Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 1, January 2018, pg.53 – 60

A STUDY ON PREDICTION OF DIABETIC DISORDER USING CLASSIFICATION BASED APPROACHES

P.Hema[1], K.Palanivel[2]

[1]Research Scholar, Department of Computer Science, AVC College, Mayiladuthurai, India

[2]Associate Professor, Department of computer Science, AVC College, Mayiladuthurai, India

[1] hemapalanivel@gmail.com

Abstract— Data mining has a considerable potential in health cooperation industry to enable health position by systematically handling data, look the composure and improve care with minimized cost. Medical professionals have passion in developing reliable prediction methodologies to recognize various diseases. Medical data mining helps to identify the health problem and get recovery quickly. The diabetic disease is a common problem found in disease most of the countries and people are suffering a lot, because of this disease. This research work focuses on the classification techniques, namely Random Forest Tree, Rep Tree, and Decision Stump which are applied to diabetic datasets to predict the possibility of the disease efficiently by analysing the relationship of diabetic data. The objective here is to study the performance of the three classification algorithm and identify the best classifier technique with good accuracy. From the Experimental results of three algorithms, Rep Tree provided best result when compared with other two algorithms. The result will help the doctors in that considered diagnosis process.

Keywords— Data mining, Diabetic Disease, Random Forest Tree, Rep Tree, Decision Stump, Classification, WEKA.

I. INTRODUCTION

Diagnosis of diabetic infection especially relies on clinical and pathological data. System can help predict diabetic disease medical professional predicts data based on the patient`s clinical position of diabetic disease. Extracting or mining knowledge from large amount of data is known as data mining. This research paper is presented as follows; section 2 lists on top of each other work. Section 3 presents the methodology and the aspect of detailed list algorithm. Section 4 elaborates experiments and finalizes the result produced by the algorithm. This paper, intensity on the data classification and the performance measure of the classifier algorithm based on the true positive rate, false positive rate, precision, recall, F-measure generated separately algorithm when applied on the dataset.

II. DATA MINING

Data mining involves the concern of detailed statement analysis tools to discover earlier unknown, reliable patterns and relationships in large message sets. These tools can include statistical models, mathematical algorithms, and machine learning

methods a well-known as neural networks or decision trees. Consequently, data mining consists of in a superior way than collecting and managing report, it also includes analysis and prediction. The design of data mining is to identify valid, contemporary, potentially convenient, and clear correlations and patterns in prompt data. Finding convenient patterns of data are supported by diverse names (e.g., knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing). The term "data mining" is primarily used by statisticians, database researchers, and the function communities. The term KDD (Knowledge Discovery in Databases) affect the everywhere behaviour of discovering convenient knowledge from data, where data mining is a particular step in this process. The steps in the KDD behaviour, such as data preparation, message selection, data cleaning, and related interpretation of the results of the message mining process, assure that convenient knowledge are derived from the data. Data mining is an opportunity for traditional data analysis and statistical approaches as it incorporates contemplative techniques drawn from various disciplines like AI, equipment learning, OLAP, data visualization, etc.

III.KNOWLEDGE DISCOVERY PROCESS

Knowledge Discovery in Databases (KDD) is an iterative process that transforms raw data into useful information. Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modelling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets.

