A comparative study of classification algorithm on blood transfusion

S.Asha Rani[1],Dr.S.Hari Ganesh[2],

[1]Department of Computer science, Bishop Heber College, Trichirapalli, India, [2]Department of Computer science, Bishop Heber College, Trichirapalli, India.

Email: bluewhirlpool999@gmail.com[1], hariganesh17@gmail.com[2.]

ABSTRACT

Requirement of blood is increasing gradually due to accidents, surgeries etc.Blood transfusion play an important role in healthcare. Blood donor prediction provide essential details to medical professionals to increase the number of voluntary blood donors in future .Data mining is the process of collecting relevant data from enormous amount of data . This paper focus on analyzing the efficiency of different classification algorithm in data mining using blood transfusion dataset

KEYWORDS: Data mining, Blood Bank, Blood transfusion dataset, Blood donor, classification algorithms

1. INTRODUCTION

Many developed countries, blood donors are mostly regular and volunteer donors and new volunteer donors are also increased gradually day by day. In developing countries, blood donors are paid and not volunteer .They are mostly donate their blood when their relations or friends are in need of blood.

2. RELATED WORK

Arvind Sharma [C], the main purpose of his research work is to predicting the blood donors through their age and blood group. In this paper , measure the accuracy of blood donors prediction using real time data set. WEKA tool used in this research. In conclusion, he said that blood donor's behaviors and attitude predicted using various data mining techniques.

Vikram singh [D] ,in this research work ,they explain how they acquire interactive KDD process in less effort with low time consumption by interaction between human and computer system. Blood transfusion dataset from UCI repository is used for this research work.

Shyam sundaram [E] ,the main purpose of this research work is to test whether the particular person is RVD (Regular Voluntary Donor) or not. DB2K7 (Donated Blood in 2007) and RVD (Regular Voluntary Donor) are two models used for testing.RVD model show better accuracy than DB2K7.

Alaa hamouda et al [F], RBC counted automatically using image processing techniques. In this RBC counting, learning through decision tree techniques shows more accuracy than other techniques.

Shyam sundaram et al. [G], predict the regular voluntary donors form various locations in India. Compare various locations in India and identify voluntary donors from which location .Blood transfusion dataset from UCI repository used geo location attribute is added randomly with this dataset.

Ivana D.Radojevic et al. [H] ,water quality analyzed by using coliforms presented in reservoirs. Classification and clustering techniques used for total coliform analysis.

Wen-Chen Lee et al.[I] created the system using classification and clustering techniques for predicting blood donors behaviours. Understanding blood donor's behaviours increase voluntary blood donor's rate gradually. Classification techniques such as Naive bayes ,NB tree are used for blood donor behavior prediction.

3. DATA SET DESCRIPTION

Dataset used for testing in this paper is taken from UCI repository. The dataset details collected from 748 donors from donor database.

This dataset consist of five attributes.

- ❖ Recency (months since last donation)
- ❖ Frequency (Total number of donation)
- ❖ Monetary (total blood donated in c.c)
- ❖ Time (months since first donation)

Whether the person donated blood in 2007
(Donated blood in 2007:1 otherwise 0)

Table 1: Attributes used in dataset

Sl.No	Attributes	Types
1	Recency	Numeric
2	Frequency	Numeric
3	Monetary	Numeric
4	Time	Numeric
5	Whether he/she donated blood in 2007(0 or 1)	Nominal

4. TECHNIQUES

We apply different classification techniques to diabetes dataset and the error results obtained is tabulated in table given below.

Table 2: Algorithm Comparison based on time taken

Classification Technique	Time taken to Build (sec)
Naive Bayes	0.03
J48	0.01
Random Tree	0.01

The graphical comparison of various classification algorithms can be given as below

