A Novel Technique for Prediction of Diabetic Patients Data Using Naives Bayes Classification in WEKA Tool

S.Nivetha
Assistant Professor, Department of Computer Science,
Kamban Arts and Science College,
Pollachi, Tamil Nadu, India.
Email: nive0501@gmail.com
Dr.A.Geetha
Assistant Professor, PG & Research Department of Computer Science,
Chikkanna Govt. Arts College,
Tirupur, Tamil Nadu, India.
Email: gee_sam@yahoo.com

Abstract - Data mining is a process of extracting information from a dataset and transform it into understandable structure for further use, also it discovers patterns in large data sets. Data mining has number of techniques such as pre-processing, classification. Classification is a technique used for predicting group for the diabetic dataset instance. In this paper, classification on diabetes database are developed classifiers are compared with the result based on certain parameters using WEKA tool. India is suffering from diabetes patients of the population with equal rates in both women and men resulted in deaths with worldwide. A comparative analysis has been performed with the classifiers which result in the chance of diabetic patients getting heart disease. The performances compared with precision, recall, F-measure, Kappa statistic, root mean square and time seconds build the model has exhibited a great overall performance.

Keywords- Data Mining, Diabetic Data, RF, NB, WEKA.

1. INTRODUCTION

Today, information has a great value and the amount of information has been expansively growing during last few years. Especially, text databases are rapidly growing due to the increasing amount of information available in electronic forms.

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. That type of Information can be used to increase revenue, cuts costs or both. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified.

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information.

2. REVIEW OF LITERATURE

In this research work, discussed about the methodology for feature extraction and document classification. In order to propose this works are have analyzed various Literatures which are very much relevant and helpful to do this work. The literature where are have retrieved and analyzed are presented in the following section.

Thangaraju [1] et al., proposed a preclusion and discovery of skin melanoma risk using clustering techniques. The skin melanoma patients data are gathered from different diagnostic centre which contains both cancer and non-cancer patient's information. The gathered data are pre-processed and then clustered using K-means algorithm for separating relevant and non- relevant data to skin melanoma. It is implemented in c#.net to predict skin melanoma risk level with suggestions which is easier, cost reducible and time savable.

Mohd Fauzi bin Othman [2] et al., examined the performance of different classification and clustering methods for a set of bulk data. The algorithm and methods tested are Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm.

Bharat Chaudharil [3] et al., analyzed the comparison of three major clustering algorithms are K- Means, Hierarchical clustering and Density based clustering algorithm. The performances of these three algorithms are compared based on the feature of correctly class wise clustering. The performance of these three clustering algorithms is compared using a Data mining tool WEKA.

Amandeep Kaur Mann [4] et al., presented a clustering and its different techniques in data mining is done. Kawsar Ahmed [11] et.al proposed a system to detect the Lung cancer risk. Their proposed system was easy, cost effective and time saving. The data are collected from different diagnosis centres. The collected data are pre-processed and clustered using K-means algorithm. Then Apriori and Decision tree algorithm are used to find significant frequent pattern. Then they developed a significant frequent pattern tool for lung cancer prediction system.

Rajalingam [6] et al., presented a comparative study of implementation of hierarchical clustering algorithms agglomerative and divisive clustering for various attributes. The Visual Programming Language is used for implementation of these algorithms. The result of this paper study is the performance of divisive algorithm works as twice as fast as the agglomerative algorithm.

Khaled Hammouda [5] et al., presented the reviews of four off-line clustering algorithms are K-means clustering, Fuzzy C-means clustering, Mountain clustering, and Subtractive clustering. The algorithms are implemented and tested against a medical problem of heart disease diagnosis. The accuracy and performance are compared.

Aastha Joshi [7] et al., proposed a brief review of six different types of clustering techniques are K-means clustering, Hierarchical clustering, DBSCAN clustering, OPTICS, and STING.

Manish Verma et al., [8] proposed a analysis of six types of clustering techniques are k-Means Clustering, Hierarchical Clustering, DBSCAN clustering, Density Based Clustering, Optics and EM Algorithm. WEKA tool is used for implemented and analyzed.