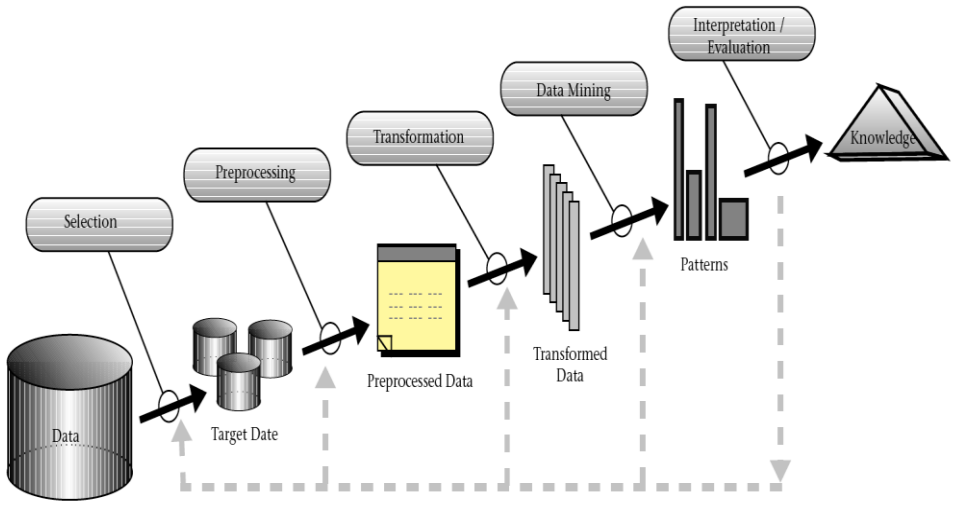

Figure 1. Steps of Knowledge Discovery in Databases

IV. THE STEPS IN THE KDD PROCESS

Data cleaning: It is also known as data cleansing; in this phase spread data and trivial data are displaced from the collection.

Data integration: In this second, multiple data sources, routine heterogeneous, are mutually in a common source.

Data selection: The statement relevant to the analysis is decided on and retrieved from the message collection.

Data transformation: It is also known as data consolidation; in this phase the selected data is transformed directed toward forms decent for the mining procedure.

Data mining: It is the severe lead everywhere efficient techniques are applied to memorize potentially relaxed patterns.

Pattern evaluation: In this step, delightful patterns representing knowledge are identified based on given measures.

Knowledge representation: It is the final phase in which the discovered knowledge is visually spotted to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

V. DIABETIC DISEASE

Diabetes is a infection that occurs when your ties of blood glucose, also called blood sugar, is too high. Blood sugar is your dominant source of energy and comes from the foot to eat. Insulin, a hormone made by the pancreas, helps glucose from food get

into your cells to be used for energy. Sometimes your body doesn`t collect enough or any insulin or doesn`t use insulin well. Glucose then stayed in your blood and doesn`t reach your cells. Overtime, having at length glucose in your blood can cause health problems. Although diabetes has no cure, you can step to manage your diabetic and stay healthy. Sometimes people call diabetes "a touch of sugar" or "borderline diabetes". These terms represent that someone doesn`t sure thing have diabetes or has a slight serious case, for all that every situation of diabetes is serious. Diabetes is often called a modern-society disease because widespread desire for regular exercise and rising obesity rates are some of the main contributing factors of it. Overtime, high blood glucose leads to problem such as 1) heart disease 2) stroke 3) kidney disease 4) eye problem 5) dental disease 6) nerve damage 7) foot problems.

VI. LITERATURE SURVEY

The cause of diabetes is a mystery, although obesity and lack of exercise appear to possibly play significant roles. Huy Nguyen A.P. et.al [4] proposed a new algorithm Homogeneity-Based Algorithm to determine over fitting and over generalization behaviour of classification. The algorithms used in this paper are Support Vector Machine, Decision Tree and Artificial Neural Networks. They predict whether a new patient would test positive about diabetes.

Joseph L. Breault [5], used the publicly available Pima Indian diabetic database (PIDD) at the UCIrvine Machine Learning Lab. They tested data mining algorithms to predict their accuracy in predicting diabetic status from the 8 variables given. Out of 392 complete cases, guessing all is non-diabetic gives an accuracy of 65.1%. Rough sets as a data mining predictive tool applied rough sets to PIDD using ROSETTA software. The test sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracies ranged from 69.6% to 85.5% with a mean of 73.8% and a 95% CI .The accuracy of predicting diabetic status on the PIDD was 82.6% on the initial random sample, which exceeds the previously used machine learning algorithms that ranged from 66-81%. Using a group of 10 random samples the mean accuracy were 73.2%. A study conducted in [6] intended to discover the hidden knowledge from a particular dataset to improve the quality of health care for diabetic patients. In [7] Fuzzy Ant Colony Optimization (ACO) was used on the Pima Indian Diabetes dataset to find set of rules for the diabetes diagnosis. There are diverse kinds of studies for DM techniques in medical databases. J.W.Smith et al [8] dealing with this data base uses an adaptive learning routine that generates and executes digital analogy between perceptions-like devices, called ADAP. They used 576 training instances and obtained a classification of 76% on the remaining 192 instances.