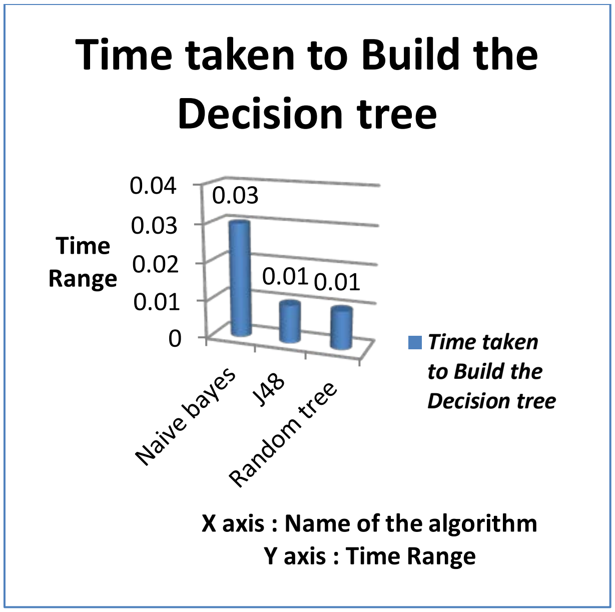


Fig 4.1: Graph for comparison of algorithm by time taken to build

Table 3: Algorithm comparison based on accuracy

Classification algorithm	Error rate	accuracy
Naive Bayes	0.2906	75%
J48	0.2824	80.88%
Random tree	0.0895	93.18%

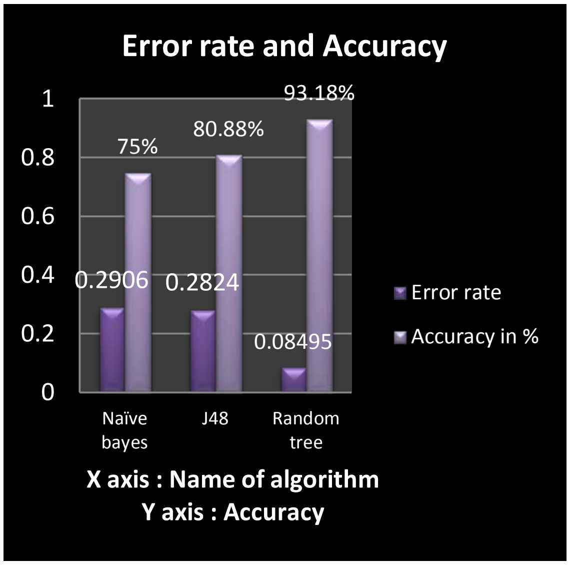
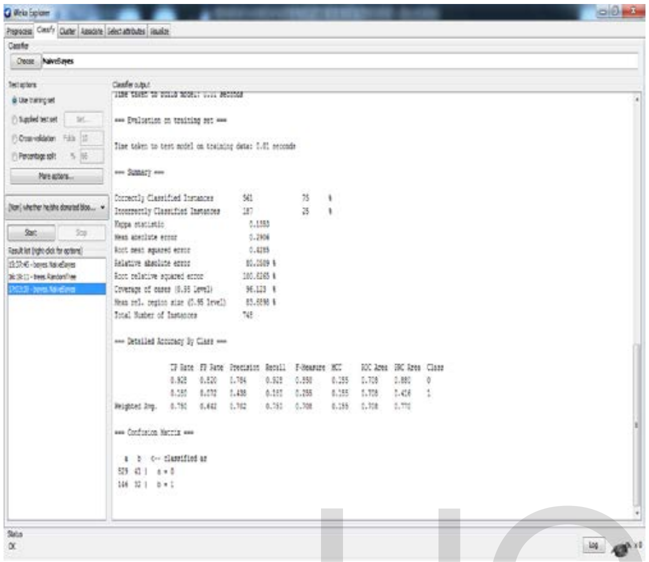


Fig 4.2: Graph for comparison of algorithm by Error rate and accuracy

Naive Bayes:

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. It is particularly suited for large dataset contain enormous amount of Inputs. Naive bayes classifiers uses all attributes in the dataset and analyzes individually and make it independent of each other. Despite of its simplicity it outperforms many complex algorithms.



Classifier: NaiveBayes

Result list:
15:37:45 - bayes.NaiveBayes
16:18:11 - trees.RandomTree
17:20:39 - bayes.NaiveBayes

