IFERP
connecting engineers... developing research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Performance Analysis of Bayes Classification Algorithms in WEKA Tool using Bank Marketing Dataset

[1] M. Purnachary, [2] B. Srinivasa S P Kumar, [3] Humera Shaziya
[1][3] Assistant Professor, Dept. of Informatics, Nizam College, OU, Hyderabad
[2] Assistant Professor, Dept of MCA, CBIT, OU, Hyderabad

***Abstract:* - Data Mining is an interdisciplinary field that aims to extract knowledge or insights from data in various forms, either structured or unstructured. Classification is a supervised learning approach of data mining and It is used to classify huge data. WEKA is powerful machine learning tool that contains many inbuilt algorithms to extract knowledge. In this paper we tried to analyze the performance of two built in Bayes type of Classification algorithms (Bayes Net, Naïve Bayes) in WEKA tool using Bank Marketing Dataset which is extracted from UCI Repository. It has been observed that Bayes Net classification algorithm performed better compared to Naïve Bayes algorithm.**

***Index Terms*- Classification, Knowledge, Naïve Bayes, WEKA, Confusion Matrix.**

I. INTRODUCTION

A. WEKA TOOL

The Waikato Environment for Knowledge Analysis (Weka) is a machine learning toolkit introduced by Waikato University, New Zealand. It is open source software written in Java. It contains collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

B. Data Mining

We are in an age often referred to as the information age.[1][2] We believe that information leads to power and success. With the enormous amount of data stored in files, databases, and other repositories, it is important to develop powerful means for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD) is defined as extraction of nontrivial, implicit, previously unknown and potentially useful information from data in databases.

C. Classification

Classification is a supervised learning model of data mining function that create the classification model by using training dataset and classify the test data using that model [1][2]. The Classification process involves following steps:

a. Create training data set.

b. Identify class attribute and classes.

c. Identify useful attributes for classification.

d. Learn a model using training examples in Training set.

e. Use the model to classify the unknown data

There are many built-in Bayes classification algorithms in Weka tool(Bayes Net, Naïve Bayes)

1. Bayes Net Algorithm

It is popular algorithm in data classification. A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG)[1][2].

A Bayesian network B =< N, A, 0 > is a directed acyclic graph (DAG) with a conditional probability distribution (CP table) for each node, collectively represented by e. Each node n belongs to N represents a domain variable, and each arc a belongs to A between nodes represents a probabilistic dependency. In general, a BN can be used to compute the conditional probability of one node, given values assigned to the other nodes; hence, a BN can be used as a classifier that gives the posterior probability distribution of the classification node given the values of other attributes.

When learning Bayesian networks from datasets, we use nodes to represent dataset attributes. In this algorithm, 10 Fold Cross Validation is used as a test option in weka tool to analyze bank dataset.

2. Naïve Bayes Algorithm

The Naive Bayes classification algorithm is a probabilistic classifier[1][2]. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive.You can derive probability models by using Bayes' theorem (credited to Thomas Bayes). Depending on the nature of the probability model, you can train the Naive Bayes algorithm in a supervised learning setting.In this algorithm, 10 Fold Cross Validation is used as a test option in weka tool to analyze bank dataset.

D. Bank Marketing Dataset [3]

Bank dataset is collected from UCI Web resource.This data is related with direct marketing campaigns of a Portuguese banking institution.Here The classification goal is to predict whether the client will subscribe a term deposit or not.

Number of Instances: 45211

Number of Attributes: 16 + output attribute.

