Performance Evaluation of Classification Algorithms Based on Different Data Sets

J.P.Medlin Julia M.C.A;M.Phil, Scholar, Research and Development Centre, Bharathiar University,Coimbatore-641046, medlinjp@yahoo.com.

Dr.D.Bennet M.C.A;M.Phil;Ph.D, Research Supervisor, Department of computer Applications, Narayana Guru College of Engineering, Manjalumoodu, K.K.Dist, profdbennet@gmail.com

Abstract: Data **Mining refers to the extraction of hidden knowledge from a huge pool of data. Classification, clustering and association rule generation are the very important data mining techniques. This research analyzed the performance evaluation of popular classification algorithms used by machine learning systems in terms of their accuracy by applying them on three completely different datasets: medical, weather forecasting and bank marketing. The main objective of classification algorithm is to produce a good classification model which classifies the data accurately based on the training data set. Each classifier was tested with respect to accuracy, performance and execution time using average of the classified values. The experiments were carried out using WEKA, a free software data mining tool which includes implementation of machine learning algorithms. Using WEKA, the algorithms, Naive Bayes, Bayes net, J48, KNN and JRip were compared. These classification algorithms produced different results for the datasets. The quality of the result was measured depending on the correct and incorrect instances that were classified by different classifiers.**

Keywords — **Data Mining, Classification, Naive Bayes, Bayes net, J48, KNN, JRip, Weka Tool.**

I. INTRODUCTION

Data mining technique is a process of discovering useful information from a huge volume of structured and unstructured datasets and predicting the future based on the historical data. The discovered information must be meaningful. Data mining technique is being used in various sectors such as Finance, Healthcare, Intelligence, Telecommunication, Sales and Marketing, E-commerce, Biological Data Analysis, Crime Agencies etc. Different types of algorithms and strategies used in data mining include data classification, clustering analysis, association rules etc. Data classification plays a vital role in data mining technique. In this research different characters of Naive Bayes, Bayes net, J48, KNN and JRip algorithms were compared using the three different types of datasets available, medical data set from Apollo hospital group, Weather forecasting dataset from Indian Meteorological Department repository and bank marketing data sets from data.gov.in website. Comparison of Different Classification algorithms was done using WEKA Tool. These algorithms were studied and compared on the parameters like Correctly Classified Instances (CCI), Incorrectly Classified Instances (ICI), TP rate, FP rate, Precision, Recall, F-Measure, Root Mean Squared Error and build time. Figure 1 shows the Architecture of the proposed research.

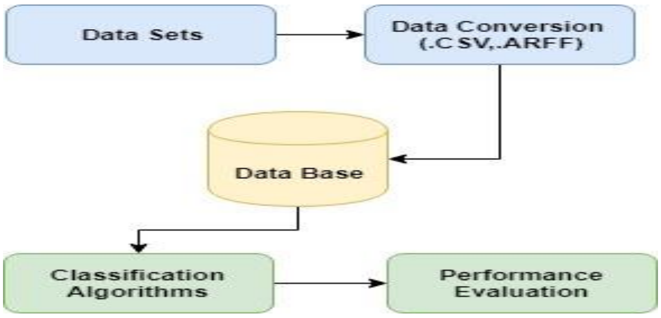

Figure 1: Architecture of Proposed Methodology

The rest of the paper is organized as follows. Section II covers proposed methodology. In Section III covers experimental results. Finally in section IV, we conclude the comparative results.

II. PROPOSED METHODOLOGY

Open source WEKA tool is used for this research. Three drastically differently datasets collected from different data repositories are used. Five different types of classifiers are used to perform comparative study: Naive Bayes, Bayes net, J48, KNN and JRip. TP Rate, FP Rate, Precision (P), Recall (R), FMeasure (F) and accuracy of each classifier is calculated. Figure (2) shows the process flow of this research.

