Data Mining Classification Algorithms for Hepatitis and Thyroid Data Set Analysis

S.Vijayarani,[1] R.Janani,[2]S.Sharmila[3]
[1] Assistant Professor, Dept. of CS, Bharathiar University, Coimbatore, India
[2] Ph.D Research Scholar, Dept. of CS, Bharathiar University, Coimbatore, India
[3]Ph.D Research Scholar, Dept. of CS, Bharathiar University, Coimbatore, India

Abstract—Data Mining extracts the knowledge or interesting information from large set of structured data that are from different sources. Data mining applications are used in a range of areas; they are financial data analysis, retail and telecommunication industries, banking, health care and medicine. In health care, the data mining is mainly used for disease prediction. In data mining, there are several techniques have been developed and used for predicting the diseases that includes data preprocessing, classification, clustering, association rules and sequential patterns. This paper analyses the performance of two classification techniques such as Bayesian and Lazy classifiers for hepatitis and thyroiddataset.This classification task helps to classify the hepatitis dataset into two classes namely live and die and also to classify the thyroid dataset into two classes hyperthyroid or hypothyroid. In Bayesian classifier, two algorithms namely Bayes Net and Naive Bayes are considered. In Lazy classifier we used two algorithms namely IBK and KStar. Comparative analysis is done by using the WEKA tool. It is open source software which consists of the collection of machine learning algorithms for data mining tasks.

Keywords—Disease Prediction, Bayesian, Lazy, BayesNet, NaiveBayes, IBK, KStar.

I. INTRODUCTION

Data mining refers to extracting knowledge from massive amount of information. It is the set of activities used to find new, hidden or unexpected patterns or unusual patterns in data. Compared with other data mining application fields, medical data mining plays an important role and it has some unique characteristics. The medical data processing has the high potential in medical domain for extracting the hidden patterns within the dataset [15]. These patterns are used for clinical diagnosis and prognosis. The medical data are generally distributed, heterogeneous and voluminous in nature. An important problem in medical analysis is to achieve the correct diagnosis of certain important information. This paper describes classification algorithms and it is used to analyze the performance of these algorithms. The accuracy measures are True Positive (TP) rate, F Measure, Receiver Operating Characteristics (ROC) area and Kappa Statistics.

The error measures are Mean Absolute Error (M.A.E), Root Mean Squared Error (R.M.S.E), Relative Absolute Error (R.A.E) and Relative Root Squared Error (R.R.S.E) [5].Section 2 explains the literature review; Section 3 describes the classification algorithms. Experimental results are analyzed in section 4 and section 5 illustrates the conclusion of this paper.

II. LITERATURE REVIEW

S.Vijayarani et al., [11] determined the performance of various classification techniques in data mining for predicting the heart disease from the heart disease dataset. The classification algorithms are used and tested in this work. The performance factors evaluate the efficiency of algorithms, clustering accuracy and error rate. The result illustratedtheLOGISTICS classification function efficiency is better than multilayer perceptron and sequential minimal optimization.Kaushik H. Raviya et al., [3] characterize the comparison on three classification techniques such as K-nearest neighbour, Bayesian network and Decision tree. The main purpose of this analysis is to enumerate the best technique from all the three techniques. This paper describes the direct relationship between execution time and the quantity of data records. It also determines an indirect relationship between execution time and attribute size of the data sets.G.RaviKumaret.al,[16] have examined the comparative study between J48, Naivebayes, KNN, SVM, MLP, and Logistic and finds the performance, accuracy, progression of error, execution time and the effective algorithm. In this research work breast cancer data set has been collected from the UCI repository dataset and result has been produced in WEKA.AnshulGoyal et al., [17] determined a performance evaluation of naïve bayes and J48 classification algorithms. The experimental results illustrated classification accuracy and cost analysis. Comparison is made on both the algorithms and J48 gives more classification accuracy for class gender in bank dataset which has two values male and female. The result shows the efficiency, cost and the accuracy of j48 algorithm isgood compared to naïve bayes algorithm.

III. RESEARCH METHODOLOGY

Classification is used to classify the data into predefined class labels. The main objective of this paper is to find the best classification algorithm among Bayesian and Lazy classifiers for classifying hepatitisand thyroid data set. Figure 1 shows the proposed methodology.