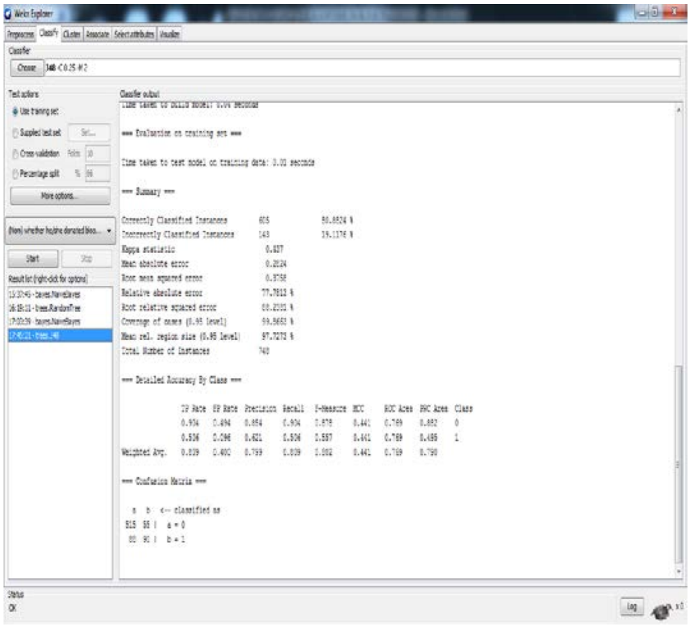
```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         561               75      %
Incorrectly Classified Instances       187               25      %
Kappa statistic                          0.1383
Mean absolute error                      0.2906
Root mean squared error                  0.4285
Relative absolute error                 80.2509 %
Root relative squared error            100.5265 %
Coverage of cases (0.95 level)          96.123  %
Mean rel. region size (0.95 level)      85.5896 %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.928    0.820    0.784      0.928    0.850      0.155    0.708     0.880     0
                 0.180    0.072    0.438      0.180    0.255      0.155    0.708     0.416     1
Weighted Avg.    0.750    0.642    0.702      0.750    0.708      0.155    0.708     0.770

=== Confusion Matrix ===

   a   b   <-- classified as
 529  41 |   a = 0
 146  32 |   b = 1
```

Fig 4.3: Naive Bayes

J48:

J48 decision tree is the implementation of algorithm ID3(Iterative Dichotomiser 3).Decision tree is the predictive machine learning model.J48 build decision tree based on the set of training data .

Decision tree is the most useful and important in classification . Each attribute in the dataset is used. The fundamental idea that used in decision tree is to divide the data into smaller subset that used for easy way of classification. Decision tree that always provides an easy way of understanding the classification by splitting the data.J48 accept both continuous and discrete attributes for classification.



Classifier: J48 -C 0.25 -M 2

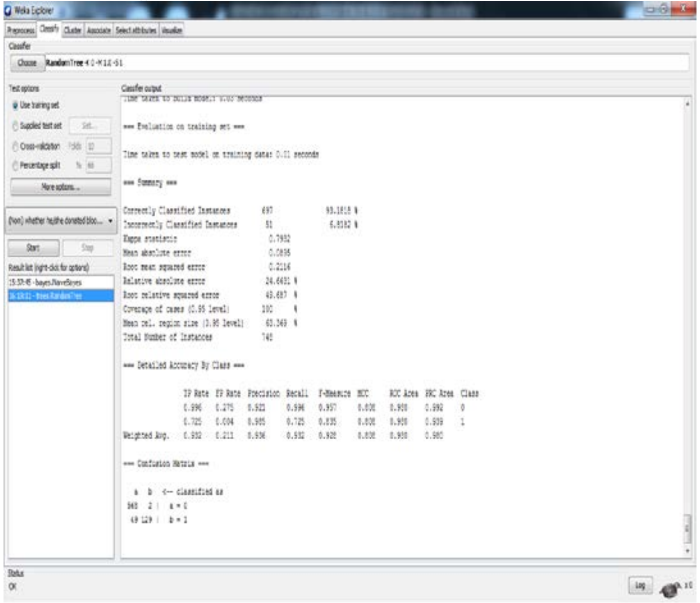
Result list:
15:37:45 - bayes.NaiveBayes
16:18:11 - trees.RandomTree
17:20:39 - bayes.NaiveBayes
17:40:21 - trees.J48

```
Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         605               80.8824 %
Incorrectly Classified Instances       143               19.1176 %
Kappa statistic                          0.457
Mean absolute error                      0.2824
Root mean squared error                  0.3758
Relative absolute error                 77.7812 %
Root relative squared error             88.2301 %
Coverage of cases (0.95 level)          99.8662 %
Mean rel. region size (0.95 level)      97.7273 %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.904    0.494    0.854      0.904    0.878      0.441    0.769     0.882     0
                 0.506    0.096    0.621      0.506    0.557      0.441    0.769     0.495     1
Weighted Avg.    0.809    0.400    0.799      0.809    0.802      0.441    0.769     0.790

=== Confusion Matrix ===

   a   b   <-- classified as
 515  55 |   a = 0
  88  90 |   b = 1
```

Fig 4.4: J48

RANDOM TREE:

Random tree that both work with classification and regression problems. It construct multiple decision tree randomly. while in building each tree, this algorithm get remaining feature randomly at each node without any check such as information gain,gini index etc. It is a collection of tree predictors called forest. This tree algorithm stop it growth when no more node to split or node is empty. This algorithm not require any accuracy estimation procedures like Bootstrap.