Rajesh et al., [9] used various classification algorithms like ID3, C4.5, LDA, Naïve Bayes, K-NN for diagnosing diabetes for the given dataset. The author concluded that C4.5 is the best algorithm with less error rate of 0.0938 and more accuracy value of 91%. Pardha Repalli [10], In their research work predicted how likely the people with different age groups are affected by diabetes based on their life style activities. They also found out factors responsible for the individual to be diabetic. Statistics given by the Centres of Disease Control states that 26.9% of the population affected by diabetes are people whose age are greater than 65, 11.8% of all men aged 20 years or older are affected by diabetes and 10.8% of all women aged 20 years or older are affected by diabetes. The dataset used for analysis and modelling has 50784 records of 37 variables. They computed a new variable age_new as nominal variable, dividing in to three group's young age, middle age and old age and the target variable diabetes_diag_binary is a binary variable. They found 34% of the population whose age was below 20 years were not affected by diabetes. 33.9% of the population whose age was above 20 and below 45 years was not affected by diabetes. 26.8% of the population whose age was above 45 years was not diabetic. Recently Karthikeyini et.al [11 & 12] discussed comparison a performance of data mining algorithms for diabetes disease based on computing time, precision value , the data evaluated using 10 fold Cross Validation error rate, error rate focuses True Positive, True Negative, False Positive and False Negative, bootstrap validation and accuracy. Mohd Fauzi bins Othman and Thomas Moh Shan Yau [13], examined the performances of different classification and clustering methods of a large set of data.

VII. METHODOLOGY USED

The following steps are included in the classification process of this paper. Three different classifier is chosen Random Forest Tree, Rep Tree, Decision stump. WEKA tool is used to analyse the predicted values by each of the classifier. The Precision, Recall & F-Measure of each classifier is calculated. Finally the result is analysed and the best performance algorithm identified.

Random Forest Classifier

The term came from random decision forests that were first proposed by Tin Kam HO of Bell lab1s in 1995. Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the made of the classes output by individual trees. It is presented independently with some controlled modification. Trees and the results included at random forest is based on majorities of accurate output. In dataset, where M is the total number of input attributes to the dataset, only m attributes to chosen at random for each tree where m<M [18].

Advantages of Random Forest

- Works for both classification and regression.
- Handles categorical predictors naturally.
- No formal distributional assumptions
- Can handle highly non-liner interactions and classification boundaries.

Random Forest are broadly to be the finest "off-the-shelf" classifier for high-dimensional data. Random Forest are mixture of tree predictors such that each depends on the values of random vector sampled autonomously and with the same distribution for all tree in the forest.

Rep Tree Classifier

This algorithm is first recommended to [15]. Reduces Error Pruning (Rep) Tree classifier is a fast decision tree learning algorithm and is based on the principle of computing the error arising from variance [14]. In pruning tree the measure used in the Mean Square Error on the prediction made by the tree. Basically reduced error pruning tree "REPT" is fast decision tree learning and its builds a decision tree based on the information gain or reducing the variance[16][17].

Decision Stump Classifier

Decision stump is a Decision tree, which uses only a single attribute for splitting. Decision tree is one of best classification technique in data mining [19]. The internal node denotes a substantiate on attribute, each piece of action represents a risk of the verify and the palm blade node delineate classes. It is a graphical representation of possible solutions based on condition on these solutions optimum courses of action is carried out. In our work, we have used decision tree classifier such as decision stumps, Random Forest, Rep tree to classify the diabetic data set.