Attribute information: (bank client data)

1 - age (numeric)

2 - job : type of job (categorical: "admin." ,"unknown" ,"unemployed","management","housemaid","entrepreneur", "student","blue-collar","self employed", "retired" "technician" ,"services")

3 – marital : marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced / widowed)

4 – education (categorical: "unknown", "secondary", "primary", "tertiary")

5 - default: has credit in default? (binary: "yes", "no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes", "no")

8 - loan: has personal loan? (binary: "yes", "no")

9 - contact: contact communication type (categorical: "unknown", "telephone","cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

Output variable (desired target):

17 - y - has the client subscribed a term deposit? (binary: "yes", "no")

II. LITERATURE SURVEY

In paper [5], authors explore the performance evaluation of Naïve Bayes, Logistic Regression and Decision tree, Random forest using datasets (Pima Indian Diabetes data from UCI Repository). Naïve Bayes algorithm is depending probability; it is fast and stable to data changes. Logistic Regression, calculate the relationship of each feature and weights them based on their impact on result. Random forest algorithm is an ensemble algorithm, fits multiple trees with subset of data and averages tree result to improve performance and control over-fitting. Decision tree can be nicely visualized uses binary tree structure with each node making a decision depending upon the value of the feature [1][2]. Finally they concluded with a comparative evaluation of Naïve Bayes, Logistic Regression, Decision tree and Random Forest in the context of Pima Indian Diabetes Dataset

In this paper [4], authors presented a systematic analysis of twenty four performance measures used in thecomplete spectrum of Machine Learning classification tasks, i.e., binary, multi-class,multi-labelled, and hierarchical.For each classification task, the study relates a set of changes in a confusion matrix to specific characteristics of data. In this paper [6], authors have comparatively tested four classification algorithms to find the optimum algorithm for classification. The Credit Card Approval dataset has been used for experimental purposes that contain 690 instances with 15 attributes and 1 class attribute to test and justify the differences among classification algorithms. The four classification algorithms are – Decision Tree (DT), Naïve Bayes‟ (NB), Artificial Neural Network (ANN) and Support Vector Machine (SVM).

In this paper[7], authors explains the analysis of classification and clustering using some terms like Kappa Statistics ,Mean Absolute Error , Confusion Matrix ,Classification Accuracy correctly classified, incorrectly classified ,root mean square error for different algorithms of classification and clustering. This paper considers the most extensively used tools, WEKA tool for this analysis purpose. The training and testing is performed for this analysis. In this paper[8] Authors compared well performing classification algorithms such as Naïve Bayes , decision tree (J48), Random Forest, Naïve Bayes Multiple Nominal, K-star and IBk. Data that they have used is Student dataset and gauge students' potential based on various indicators like previous performances and in other cases their background to give a comparative account on what method is the best in achieving that end. They discussed about various statistical measure used to calculate the performance of each classifier.

III. IMPLEMANTATION

In this implementation, various Bayes Classification Algorithms performance is analyzed by using WEKA Tool. WEKA version 3.8.2 has been used to implement and execute the classification algorithms. The Bank Marketing Dataset from UCI repository has been utilized to pursue the analysis and this dataset is in .csv format.

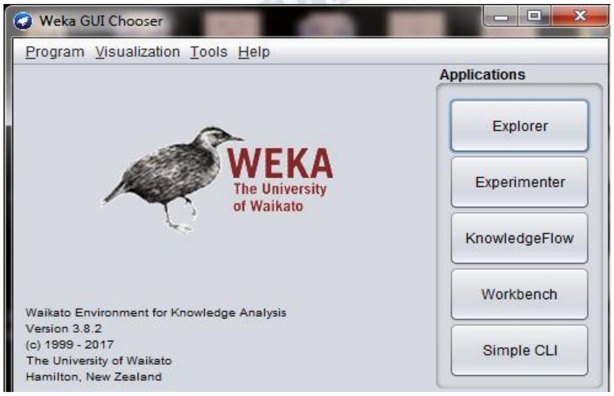
When we startup Weka 3.8.2 software, starting window shows multiple options like Explorer, Experimenter, Knowledge Flow, Workbench, Simple CLI. In this Test, The Explorer has been used.

