ISSN:2320-0790

An International Journal of Advanced Computer Technology

Performance Analysis of Various Data mining Classification Algorithms on Diabetes Heart dataset

G.G.Gokilam[1], Dr K.Shanthi[2]

[1]Research scholar, [2]Research Guide

Department of Computer Science and Engineering, PRIST University, Thanjavur, TamilNadu

Abstract: This Data mining techniques are applied in building software for fast and easy classification models. Early identification has high-risk modules also likely to have a high number of faults. Classification tree models are simple and effective as software quality prediction models while predictions of defects from such models can be used to achieve high software reliability. In this paper, the performance of some data mining classifier algorithms named J48, Random Forest, Random Tree, REP and Naïve Bayesian classifier (NBC) are evaluated based on 10 fold cross validation test. Diabetes is the most rapidly growing chronic disease of our time. People with diabetes are more likely to cause of new blindness, kidney disease, amputation and cardiovascular disease (heart disease and stoke). In this paper we take diabetes and heart datasets relate with their matching fields then apply the classification algorithm in diabetes heart dataset in WEKA (software tool) finding weather people affected by diabetes are getting chance to get heart disease or not, output are evaluated as Tested Negative (No Diabetes), Tested Normal(Not affected), Tested High(affected).

Keywords: Data mining, classification algorithms, Diabetes, Heart problem, WEKA Tool.

I. INTRODUCTION

Data mining is a process of finding the useful patterns from huge set of data. Data mining is the analysis of data sets to find unsuspected relationships to summarize the data in novel ways that are both understandable and useful to the data owner. For Data mining analysis data are deals with already collected data mining [1]. It have some techniques like classification, clustering, Pattern Evaluation, Association, etc., Here classification is a supervised learning approach also a tree based structure. It is assigning an object to a certain class based on its similarity to the previous examples of other object. Classification according to the kinds of knowledge mined, Database mined, Technique utilized, application adapted. The used classification techniques commonly build models that are used to predict future data trends. There are several algorithms for data classification such as decision tree, j48, Random Tree, Random Forest and Naïve Bayes classifiers. With classification, the structural model shown in (Figure: 1) can be designed and applied algorithms for their process.

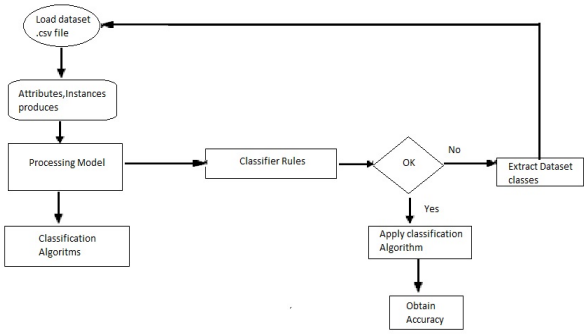


Figure 1: Structural Design for the process.

Diabetes is a disease in which the body's blood glucose (sugar) level is too high. Normally, the body breaks down food into glucose and carries it to cells throughout the body. The cells use a hormone called insulin to turn the glucose into energy [2]. Diabetes is generally of 2 kinds: type 1(insulin dependent diabetes) and type 2(non-insulin-dependent diabetes). Diabetes is a prolonged medical disease. In diabetes, the cells of a person produce insufficient amount of insulin or defective insulin or may unable to use insulin properly and efficiently that further leads to hyperglycemia and type-2 diabetes. People with

diabetic have chance to get some other problem like Heart Disease, Eye Complications ,Kidney Disease , Nerve Damage (neuropathy), Foot Problems, Skin Complications ,Dental Disease. People with diabetes also tend to develop heart disease or have strokes at an earlier age than other people. If you are middle-aged and have type 2 diabetes, some studies suggest that your chance of having a heart attack is as high as someone without diabetes who has already had one heart attack [3]. Women who have not gone through menopause usually have less risk of heart disease than men of the same age [4].

WEKA is most powerful Data mining Tool Created by researchers at the University of Waikato in New Zealand. It is open source, free, Extensible. It's GUI (Graphical User Interface) relatively easy to use. Its features like run an individual experiment or build the KDD (Knowledge Discover Data) phases. It functions like Preprocessing Filters, Attribute selection, Classification/Regression, Clustering, Association discovery, Visualization. Different types of classification algorithms are compared by using diabetes heart dataset in WEKA Tool.