VIII. STATISTICAL MEASURE

The accuracy of the classifier is given by TP rate, FP Rate, Re-Call, F-Measure, Precision using WEKA tool. WEKA tool is a powerful software platform that gives an integrated environment for machine learning, data learning, text learning and other business & prediction analysis.

The honest truth about the classifiers is given by false positive rate, true positive rate, recall, precision and F-measures using WEKA tool. WEKA is a powerful software statement of belief that gives an exhaustive environment for equipment learning, data mining, text mining and other trade and prediction analysis. The decent of measures from generally told the classes has been taken to try the during measure for classifiers. For concrete illustration, to try the around precision for a classifier for a subject to dataset, adequate of precisions of both (true/false) classes is calculated.

1) Precision

Precision is the preciseness or exactness of approximately classified class, properly known as positive predictive value. It is the symmetry of instances which originally have class x / Total classified as class x. So basically valuable precision directed the accurate results and it takes all told relevant data yet returns only topmost results. In quickly, it is the number of selected items which were related.

Precision = (True Positive / (True Positive + False Positive))*100

2) Recall

Recall to give sensitivity of problem and it processes values or product quantity or completeness. It returned the most relevant and part from the documents that are relevant as result from the query. In other words, modules that are really recognized as difficult to maintain from the total number of modules. In short, it is the number of related objects that were chosen.

Recall = (True Positive / (True Positive + False Negative))*100

3) True Positive (TP)

True positive are the positive tuples which were appropriately labelled every classifier. It is the proportion categorized as class x / Actual total in class x. True positive projected individually modules that are predicted genuinely as the results specified at the end.

True Positive rate = (True Positive / (True Positive + False Negative))*100

4) False Positive (FP)

False positive, proportion incorrectly categorized as class x / Actual total of all classes, except x. It is incorrectly predicted compared to original results.

False Positive rate = (False Positive / (False Positive + True Negative))*100

5) F-Measure

F-Measure categorized as (2*Precision*Recall / (Precision+ Recall))*100. It is a combined measure for precision and recall.

IX. EXPERIMENTAL RESULTS

A few classification algorithms for experiment in order to evaluate with their time performance and efficiency. For this purpose have used Weka tool to evaluate performance. The dataset for the experiment has been collected from the UCI repository. In diabetic dataset comprise 9 attributes and 768 instances. Classification was once carried out having all the 9 attributes and their respective time is noted.

Parameters : Preg - No. of times pregent; Plas – Plasma Glucose Concentration; Pres – Diastolic Blood Pressure; Skin – Skin Fold Thickness; Insu – Insulin; Mass – Body Mass index; Pedi – Diabetic Pedigree Function; Age – Age; Class – Class variable;

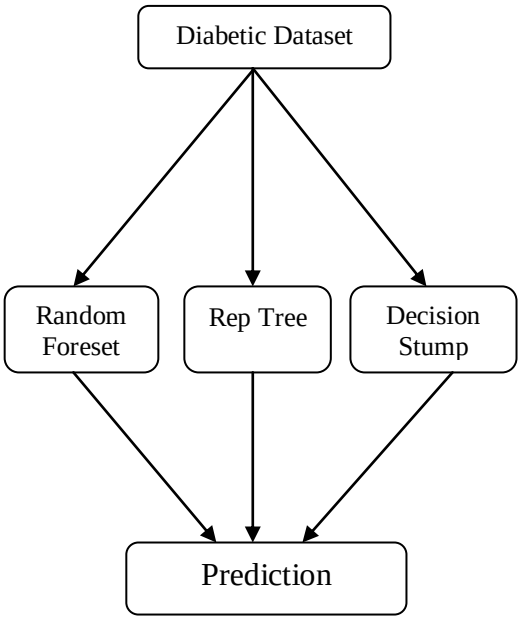

Figure 2. Working Architecture for Proposed Work

The result have been tabulated

Method	Correctly Classified Instance	Incorrectly Classified Instance	Time in sec taken with 9 attributes
Random Forest	564	204	0.17 sec
Decision Stump	552	216	0.02 sec
Rep Tree	578	190	0.08 sec

Table 1. Performance for comparison for different algorithm.