DOI : 10.18231/2454-9150.2018.1227

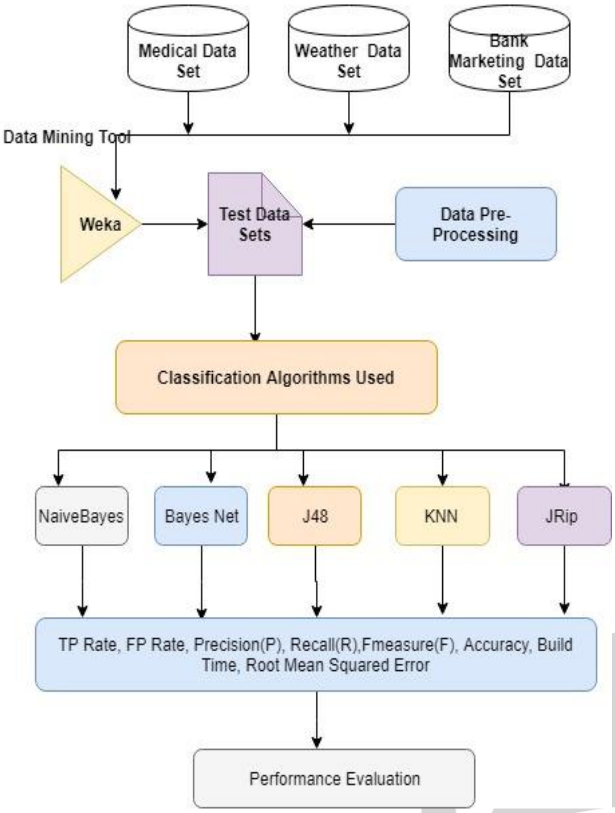

Figure 2: Process flow diagram

A. *WEKA and its functions*

WEKA is an open source which comes under GNU [1]. It was developed by university of Waikato in New Zealand. We can download this software free of cost from (https://www.cs.waikato.ac.nz/~ml/weka/downloading.html).This software is developed by using object oriented programing language. WEKA is mainly used for data mining that uses a collection of machine learning algorithms. WEKA is a collection of tools for: Regression, Clustering, Association, Data pre-processing, Classification, and Visualization. Students and researchers mainly using WEKA for study and research purposes. There are more than one million lines of codes in WEKA.

WEKA provides the following vital features.

- As it provides a very user friendly interface even a person with basic computer novices can access this. Users can choose and analyze different algorithms easily and apply them to the data sets. The results of the analysis are shown by the system intuitively in different ways.
- WEKA gives partial exposure to researchers who study theory and application of data mining.
- WEKA has an options to preprocess the datasets even if they have lot of noise.
- WEKA offers an options to check the programs written by us.
- WEKA has been providing a platform to compare the performance of the classification algorithms in the recent years.

B. *Classification*

Data mining is the process of analyzing the historical data and extracting the hidden information out of it [3]. There are various kinds of methods used for data mining such us classification, clustering, regression, rule generation etc. classification concept is useful in all walks of our life. Classification is a very important data mining technique which has numerous applications and it is a very challenging task in data mining [2]. It classifies the data in to different classes based on training set. It has two important parts: the first one is training set and the second one is test set. Training set is used to construct a classification model. Training model has been created by using historical information's. The accuracy of the prediction output is proportional to the perfectness and strengths of the training model. Even if the class of the input is hidden the training model will predict the class accurately. The classification algorithm used to conduct the comparative study are Naive Bayes, Bayes net, J48, KNN and JRip:

Naïve Bayes Classification: Naive Bayes is a simple and powerful algorithm for data mining which is based on Bayes' theorem.