Shraddha K.Popat et.al [9] focused on different clustering techniques. They are Partition algorithms, Hierarchical algorithms, Density based clustering algorithm. The result was hierarchical clustering can be perform better than the other techniques.

Pradeep Rai et al., [10] presented a survey is to provide a comprehensive review of different clustering techniques in data mining.

3. DATA MINING TOOL

An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences. The open source tools available for data mining.

3.1 WEKA

Waikato Environment for Knowledge Analysis. Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. Weka contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.

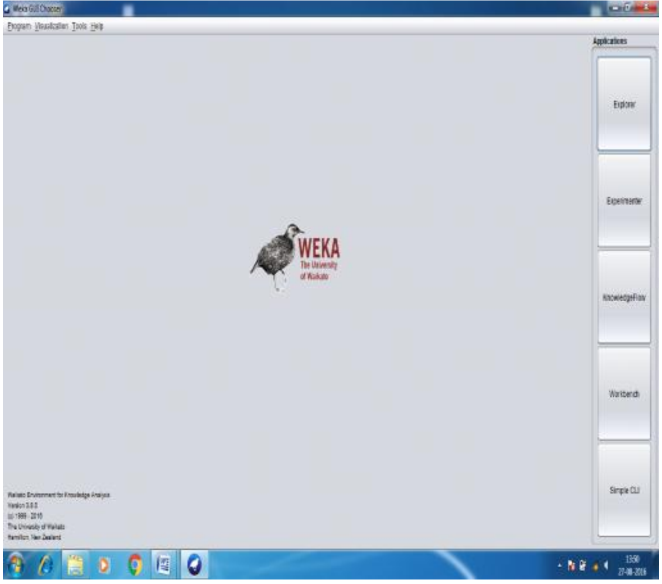
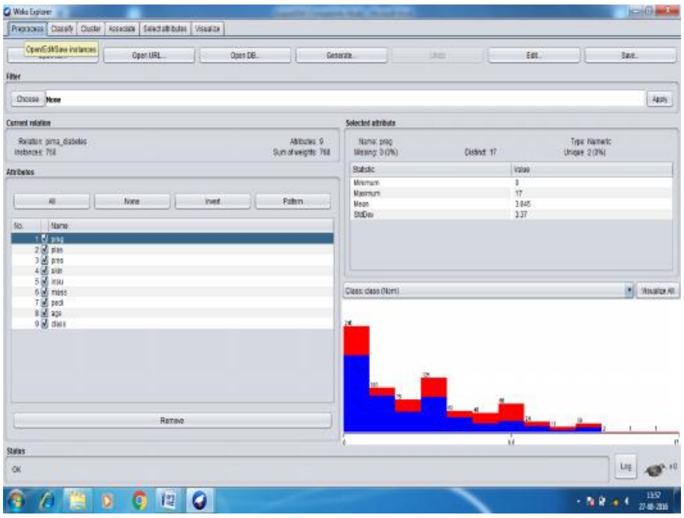


Fig 3.1 Explorer in WEKA



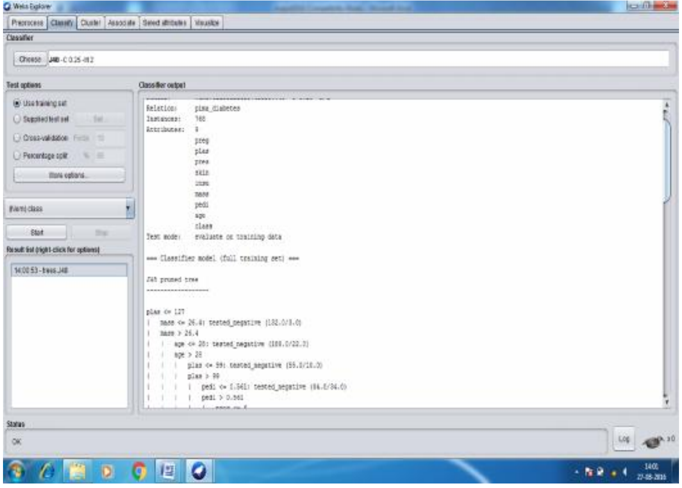
```
Filter: None
Relation: pima_diabetes          Attributes: 9
Instances: 768                   Sum of weights: 768

No.  Name
1  ☑ preg
2  ☑ plas
3  ☑ pres
4  ☑ skin
5  ☑ insu
6  ☑ mass
7  ☑ pedi
8  ☑ age
9  ☑ class

Name: preg                                Type: Numeric
Missing: 0 (0%)        Distinct: 17       Unique: 2 (0%)

Statistic    Value
Minimum      0
Maximum      17
Mean         3.845
StdDev       3.37

OK
```