A. Dataset

In order to compare the data mining classification techniques, the hepatitis and thyroid data is collected from the UCI repository. The hepatitis dataset has 156 instances and 20 attributes and thyroid dataset has 9172 instances and 28attributes. Weka (Waikato Environment for Knowledge Analysis) tool is used for analyzing the performance of the classification algorithms.

B. Data Preprocessing

Missing Data-Missing data or information would possibly occur because the value is not relevant to a particular case, could not be recorded once the data was collected, or is neglected by users due to privacy concerns. Missing values may create the difficulty of extracting useful information from the dataset. If the attributes are missing in the training dataset, the system can either ignore that object totally, or try to take it into account by, finding what is the missing attribute's most feasible value is, or use the value 'missing', 'unknown' or 'null' as a separate value for the particular attribute [2] [9]. Missing data are the lack of data items that hide some information that may be vital. Most of the real world database is categorized by associate inevitable problem of incompleteness, in terms of missing or inaccurate values.

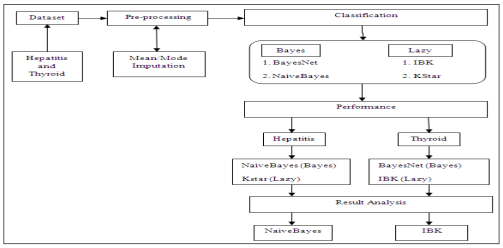


Figure 1: Proposed Methodology

Types of Missing Data- Basically there are three types of missing data, these are

- MCAR: The term Missing Completely at Random refers to data where the missingness mechanism does not rely upon the variable of interest, or the any other variable, which is examined within the dataset [10]. It is possibility of missing data on any attribute does not depend on any value of attribute.
- MAR: Sometimes data may not be missing at random but may be named as missing at Random. We consider an entry V is missing at random, if the data meets the requirements that missingness should not depend on the variable of V after controlling for another variable [9].
- NAMR: If the data is not missing at random or instructively missing it is termed as Not Missing at Random. This situation occurs once the missingness mechanism depends on the actual value of missing data. Modeling such a condition could be a terribly tough task to achieve [2]. This means we need to write a new model for missing data and then incorporate it into a complex model for finding missing values.

Missing Data Imputation Techniques

- Lit wise Deletion: It is the easiest way of handling missing data is to delete the topic that has missing values. This technique consists of discarding all instances with missing values for at least one feature. A variation of this technique is to delete the instances and/or attributes with high levels of missing data. The advantage of this technique is it decrease the sample size file used for analysis [9] [10].
- Mean/Mode Imputation (MMI):This is one of the most frequently used techniques for replacing missing values. It consists of replacing a missing data with the mean for numeric attribute or mode for nominal attribute. Many machine learning systems uses a simple imputer, called as mean imputation, which replace the missing value with the mean value of overall instances or overall instances in the same class or with the most frequently estimated value of attribute [2].In this paper we have used this technique to replace the missing values.
- K-Nearest Neighbor Imputation (KNN):This technique uses k-nearest neighbor algorithms to estimate and replacing the missing data. In this technique the similarity of two instances is determined using distance function. The main advantage of this technique is i) it can estimate both qualitative attributes and quantitative attributes ii) It is not necessary to build a predictive model for each and every attribute with missing data, even no need to build visible models [9]. The algorithm for KNN is as follows,
- Determine the K value (Nearest neighbors). K value will be chosen randomly.
- Determine the distance between the missing value instance and other training instance. The Euclidean distance is used to calculate the distance. The equation is given as follows,

$$\boxed{D(x, y) = \sum_{i=1}^{n} \sqrt{x_i^{2} - y_i^{2}}}$$

- After calculating the distances, the data values which have minimum distance are selected.Ifthe value of K is 5 then we have to choose 5 values that having minimum distance.
- Calculate the mean value of these chosen values. The equation is to calculate the mean value as follows,

$$\boxed{M = 1/n \sum_{i=1}^{n} mi}$$

Return M as the output value for the missing data.