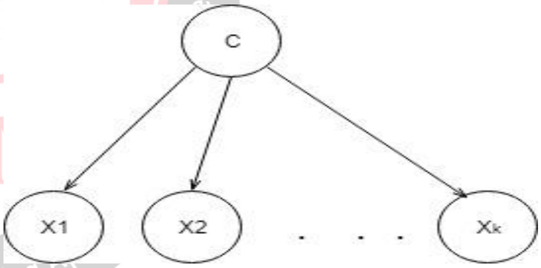

Figure 3: Navie Bayesian classifier

Figure 3: Navie Bayesian [9] classifier depicted as a Bayesian in which the predictive attributes (X1,X2,….Xn) are conditionally independent given the classes attributes (C). Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$.

$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$$

- 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 the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

DOI : 10.18231/2454-9150.2018.1227

Bayesian Network Classifiers: Probabilistic graphical model that represents random variables and conditional dependencies in the form of a directed acyclic graph. A Simple Estimator algorithm has been used for finding conditional probability tables for Bayes net [4]. A K2 search algorithm was used to search network structure [10, 11]. These networks are factored representations of probability distributions that generalize the naive Bayesian classifier and explicitly represent statements about independence.

J48 (C4.5): J48 is an open source Java implementation algorithm in the WEKA data mining tool. This algorithm was developed by Ross Quinlan. The decision trees generated byC4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [5].

1. Advantages: Decision trees capable to learn disjunctive expressions and their robustness to noisy data seem convenient for document classification.
2. Disadvantages: learning of decision tree algorithms cannot guarantee to return the globally optimal decision tree [6].

K-Nearest Neighbors: k-nearest neighbor's [12] algorithm (k-NN) is a non-parametric method used for classification and regression which is effective classifier of text categorization. The KNN classifier has three defects: the complexity of computing its sample similarity is huge; its performance is easily affected by single training sample and KNN doesn't build the classification model since it is a lazy learning method. The complexity of KNN can be reduced by utilizing three ways, reducing dimension of vector text, reducing amount of training samples and fasting process of finding K nearest neighbors [7].

JRip: Optimized version algorithm proposed by William W. Cohen. This algorithm will try to add every possible rule until it becomes accurate. It Optimizes rule set using discretion length [8].

C. *Parameters used*

- TP = True Positives: All the values it predicts are truly positive. It predicts the positive instances rightly positive.
- FP = False Positives: All the values it predicts are truly negative. It predicts the negative instances rightly negative.
- TN = True Negatives: It predicts the instances as positive, but truly they are positive.
- FN = False Negatives: It predicts the instances as negative but truly they are positive.
- Precision - % of selected items that are correct and are calculated as Precision (P) = TP / (TP+FP).
- Recall - % of correct items that are selected and the calculation for it is Recall (R) = TP / (TP+FN).
- F-Measure (F) - the Harmonic mean of precision and recall, calculated as F=2*R*P/(R+P).

D. *Data sets*

In order to measure the classification three data sets are used: Medical Dataset, Weather forecasting and Bank Marketing. The number of instances of three datasets are 14068, 11021and 20496. Table1 shows the data sets details.

Table1: Data sets details

Datasets specifications	**Medical**	**Weather**	**Bank Marketing**
Number of instances	14068	11021	20496
Number of attributes	7	9	8
Missing values	Yes	Yes	No
Field	Social	Social	Financial

E. *Medical Data Set*

Dataset is a collection of data or a single statistical data where every attribute of data represents a variable and each instance has its own description. The dataset used for this experiment consists of randomly selected 14068 instances from Apollo hospital group. This dataset contains the data of patients affected by diabetes, blood pressure and cholesterol in India. Patients over 40 years old were only selected. The dataset uses 6 variables as regular attributes and one as a class attribute which specify the descriptive properties of patients. The dataset helps us to conclude whether a person is affected or not affected by a particular disease. Table2 shows the attributes and details.