Fig 3.2 Diabetic Attributes Selected in WEKA

4.1 EXISTING METHODOLOGY

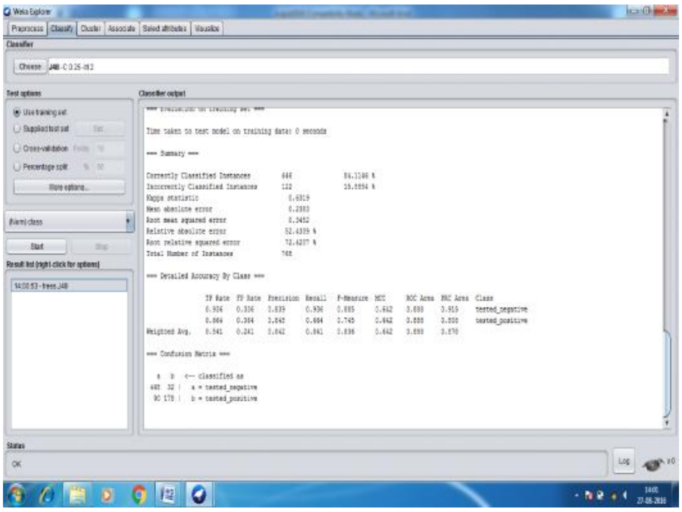
4.1.1 J48 Pruned Tree

J48 is a module for generating a pruned or unpruned C4.5 decision tree. When we applied J48 onto refreshed data, got the results shown as below on Figure.



```
Classifier: J48 -C 0.25 -M 2
◉ Use training set
○ Supplied test set
○ Cross-validation  Folds 10
○ Percentage split  % 66
(Nom) class
14:00:53 - trees.J48

Relation:     pima_diabetes
Instances:    768
Attributes:   9
              preg
              plas
              pres
              skin
              insu
              mass
              pedi
              age
              class
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561

OK
```

Fig 4.1 J48 Diabetic Dataset Classifier Output

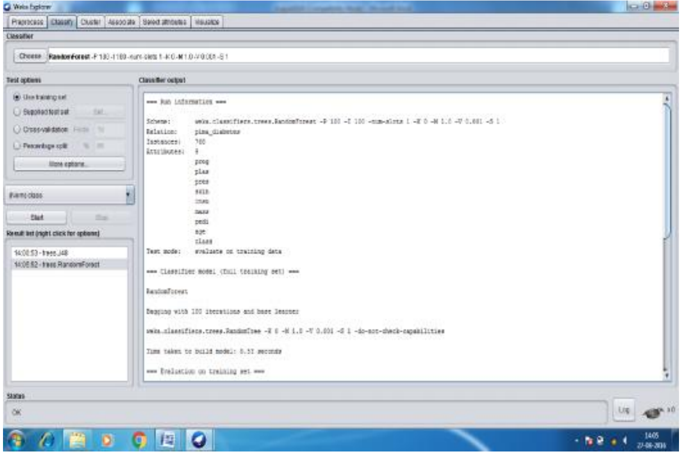


```
Classifier: J48 -C 0.25 -M 2
◉ Use training set
○ Supplied test set
○ Cross-validation  Folds 10
○ Percentage split  % 66
(Nom) class
14:00:53 - trees.J48

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         646               84.1146 %
Incorrectly Classified Instances       122               15.8854 %
Kappa statistic                          0.6319
Mean absolute error                      0.2383
Root mean squared error                  0.3452
Relative absolute error                 52.4339 %
Root relative squared error             72.4207 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.936    0.336    0.839      0.936    0.885      0.642    0.888     0.915     tested_negative
                 0.664    0.064    0.848      0.664    0.745      0.642    0.888     0.808     tested_positive
Weighted Avg.    0.841    0.241    0.842      0.841    0.836      0.642    0.888     0.870

=== Confusion Matrix ===

   a   b   <-- classified as
 468  32 |   a = tested_negative
  90 178 |   b = tested_positive

OK
```