C. Classification

Classification is an important data mining technique with extensive applications. It is used to classify each item in a data set into predefined set of classes or groups [1].In this paper we have analysed two classifiers namely Bayesian and Lazy classifier. In Bayesian classifier we have analysed two classification algorithms such as BayesNet and NaiveBayes. In Lazy classifier we have analysed two classification algorithms such as IBK and KStar.

a) Bayesian Classifier

Bayesian classifiers are powerful illustration, and their use for classification has received substantial attention. This algorithm predicts the class depending on the probability of fitting to that class. A Bayesian classifier is a graphical model for probability relationship among a group of variable features [1]. This classifier consists of two components. First component is especially a Directed Acyclic Graph (DAG) in which the nodes within the graph are called random variables and the edges between the nodes or random variables represent the probabilistic dependencies among the related random variables. The next component is a set of parameters that describe the chance of each variable given its parents. The conditional dependencies within the graph are calculated by statistical and computational methods [14].

b) BayesNet

BayesNet learns Bayesian networks created in nominal attributes and no missing values. Bayes Nets are graphical illustration for probabilistic relationships among a collection of random variables. Given a finite set X={X_1….X_n} of separate random variables where each variable X_i may take values from a finite set represented by Val (X_i).A BayesNet is an annotated directed acyclic graph G that encodes a probability distribution over X. The nodes of the graph resemble to the random variables $X_1 \ldots X_n$. The links of the graph represent to the direct influence from one variable to the other variable. If there is a directed relationshipfrom variable X_i to variableX_j, variableX_i is going to be a parent of variableX_j [3]. Every node is annotated with a contingent probability distribution (CPD) that represents P (X_i | Pa (X_i)) where Pa (X_i) denotes the parentsof X_i in G. the pair (G, CPD) encodes the joint distribution P ($X_i \ldots . X_n$). A unique joint probability distribution over X from G is factorized as:

$$\mathrm{P}(X_1 \ldots . X_n) = \Pi_i (P(X_i \mid \mathrm{Pa}\ (X_i)))$$

c) NaiveBayes

NaiveBayes implements the probabilistic NaiveBayes classifier. It uses the normal distribution to model numeric attributes. It can use kernel density estimators, which develop performance if the normality assumptionis correct; it can also handle numeric attributes using supervised discretization. The NaiveBayes algorithm is based on conditional probabilities. NaiveBayes uses Bayes theorem that is a formula that calculates a probability by counting the frequency of values and mixtures of values within the historical data [15].

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

$$\mathrm{P(c|x) = P}(x_1|\mathrm{c})\ \mathrm{X\ P}(x_2|\mathrm{c})\ \mathrm{X}\ \ldots\ \mathrm{X\ P}(x_n|\mathrm{c})\ \mathrm{X\ P(c)}$$

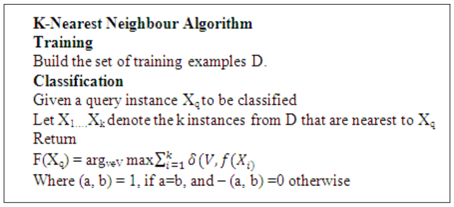
- P (c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P (x|c) is the chance which is the probability of predictor given category.
- P(x) is the previous probability of predictor.

d) Lazy Classifier

Lazy learners store the training instances and no real work until classification time. Lazy learning is a learning method within which generalization beyond the training data is delayed until a query is created to the system wherever the system tries to generalize the training data before receiving the queries [13]. The main advantage gained in using a lazy learning method is that the target function will be approximated regionally such as in KNN algorithm. Because the objective function is approximated regionally for each query for the system, this Lazy learning systems will concurrently solve multiple issues and deal successfully with changes in the problem field [1] [3]. The disadvantages with this method include the massive space requirement to store the entire training dataset. Mostly noisy training data increases the case support unnecessarily, as s result of no concept is made during the training phase.

e) IBK (K- Nearest Neighbour)

IBK is a k-nearest neighbor classifier that uses the same distance metric. The number of nearestneighbours will be specified explicitly in the object editor or determined automatically using leave-on-out cross-validation focus to an upper limit given by the specific value. There are different search algorithms will be used to speed up the task to find the nearest neighbours [1] [3].