II. DATA PREPARATION

Normally for diabetes dataset can be processed by pima Indian. It includes name of the attribute as well as the explanation of the attributes. Indian Council of medical Research–Indian Diabetes (ICMR-INDIAB). Attribute such as 1)Number of times pregnant (preg),)plasma glucose concentration a two hours in an oral glucose tolerance test (plasma), 3)diastolic blood pressure(pres). 4)Triceps skinfold thickness(skin), 5) Two hour serum insulin(insu) 6)Body mass index(mass) 7)pedigree function(pedi), 8Aage limit(age), 9)Class variable(class) Heart dataset also have some attributes like 1)Age limit(age))obesity (obes), 3)Heart rate(heart), 4)chest pain (chest), 5)Blood pressure(pres), 6)Blood sugar(insu), 7)Cholesterol[5] . From the both dataset several fields are common like age, pressure, sugar, so new dataset was create with the fields like preg, plas, pres, skin ,insu ,mass, pedi, age, heart, pain, chols, class and named as diabetes_heart.Using the dataset some decicion making algorithms in classification techniques are applied and find weather the person have the problem of diabetes or not, if diabetes then there is chance of heart problem or not. This diabetes_heart dataset are applied in WEKA tool ,It produce 12attributes and 112 Instances in Figure 2.

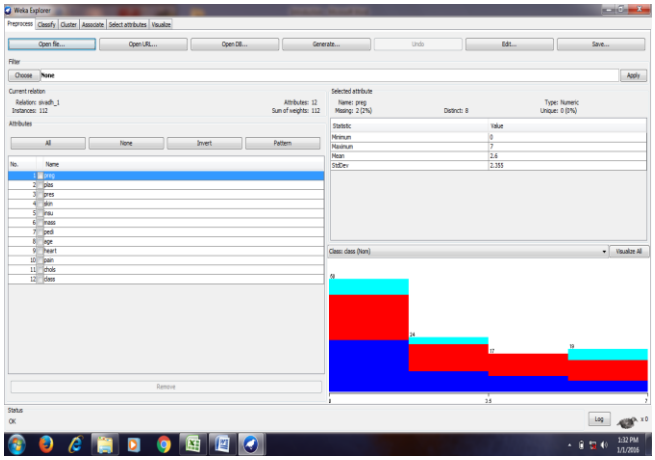


Figure 2: Instance and Attributes produced in WEKA Tool

III. CLASSIFICATION ALGORITHMS

A. Decision Tree.

Decision tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or *terminal node*) holds a class label. The topmost node in a tree is the root node. Decision trees can handle high dimensional data [6]. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand.

Procedure

Step 1: Create a node (where every database have some important) i:e base among all element called root.

Step 2: Based on the condition split the elements in database.

Step 3: Check weather all tuples are relevant to the base

Step 4: Repeat the step until find leaf node, then tree will obtain.

Step 5:Here diabetes heart dataset handle three outcomes, first person check diabetic if its outcome may yes or no, if yes means check the person affected by heart disease or not

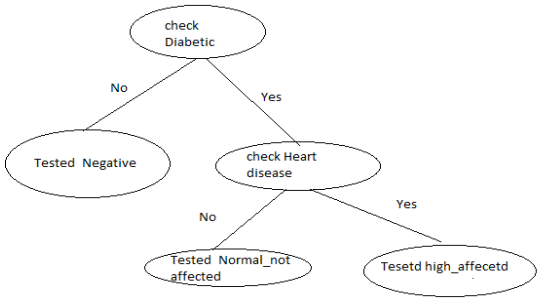
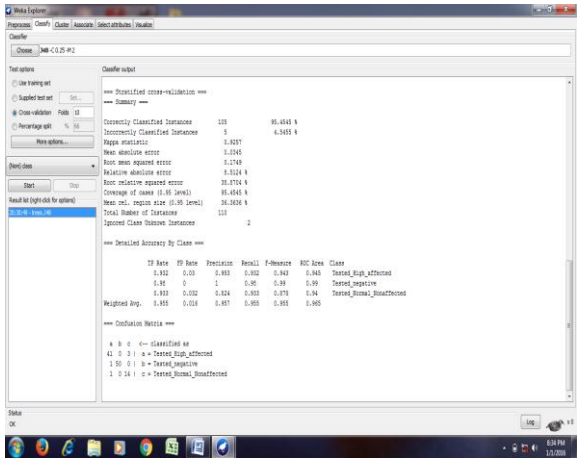


Figure 3: Tree structure for analysis the problem.

