ISSN: 0975-8585

**Research Journal of Pharmaceutical, Biological and Chemical Sciences**

# Comparative Study on Efficiency of Classifier Models.

**Arvinda Soundararajan*, and Santhi B.**

Department of Information and Communication Technology, Sastra University Thanjavur, Tamil Nadu, India.

## ABSTRACT

This work deals with the implementation of various classifier models, each run on a training set, cross-validation technique and a percentage split with best accuracy value and the accuracy values obtained by each of the classifier models under the “cross-validation” technique has been recorded. Data classification refers to the process in which data can be separated into distinct categories or levels so that essential data can be made easier to find and be retrieved. The proposed work involves analysis of the techniques of each of the classifier models and their accuracy values are compared so as to determine the best algorithm while working with all the records and attributes of the publicly available data set “Statlog Heart” of the UCI repository.

**Keywords:** Classification Algorithms, Machine Learning, Data Mining,**Efficiency**.

**Corresponding author*

## INTRODUCTION

The classification problem refers to predicting the class label of newly received data based on a set of data that has already been grouped into classes. Some of the established approaches used for this purpose are-Support Vector Machines, Logistic Regression, Decision Trees, Bayesian Approaches, etc. [1- 4].

Generally, the training set is supposed to contain more information so that unknown data can be classified accurately. But, labeled data is difficult to be obtained in a large number of applications. Cross-validation assesses the results of statistical analyses. Each round of this technique comprises of partitioning a data sample into two sets-a training set and a validation set(testing set). Multiple rounds of splitting and analysis is done on various subsets to reduce variability.

Percentage splits involve creating two sets of a single data set – training set and testing set. In one round, the accuracy of the model is determined based on the percentage allotted for each of the subsets.

A set of various classifiers, have been used. These include Bagging, RandomForest, RandomTree, BayesNet and NaïveBayes classifier models [5- 6].

Results of various classifiers applied on the data set "Statlog Heart" are analyzed. Based on the results obtained from classification, it is determined if a person suffers from a heart disease or not. The effect in the use of training data, cross validation and percentage split options are noted.

## MATERIALS AND METHODS

In a given data set, the Record Scheme 'R' refers to the structure of the records present in it. The data set comprises of a set of fields or attributes, '$f_i$', with i=1, 2…., m. Each record 'r' is described by a set of values '$v_i$' for each of the corresponding fields. A record is said to be classified when assigned a class label from a possible set of classes 'C'.

For developing a good classification, a training set 'T', consisting of records that have already been classified, is provided. The Test set(also called Validation set) 'V', is used to check the performance of the classifier. For this purpose, the real class label of each record 't' of the test set should be known prior. The classification performed by that classifier model is then compared to the actual classification for each record in 'V'.

Cross-validation is a technique to validate and assess how accurate the prediction can be performed. Cross-validation defines a dataset to be tested while training the classifier model. Each round of this validation technique consists of the following steps- partitioning the data sample into two subsets, analysis of the training set and validating the results of the analysis on the test set. To reduce the variability in results obtained from each round, this process is applied multiple times on different partitions of the data set. The results obtained from each of the rounds are then averaged. Generally, cross validation is a 10-fold process. Since this technique can be used to provide an averaged value of accuracy, the results obtained by this technique alone have been recorded in the comparative study.

Classification can also be done by separating the same dataset into training data and testing data through percentage-split option to arrive with best accuracy values. Common percentage-split values are 80%-20%, 70%-30%, 66%-34%.

Before performing classification, the data need to be preprocessed. Preprocessing is done to make it more suitable for mining purposes after checking for noise and outliers, missing data and duplicate data through various techniques like sampling, aggregation, discretization, etc. In this work, a supervised discretize filter is used as the preprocessing technique prior to being implemented by classification models.

Also, the results obtained in every step of the classification process would vary depending on the attributes included, the number of records passed for testing as well as the various options on each of the classifier models. However, in this work, all attributes of the data set as well as all the records have been included in every step of the classification process.