K-Nearest Neighbour Algorithm

Training

Build the set of training examples D.

Classification

Given a query instance X_q to be classified

Let $X_{1 \ldots} X_k$ denote the k instances from D that are nearest to X_q

Return

$F(X_q) = \arg_{v \in V} \max \sum_{i=1}^{k} \delta (V, f(X_{i)}$

Where (a, b) = 1, if a=b, and – (a, b) =0 otherwise

Predictions from more than one neighbor are often weighted according to their distance from the test instance and two different formulas are implemented for changing the distance into a weight. The number of training instances kept by the classifier will be restricted by setting the window size option. As new training instances are added, the old one is removed to maintain the amount of training instances at this size [5].

f) Kstar

K* algorithm can be defined as a methodology of cluster analysis that mainly aims at the partition of ‘n observations into k’ clusters in which each observation belongs to the cluster with the nearest mean. K* is a simple, instance based classifier, similar to K- Nearest Neighbour (KNN). We can describe the K* algorithm as an instance based learner that uses entropy as a distance measure [7]. The advantages are it provides a consistent approach for handling the real value attribute, symbolic attributes and missing values. The K* function can be calculated as follows,

$$\mathrm{K}^*(y_i, \mathrm{x}) = -\ln \mathrm{P}^*(y_i, \mathrm{x})$$

Where P* is the probability of all transformational ways from instance x to y. It may be helpful to understand this as a probability that x will arrive at y via a random walk in IC feature space [1].

IV. EXPERIMENTAL RESULTS

A. Accuracy Measure

In this paper wehave used 10-fold cross-validation method to estimate the performance of these different classification methods.The following tables 1and 2 shows the accuracy measuresfor the classification techniques. The term accuracy refers the correctly classified instances by the total number of instances present in the dataset.

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

WhereTP-True Positive, FP-False Positive, TN-True Negative, FN- False Negative.TP Rateis the ability which is used to find the high true-positive rate.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

F Measure is a way of combining recall and precision scores into a single measure of performance. Recall is the ratio of relevant documents found in the search result to the total of all relevant documents [1][11].

$$\text{F} - \text{Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

ROC Area is a traditional to plot this same information in a normalized form with false negative rate plotted against the false positive rate.

Table 1. Accuracy Measure For Hepatitis Dataset

Algorithms → Performance Factors ↓	BayesNet	Naive Bayes	IBK	KStar
Correctly Classified Instances	82.58	83.87	80.64	81.29
Incorrectly Classified Instances	17.41	16.12	19.35	18.70
Kappa Statistic (%)	50.82	52.42	40.93	37.87
TP Rate (%)	82.60	83.90	80.60	81.30
FP Rate (%)	84.20	84.50	39.70	46.50
F-Measure (%)	83.20	84.10	80.60	80.40
ROC Area (%)	88.70	85.20	68.70	80.30

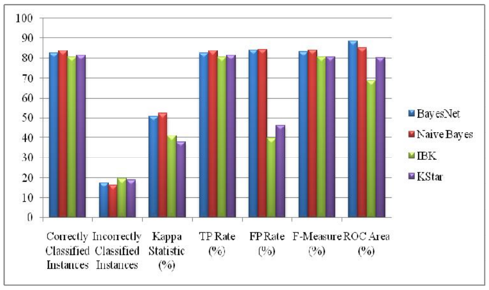


Figure 2: Accuracy Measure for Hepatitis Dataset

From the analysis of Accuracy Measures of Bayesian and Lazy classifier from the Table 1, Naive Bayes and KStar performs well when compared to all accuracy measures namely TP rate, F Measure, ROC Area and Kappa Statistic. As a result NaiveBayes and KStarperform well when compared to other Bayesian and Lazy algorithm for hepatitis dataset.

