International Journal of Scientific Research in Multidisciplinary Studies
Vol.3, Issue.6, pp. pp.48-53, June (2017)

Research Paper
E-ISSN: 2454-9312
P-ISSN: 2454-6143

Clustering Classification for Diabetic Patients using K-Means and M-Tree prediction model

Prateeksha Tomar[1*], Amit Kumar Manjhvar[2]

[1*]Department of Computer Science and Engineering, Madhav Institute of Technology and Science, RGPV, Gwalior, India
[2] Department of Computer Science and Engineering, Madhav Institute of Technology and Science, RGPV, Gwalior, India

Corresponding Author e-mail: tomarprateeksha09@gmail.com, Mob. : 9617318979

Available online at: www.isroset.org

Received 10th May 2017, Revised 24th May 2017, Accepted 14th Jun 2017, Online 30th Jun 2017

***Abstract*—** Medicinal Data mining is the way in the direction of removing concealed examples from therapeutic data. This paper shows the advancement of a crossover model for ordering Pima Indian diabetic database (PIDD). The model comprises of two phases. In the primary stage, the K-means bunching is utilized to distinguish and take out erroneously grouped examples. The nonstop data is changed over to all out frame by rough width of the coveted interims, in light of the conclusion of restorative master. In the second stage an adjusted arrangement is finished utilizing M tree C4.5 by taking the accurately bunched event of first stage. Test comes about imply the fell K-means grouping and M tree C4.5 has upgraded arrangement precision of C4.5. Additionally administers produced utilizing fell C4.5 tree with clear cut data are less in numbers and simple to translate contrasted with principles created with C4.5 alone with persistent data. The proposed fell model with all out data got the arrangement precision of 93.33 % when contrasted with exactness of 73.62 % utilizing C4.5 alone for PIMA Indian diabetic dataset.

General Terms*— Medical data mining, clustering, rule based classification using M-tree , K-means, Weka.*
Keywords*— *K-means clustering, Categorical data, rule based classification, M-tree, Pima Indian Diabetics.*

I. INTRODUCTION

The data mining functionalities are used to specify the kind of patterns to be found in the data-mining task. The data mining functionalities mainly include association rule mining, classification, prediction & clustering. Association analysis is used for discovering interesting relations between variables in large databases, which in given in the form of rules to user. Classification predicts the class labels. Prediction is used to access the value of an attribute that a given sample is likely to have. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Classification is supervised learning algorithms in contrasts with clustering, which are unsupervised learning algorithm [1]. Classification is a supervised model, which maps or classifies a data item into one of several predefined classes. Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The most common classification data mining techniques are Case-Based Reasoning, M tree, Back propagation neural network, Radial basis neural network, Bayesian classification, Rough set Approach, Fuzzy Set Approaches, and K-nearest neighbor classifiers.

In this paper a cascaded K-means clustering and M tree has been used to categorize diabetic's patients. Literature survey of classification of diabetic data set is briefed in section 2. For the sake of completeness M -tree and K-mean clustering have been briefly explained in section 3 and 4. Preprocessing of diabetic data set and working of cascaded K-means clustering and M tree classifier is explained in section 5, followed by results and conclusion in section 6 and 7 respectively.

II. RELATED WORK ON DIABETIC DATA SET CLASSIFICATION

Diabetics:-

Diabetes mellitus is a disease in which the body is unable to produce or unable to properly use and store glucose (a form of sugar). Glucose backs up in the bloodstream causing one's blood glucose or "sugar" to rise too high. There are two major types of diabetes. In type 1 diabetes, the body completely stops producing any insulin, a hormone that enables the body to use glucose found in foods for energy.

People with type 1 diabetes must take daily insulin injections to survive. This form of diabetes usually develops in children or young adults, but can occur at any age.

Type 2 (also called adult-onset or non insulin-dependent) diabetes results when the body doesn't produce enough insulin and/or is unable to use insulin properly (insulin resistance). This form of diabetes usually occurs in people who are over 40, overweight, and have a family history of diabetes, although today it is increasingly occurring in younger people, particularly adolescents[2], [3].