B. J48 Algorithm.

J48 is an extension of ID3. J48 is an open source Java implementation of the C4.5 algorithm. The J48 classifier algorithm works, In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain[7]. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained. We then look for another attribute that gives us the highest information gain. Hence we continue in this manner until we either get a clear decision of what combination of attributes gives us a particular target value, or we run out of attributes. In the event that we run out of attributes, or if we cannot get an unambiguous result from the available information, we assign this branch a target value that the majority of the items under this branch possess, using the diabetes_heart dataset apply technique in WEKA Tool for finding is accuracy. Apply 10fold cross validation in training dataset then obtain result by confusion matrix it shown in the Figure 4.

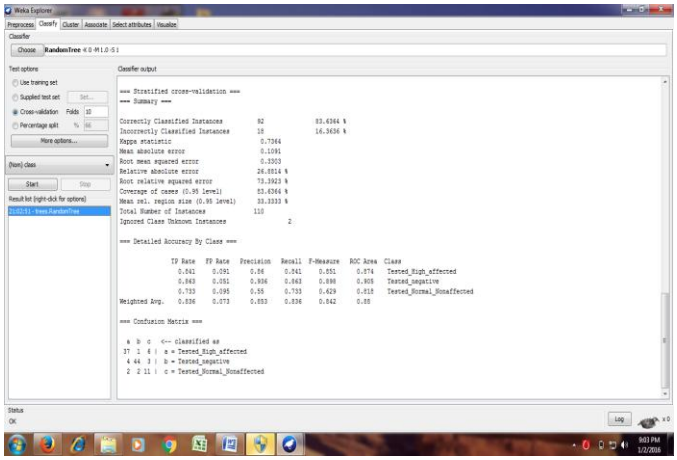


```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 41  0  3 |  a = Tested_High_affected
  1 50  0 |  b = Tested_negative
  1  0 14 |  c = Tested_Normal_Nonaffected
```

Figure 4: Output Matrix for J48.

C. Random Tree.

A random tree is a tree constructed randomly from a set of possible trees having K random features at each node. "At random" in this context means that each tree has an equal chance of being sampled among the set of trees. Or we can say that trees have a "uniform" distribution. Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models. There has been an extensive research in the recent years over Random trees in the field of machine Learning. Apply this technique in WEKA Tool using diabetes_heart dataset of 10fold cross validation and result was shown in Figure 5.



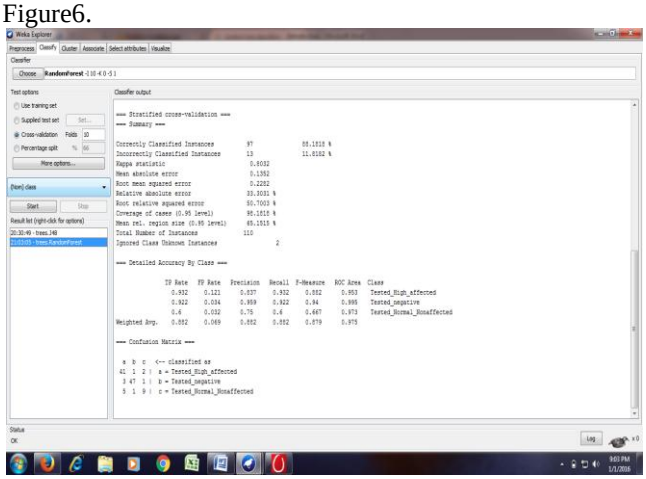
```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 37  1  6 |  a = Tested_High_affected
  4 44  3 |  b = Tested_negative
  2  2 11 |  c = Tested_Normal_Nonaffected
```

Figure 5: Output Matrix for Random Tree.

E. Random Forest.

Random tree is a collection of tree predictors called *forest*. Random Forest algorithm was initially developed by Leo Breiman, a statistician at the University of California Berkeley [8]. Random Forests is a method by which one can calculate accuracy rate in better way. The random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". All the trees are trained with the same parameters but on different training sets. These sets are generated from the original training set using the bootstrap procedure: for each training set, you randomly select the same number of vectors as in the original set. Using the diabetes_heart dataset apply 10fold cross validation in training dataset then obtain result by confusion matrix in WEKA Tool and shown in the

Figure6.