In the proposed work, the data set- Statlog Heart, that is publicly available on the UCI Repository is used for analysis of results obtained by various classifiers-Bagging, BayesNet, NaiveBayesian, RandomForest and RandomTree. This data set has 270 record instances and 13 attributes of real and categorical types. This data set includes specific attributes such as resting electrocardiographic results, maximum heart rate achieved, exercise induced angina,old peak and slope of the peak exercise ST segment and other usual attributes like age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, the number of major vessels and thal.

**Bagging-**Also called bootstrap aggregating, this ensemble meta-algorithm method uses multiple machine learning algorithms to improve its stability and accuracy. This technique also helps reduce variance and avoid over fitting.

Consider a training set 'T' of size 'n' from the given data set. 'k' new training sets '$T_i$ ' of size 'n'' are generated from 'T' after sampling it with replacement. This is called a bootstrap sample. If size 'n'' of the generated training set is equal to the original training set size 'n', '$T_i$' is said to have about 63.2% unique records from 'T'. These 'k' bootstrap samples are then used to fit the model and combined by voting for classification problems and by averaging the results for regression problems [7].

Bagging leads to "improvements in unstable procedures"[8-10]. However, the performance of stable methods such as K-nearest neighbors degrades mildly.

There are two major disadvantages of bagging:

i) Loss in interpretation: The final classifier obtained from Bagging is not a tree. Hence, it is difficult for interpreting the classification obtained on the given data set.
ii) Complex Computation: We essentially multiply the work of growing a single tree by the number of independent classifiers, especially while pruning and validating on the original training data. We go from fitting a single tree to a large group of trees. Also, there is no single tree representation that can provide the final prediction rule.

Two strategies are used for aggregating the predictions: selecting class label based on majority vote (also called consensus strategy) or by estimating each class' probabilities, averaging them and then voting( also called probability strategy). Unlike the voting strategy, the consensus strategy does not provide good estimated class probability values. The voting strategy also improves the accuracy of the prediction rule.

Bagging works only if the base classifier has an acceptable range of prediction error to begin with as bagging bad classifiers can further degrade their performance.

**BayesNet-**This classifier model is also called Bayesian Network, Belief Network or Bayes Network. A Directed Acyclic Graph (DAG) is used to represent a set of random variables along with their probabilistic relationships or conditional dependencies.

The nodes in a DAG represent the random variable while their conditional dependencies are represented along the edges. When a pair of nodes is not connected, it implies that the variables represented by the nodes are independent of each other. A set values are associated with the parent variables of each node, as input and the output obtained is the probability of the node representing the associated variable.

Bayesian networks can also be considered as dynamic Bayesian Networks as they can model a sequence of variables like in a protein sequence or in speech signals. They can also solve decision problems under uncertainty and be represented through influence diagrams.

Bayesian Networks perform three main inference tasks [11]-

i) Inference of unobserved variables-Since a Bayesian network represents a set of nodes and their relationships, it can also determine queries related to their probability distribution. This also helps choose values for a subset of variables that can alter some function derived out of the graphical representation, like probability of decision error. Some common algorithms for this purpose are mini-bucket elimination, variational methods, importance sampling, etc.

ii) Parameter learning- To understand and represent the joint probability distribution, every node's conditional dependence on its parent nodes must be specified through a probability distribution. The conditional distributions may be unknown and hence need to be estimated.

iii) Structure learning- The network representation and distribution of parameters can sometimes be complex and understood through the data based on the parameters and their dependencies. Other strategies include optimization based search or through study of decomposable models.

**Naïve Bayes-** This classification model is based on the strong independence relations between the features. It is a very popular text categorization model, attaches a class label to the documents based on word frequencies. This classifier model can also be applied for numeric data. One technique is to discretize the continuous variables into fewer categories. However, this is subjective as the cutoff values of each category may differ for every implementation. This leads to loss of information. Another technique is to determine the probability density function values for an attribute and use it in the naïve Bayes formula to obtain the probability of each class label [12].