Fig 4.2 J48 Diabetic Dataset Classifier Accuracy

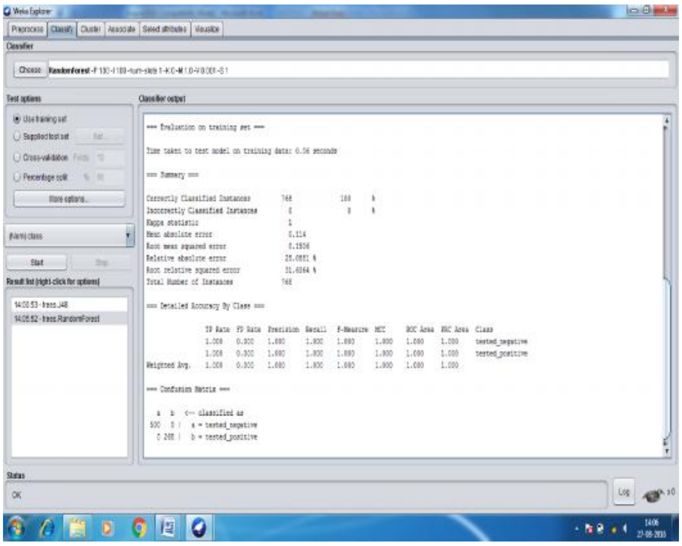
4.1.2 Random Forest

Random forest is an algorithm that consists of many decision trees. It was first developed by Leo Breiman and Adele Cutler [Bre01]. The idea behind it is to build several trees, to have the instance classified by each tree, and to give a vote at each class. The model uses a bagging approach and the random selection of features to build a collection of decision trees with controlled variance. The instances class is to the class with the highest number of votes, the class that occurs the most within the leaf in which the instance is placed. By using trees that classify the instances with low error the error rate of the forest decreases. The correlation and strength of the forest increases with the number m of variables selected. A smaller m returns a smaller correlation and strength. To improve the prediction's accuracy, a bootstrap method is used to create different trees. Every time a tree is created, one-third of the bootstrap sample is kept aside to estimate the test error. The subset that is not included in the tree construction is used as a test set for the same tree. Once a tree is built, it is used to predict all data and this allows computing proximities. Every time two instances fall into the same node, it increases the proximities. This computation is done for each tree and proximities can be used to replace missing values.



```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:     pima_diabetes
Instances:    768
Attributes:   9
              preg
              plas
              pres
              skin
              insu
              mass
              pedi
              age
              class
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.37 seconds

=== Evaluation on training set ===
```

Fig 4.3 Random Forest Diabetic Dataset Classifier Output



```
=== Evaluation on training set ===

Time taken to test model on training data: 0.06 seconds

=== Summary ===

Correctly Classified Instances         768              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.114
Root mean squared error                  0.1536
Relative absolute error                 25.0851 %
Root relative squared error             31.6064 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     tested_negative
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     tested_positive
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 500   0 |   a = tested_negative
   0 268 |   b = tested_positive
```

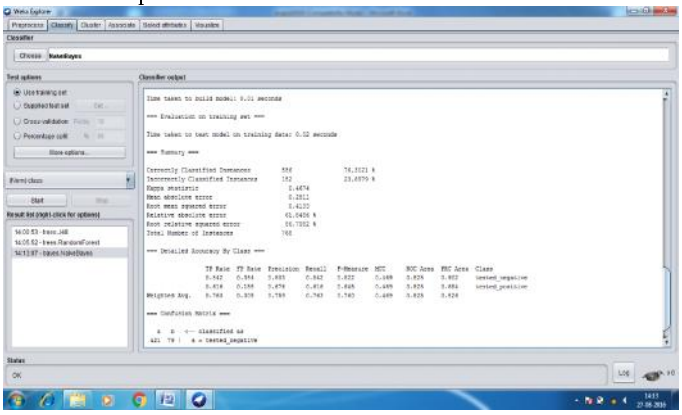
Fig 4.4 Random Forest Diabetic Dataset Classifier Accuracy