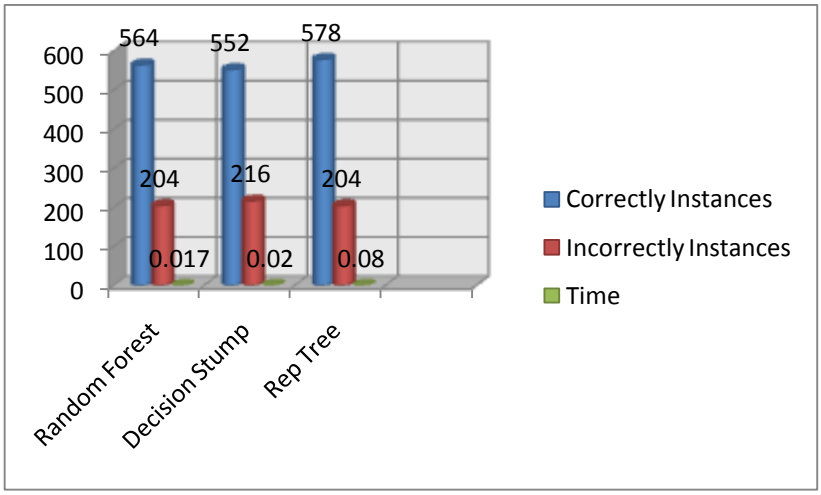

Figure 3. Performs diagram of different algorithm

Table 1 shows the end result of the algorithm quite incredible in contrast to other algorithms in data mining. The comparing with different two algorithms Decision Stump will performs better of 0.02 sec time taken. It actually shown Decision Stump will be higher than the Random Forest and Rep Tree algorithm time.

METHOD	RMSE	MAE	KAPPA STATISTIC
Random Forest	0.4463	0.3155	0.3878
Decision Stump	0.4418	0.3802	0.3745
Rep Tree	0.4282	0.3272	0.4380

Table 2. Comparison of different algorithm in RMSE, MAE & KAPPA STATISTIC

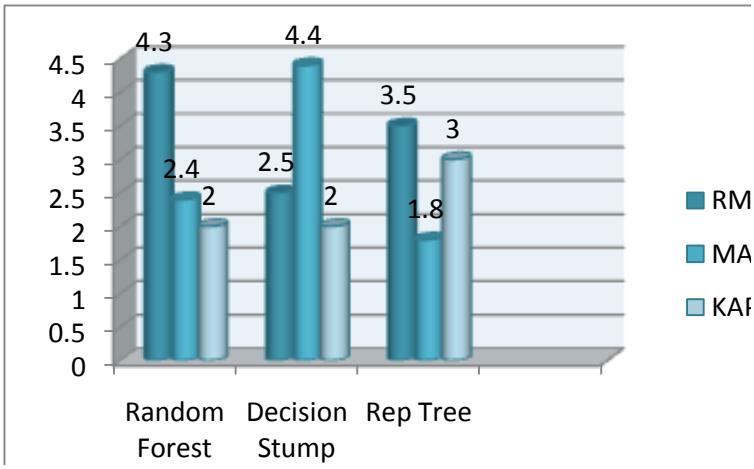

Figure 4. Diagram of comparison RMSE, MAE & KAPPA STATISTIC

From the experiment we have considered that Decision stump takes the minimal time with middle value of RMSE where as Random Forest tree takes maximum RMSE value. Table 2 proven the result of Kappa Statistic evaluating with two algorithms in Decision Stump will perform better of lowest accuracy in Kappa Statistic. Experiments are performed on the diabetic dataset by using Random Forest, RepTree and Decision Stump using WEKA tool.