Naïve Bayes classifiers are highly scalable and only require a small amount of training set records to build the classifier model and estimate the parameters necessary for classification[13-14].

Essentially, NaiveBayes is not a single algorithm for classification, but a family of algorithms based on a common principle that the value of a feature is independent of any other feature, regardless of any correlations between the features.

**RandomForest**- This model involves building a forest of uncorrelated trees, combining the concept of node optimization and bagging. This practice also involves estimation of generalization error using out-of-bag (oob) error and also calculating variable importance.

Generally, tree learning is robust to scaling and transformation of feature values, inclusion of irrelevant features and produces inspectable models. However, irregular patterns are generated when trees are grown very deep as they overfit the training data due to low bias and high variance.

In this model, various decision trees are trained on sections of the same training set so as to reduce variance [15]. Increase in bias occurs along with loss of interpretability which further improves the performance of the resulting model. The training algorithm applies the basic principle of bagging. A modified tree learning algorithm is used to randomly select a subset of features for each candidate split. This process is called feature bagging. It is applied to check for correlation of trees in a sample and also to check for strong predictors to obtain the targeted output.

This classifier model also ranks the importance of variables. Breiman's original paper described the technique as follows [9].Initially, a random forest is obtained by fitting the data. During this process, the average of the out-of-bag error, recorded at each data point over the forest is computed. To measure each feature's importance after training, the out-of-bag error is computed again by permuting the feature's values among the training data. The average of the differences between the out-of-bag error before and after performing the permutation along every feature is considered as its importance score. The final score is the normalization of the standard deviation of the differences obtained. Features with the higher score are ranked as more important.

The process of determining the importance of variables has certain drawbacks- In data sets where categorical variables of multiple levels exist, features with more levels are biased. Also, if there exist groups of related features with similar relevance in outputs, smaller groups are favored

**Random Tre**e-Random Tree was also introduced by Leo Breiman and Adele and deals with classification and regression problems [10]. Random Tree is an ensemble of tree predictors called a forest. The process of classification occurs as follows- input of a Random Tree classifier is a set of vectors representing the features and the output is a class label that received a majority of the votes. For regression problems, responses over all trees is given as output [16]. In this classifier model, all trees are trained with the same set of parameters on different sets of data allocated as training sets that are generated using the bootstrap method. Each training set consists of the same number of vectors 'N' as the original set. These vectors can be chosen with replacement. For each new node, a subset is generated whose size is fixed for all nodes. It is a training parameter set by default $\sqrt{p}$ , where 'r' represents the number of variables. The size of the 'oob' data, due to sampling the vectors with replacement, is 'N/3'. This algorithm constructs multiple decision trees.

Multiple decision trees are randomly generated using the Random decision tree algorithm. During the tree construction, a "remaining" feature is randomly selected at each node expansion without checking for its purity (such as Gini index, information gain, etc.). A categorical feature that has not been chosen previously along a particular decision path is referred to as a “remaining" feature. This path begins at the root of a tree and proceeds to the current node. It is useless to pick an already chosen feature along the same decision path as it will result in the same value. However, a continuous feature can be chosen multiple times along the same decision path with a random threshold during every selection.

A tree stops growing if either of the conditions are met:

i) A node becomes empty or there is no available feature based on which the current node can be further split.
ii) The depth of the tree goes beyond the set limits.

This algorithm also removes unnecessary nodes (i.e.) if none of the descendants of the tree have significantly different class distribution. During classification, if a leaf node is empty, instead of providing the output as NaN (Not a Number) or 0 for that node's probability distribution, the classifier instead outputs the parent node's class distribution.