4.2 PROPOSED METHODOLOGY

Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the able purpose of being to use the model to predict the class of objects whose class label is unknown. It is a technique which is used to predict group membership for data instances. Classification is a two step process, first, it builds classification model using training data. Every object of the dataset must be pre-classified and the second the model generated in the preceding step is tested by assigning class labels to data objects in a test dataset. Here using diabetes dataset now a days the percentage of diabetes patient is growing very fast. India accounts for the largest number of people suffering from diabetes in the world. The diabetes in country's population is likely to be affected from the disease. It is estimated that every five person with diabetes will be an Indian. It means that India has highest number of diabetes in any one of the country in the world. The attributes predict whether a person having diabetes or not.

4.2.1 Naive Bayes Approach

Naive Bayes classifier as a term dealing with a simple probabilistic classifier based on application of Bayes theorem with strong independence assumptions. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature. It is based on conditional probabilities. It uses Bayes' theorem which finds the probability of an event occurring, given the probability of another event that has already occurred. An advantage of the Naive Bayes classifier is that it requires only a small amount of training data to estimate the parameters necessary for classification. Since independent variables are assumed, only the variances of the variables for each class need to be determined. It can be used for both binary and multi class classification problems. Naive Bayes data mining classifier technique has been applied which produces an optimal prediction model using minimum training set to predict the chances of diabetic patient getting heart disease. The diagnosis of diseases plays vital role in medical field. Using diabetic's diagnosis, the proposed system predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. It should be noted that the attributes used in our proposed method are those used for diagnosis of diabetes and are not direct indicators of heart disease. Each algorithm requires submission of data in a specified format.



```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4674
Mean absolute error                      0.2811
Root mean squared error                  0.4133
Relative absolute error                 61.8406 %
Root relative squared error             86.7082 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.842    0.384    0.803      0.842    0.822      0.469    0.825     0.902     tested_negative
                 0.616    0.158    0.678      0.616    0.645      0.469    0.825     0.684     tested_positive
Weighted Avg.    0.763    0.305    0.759      0.763    0.760      0.469    0.825     0.826

=== Confusion Matrix ===

   a   b   <-- classified as
 421  79 |   a = tested_negative
```

Fig 4.6 Naive Bayes Diabetic Dataset Classifier Accuracy

5. EXPERIMENTATION & RESULTS

5.1 Performance evaluation

To measure the performance sensitivity, accuracy and specificity are used. TP is true positive, FP is false positive, TN is true negative and FN is false negative. TPR is true positive rate, which is equivalent to Recall.

$$\text{Sensitivity}=\frac{\text{True Positive Rate}}{(\text{True Positive + False Negative})} \quad \text{...... 1}$$

$$\text{Specificity}=\frac{\text{True Negative Rate}}{(\text{False Positive + True Negative})} \quad \text{...... 2}$$

5.2 False Positive Rate

The ratio of false positives to false positive plus true negatives. It can be defined as

$$\text{False Positive Rate}=\frac{\text{False Positive}}{(\text{False Positive + True Negative})} \quad \text{...... 3}$$

5.3 Precision

Precision is calculated as number of correctly classified instances can be defined as

$$\text{Precision}=\frac{\text{True Positive}}{(\text{True Positive + False Positive})} \quad \text{...... 4}$$

5.4 F-Measure

F-measure is combining recall and precision scores into a single measure of performance.

$$\text{F-Measure}=\frac{\text{2*recall*precision}}{(\text{recall + precision})} \text{...... 5}$$

Methods / Parameters	J48 tree	Random Forest	Naive Bayes Classifier
Number of Instances	768	768	768
TP Rate	0.936	1.000	0.842
FP Rate	0.336	0.001	0.384
Precision	0.839	1.000	0.803
Recall	0.936	1.000	0.842
F-Measure	0.885	1.000	0.822
Kappa statistic	0.6319	1	0.4674
Mean absolute error	0.2383	0.114	0.2811
Root mean squared	0.3452	0.1506	0.4133
Time taken	0.02 sec	0.53 sec	0.01 sec

Table.5.1 Comparison Results

From the above table 5.1 shows the performance of naive bayes classifier. The fig.5.1 shows comparison graphical representation of methods. The method can over perform the traditional method with classify recall rate of 0.842.