The Explorer[9] is used to

- Gives access to all facilities of Weka using menu selection and form filling
- Prepare the data, open the Explorer and load the data
- Flip back and forth between results, evaluate models built on different datasets and visualize graphically both and models and datasets, including classification errors

A. Execution Procedure:

In Weka tool The Explorer option has been chosen for this performance analysis. Bank Marketing dataset is analyzed in two steps namely Preprocess and Classify. Fig.1 shows the weka Explorer environment in a clear manner.

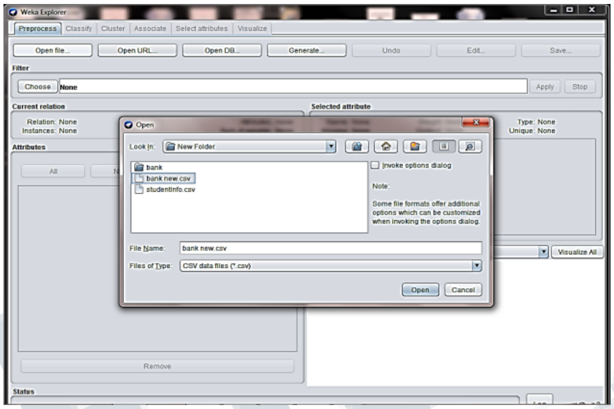


Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

Fig.1. Weka Explorer Interface

1. *Preprocess*

After starting Weka Explorer, first dataset location will be selected and loaded and then filter the attributes by using supervised attribute selection option. After filtering, it has been identified that only 6 among the 17 attributes have chosen for classification. This step is common for all classification algorithms. First dataset will be selected and loaded into weka tool form computer. Fig.2 shows the selecting and loading of dataset into weka form computer.



Filter: None

Relation: None
Instances: None

Type: None
Unique: None

Look In: New Folder

☐ Invoke options dialog

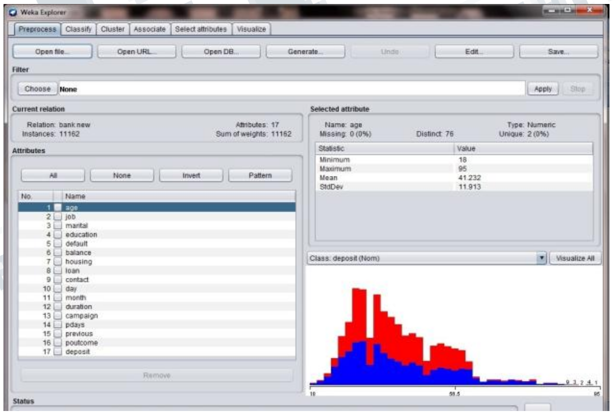
Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: bank new.csv

Files of Type: CSV data files (*.csv)

Fig.2 Selecting and loading Dataset from Computer



Filter: None

Relation: bank new
Instances: 11162
Attributes: 17
Sum of weights: 11162

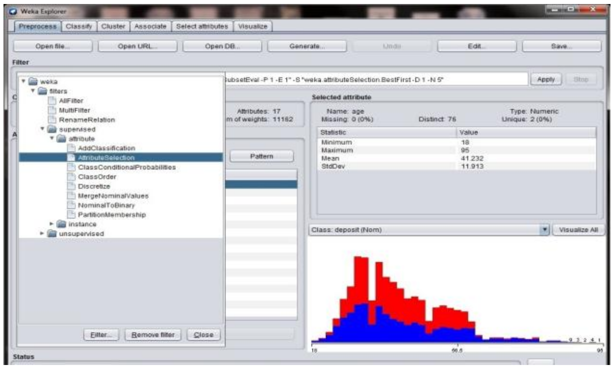
Name: age
Missing: 0 (0%)
Distinct: 76
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	18
Maximum	95
Mean	41.232
StdDev	11.913

No.	Name
1	age
2	job
3	marital
4	education
5	default
6	balance
7	housing
8	loan
9	contact
10	day
11	month
12	duration
13	campaign
14	pdays
15	previous
16	poutcome
17	deposit