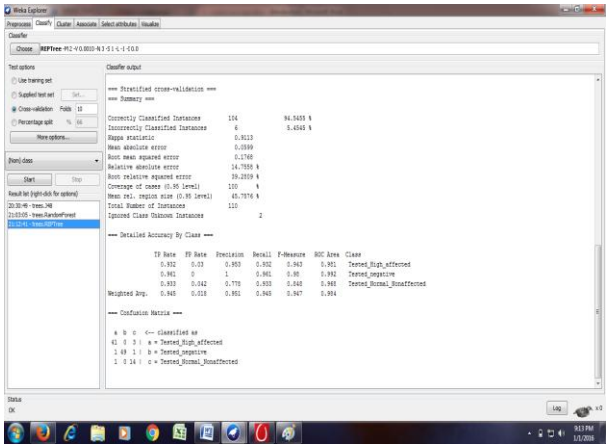


```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 41  1  2 |  a = Tested_High_affected
  3 47  1 |  b = Tested_negative
  5  1  9 |  c = Tested_Normal_Nonaffected
```

Figure 6: Output Matrix for Random Forest.

F. Reduced Error Pruning (REP) Tree

REP Tree algorithm was introduced by Quinlan [9]. It is a fast decision tree learner which builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). It is the simplest and most understandable method in decision tree pruning. For every non leaf sub tree of the original decision tree, the change in misclassification over the test set is examined. Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces and prune the set then node is pruned. This procedure will continue until any further pruning would decrease the accuracy. Apply this technique in WEKA Tool using diabetes_heart dataset of 10fold cross validation and result was shown in Figure 7.



```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 41  0  3 |  a = Tested_High_affected
  1 49  1 |  b = Tested_negative
  1  0 14 |  c = Tested_Normal_Nonaffected
```

Figure 7: Output Matrix for REP Tree.

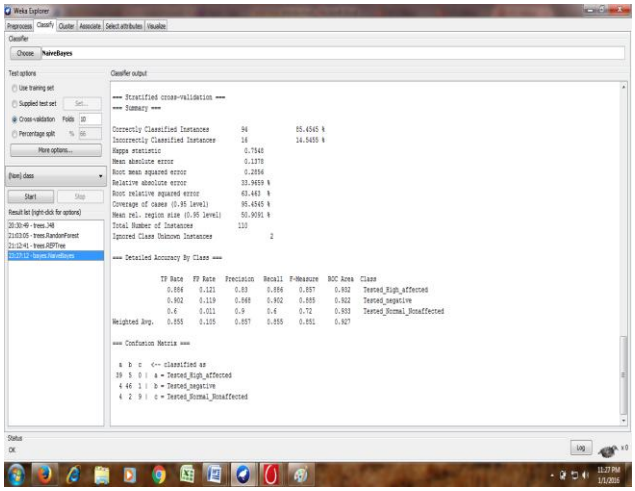
G. Naïve Bayesian

The Naïve Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. Bayesian reasoning is applied to decision making and inferential statistics that deals with probability inference. It is used the knowledge of prior events to predict future events. It builds, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c) .Naive Bayes classifier assume that the effect of the value of a predictor(x) on a given class(c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c|x) = \frac{P(x|c)\,P(c)}{p(x)}$$

- P[(c/x) is the posterior probability of class(target) given predictor(attribute).
- P(c) is the prior probability of class.
- P(x/c) is probability of predictor given class.
- P(x) is the prior probability of predictor class.

Above algorithm are applied in WEKA Tool using diabeties_heart dataset then finding its confusion matrix shown in Figure 8.



```
=== Confusion Matrix ===

   a   b   c   <-- classified as
  39   5   0 |   a = Tested_High_affected
   4  46   1 |   b = Tested_negative
   4   2   9 |   c = Tested_Normal_Nonaffected
```

Figure 8: Output Matrix for Naïve Bayesian.

IV. PERFORMANCE EVALUATION

This section has shown the comparison of the different data mining algorithms. The formula to calculate accuracy is:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Correctly classified instance = TP + TN
Incorrectly classified instance = FP + FN

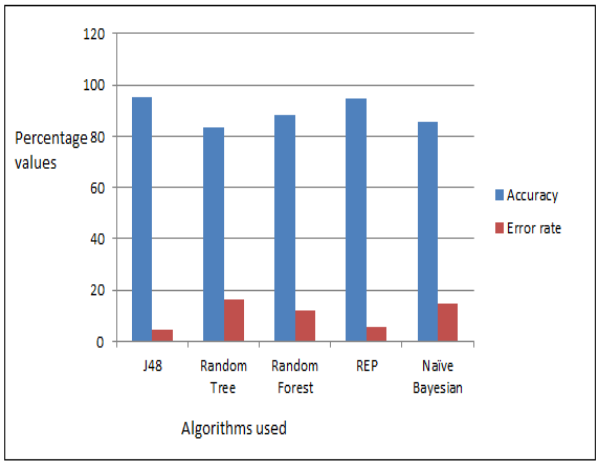
Formula represents TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative applied for the proposed dataset result shown in Table 1. Cross-validation (CV) method used in order to validate the predicted model. This test basically divides the training data into a number of partitions or folds. The classifier is evaluated by accuracy on one phase after learned from other one. This process is repeated until all partitions have been used for evaluation [10]. The most common types are 10-fold applied in the dataset and result for confusion matrix shown in Table 2 and Graph 1, Graph 2 shows its accuracy and time takes to build the classifiers.