World Health Organization (WHO) report had shown a marked increase in the number of diabetics and this trend is expected to grow in the next couple of decades. In the International Diabetes Federation Conference 2003 held in Paris, India was labeled, as "Diabetes Capital of the World," as of about 190 million diabetics worldwide, more than 33 million are Indians. The worldwide figure is expected to rise to 330 million, 52 million of them Indians by 2025, largely due to population growth, ageing, urbanization, unhealthy eating habits and a sedentary lifestyle. Poorly managed diabetes can lead to a host of long-term complications among these are heart attacks, strokes, blindness, kidney failure, blood vessel disease.

Literature review of classification of Diabetic Dataset:-

A lot of research work has been done on various medical data sets including Pima Indian diabetes dataset. Classification accuracy achieved for Pima Indian diabetes dataset using 22 different classifiers is given in [4] and using 43 different classifiers is given in [5]. The performance of proposed cascaded model using k-means and M tree is is compared with [4] and [5]. The results of [5] and [4] are shown in Table 1 and Table 2 respectively. The accuracy of most of these classifiers is in the range of 66. 6% to 77.7%. Hybrid K-means and M tree [6] achieved the classification accuracy of 92.38% using 10 fold cross validations for continuous data. Further cascaded learning system based on Generalized Discriminate analysis (GDA) and Least Square Support Vector Machine (LS_SVM), showed accuracy of 82.05% for diagnosis of Pima dataset [7]. Further authors have achieved classification accuracy of 72.88 % using ANN, 78.21% using DT_ANN where M tree C4.5 is used to identify relevant features and given as input to ANN [8], 79.50% using Cascaded GA_CFS_ANN, relevant feature identified by Genetic algorithm with Correlation based feature selection is given as input to ANN [9], 77.71% using GA optimized ANN, 84.10% using GA optimized ANN with relevant features identified by M tree and 84.71% with GA optimized ANN with relevant features identified by GA_CFS[10]. Authors have achieved an accuracy of of 96.68% for diabetic dataset using cascaded k-means and K-nearest neighbor [11,12,13].

III. M-TREE

M-tree represents a supervised approach to classification. A M tree is a simple tree structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect M outcomes. The basic M-tree induction algorithm was enhanced The WEKA classifier package has its own version of known as J4.8. Information gain and gain ratio measures are used by as splitting criterion respectively. The summary of M tree algorithm is given:

1. Choose an attribute that best differentiates the output attribute values.
2. Create a separate tree branch for each value of the chosen attribute.
3. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
4. For each subgroup, terminate the attribute selection process if:

(a) The members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.

(b) The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.

1. For each subgroup created in (iii) that has not been labeled as terminal, repeat the above process.

IV. K-MEANS CLUSTERING

K-means is one of the simplest unsupervised learning algorithms and follows partitioning method for clustering. K-means algorithm takes the input parameter, k as number of clusters and partitions a dataset of n objects into k clusters, so that the resulting objects of one cluster are dissimilar to that of other cluster and similar to objects of the same cluster. In k-means algorithms begins with randomly selected k objects, representing the k initial cluster center or mean. Next each object is assigned to one the cluster based on the closeness of the object with cluster center. To assign the object to the closest center, a proximity measure namely Euclidean distance is used that quantifies the notion of closest. After all the objects are distributed to k clusters, the new k cluster centers are found by taking the mean of objects of k clusters respectively. The process is repeated till there is no change in k cluster centers. K-means algorithm aims at minimizing an objective function namely sum of squared error (SSE). SSE is defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \qquad (1)$$

Where E is sum of the square error of objects with cluster means for k cluster. p is the object belong to a cluster Ci and mi is the mean of cluster Ci .The time complexity of K-means is O (t*k*n) where t is the number of iterations, k is

number of clusters and n is the total number of records in dataset.

K-means partitioning algorithm: (Input is k is the number of clusters, D is input data set. Output is k clusters).

i.Randomly choose k objects from D as the initial cluster centers.
ii. Repeat
iii. Assign each object from D to one of k clusters to which the object is most similar based on the mean value of the objects in the cluster.
iv. Update the cluster means by taking the mean value of the objects for each of k cluster.
v. Until no change in cluster means/ min error E is reached.