Using all the above mentioned classifier models on training set, cross-validation and percentage split, a confusion matrix is formed. Confusion matrix is a table that describes the performance of the classifier model used on a particular data set. Each column of the matrix represents the number of instances that have been predicted to belong to particular class while each row represents the number of instances that actually belong to that class. The number of correctly classified instances is represented along the major diagonal of the confusion matrix.

In general, the following references are made:
Correctly identified instance → True positive
Incorrectly identified instance → False positive
Correctly rejected instance → True negative
Incorrectly rejected instance → False negative

In a 2x2 confusion matrix, the first column contains the number of instances that the classifier has assumed to be of one class, say A, while the second column contains those that are assumed to be of another class, say B.

The first row contains the number of elements that actually belong to a class 'A' while the second row contains the number of elements that actually belong to another class 'B'.
Model of a confusion matrix:

| Predicted Class A ↓ | Predicted Class B ↓ | |
|---|---|---|
| Correct value | Confused value | ← Actual Class A |
| Correct value | Confused value | ← Actual Class B |

From the confusion matrix generated, the following measures are determined:
True Positive Rate (or) Sensitivity = True Positive/(True Positive+False Negative)
False Positive Rate (or) Fallout = False Positive/(False Positive+True Negative)

Precision (or) Positive Prediction Rate = True Positive/(True Positive+False Positive)
Recall = True Positive/(True Positive+False Negative)

Accuracy =sum of true positive and true negative instances/ total number of instances in the confusion matrix Root Mean Squared Error=$\sqrt{}\ (\Sigma^{n}_{i=1}(y'_{i}-y_{i})^{2}\ /n)$

Receiver Operating Characteristic (ROC) curve refers to a plot between TP rate and FP rate. A classifier is said to have better efficiency if its ROC area value is approximately equal to 1.

## EXPERIMENTAL SETUP AND RESULTS

The following results are obtained while working on an Intel Celeron CPU B830 with 1.80GHz processor and 2Gb RAM. The various classifiers are used through Weka, a data mining tool. The data set used to carry out these tests is “Statlog Heart”, which is publicly available in the UCI Repository.

This data set has 270 records and 13 fields: six of which are real valued(age, resting blood pressure, serum cholesterol, maximum heart rate achieved, old peak and number of major vessels coloured by fluorosopy),one is ordered(slope of peak exercise ST segment), three are binary(sex, fasting blood pressure and exercise induced angina) and the rest three are nominal(chest pain type, resting electrocardiographic results and thal) . All the fields indicate several medical information about patients.

Based on the attribute values, a class label must be attached to each record, stating whether the patient, whose details are associated with that record, suffers from a heart disease or not. Absence of heart disease is indicated as 1while presence is indicated as 2.

In this work, the complete data set was tested under 3 techniques- training set, cross-validation and percentage split. Even though training set and percentage splits provide better accuracy under certain circumstances, the results thus obtained cannot be used for comparative purposes as the results are variable when used on the complete data set or on selected records or attributes alone. Hence, only cross-validation results are recorded as this provides the average values of accuracy and can be used for the comparative study of the classifier models.

**Comparing various efficiency factors for each classifier mode**

| Classifier name | Accuracy | TP Rate | FP Rate | Precision | Recall | Root Mean Squared Error | ROC area |
|---|---|---|---|---|---|---|---|
| Bagging-Cross validation | 82.96% | 0.83 | 0.18 | 0.83 | 0.83 | 0.36 | 0.89 |
| RandomForest-cross validation | 83.33% | 0.83 | 0.18 | 0.83 | 0.83 | 0.36 | 0.88 |
| RandomTree-Cross validation | 80.74% | 0.80 | 0.21 | 0.81 | 0.81 | 0.42 | 0.80 |
| BayesNet- Cross validation | 83.33% | 0.83 | 0.18 | 0.83 | 0.83 | 0.3589 | 0.91 |
| Naïve Bayes-Cross validation | 83.33% | 0.83 | 0.18 | 0.83 | 0.83 | 0.3584 | 0.91 |

## RESULTS AND DISCUSSIONS

It is difficult to obtain one best classifier algorithm for a particular data set. Parameters like accuracy percentage, ROC value, time for evaluation, ease of use, etc. are evaluated to determine the best one. Classifier performance also depends on the characteristics of the data as well as the features and the records of the data set included during every step of the classification process. These various parameters that have been obtained are applicable only on this data set when all its attributes as well as records have been passed for testing.