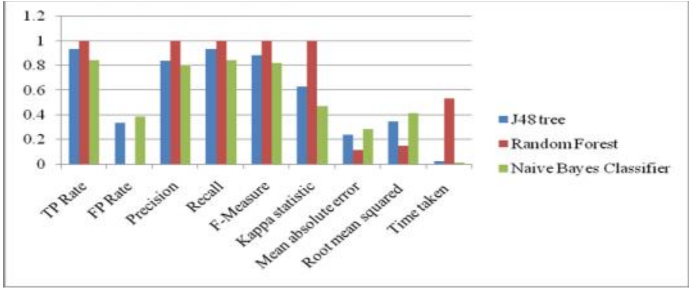


Fig.5.1 Comparison Results

6. CONCLUSION

Data mining plays a important role in extracting the hidden information in the diabetic data base. The preprocessing is used in order to improve the quality of the data. This model is built based as a test case on the diabetic dataset. The experiment has been successfully performed with several data mining classifier techniques and naives bayes gives a better performance than other existing methods. To improve the quality of health care of diabetes patients can be implemented using WEKA tool. In this paper, considered time taken to build a model by the algorithms as a parameter.

REFERENCES

1. P.Thangaraju, B.Deepa, “A Case study on Perclusion and Discovery of Skin Melanoma Risk using Clustering Techniques”, International Journal of Advanced Research in Electronics & Communication Engineering, Vol.3,Iss.7, 2014.
2. Mohd Fauzi bin Othman, Thomas Moh Shan Yau, “Comparison ofDifferent Classification Techniques using WEKA for BreastCancer”, F.Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri: Biomed 06, IFMBE Proceedings 15, pp.520-523, 2007.
3. Bharat Chaudharil, Manan Parikh, “A Comparative Study of clustering algorithms using weka tools”, International Journal of Application or Innovation in Engineering & Management, Volume 1, Issue 2, 2012.
4. Amandeep Kaur Mann and Navneet Kaur, “Survey Paper on Clustering Techniques”, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, 2013.
5. Khaled Hammouda, Prof. Fakhreddine Karray, “A Comparative Study of Data Clustering Techniques”, University of Waterloo, Ontario, Canada.
6. N. Rajalingam, K. Ranjini, “Hierarchical Clustering Algorithm, A Comparative Study”, International Journal of Computer Applications, Vol.19, No 3, 2011.
7. Aastha Joshi, Rajneet Kaur,“A Review: Comparative Study of Various Clustering Techniques in Data Mining”, International journal of Advanced Research in Computer Science & Software Engineering, Vol.3, Iss.3, 2013.
8. Manish Verma, “A Comparative Study of Various Clustering Algorithms in Data Mining” International Journal of Engineering Research and Applications, Vol. 2, Issue 3, pp.1379-1384,2012.
9. Shraddha K. Popat , “Review and Comparative Study of Clustering Techniques” International Journal of Computer Science & Information Technologies,Vol.5(1), 805-812, 2014.
10. K. Ruth Ramya, “A Class Based Approach for Medical Classification of Chest Pain”, International Journal of Engineering Trends and Technology, Vol. 3, Issue.2, pp.89-93, 2012.
11. Ru Xu, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Networks, Vol.16, No 3, 2005.
12. A.Geetha,G.M.Nasira, “Rainfall Prediction using Logistic Regression Technique”,CiiT International Journal of Artficial Intelligence Systems and Machine Learning,Vol.6,No.7,pp.246-250,2014.
13. Sripriya, A.Geetha, “Cyclone Storm Prediction using KNN Algorithm”, Indian Journal of Engineering, Discovery Publication,12(30), pp.350-354,2015.