Class: deposit (Nom)

Fig.3 Showing Dataset in Weka



ubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5"

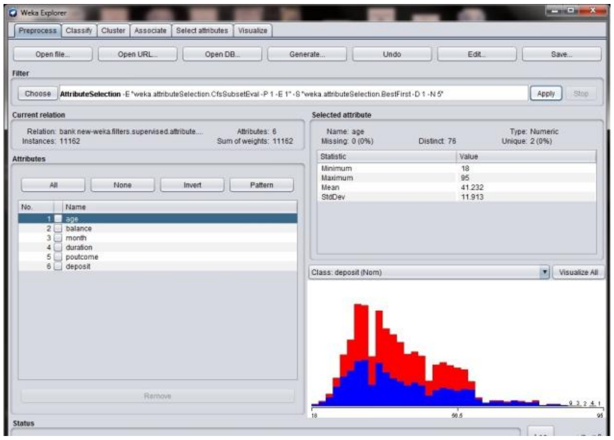
Attributes: 17
m of weights: 11162

Name: age
Missing: 0 (0%)
Distinct: 76
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	18
Maximum	95
Mean	41.232
StdDev	11.913

Class: deposit (Nom)

Fig.4. Selecting Filter Option to Preprocess dataset in Weka



Weka Explorer

Filter: Choose **AttributeSelection** -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Current relation
Relation: bank new-weka.filters.supervised.attribute.... Attributes: 6
Instances: 11162 Sum of weights: 11162

Selected attribute
Name: age Type: Numeric
Missing: 0 (0%) Distinct: 76 Unique: 2 (0%)

Statistic	Value
Minimum	18
Maximum	95
Mean	41.232
StdDev	11.913

Attributes

No.	Name
1	age
2	balance
3	month
4	duration
5	poutcome
6	deposit

Class: deposit (Nom)

Fig.5. List of attributes after Filtering in Weka

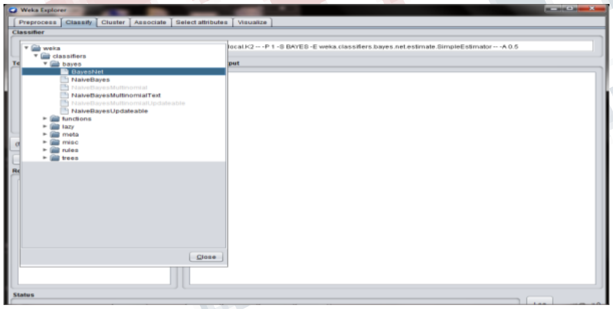
2. *Classify*

In this step, preprocessed dataset has been used and classified by using various classifiers available in Weka Tool.

Here Bayes Net and Naïve Bayes Classifiers are used to classify preprocessed data.

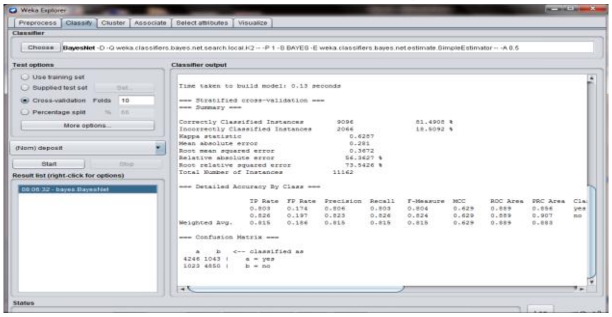
B. Bayes Net Algorithm in Weka :

In weka Tool, select the Classify option in Explorer and choose the Bayes type of classifiers and select Bayes Net algorithm and choose the 10 fold Cross validation as test option and start the classifier. Detailed process of choosing and starting the classifier has shown in Fig.6,7



local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Fig.6. Choosing Classifier



Classifier: Choose **BayesNet** -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Test options: ● Cross-validation Folds 10

