ROC Curve For KNN, DT, NB, LR, SVM, ANN for (After Balancing)

Results in Fig.4 show the imporovement of the model performance after preprocessing data and balancing. It can be seen that there are dramatic change happening in the performance of some algorithms. For example, in K-Nearest Neighbors the area under curve was significantly increased from only 67% (in Fig.3) into 84% (Fig.4). In the same token, the performance of Support Vector Machine classifier was increased from 50% into 70%. Similar results can be noticed for all other algorithms. This indicates that the resampling techniques used was useful, and the performance of all algorithms were improved. The amount of improvement for each classifier used after balancing data are presented in Table III.

Table III and Fig. 5 show that the amount of improvement of resampling (balancing) method applied was always positive for all algorithms. The highest improvement was occurred for Support Vector Machine model of 20%, while the lowest improvement was for Decision Tree model with only 3%.

In almost all the experiments of the balanced data framework used in this paper, the SVM has performed better in preprocessing and resampling methods used. This is due to the fact that the SVM has is sensitive to the kind of features used as input to the model. The 10-fold cross validation was also performed better which gives higher accuracy compared to other classifiers.

TABLE III. AMOUNT OF MODEL PERFORMANCE IMPROVEMENT AFTER BALANCING DATA

	KNN	DT	NB	LR	SVM	ANN
Imbalanced	0.67	0.67	0.73	0.72	0.50	0.69
Balanced	0.82	0.70	0.78	0.79	0.70	0.79
Difference	+0.15	+0.03	+0.05	+0.07	+0.20	+0.10

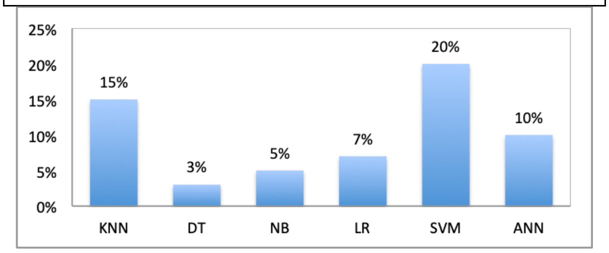

Fig. 5. The amount of improvement for each model after resampling

The performance of classification algorithims under balanced data varies from one model to another. This can be clearly seen in Fig.5. The class imbalanced issue relies on complexity of the data, level of imbalanced class, number of instances in the data and the classification algorithm used. For example, in KNN algorithm the distance metric used and the number of k selected can affect the performance of the sampling method used. Therefore, the development resampling method for handling the class imbalance depends on different factors which must be taken into consideration when dealing with this problem. The selection of a propriate classification algorithm is also important for obtaining better results.

V. CONCLUSION

Diabetes is a very common and dangerous disease in the world. Detection of diabetes at its early stage is extremely crucial. In this paper, we developed an approach for diabetes classification using supervised machine learning algorithms for imbalanced data. We applied SMOTE method for resampling the minority class in the output feature, and the data were preprocessed for rescaling and noise removal. In order to evaluate the performance of our approach, several experiments were implemented using Pima India diabetes dataset. We applied our method on six different classifiers (KNN, SMV, NB, DT, LR, ANN). These six machine learning algorithms were studied and their performance were evaluated and compared using different accuracy measures. The results indicated a significant improvement based on the resampling method used as preprocessing before classification. The highest improvement was noticed in SVM classifier. Experimental results determine the performance of our approach with an achieved ROC area under curve of 84% for KNN, 78% for Naive Bayes and 79% for both LR and ANN algorithms. The evaluation measure for balanced data give much better results with the measures on confusion matrix such as accuracy, precision, recall, F-value and ROC. In the future, the proposed method can be used to classify and diagnose other diseases. One can extend and improve the automation of diabetes prediction using some other classification and clustering algorithms.

References

[1] Nilasi, M., Ibrahim O., Dalvi M., Ahmedi H. and Shahmoradi L., “Accuracy improvementfor diabetes disease classification: a case on a public medical dataset” Fuzzy Information Engineering, vol. 9, pp. 345-357, 2017.

[2] Sisidia D. and Sisidia D.S. “Prediction of diabetes using classification algorithms” International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Procedia Computer Science, vol. 132, pp 1578-1585, 2018.

[3] Swapna G., Vinayakumar R. and Soman K.P. “Diabetes detection using deep learning algorithms” ICT Express 4, pp 243–246, 2018.

[4] Sujni P., C. Beulah Christalin Latha “Prediction of diabetes using a classification model” Al Dar Research Journal for Sustainability, 2, May 2017.

[5] Hina S., Anita S. and Sohail Abul Sattar “Analyzing diabetes datasets using data mining” Journal of Basic & Applied Sciences, 13, pp 466-471, 2017.

[6] Prema N. S., Varshith V. and Yogeswar J. “Prediction of diabetes using ensemble techniques” International Journal of Recent Technology and Engineering (IJRTE), 2277-3878, Volume-7, Issue-6S4, April 2019.

[7] Kadhm M. S., Ikhlas W. G. and Duaa E. M. “An accurate diabetes prediction system based on K-means clustering and proposed classification approach” International Journal of Applied Engineering Research, Volume 13, pp. 4038-4041, Number 6, 2018.

[8] Wu H., Shengqi Y., Zhangqin H., Jian H. and Xiaoyi W. “Type 2 diabetes mellitus prediction model based on data mining” Informatics in Medicine Unlocked, Vol. 10, pp 100-107, 2018.

[9] Gopinath M.P. and Murali S. “Comparative study on classification algorithm for diabetes data set” International Journal of Pure and Applied Mathematics, Volume 117, No. 7, pp 47-52, 2017.

[10] Sa’di S., Maleki A. , Hashemi R, Panbechi Z. and Chalab C. “Comparisonof data mining algorithms in the diagnosis of type II diabetes” International Journal on Computational Science & Applications, Vol.5, No.5, October 2015.

[11] Zou Q., Qu K.., Luo Y., Yin D., Ju Y. and Tang H. “Predicting diabetes mellitus with machine learning techniques” Front. Genet. 9:515. doi: 10.3389/fgene.2018.00515, 2018.

[12] Zia, U. A. and Khan N. “Predicting diabetes in medical datasets using machine learning techniques” International Journal of Scientific & Engineering Research, Volume 8, Issue 5, May 2017.

[13] Jeevanandhini D., Raj D. G. , Kumar V. D. and Sasipriyaa D. “Prediction of type2 diabetes mellitus based on data mining” International Journal of Engineering Research & Technology, ETEDM - 2018 Conference Proceedings, Special Issue – 2018.

[14] http://archive.ics.uci.edu/ml/datasets/Pima.Indians.Diabetes.

[15] Santoso B., Wijayanto H., Notodiputro K. A. and Sartono B. “Synthetic over sampling methods for handling class imbalanced problems: a review” IOP Conf. Series: Earth and Environmental Science 58, doi:10.1088/1755-1315/58/1/012031, 2017.

[16] Adel S. Eesa and Wahab Kh. Arabo “A normalization methods for backpropagation: a comparative study” Science Jounal of University of Zakho, Vol. 5, No. 4, pp. 319 –323, Dec. 2017.