In recent analyses, accuracy values and ROC area values are popularly used to select the most suitable a classifier model.

In this work, the most suitable classifier model is chosen based on the accuracy and ROC area values implemented with cross validation.

To classify with a good degree of accuracy, a training set is required in many applications. The best accuracy is generally obtained with a training set, and in this data set, it is obtained using RandomForest and RandomTree models with a result of 95.56%. However, if a percentage split is to be made for creating a training set as well as test set from a single dataset, a split of 82% provides the best accuracy of 87.04%. Since the 10-fold cross-validation technique is used for comparative purposes, it has been observed that NaiveBayes, BayesNet and RandomForest provide the best accuracy of 83.33%.

## CONCLUSION

Since the data set contains values of the various factors associated with the working of the human heart, it is necessary to select a classifier model with highest accuracy value so as to accurately identify the presence of a heart disease while testing on other records with similar factors.

It has been observed that both, NaiveBayes and BayesNet classification algorithms have the same accuracy and ROC area values. However, on comparison of the two models, value of Root Mean Squared Error is marginally lesser for NaiveBayes classifier.

Hence, it can be concluded that NaiveBayes is the most suited classifier model for the dataset "Statlog Heart'.

While this work depicts the parameters for accuracy and error only on cross-validation technique of the complete data set, this work can also be extended to work on a data set that has either a selected set of features for classification or a set of features added to the already existing data set. It is to also be noted that results can vary based on pre-processing techniques and various classifier models.

## REFERENCES

[1] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA, 1994; 16(3): 235-240.

[2] Mitchell, Tom M . Machine Learning. McGraw-Hill, Singapore,1997.

[3] Vapnik, V. Statistical Learning Theory. Wiley, New York, NY, USA, 1998.

[4] Hand, D. J., Mannila, H. and Smyth, P. Principles of Data Mining. MIT Press, London, U.K,2001.

[5] Hastie, T., Tibshirani, R. and Friedman, J. The Elements of Statistical Learning. Springer-Verlag, New York, USA, 2002.

[6] Klosgen, W. and Zytkow, J.M. (Eds). Handbook of Data Mining and Knowledge Discovery. Oxford Univ. Press, London, U.K,2002.

[7] http://www.stat.cmu.edu/~ryantibs/datamining/lectures/24-bag.pdf

[8] Aslam, Javed A., Popa, Raluca A. and Rivest , Ronald L. On estimating the size and confidence of a statistical audit. Proceedings,Electronic Voting Technology Workshop, Boston, MA,2007.

[9] Shinde, A. , Sahu, A. , Apley, D. and Runger, G. Preimages for Variation Patterns from Kernel PCA and Bagging. IIE Transactions, 2014; 46(5).

[10] Breiman, Leo.Bagging Predictors. Machine learning,1996;24(2):123-140.

[11] https://www.cs.cmu.edu/~dmarg/Papers/PhD-Thesis-Margaritis.pdf

[12] Rish, Irina. An empirical study of the Naïve Bayes classifier (PDF). IJCAI Workshop on Empirical Methods in AI,2001.

[13] Hand, D. J., Yu, K. Idiot's Bayes - not so stupid after all?. International Statistical Review,2001; 69(3): 385–399.

[14] Caruana, R. and Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. Proceedings, 23rd International Conference on Machine Learning, 2006.

[15] Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning (2nd ed.). Springer,2008.

[16] http://stat-www.berkeley.edu/users/breiman/wald2002-1.pdf