IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 1, Ser. II (Jan - Feb 2020), PP 49-56
www.iosrjournals.org

A Comparative Analysis of Different Classification Algorithms based on Students' Academic Performance Using WEKA

Er. AmlanJyoti Baruah[1], Dr. Siddhartha Baruah[2], Dr. JyotiProkash Goswami[3]
[1](Assistant Professor, Department of Computer Science and Engineering, Assam Kaziranga University, India)
[2](Professor, Department of Computer Applications, Jorhat Engineering College, India)
[3](Associate Professor, Department of Computer Applications, Assam Engineering College, India)

Abstract: *Education is the most important factor for shaping the personality of an individual. Development of a country depends on the upgradation of education system.Due to the large volume of educational data in higher education, it is challenging to predict the academic performance of students.In this paper authors presentseven built in classifiers namely J48, Random Forest, Rap Tree, LMT, Naïve Bayes, BayesNet and PARTwith the questionnaires filled up by final year students of Computer Science and Engineering Department,Assam Kaziranga University, Jorhat, Assam and suggests the efficient algorithm based on certain parameters. The survey was done based on total 22 questionnaires and 152 responses. In the whole analysis process WEKA 3.8 was used and Random Forest algorithm was found as the most efficient algorithm among all the considered algorithms.*

Key Word*:WEKA, Prédiction, J48, Random Forest, Rap Tree, LMT, Naïve Bayes, BayesNet, PART*

Date of Submission: 28-01-2020 Date of Acceptance: 13-02-2020

I. Introduction

Now a day's Data Mining is an emerging field of computer science .Due to the increasing volume of data, there is a need of converting that data to useful information and knowledge. There are various applications of data mining such as health care, banks, customer segmentation, fraud detection etc. Currently Educational Data Mining has been used in educational sector.

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in [1].

Students' performance is an important part in higher education system of educational institutions. One of the criteria for a high quality institution is its excellence in academic achievements. Most of the higher educational institutions use the final grades or marks to evaluate student's performance, which is based on course structure, assessment mark, attendance etc. The analysis of students' performance obtaining higher education is the need of the hour to upgrade the current education system. This analysis is important for maintaining the effectiveness of learning methods and planning a strategic program during their educational period in an institution.

For improving education system, various Data Mining tools and techniques can be used effectively in Educational Data Mining. There are many classification algorithms like J48, Random Forest, Rap Tree, LMT, Naïve Bayes, BayesNet and PARTetc [2]. In this paper an attempt has been made to compare the classification algorithms J48, Random Forest, Rap Tree, LMT, Naïve Bayes, BayesNet and PART using students' academic performance dataset. The classification algorithm of data mining can be used to classify students based on certain fields such as evaluation of marks, attendance etc.

II. Literature Review

Khasanah et. al. [2] had proposed a method where authors tried to predict the performance of students based on selecting highly influential attributes .In this paper, authors collected the data from Department of Industrial Engineering Universitas Islam Indonesia and used Bayesian Network and Decision Tree algorithms for classification and prediction of student performance. In the process of attribute selection student's attendance and Grade Point Average showing the highest value.

Sheik et. al. [3] conducted a study to analysis student learning behavior by using different data mining models, like classification, clustering, decision tree, sequential pattern mining and text mining. They used tools like KNIME (Konstanz Information Miner), RAPIDMINER, WEKA, CARROT, ORANGE, R Programming, and iDA. These tools have different capacities for prediction and evaluation.

Ankita A Nichatet. al. [4] used decision tree and artificial neural network techniques to propose classification models. They used questionnaires to get positive and negative points for evaluation of the students' performance.

V. Hegdeet. al [5] proposed methodologies like computational techniques, feature selection , preparing survey based on questionnaires, Collecting facts of academic records, pre-processing techniqueand prediction. Vanaja, et al. [6] had applied feature selection techniques on the medical dataset. This technique was applied on high dimensional dataset to select the appropriate features and it was helpful to produce the best accuracy as well as reduce the time and space. They had also used Filter method, Wrapper method and embedded method for feature selection

III. Methodology

Data mining is the knowledge discovery process from a huge data volume. The mechanism works in large dataset where the student performance is evaluated.