(Nom) deposit

Result list (right-click for options): 08:06:32 - bayes.BayesNet

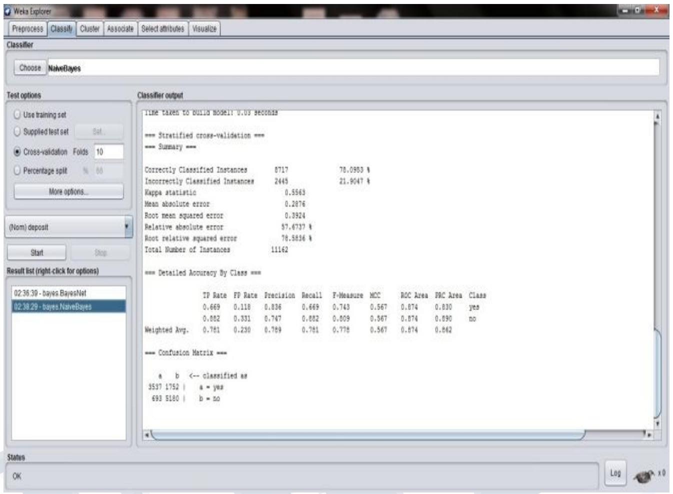
Classifier output

```
Time taken to build model: 0.13 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9096               81.4908 %
Incorrectly Classified Instances      2066               18.5092 %
Kappa statistic                          0.6287
Mean absolute error                      0.281
Root mean squared error                  0.3672
Relative absolute error                 56.3627 %
Root relative squared error             73.5426 %
Total Number of Instances            11162

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.803    0.174    0.806      0.803    0.804      0.629    0.889     0.856     yes
                 0.826    0.197    0.823      0.826    0.824      0.629    0.889     0.907     no
Weighted Avg.    0.815    0.186    0.815      0.815    0.815      0.629    0.889     0.883

=== Confusion Matrix ===

    a    b   <-- classified as
 4246 1043 |    a = yes
 1023 4850 |    b = no
```

Fig.7:Starting Bayes Net Classifier and Observing Result

C. Naïve Bayes Algorithm in Weka

In weka Tool, select the Classify option in Explorer and choose the Bayes type of classifiers and select Naive Bayes algorithm and choose the 10 fold Cross validation as test option and start the classifier. Detailed process of choosing and starting the classifier has shown in Fig.6,8



Classifier: Choose NaiveBayes

Test options: ● Cross-validation Folds 10

(Nom) deposit

Result list (right-click for options):
02:36:39 - bayes.BayesNet
02:38:29 - bayes.NaiveBayes

Classifier output

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8717               78.0953 %
Incorrectly Classified Instances      2445               21.9047 %
Kappa statistic                          0.5563
Mean absolute error                      0.2876
Root mean squared error                  0.3924
Relative absolute error                 57.6737 %
Root relative squared error             78.5836 %
Total Number of Instances            11162

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.669    0.118    0.836      0.669    0.743      0.567    0.874     0.830     yes
                 0.882    0.331    0.747      0.882    0.809      0.567    0.874     0.890     no
Weighted Avg.    0.781    0.230    0.789      0.781    0.776      0.567    0.874     0.862