Table 2: Accuracy Measure For Thyroid Dataset

Algorithms → Performance Factors ↓	BayesNet	Naive Bayes	IBK	KStar
Correctly Classified Instances	91.99	83.87	80.64	81.29
Incorrectly Classified Instances	08.00	16.12	19.35	18.70
Kappa Statistic (%)	47.40	29.70	40.93	37.87
TP Rate (%)	92.00	83.90	80.60	81.30
FP Rate (%)	48.70	42.90	39.70	46.50
F-Measure (%)	91.60	85.90	80.60	80.40
ROC Area (%)	88.40	86.70	68.70	80.30

From the analysis of Accuracy Measures of Bayesian and Lazy classifier from the Table 2, Bayes Net and IBK performs well when compared to all accuracy measures namely TP rate, F Measure, ROC Area and Kappa Statistic. As a result Bayes Net and IBK perform well when compared to other Bayesian and Lazy algorithm for thyroid dataset.

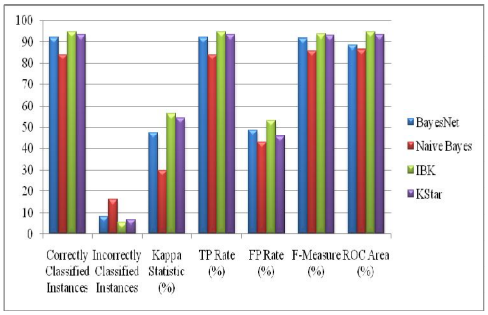


Figure 3: Accuracy Measure for Thyroid Dataset

B. Error Rate

A table 3 and 4 shows the error rate for the classification techniques. They are the Mean Absolute Error (M.A.E), Root Mean Square Error (R.M.S.E), Relative Absolute Error (R.A.E) and Root Relative Squared Error (R.R.S.R) .The mean absolute error (MAE) is defined as the quantity used to measure how close predictions or forecasts are to the eventual outcomes [4]. The root mean square error (RMSE) is defined as frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. It is a good measure of accuracy, to compare the forecasting errors within a dataset as it is scale-dependent. Relative error is a measure of the uncertainty of measurement compared to the size of the measurement [1].

Table 3:Error Rate For Hepatitis Dataset

Algorithm → Error Measures ↓	*BayesNet*	*Naïve Bayes*	*IBK*	*KStar*
MAE	0.17	0.17	0.19	0.19
RMSE	0.37	0.36	0.43	0.38
RAE	52.55	51.48	59.92	57.99
RRSR	91.88	90.62	107.88	94.69

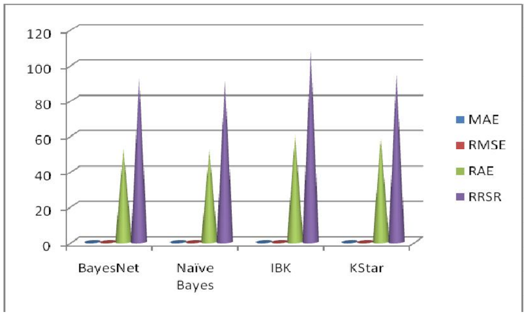
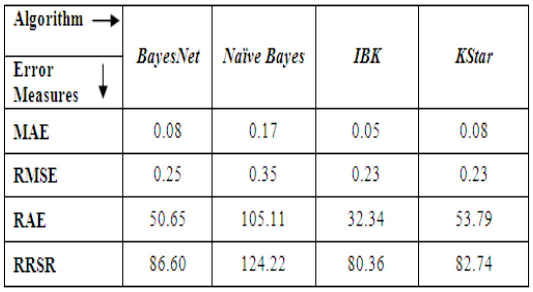


Figure 4: Error Rate for Hepatitis Dataset

From the graph, we observed that, In Bayes Classifier BayesNet attains highest error rate and IBK attains highest error rate in Lazy Classifier. Therefore the NaiveBayes and KStar classification algorithms performwellbecause it contains least error rate when compared to other algorithm.