Table2: Attributes and data description of medical dataset

Attribute	**Description**
ID	Patient's Hospital Identification number
Gender	F for female and M for Male
Age	Patient's Age
Sugar	Sugar level in blood
cholesterol	cholesterol level in blood
Blood pressure	Blood pressure level
Class label	The class each patient belongs to according to his diagnoses

F. Weather forecasting dataset

A Meteorological data is indispensable for weather prediction and water resource planning. In this research work, the data were collected from Indian Meteorological Department Website on a monthly basis between 2016 and

DOI : 10.18231/2454-9150.2018.1227

2017 of New Delhi region. Monthly data were collected on day to day basis and different parameters such as Temperature, Humidity, Windy, Wind Direction, and Wind Chillness were collected and stored in the data file with .csv format. Weather forecasting dataset has 20496 number of instances and 7 attributes with last one as a class attribute. Table3 shows the attributes and details of weather forecasting dataset.

Table3: Attributes and data description of medical dataset

Attribute	**Description**
Date	Current date and time
Temperature Numeric	The temperature values in Fahrenheit.
Temperature Nominal	This Attribute indicates the weather being below 60F as cool and above 60F as hot.
Humidity Numeric	Humidity is normally expressed as a percentage. A higher percentage means that the air–water mixture is more humid.
Humidity Nominal	This Attribute indicates humidity level where above 80% is high value and below 80% is normal.
Windy	This Attribute explains the windy weather conditions.
Wind direction	Wind direction is the direction from which it originates.
wind chill	This Attribute explains the amount of chillness in the wind.
Class label	Sunny, Rainy, Overcast

G. Bank Marketing Dataset

The Bank Marketing dataset could be vital, depending upon the kind of variables that are included in the database. It can be used to predict and forecast the revenue model, subscribers for various bank offerings. Promotions can be handled effectively by deciding where to invest to gain maximum conversion. The data is collected from www.kaggle.com. The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyze in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank. This dataset contains in total 20496 instances, 7 attributes and binary class. Table4 shows the attributes and details of bank marketing dataset.

Table4: Attributes and data description of bank marketing dataset.

Attribute Name	**Description**
age	This attribute indicates the customer's age.
Marital	This Attribute indicates whether the customer is married or unmarried.
Education	This Attribute explains about the educational qualification of the customer.
Balance	This Attribute shows the current balance of the customer.
Housing	This Attribute indicates whether customer owns a House or not.
Loan	This Attribute helps to show whether the Customer has taken loan.
Contact	This Attribute indicates the customer's contact details.
Deposit	This Attribute is the binary class attribute which indicates whether he has deposit or not.

H. Data Pre-Processing

Data pre-processing is a very important step in the data mining process that converts the raw data into an understandable format. Real-world data is often lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. These type of data can produce erroneous results in the process of analysis. Hence to generate the accurate result the data should be preprocessed. The medical datasets contain many imperfect data like missing certain fields to be filled by patients due to emergency cases, in some context collected data will be noisy. Filling the missing value in medical data is very difficult task. Improperly handling the missing values will produce inaccurate results. In this research Bayesian formalism algorithm is used to fill the missing values. Noisy data is another important concern in data mining. In order to reduce noise, use noise filters which identify and remove the noisy instances in the training data.

III. RESULTS AND DISCUSSION

Results obtained this research are based on different test options: k-fold cross-validation and use the average perdition value of three different datasets in the proposed work in order to get a perfect comparison.

A. Prediction: k-fold validation

This research has used K-fold cross validation (k=10) method. This method divides a dataset into 10 folds, 9 folds of which are used for training, and the final fold is for testing.

B. Correctly and incorrectly classified incidents of each classifier method

Classification accuracy is the degree of correctness in classification. The degree of correctness can be evaluated using various classifiers for individual instances in the data set. Larger test set provides a good assessment on classifier accuracy.

Table 5.Classifiers accuracy on the dataset based on 10-fold cross validation.