Algorithms	Accuracy	Error rate	Time Taken to Build
J48	95.4545	4.5455	0.02
Random Tree	83.6364	16.3636	0.02
Random Forest	88.1818	11.818	0.02
REP	94.5454	5.4545	0.02
Naïve Bayesian	85.4545	14.5455	0

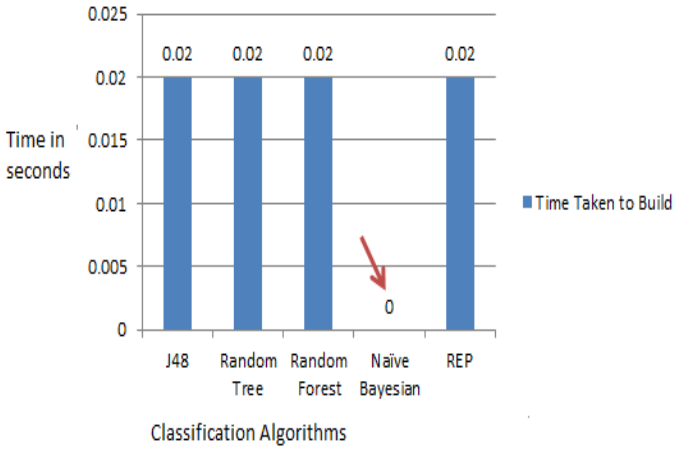
Table 1: Accuracy, Error, Time taken to build the correctly and Incorrectly Classifiers.

Algorithms	Results								
	Tested_ High_ affected			Tested_ Negative			Tested_ Nornal_ Nonaffected		
	a	b	c	a	b	c	a	b	C
J48	41	0	3	1	50	0	1	0	14
Random Tree	37	1	6	4	44	3	2	2	11
Random Forest	41	1	2	3	47	1	5	1	9
REP	41	0	3	1	49	1	1	0	14
Naïve Bayesian	39	5	0	4	46	1	4	2	9

Table 2: Result for confusion Matrix by 10 fold cross validation using dataset.



Graph 1: Shows Accuracy, Error rate for classification Algorithms.



Graph 2: Shows Time Taken to build the classifiers.

V. CONCLUSION

This research work has proposed a new approach for efficiently predicting the diabetes_heart disease from some medical records of patients. Dataset has designed with matching attributes applied in classification algorithms like J48, Random Tree, Random Forest, REP, Naïve Bayesian in WEKA Tool. On this experiment classification wise J48 Produces highest accuracy (95%) apart from decision tree Naïve Bayesian take minimum time (0.00 Seconds) to classify.

VI. REFERENCES

[1] . Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", second edition, Morgan Kaufmann Publishers an imprint of Elsevier.

[2]. V.Karthikeyani**,**"Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction" *International Journal of Computer Applications (0975 – 8887) Volume 60– No.12, December 2012.*

[3]. M. Anbarasi "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm" International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376.

[4]. http://www.niddk.nih.gov/health-information/health-topics/Diabetes/diabetes-heart-disease-stroke/Pages/index.aspx#connection.

[5] . Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction International" Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[6]. Chau, M., Shin,D., "A Comparative study of Medical Data classification Methods Based on Decision Tree and Bagging algorithms",Proceedings of IEEE International Conference on Dependable,Autonomic and Secure Computing "2009, pp.183-187.

[7]. Gaganjot Kaur "Improved J48 Classification Algorithm for the Prediction of Diabetes**"** *International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014*

[8]. Random Forest by Leo Breiman and Adele Cutler: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[9]. J.R. Quinlan, "Simplifying decision trees", Internal Journal of Human Computer Studies,Vol.51, pp. 497-491, 1999.

[10]. N. Laves son and P. Davidson, "Multi-dimensional measures function for classifier performance", 2nd. IEEE International conference on intelligent system, pp.508-513, 2004.