V. K-MEANS AND M-TREE

Data preprocessing:-

The PIMA diabetic database consist of two categories in the data set (i.e. Tested positive , Tested Negative) each having 8 features :Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function and Age (years]. The data is availed from UCI Machine Learning the data processing techniques, when applied prior to mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining. Data preprocessing is a significant step in the knowledge discovery process, since quality Ms must be based on quality data. A total of 768 cases are available in PIDD. 5 patients had a glucose of 0, 11 patients had a body mass index of 0, 28 others had a diastolic blood pressure of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0. After deleting these cases there were 392 cases with no missing values (130 tested positive cases and 262 tested negative) [16].

Working of Proposed method:-
In the first stage of proposed model, simple K-means clustering (with k = 2) of Weka tool, is applied to 392 diabetic patients samples as obtained in section 5.1. The wrongly classified samples are eliminated to get final 299 samples.

As a part of preprocessing the continuous data is converted to categorical form by approximate width of the desired intervals, based on the opinion of medical experts as shown in table 3. Finally in the second stage, the correctly classified samples from first stage and the categorical data is given as input to M tree C4.5 (weka J4.8). The data is partitioned using (a) 60-40 ratio partitioning method (training-test) and (b) 10 fold cross validation. For the sake of completeness few of the performance metrics have been discussed. True positive (TP) corresponds to the number of positive examples correctly predicted by the classifier. False negative (FN) corresponds to the number of positive examples wrongly predicted as negative by the classifier. False positive (FP) corresponds to the number of negative examples wrongly predicted as positive by the classifier. True negative (TN) corresponds to the number of negative examples correctly predicted by the classifier. The true positive rate (TP rate) or sensitivity is the fraction of positive examples predicted correctly by the model. TP Rate = TP/ (TP

+ FN). The false positive rate (FP rate) or Specificity the fraction of negative examples predicted as a positive class. FP Rate = FP/ (TN + FP).Precision is the fraction of records that actually turns out to be positive in the group the classifier has declared as a positive class. Precision = TP / (TP + FP). Recall is the fraction of positive examples correctly predicted by the classifier. Recall = TP / (TP + FN). F-measure is used to examine the tradeoff between recall and precision. Measure = 2*TP/ (2*TP+FP + FN).

VI. RESULTS

In this paper an eager learner: M-tree classifier has been used for classification of standard medical database namely PIMA diabetic. Incorrect labeled instance are eliminated using K-means clustering followed by converting the continuous data to categorical data by consulting medical experts. The resultant dataset is used to train and test the diabetic data set using two method (a) dividing training data and test data using 60-40 ratio (b) 10 fold cross validation method. The performance of M tree with unprocessed data compared to proposed cascaded k-means with M tree using processed data by means of TP rate, FP rate, Precision, Recall and F-measure is also computed for tested positive and tested negative class is shown in table 4. For unprocessed diabetic data, the classification accuracy of M tree using 10 fold cross validation and 60-40 training–testing partitioning of data was found to be almost same. However for the proposed cascaded model with categorical data, the performance of M tree with 60-40 training-testing partitioning of data outperforms the M tree with 10 fold cross validation by an order of 1.38 %. Kappa value , mean absolute value and classification accuracy for unprocessed and processed data using C4.5 using 60-40 ratio (partitioning of training and test data) and 10 fold cross validations is shown in figure 1 and 2 respectively. Experimental results show the improvement in accuracy Diabetic data set using proposed cascaded method: k-means with M tree (with categorical data) by an order of compared to M tree with unprocessed data.

Cluster instances	%	accuracy
M-TREE	69.0	8.990
K-MEANS	255	33.2031

Further the rules generated by M tree with categorical data are less in number and are easy to interpret compared to rules generated by M tree with continuous data.. The rules generated by the proposed cascaded model are given below.

1. If Plasma=low then class=> Tested Negative
2. If Plasma =medium & Age=low & Pedigree =low then Class => Tested Negative
3. If Plasma =medium & Age=low & Pedigree =medium & Diastolic BP=medium then Class=> Tested Negative

The result show the value of goodness values of M-TREE algorithm

4.