Table 4: Error Rate Of Thyroid Dataset



Algorithm → Error Measures ↓	BayesNet	Naïve Bayes	IBK	KStar
MAE	0.08	0.17	0.05	0.08
RMSE	0.25	0.35	0.23	0.23
RAE	50.65	105.11	32.34	53.79
RRSR	86.60	124.22	80.36	82.74

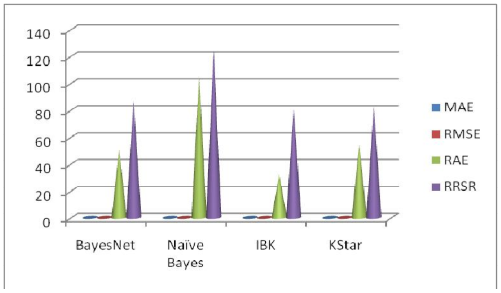


Figure 5: Error Rate for Thyroid Dataset

From the graph, we observed that, In Bayes Classifier Naïve Bayes attains highest error rate and KStar attains highest error rate in Lazy Classifier. Therefore the Bayes Net and IBK classification algorithms perform well because it contains least error rate when compared to other algorithm.

V. CONCLUSION

Data mining is the extraction of useful information or knowledge from huge data repositories. In this paper, two classification techniques in data mining are compared to find the better classification algorithm. The classification algorithms namely Bayesian and Lazy classifier are used for classifying the hepatitis and thyroid dataset. The Bayesian Algorithm includes two techniques namely BayesNet and NaiveBayes, the Lazy algorithms includes two techniques namely IBK (K-Nearest Neighbour) and KStar.By analysing the experimental results it is observed that the Bayesian classifier's NaiveBayes classification technique has earns better result than other techniques for hepatitis dataset and BayesNet classification technique gives the best accuracy for thyroid dataset. Generally, the overall results specify that the performance of the classifier based on the dataset.

REFERENCES

[1] Dr.S.Vijayarani, Mrs.M.Muthulakshmi, Comparative Analysis of Bayes and LazyClassification Algorithms, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE),Vol. 2, Issue 8, August 2013, ISSN (Print) : 2319-5940ISSN (Online) : 2278-1021

[2] BhavikDoshi, Handling Missing Values in Data Mining, Data Cleaning and Preparation Term Paper.

[3] Kaushik H. Raviya, BirenGajjar, Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA, Indian Journal of Research, Volume: 2, Issue: 1, January 2013, ISSN - 2250-1991

[4] Wikipedia

[5] Dr. S.Vijayarani, S.Sudha, Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering, Vol.1,Issue 3, May 2013, ISSN (Print) : 2320 – 9798, ISSN (Online): 2320 – 9801

[6] B S Harish, D S Guru, S Manjunath, "Representation and Classification of Text Documents: A Brief Review", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" ,RTIPPR, 2010

[7] Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering,Vol. 2, Issue 4, April 2013

[8] RohitArora, Suman, Comparative Analysis of Classification Algorithms on Different datasets using WEKA, International Journal of Computer Applications (0975 – 8887) Vol.54, No.13, September 2012

[9] Gimpy, Dr. RajanVohra, Minakshi, Estimation of Missing Values Using Decision Tree Approach , International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (4) , 2014, 5216-5220

[10] Gimpy, Dr. RajanVohra, Minakshi ,Missing Value Imputation in Multi Attribute Data Set, International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (4) , 2014, 5315-5321

[11] Dr. S.Vijayarani, S. Sudha, "An Effective Classification Rule Technique for Heart Disease Prediction", International Journal of Research in Engineering and Technology (IJRET), vol.2, Issue-10, Oct-2013, ISSN.2319- 1163

[12] HetalBhavsar, AmitGanatra, A Comparative Study of Training Algorithms for Supervised Machine Learning, International Journal of Soft Computing and Engineering (IJSCE), Vol-2, Issue-4, September 2012, ISSN: 2231-2307

[13] Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, NagarajuOrsu, Suresh B.Mudunuri , Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification, International Journal of Advanced Research in Artificial Intelligence (IJARAI), Vol. 2, No.5, 2013

[14] Abdullah H. Wahbeh, Mohammed Al-Kabi, "Comparative Assessment of the performance of three WEKA text classifiers applied to Arabic Text"

[15] Ian H. Witten, Eibe Frank. Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2005

[16] G.RaviKumar, Dr.G.A.Ramachandra, K.Nagamani "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques" International Journal of Innovations in Engineering and Technology (IJIET)

[17] Anshul Goyal, Rajni Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms". International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012)