Classification Method	**Domain Name**	**Correctly Classified Incidents**	**Incorrectly Classified Incidents**
	Medical Dataset	**92.3247**	**7.6753**

DOI : 10.18231/2454-9150.2018.1227

NaiveBayes	Weather forecasting	**69.7569**	**30.2431**
	Bank Marketing	**77.4444**	**22.5556**
Average of NaiveBayes		**79.842**	**20.158**
Bayes net	Medical Dataset	**95.572**	**4.428**
	Weather forecasting	**84.5139**	**15.4861**
	Bank Marketing	**79.5326**	**20.4674**
Average of Bayes net		**86.5395**	**13.4605**
J48	Medical Dataset	**100**	**0**
	Weather forecasting	**99.3403**	**0.6597**
	Bank Marketing	**92.1838**	**7.8162**
Average of J48		**97.1747**	**2.8253**
KNN	Medical Dataset	**96.9004**	**3.0996**
	Weather forecasting	**73.7847**	**26.2153**
	Bank Marketing	**81.5427**	**18.4573**
Average of KNN		**84.07593**	**15.9240**
JRip	Medical Dataset	**100**	**0**
	Weather forecasting	**98.9583**	**1.0417**
	Bank Marketing	**86.6462**	**13.3538**
Average of JRip		**95.2015**	**4.798**

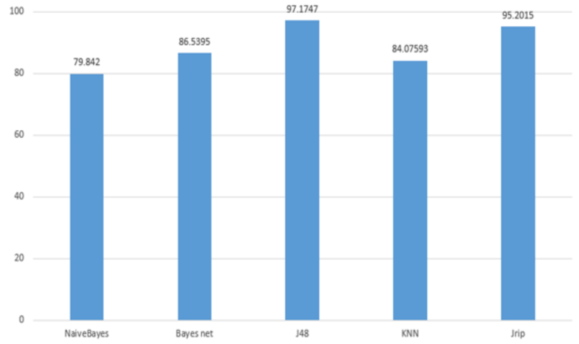

Figure: 4 Graphical view of accuracy for various classifiers

J48 classifier has identified a number of incidents correctly with 97.17%, followed by J48 having correct classification rate of 95.20% compared with other classifiers and NaiveBayes has determined the least correct instances with 79.84%.

C. Performance measures calculated based on confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Confusion matrix contains some important parameters in order to calculate the performance of the classifications model. Figure 5 shows the WEKA confusion matrix. The analysis of Medical Dataset, Weather forecasting, Bank Marketing and average on the basis of TP rate, FP rate, precision and F-Measure parameters are done. Table 5 shows the classification of testing data for different classes on TP rate, FP rate, Precision, Recall (R) and F-measure.

Table 6: Performance measures calculated based on confusion matrix

Classification Method	**Domain Name**	**TP Rate**	**FP Rate**	**Precision (P)**	**Recall (R)**	**FMeasure (F)**
NaiveBayes	Medical Dataset	**0.923**	**0.045**	**0.045**	**0.923**	**0.925**
	Weather forecasting	**0.698**	**0.070**	**0.785**	**0.698**	**0.720**
	Bank Marketing	**0.774**	**0.227**	**0.775**	**0.774**	**0.774**
Average of NaiveBayes		**0.795**	**0.114**	**0.535**	**0.801**	**0.806**
Bayes net	Medical Dataset	**0.956**	**0.015**	**0.962**	**0.956**	**0.957**
	Weather forecasting	**0.845**	**0.028**	**0.915**	**0.845**	**0.859**