Accuracy	k-means	Proposed model
Goodness of fit(%)	67.318	92.032

M-tree implementation by weka tool

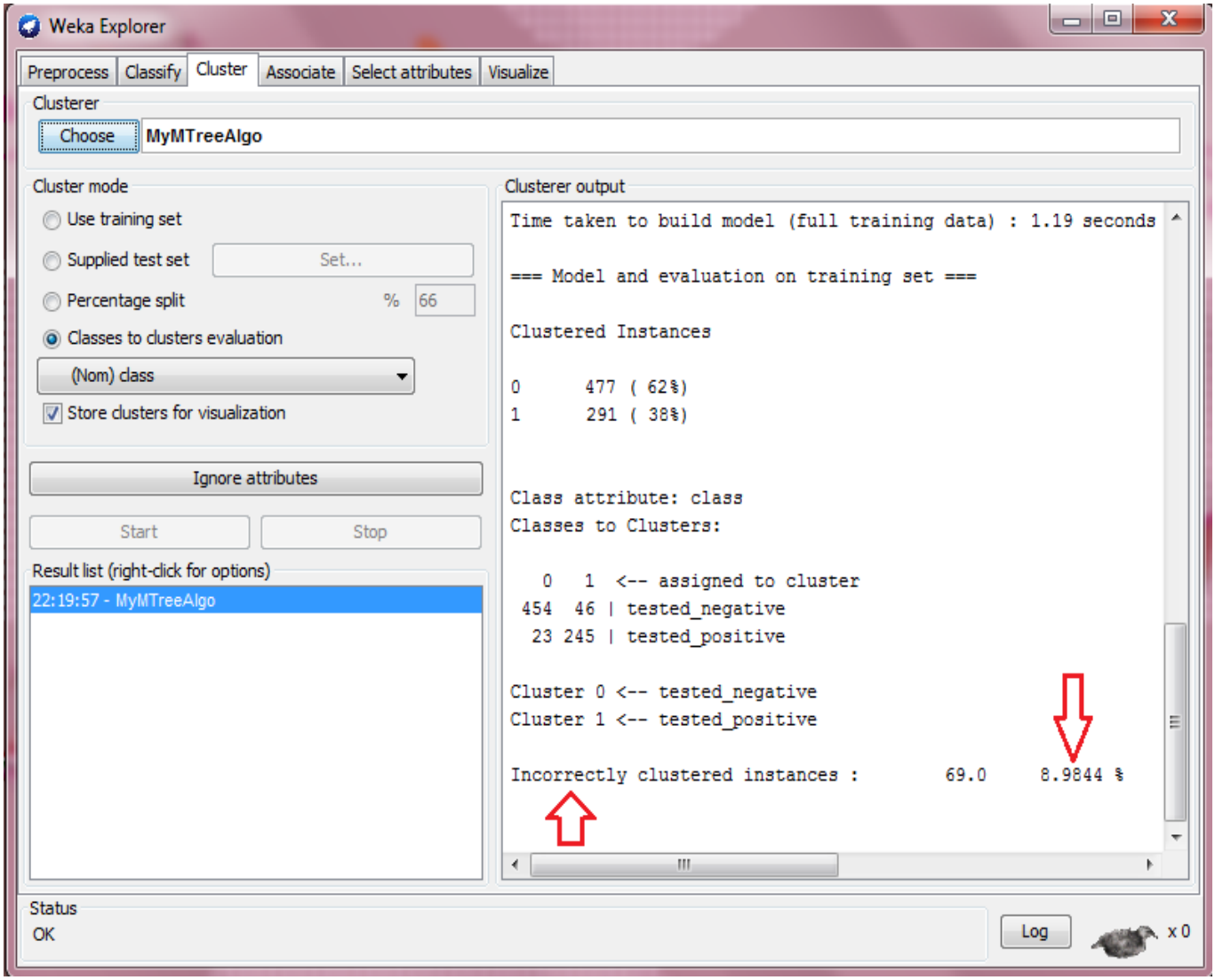

Fig.1

Bench mark of k means algorithm implemented by weka tool

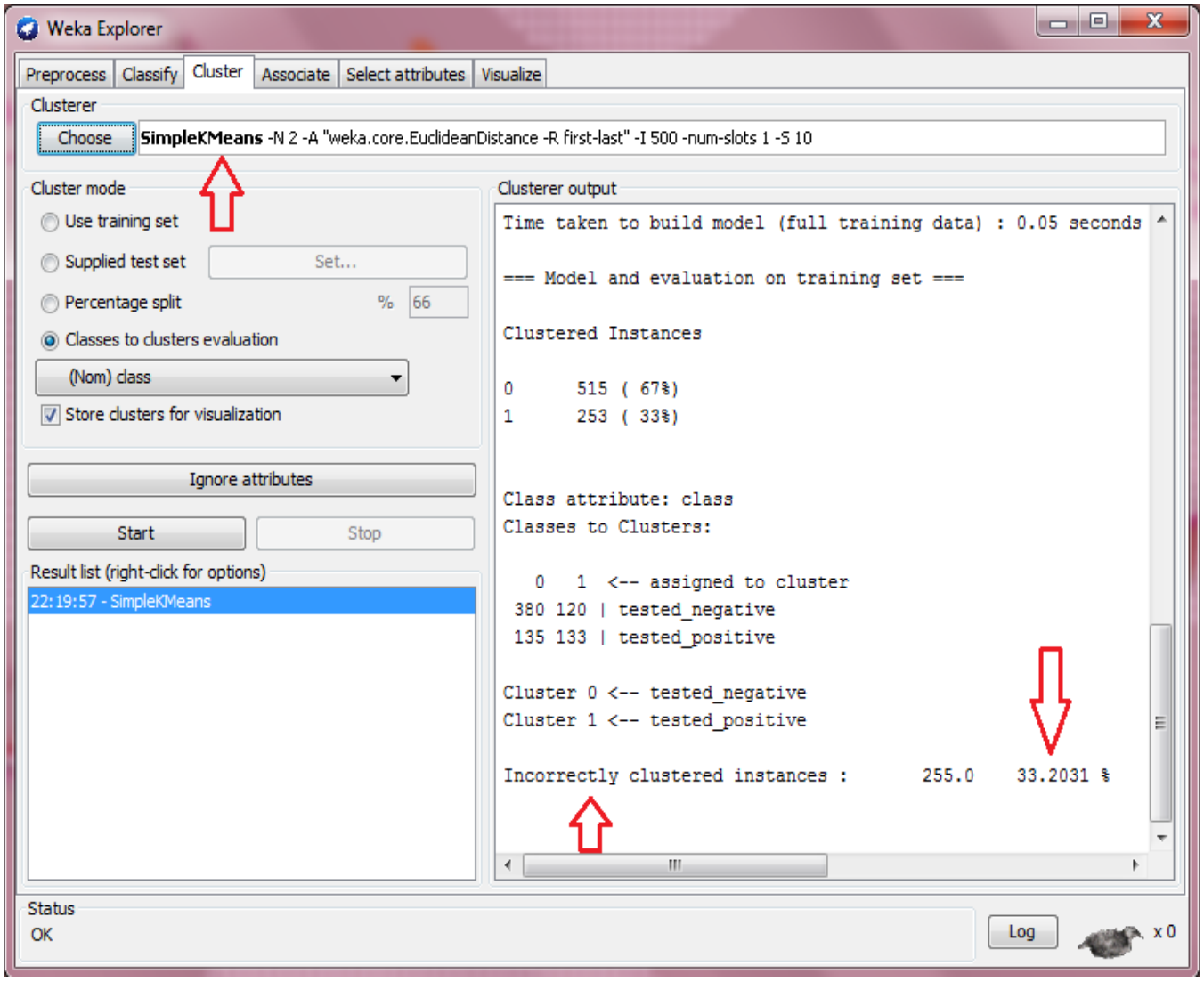

Fig.2

VII. CONCLUSION

The performance of classification algorithm depends on the quality of data. The K-means clustering is used to identify and eliminate incorrectly classified instances. Further the continuous data is converted to categorical data by consulting medical expert's advice. The correctly classified instance by k-means is used as input to M tree after conversion of continuous data to categorical data. The proposed cascaded shows improved classification of 92.33% for PIMA diabetic dataset using 60-40 % training–testing partitioning method with preprocessed data. Further results showed that the performance of cascaded model with categorical data generated comparatively less number of rules which are easy to interpret compared to rules generated by M tree with unprocessed data. The classification accuracies obtained by the proposed cascaded K_ means clustering and M tree classifier is one of the best results compared with the results of M tree reported in the literature.

The M-tree algorithm proposed model of diabetes is highly accuracy of positive test of diabetes patient.

REFERENCES

[1] Polat K., Gunes S., Aslan A., "*A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications",* Expert systems with applications, Vol.34, Issue.1, pp. 214-221, 2008.