=== Confusion Matrix ===

    a    b   <-- classified as
 3537 1752 |    a = yes
  693 5180 |    b = no
```

Status: OK

Fig.8: Starting Naïve Bayes Classifier and Observing Result using Bank Dataset.

IV. RESULTS AND DISCUSSIONS

A. Evaluation Metrics [4][9]

The parameters considered while evaluating the selected classifiers are:

1) Confusion Matrix is a useful tool for analyzing how well your classifier can recognize tuples of different classes. It is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix.

2) Kappa: Measures the relationship between classified instances and true classes. It usually lies between [0, 1]. The value of 1 means perfect relationship while 0 means random guessing.

3) TP Rate: Is the statistics that shows correctly classified instances.

4) FP Rate: Is the report of instances incorrectly labeled as correct instances.

5) Recall: Measures the percentage of all relevant data that was returned by the classifier. A high recall means the model returns most of the relevant data.

6) Precision: Measures the exactness of the relevant data retrieved. High precision means the model returns more relevant data than irrelevant data.

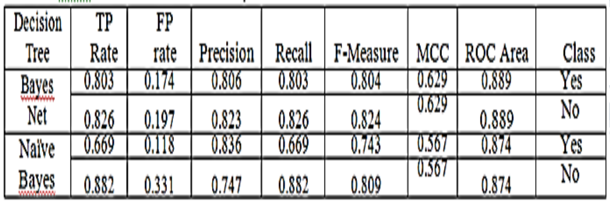
B. Observations

When Bank Marketing dataset is classified using Weka tool It has been observed that Bayes Net classifier shows the number of correctly classified instances of 9096 with the accuracy of 81.49% and the number of incorrectly classified instances of 2066 that is 18.50%. Naïve Bayes classifier shows the number of correctly classified instances of 8717 with wit the accuracy of 78.09% and the number of incorrectly classified instances of 2445 that is 21.90%. Table.1 shows the accuracy of various Bayes Classifiers.Table.2 shows the final statistics like FP Rate, TP Rate, Precision, Recall, F-Measure, MCC,ROC Area of decision tree for both the Bayes Net and Naïve Bayes classifiers in terms of two classes (YES/NO).Table.3 shows the Comparison of Weighted Avg. for Decision Tree of various Bayes Classifiers. Table.4 gives the Confusion Matrix for All Decision Trees of both the Bayes Net and Naïve Bayes classifiers.

Table.1: Comparing the Accuracy of the Bayes classification Algorithms

Name of the Algorithm	Correctly classified instances		Incorrectly classified instances	
	No. of Instances	Percentage (%)	No. of Instances	Percentage (%)
Bayes Net	9096	81.4908	2066	18.5092
Naïve Bayes	8717	78.0953	2445	21.9047

Table.2. Final Statistics of Decision Tree

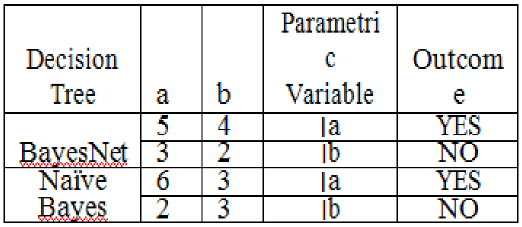


Decision Tree	TP Rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	Class
Bayes Net	0.803	0.174	0.806	0.803	0.804	0.629	0.889	Yes
	0.826	0.197	0.823	0.826	0.824	0.629	0.889	No
Naïve Bayes	0.669	0.118	0.836	0.669	0.743	0.567	0.874	Yes
	0.882	0.331	0.747	0.882	0.809	0.567	0.874	No

Table.3: Comparison of Weighted Avg. for Decision Tree

Decision Tree	Bayes Network	Naïve Bayes
TP Rate	0.815	0.781
FP Rate	0.186	0.230
Precision	0.815	0.789
Recall	0.815	0.781
F-Measure	0.815	0.778
MCC	0.629	0.567
ROC Area	0.889	0.874
PRC Area	0.883	0.862

Table 4: Confusion Matrix for All Decision Tree



Decision Tree	a	b	Parametric Variable	Outcome
BayesNet	5	4	|a	YES
	3	2	|b	NO
Naïve Bayes	6	3	|a	YES
	2	3	|b	NO

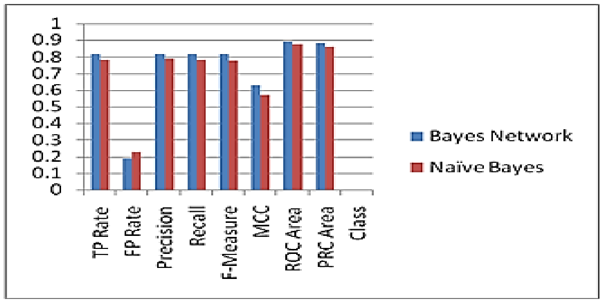


Fig.9: Graph of table.3

V. CONCLUSION