DOI : 10.18231/2454-9150.2018.1227

	Bank Marketing	**0.795**	**0.206**	**0.796**	**0.795**	**0.795**
Average of Bayes net		**0.868**	**0.086**	**0.882**	**0.868**	**0.637**
J48	Medical Dataset	**1.000**	**0.000**	**1.000**	**1.000**	**1.000**
	Weather forecasting	**0.993**	**0.005**	**0.993**	**0.993**	**0.993**
	Bank Marketing	**0.866**	**0.133**	**0.867**	**0.866**	**0.866**
Average of J48		**0.959**	**0.040**	**0.953**	**0.953**	**0.953**
KNN	Medical Dataset	**0.969**	**0.026**	**0.968**	**0.969**	**0.968**
	Weather forecasting	**0.738**	**0.197**	**0.737**	**0.738**	**0.734**
	Bank Marketing	**0.815**	**0.186**	**0.816**	**0.815**	**0.815**
Average of KNN		**0.840**	**0.136**	**0.840**	**0.825**	**0.839**
JRip	Medical Dataset	**1.000**	**0.000**	**1.000**	**1.000**	**1.000**
	Weather forecasting	**0.990**	**0.019**	**0.990**	**0.990**	**0.990**
	Bank Marketing	**0.866**	**0.133**	**0.867**	**0.866**	**0.866**
Average of JRip		**0.952**	**0.051**	**0.952**	**0.952**	**0.95**

Table 6 shows the average of TP rate, FP rate, Precision, Recall and F-Measure, obtained by using the 10-fold cross-validation approach. Decision Table and J48 have the highest TP Rate (True Positive) by 0.96 and Recall values 95%, followed by JRip having TP rate by 0.95 and recall value of 95. J48 has greater precision and FMeasure when compared with the other algorithms.

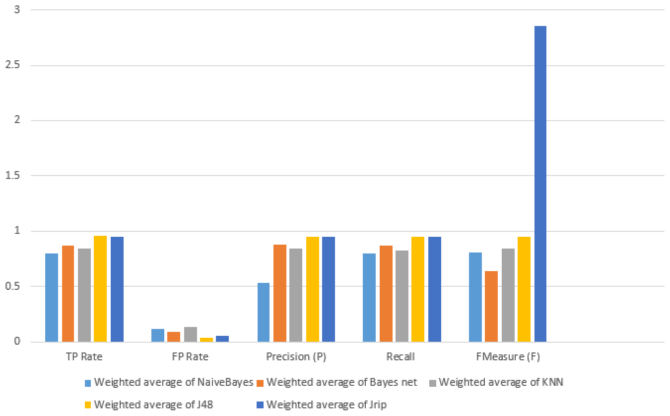

Figure: 5 Graphical view of performance measurement.

A. Classifier execution time and root mean square error on the dataset based on 10-fold cross validation test mode.

Execution time is higher for JRip with 20.35 sec and j48 with 0.53 sec, while KNN time to build the model was the least with 0.08 sec, with NaiveBayes and Bayes net time for a model build is 0.08 sec and 0.15 sec, respectively. According to our experiment JRip consume higher time of 20.35 sec to build a classification model.

Table 7: classifier execution time

Classification Method	**Domain Name**	**Time to Build the Model (Seconds)**	**Root Mean Squared Error**
NaiveBayes	Medical Dataset	0.09	0.1875
	Weather forecasting	0.03	0.3669
	Bank Marketing	0.14	0.3993
Average of NaiveBayes		0.08	0.3179
Bayes net	Medical Dataset	0.19	0.1556
	Weather forecasting	0.14	0.2758
	Bank Marketing	0.11	0.3825
Average of Bayes net		0.15	0.271

DOI : 10.18231/2454-9150.2018.1227

J48	Medical Dataset	0.13	0
	Weather forecasting	0.19	0.0663
	Bank Marketing	1.27	0.2538
Average of J48		0.53	0.212
KNN	Medical Dataset	0.03	0.124
	Weather forecasting	0.01	0.3125
	Bank Marketing	0.01	0.353
Average of KNN		0.05	0.263
JRip	Medical Dataset	0.6	0
	Weather forecasting	1.01	0.0832
	Bank Marketing	59.45	0.3365
Average of JRip		20.35	0.139