Classifier: RandomTree -K 0 -M 1.0 -S 1

Result list:
15:37:45 - bayes.NaiveBayes
16:18:11 - trees.RandomTree

```
Time taken to build model: 0.03 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         697               93.1818 %
Incorrectly Classified Instances        51                6.8182 %
Kappa statistic                          0.7982
Mean absolute error                      0.0895
Root mean squared error                  0.2116
Relative absolute error                 24.6431 %
Root relative squared error             49.687  %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      63.369  %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.275    0.921      0.996    0.957      0.808    0.958     0.992     0
                 0.725    0.004    0.985      0.725    0.835      0.808    0.958     0.939     1
Weighted Avg.    0.932    0.211    0.936      0.932    0.928      0.808    0.958     0.980

=== Confusion Matrix ===

   a   b   <-- classified as
 568   2 |   a = 0
  49 129 |   b = 1
```

Fig 4.5: Random tree

5. DISCUSSION

The final attribute used to find whether person donate blood in 2007 or not.

Table 4: confusion matrix

TP	TN
FP	FN

TP-True Positive
FP-False Positive
TN-True Negative
FN-False Negative
748 instances used in this dataset

Table 5: Confusion matrix of Random tree:

a	b	Classified as
568	2	a=0
49	129	b=1

a = not donate blood in 2007
b = donate blood in 2007
True positive:
Person not donates blood in 2007 correctly identified as not donated in 2007.
False positive:
Persons whose are not donate blood in 2007 incorrectly identified as donate blood in 2007
True negative:
Persons donate blood in 2007 correctly identified as donate blood in 2007
False negative:
Donated persons incorrectly identified as not donate blood in 2007

Table 6: Confusion matrix of J48:

a	b	Classified as
515	55	a=0
88	90	b=1

Table 7: Confusion matrix of Naive Bayes:

a	b	Classified as
529	41	a=0
146	32	b=1

6. CONCLUSION

This paper is used to comparison of various algorithms in classification and to measure the accuracy with low time consuming for testing.
The algorithm Random tree has shows 93.18% accuracy within short duration when compared with other algorithms in classification. This comparison among algorithms is used to provide proper utilization of best algorithm to give correct solution to problems in short time. This comparison should be used in further proceedings in scientific research and in prediction of blood donors.

7. REFERENCES

[A] Han, J., Kamber, M.: "Data Mining; Concepts and Techniques", *Morgan Kaufmann Publishers March 2006. ISBN 1-55860-901-6*

[B] Pieter Adrians, Dolf Zantinge: "Data Mining", *Addison Wesley, 2000*

[C] Arvind Sharma and P.C.Gupta, "Predicting the number of blood donors through their age and blood group by using data mining tool", *International journal of communication and computer technologies (IJCOTS), vol-1, and issue: 02, September 2012.*

[D] Vikram Singh and Sapna Nagpal, "Interactive knowledge discovery in blood transfusion data set", *VSRD international journal of computer science and information technology (VSRD-IJCSIT, vol-1(8), 2011, 541-517.*

[E] T.Santhanam and Shyam sundaram, "Application of CART algorithm in blood donors classification", *Journal of Computer science 6(5), 548-552, 2010.*

[F] Alaa Hamouda, Ahmed Y.Khedr and Rabie A.Ramadan ," Automated Red Blood Cell Counting", *International journal of computing science,vol.1,No.2,february 2012.*

[G] Shyam sundaram and Santhanam.T, "A comparison of blood donors classification data mining models", *Journal of theoretical and applied information technology, vol 30 No.2,*
31 august2011

[H] Ivana D.Radojevic et al, "Total coliforms and data mining as a tool in water quality monitoring", *African journal of microbiology research, vol.6 (10), 16 march 2012*

[I] wen-chanLee and bor-wen cheng, "An intelligent system for improving performance of blood donation", *Journal of quality vol.18, Issue No.11, 2011.*