[2] B.M Patil, R.C Joshi, Durga Tosniwal, "*Hybrid Prediction model for Type-2 Diabetic Patients, Expert System with*

Applications", Expert systems with applications. , Vol.37, Issue.12, pp.8102-8108, 2010.

[3] J.R. Quinlan, *"Induction of M Trees, Machine Learning",* Kluwer Academic Publishers, Boston, pp.81-106, 1986.

[4] J.Han, M. Kamber, *"DataMining: Concepts and techniques",* Morgan Kauffman Publishers, San Francisco, pp.34-118, 2001.

[5] A. Shrivastava, S. Rajawat, "*An Implementation of Hybrid Genetic Algorithm for Clustering based Data for Web Recommendation System",* International Journal of Computer Sciences and Engineering, Vol.2, Issue.4, pp.6-11, 2014.

[6] Prateeksha tomar , Amit Kumar Manjhvar, *"survey report On various decision tree classification algorithms using weka tool",* International journal of computer scienceandengineering,Vol.5, No. 3, pp.1-8, 2017.

[7] S. Joshi, F.U. Khan, N. Thakur, "*Contrasting and Evaluating Different Clustering Algorithms: A Literature Review",* International Journal of Computer Sciences and Engineering, Vol.2, Issue.4, pp.87-91, 2014.

[8] AG. Karegowda, MA. Jayaram, *"Integrating M Tree and ANN for Categorization of Diabetics Data*", International Conference on Computer Aided Engineering, pp.9-15, India, 2007.

[9] K.Selvi, "*Identify Heart Diseases Using Data Mining Techniques: an Overview*", International Journal of Computer Sciences and Engineering, Vol.3, Issue.11, pp.180-187, 2015.