In this paper two Bayes classifiers namely Bayes Net and Naïve Bayes of weka tool have been used to compare their accuracy of classification in terms of correctly classified instances with respect to Bank Marketing dataset. It has been observed that Bayes Net generates more number of correctly classified instances compared to the Naïve Bayes algorithm. Experiments were conducted on weka tool and concluded that the accuracy of Bayes Net Classifier is proved to be better.

REFERENCES

[1] Pang-Ning Tan, Vipin Kumar- Introduction to Data Mining (Second Edition) Pearson International Edition.

[2] Jiawei Han University of Illinois at Urbana– Champaign Micheline Kamber -Data Mining Concepts and Techniques Third Edition, Elsevier.

[3] UCI Machine Learning Repository https:// archive. ics. uci. Edu /ml /datasets /bank +marketing

[4] Marina Sokolova- A systematic analysis of performance measures for classification tasks, Information Processing and Management 45 (2009) 427–437 Elseviers https://www.sciencedirect.com/science/article/pii/S0306457309000259

[5] MERAJ NABI -performance analysis of classification algorithms in predicting diabetes, ijarcs

[6] Devendra Kumar Tiwary -A Comparative study of classification algorithms for credit card approval using weka GIIRJ, Vol.2, (MARCH (2014)

[7] Shivangi Gupta-Comparative Analysis of classification Algorithms using WEKA tool International Journal of Scientific & Engineering Research, Volume 7, Issue 8, August-2016 2014,ISSN 2229-5518

[8] Bhrigu Kapur-Comparative Study on Marks Prediction using Data Mining and Classification Algorithms ,IJARCS http://dx.doi.org/10.26483/ijarcs.v8i3.3066

[9]http://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf

[10] Daniel Lowd-Naive Bayes Models for Probability Estimationhttp:// aiweb. cs. washington. Edu /ai /nbe /nbe_icml. pdf

[11] Yuguang Huang-Naive bayes classification algorithm based on small sample set (Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference INSPEC Accession Number: 12305847)

[12] C. Eklan," Boosting and naive Bayes learning", Technical Report Computer Science, University of San Diego, pp97-557,Sept 1997.

[13] Earl Gose ,Richard Johnsonbaugh and Steve Jost. "Pattern Recognition and Image analysis". Prentice Hall PTR, 1996

[14] margaret h. danham,s. sridhar, “data mining, introductory and advanced topics”, person education , 1st ed.,pp.75-84,2006.

[15] aman kumar sharma, suruchisahni, “a comparative study of classification algorithms for spam email data analysis”, ijcse, vol. 3, no. 5, pp. 1890-1895,2011.

[16] barto, a. g. & sutton, r., “introduction to reinforcement learning”,mit press.m. young, the technical writer’s handbook mill valley,ca: university science, pp. 45-60,1997.

[17] Trilok Chand Sharma1, Manoj Jain2 “WEKA Approach for Comparative Study of Classification Algorithm”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013.

[18] Rashedur M. Rahman, Farhana Afroz , “ Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis”, Journal of Software Engineering and Applications, 2013, 6, 85-97 http://dx.doi.org/10.4236/jsea.2013.63013 Published Online March 2013.

[19] Mahendra Tiwari, Manu Bhai Jha, OmPrakash Yadav. “Performance analysis of Data Mining algorithms in Weka” , IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 3 (Sep-Oct. 2012), PP 32-41.

[20] Rohit Arora, Suman, “Comparative Analysis of Classification Algorithms on Different Datasets using WEKA”, International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012.

[21] WEKA, the University of Waikato, :http :// www. cs. waikato. ac. Nz /ml /weka /, (Accessed 20April 2011).