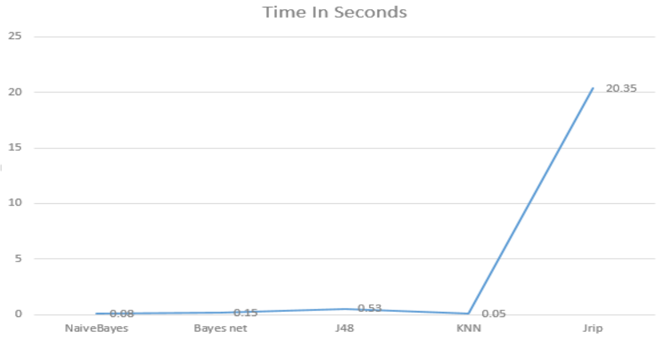

Figure: 6 Graphical representation of classifier execution time.

IV CONCLUSION

Data mining is an important research domain which focuses on knowledge discovery in databases. We use WEKA the open source data mining tool to check the result. This research examined the performance and the accuracy of Naive Bayes, Bayes net, J48, KNN and JRip classification algorithms that produced different results. Different experiment techniques were used to predict the performance of classification algorithms. The concepts used for the analysis are accuracy, performance and execution time. According to this experiment J48 and JRIP produce the highest correctly classified instances than the remaining algorithms. Bayes net and KNN also produce relatively accurate results and the time consumption is very low. Although JRIP is the most accurate classifier, it took the maximum time to build the model with 20.35 sec. This experiment results show that there is no classification algorithm which produces the best classification model, each algorithm has its own merits and demerits. Based on this comparative study we can conclude that, there is a need to develop a better classification algorithm in order to produce better classification model for different datasets.

REFERENCES

[1] Yanguang Shen, Jie Liu and Jing Shen, "The Further Development of Weka Base on Positive and Negative Association Rules", *IEEE 2010 International Conference on Intelligent Computation Technology and Automation.*

[2] Sunita B. Aher and Lobo L.M.R.J, "COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS", *International Journal of Information Technology and Knowledge Management July-December 2012, Volume 5, No. 2, pp. 239-243.*

[3] https://en.wikipedia.org, Data mining .[Online].Available:https://en.wikipedia.org/wiki/Data_mining, [Accessed:January 5, 2019].

[4] Nir FriedmanDan, GeigerMoises and Goldszmidt, "Bayesian Network Classifiers", *Kluwer Academic Publishers. Manufactured in The Netherlands, Machine Learning, 29, 131–163 (1997).*

[5] Uzair Bashir and Manzoor Chachoo, "PERFORMANCE EVALUATION OF J48 AND BAYES ALGORITHMS FOR INTRUSION DETECTION SYSTEM", *International Journal of Network Security & Its Applications (IJNSA) Vol.9, No.4, July 2017.*

[6] Korde, V., & Mahender, C. N. (2012), "Text classification and classifiers", *A survey. International Journal of Artificial Intelligence & Applications, 3(2), 85.*

[7] Rennie, J. D., & Rifkin, R. (2001). "Improving multiclass text classification with the support vector machine", *Massachusetts Institute of Technology as AI Memo 2001-026 and CBCL Memo 210.*

[8] Kailas Elekar & M.M. Waghmare. & Amrit Priyadarshi, "Use of rule base data mining algorithm for Intrusion Detection", *IEEE 2015 International Conference on Pervasive Computing (ICPC).*

[9] John GH, Langley P, "Estimating Continuous Distributions in Bayesian Classifiers", *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence; pp. 338–345.*

[10] Friedman N, Geiger D, Goldszmidt M. "Bayesian Network Classifiers". *Mach Learn. 1997;29:131–163.*

[11] Yu J, Chen X, "Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data", *Bioinformatics. 2005;21(Suppl 1):i487–i494. [PubMed].*

[12] https://en.wikipedia.org, k-nearest neighbors algorithm,2018.[Online].Available:https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm. [Accessed: 1- sep-2014].

DOI : 10.18231/2454-9150.2018.1227