[10] P. Thangarajum B. Deepa, "*A Case study on Perclusion and Discovery of Skin Melanoma Risk using Clustering Techniques*", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 3, Issue 7, pp.76-82, 2014.

[11] Siti Farhanah, Bt Jaafar and Dannawaty, Mohd Ali, "*Diabetes mellitus forecast using artificial neural networks*", Asian conference of paramedical research proceedings, Malaysia, pp.5-7, 2007.

[12] Karthika Jayprakash,Nidhi Kargathra, Pranay Jagtap,Suraj Shridhar and Archana Gupta, "*Comparison of Classification Techniques for Heart Health Analysis System*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.2, pp.92-95, 2016.

[13] Priyanka, Sana Khan, Tulsi Kour, *"Investigation on Smart Health Care Using Data Mining Methods",* International Journal of Scientific Research in Computer Science and Engineering, Vol.4, Issue.2, pp.31-36, 2016.

AUTHORS PROFILE

Ms. Prateeksha Tomar pursed Bachelor of Engineering from ITM universe in 2011. She is currently pursuing Master of Technology from Madhav Institute of Technology and Science,gwalior from branch cyber security.

Mr Amit Kumar Manjhvar pursed Bachelor of Engineering from branch Computer Engineering and Master of Technology in Software System. He is currently working as Assistant Professor in Department of Computer Science Engineering, Madhav Institute of Technology and Science, Gwalior.