Method	Precision	Re – Call	TP Rate	FP Rate	F-Measure
Random Forest	0.768	0.848	0.848	0.478	0.806
Decision Stump	0.777	0.796	0.796	0.425	0.787
Rep Tree	0.789	0.846	0.846	0.422	0.817

Table 3. Analysis on smaller dataset in tested-negative class (768 Instances)

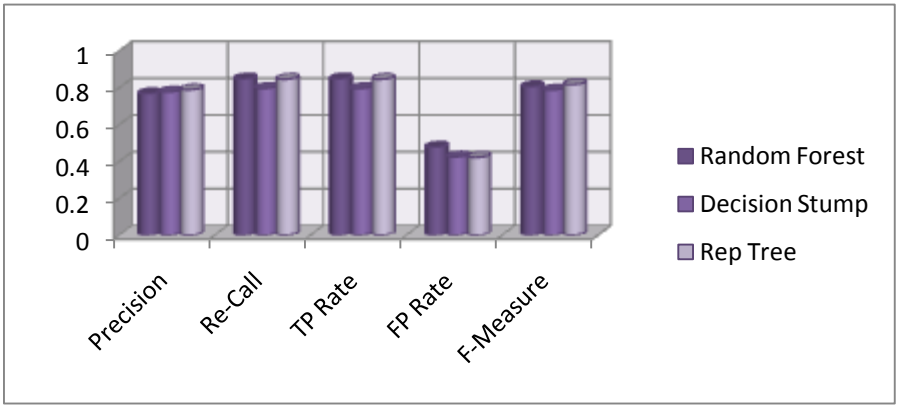

Figure 5. Describe the ratio of each classifier for each dataset based on table 3.

Method	Precision	Re – Call	TP Rate	FP Rate	F-Measure
Random Forest	0.648	0.522	0.522	0.152	0.579
Decision Stump	0.602	0.575	0.575	0.204	0.588
Rep Tree	0.668	0.578	0.578	0.154	0.620

Table 4. Analysis in smaller dataset tested positive class (768 Instances)

The result of following analysis on the dataset is clearly given by the table II, III & IV. Table II have given the RMSE, MAE & KAPPA STATISTIC in dataset using different classifier. Table III & IV listed the Precision, Recall, True Positive Rate, False Positive Rate, F-Measure to analyse the classifier.

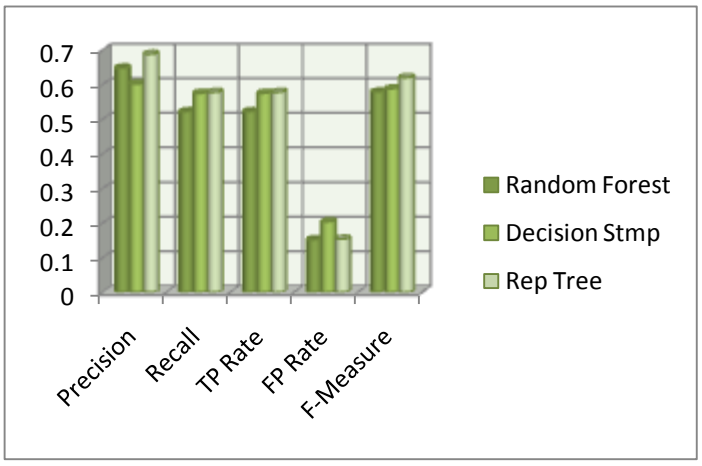

Figure 6. Describe the ratio of each classifier for each dataset based on table 4.

X. CONCLUSION AND FUTURE WORK

To study the performance of three classification algorithms, the experiments have been performed using the Weka tool. Diabetic data sets are taken from UCI repository having 768 instances and 9 attributes. This research work focuses in three algorithms namely, Random Forest, Rep Tree and Decision stump which are analysed using the Diabetic dataset to predict the symptoms of diabetic disorder. Out of the three algorithms, Decision stump consumes less time (0.02Sec) for providing result.

But Rep Tree provides maximum accuracy in classification, however it takes more time to process the data (0.08Sec). In future classifier algorithms can be combined with evolutionary algorithms such as fuzzy set or rough set, in order to handle vagueness in data.

REFERENCES

[1] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2001.

[2] Remco R. Bouckaert, Eibe Frank, Mark Hall Richard Kirkby, Peter Reutemann, Seewald David Scuse, *WEKA Manual for Version 3-7-5*, October 28, 2011.

[3] UCI Machine Learning Repository. http://www.ics.uci.edu/mlearn/MLRepository.html

[4] Huy Nguyen Anh Pham and Evangelos Triantaphyllou **"** Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization**"** Department of Computer Science, 298 Coates Hall, Louisiana State University, Baton Rouge, LA 70803.

[5] Joseph L. Breault, MD, MPH, MS, "Data Mining Diabetic Databases: Are Rough Sets a Useful Addition" http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.815&rep=rep1&type=pdf.

[6] Sankaranarayanan.S and Dr Pramananda Perumal.T "Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies", World Congress on Computing and Communication Technologies, 2014, pp. 231-233.

[7] Mostafa Fathi Ganji and Mohammad Saniee Abadeh, "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease", Proceedings of ICEE May 2010.

[8] Smith, J.,W., Everhart, J.,E., Dickson, W.,C., Knowler, W.,C. and Johannes, R.,S., **"**Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in Proceedings of the Symposium on Computer Applications and Medical Care**"** IEEE Computer Society Press, pp. 261- 265.

[9] Rajesh K, Sangeetha V. Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT). 2012; Vol.02 No.03, pp. 224–9.

[10] Pardha Repalli, **"**Prediction on Diabetes Using Data mining Approach**"**.

[11] Dr. V. Karthikeyini, I. Parvin Begum," Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction Of Diabetes Disease", International Journal on Computer Science and Engineering (IJCSE), Vol. 5 No.03, ISSN: 0975-3397, Mar 2013.

[12] Pervin begum. I., Karthikeyini.V., Tajuddin.K., Shahina Begum, **"**Comparative of data mining classification algorithm (CDMCA) in Diabetes Disease Prediction**"**, International journal of Computer Applications, 2012 Vol.60 No.12, pp. 26-31.

[13] Mohd Fauzi bin Othman, Thomas Moh Shan Yau, **"**Comparison of Different Classification Techniques using WEKA for Breast Cancer**"**, F.Ibrahim, N.A. Abu Osman, J.Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, 2007 pp.520-523.

[14] Witten IH, and Frank E, *Data mining: practical machine learning tools and techniques* – 2nd ed. the United States of America, 2005 Morgan Kaufmann series in data management systems.

[15] Quinlan J **"**Simplifying decision trees**"**, International Journal of Man Machine Studies, Vol.27, No.03, 1987 pp.221–234.

[16] Payal P.Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin, **"**Pre-processing and Classification in WEKA Using Different Classifier**"**, Int. Journal of Engineering Research and Applications, 2014 Vol.04 No.08. pp- 91-93.

[17] Dr. B. Srinivasan, P.Mekala, **"**Mining Social Networking Data for Classification Using REP Tree**"**, International Journal of Advance Research in Computer Science and Management Studies, 2014 Vol.02 No.10. pp- 155-160.

[18] Subkumar, Dr.Manish Mann, **"**E-Mail fillering for the Removal of Misclassification Error**"** International Journal of Engineering Research in Computer Science & Engineering (IJERCSE), 2015 Vol.02 No.12.

[19] A. Bellachia and E.Guvan, **"**Predicting breast cancer survivability using data mining techniques**"**, Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining.

[20] Durairaj M, Kalaiselvi G, " Prediction Of Diabetes Using Soft Computing Techniques- A Survey", International Journal of Scientific & Technology Research, March 2015, Volume 4 Issue 3.