A. Data Preprocessing

1. Data Collection

For the purpose of collecting the data, questionnaires were built in Google form and a survey was made with the students from Computer Science and Engineering Department, Assam Kaziranga University. A Total of 152 questionnaires were completed with the help of Google form [7].

2. Data selection and transformation

In this phase, only those fields are considered which are required for mining. The student Gender, High School Percentage, Higher Secondary Percentage, Internal assessment, End Semester marks, attendance, parent's education, parent's occupation etc. are taken asthe attribute values for predictions.

In this step, we have to prepare data by removing rows with empty values and transforming data for evaluation. A total of 10 rows are removed having more than one empty column. After removing these rows, we obtain a total 141 responses.

Table no 1 shows the description of all responses to the questionnaire. Response values of questions having serial no 4 to 7 are of the form {Excellent, VeryGood, Good, Pass, Fail} where Excellent defines percentage more than or equal to 90%, VeryGood defines percentage more than and equal to 80% and less than 90%, Good defines percentage more than and equal to 60% and less than 80%, Pass defines percentage more than and equal to 30% and less than 60%, Fail defines percentage less than 30%.

Response value of question having serial no 11 describe Family Income is of the form {Medium, Low, AboveMedium, High, VeryHigh} where VeryHigh defines income more than and equal to 70000 per month, High defines income more than and equal to 60000 and less than 70000 per month, AboveMedium defines income more than and equal to 40000 and less than 60000 per month, Medium defines income more than and equal to 10000 and less than 40000 per month, Low defines income less than and equal to 10000 per month.

Response value of question having serial no 12 describe study hours is of the form {Poor, Average, Good} where Poor defines less than or equal to 3 hours per day, Average defines more than 3 hours and less than or equal to 6 hours per day , Good defines more than 6 hours per day.

Response value of question having serial no 13 describe attendance of the form {Poor, Average, Good} where Good defines attendance more than or equal to 80%, Average defines attendance more than or equal to 60% and less than 80%, Poor defines attendance less than 60%.

Response value of question having serial no 14 describe hang out time of the form {More, Less, Average} where More defines hang out time more than or equal to 6 hours per day, Average defines hang out time more than or equal to 3 hours per day and less than 6 hours per day, Less defines hang out time less than 3 hours per day.

Response values of question having serial no 21 describe no of friends in the form of {More, Less, Average} More defines no of friends more than or equal to 10, Average defines no of friends more than or equal to 6 and less than 10, Less defines no of friends less than 6. Student related variables are shown in the Table no 1

Table no 1: Description of Student related attributes

SI.No	Attribute	Description	Possible values
1	Gen	Gender	{M, F}
2	Caste	Caste	{G, MOBC,OBC, ST, SC}
3	MT	Mother Tongue	{Assamese, Bengoli, Hindi, Bodo}
4	HP	High school percentage	{Excellent, VeryGood, Good, Pass, Fail }

5	HSP	Higher Secondary Percentage	{Excellent, VeryGood, Good, Pass, Fail }
6	IA	Internal Assessment	{Excellent, VeryGood, Good, Pass, Fail }
7	ESM	End Semester Marks	{Excellent, VeryGood, Good, Pass, Fail }
8	LOC	Living Location	{hostel, home, privatemess}
9	Trans	Transport	{Walking, PrivateCar, PublicTransport, UniversityBus }
10	MS	Marital Status	{Unmarried, Married}
11	FamI	Family Income	{Medium, Low, AboveMedium, High, VeryHigh}
12	SH	Study hours	{Poor, Average, Good}
13	Atd	Attendance	{Poor, Average, Good}
14	HT	Hangout time	{More, Less, Average}
15	ER	Exercise regularly	{Yes, No}
16	PMS	Parents Marital Status	{Married, Separated, Divorced, Widowed}
17	FaQ	Father's Qualification	{Elementary, Secondary, Metriculation, Degree, PostGraduate, NoEducation}
18	MoQ	Mother's Qualification	{Elementary, Secondary, Metriculation, Degree, PostGraduate, NoEducation}
19	FaO	Father's Occupation	{Business, Service , Retired, NotApplicable , Farmer}
20	MoO	Mother's Occupation	{Business, Service , Retired, Housewife}{Business, Service , Retired, Housewife}
21	NoF	No of Friends	{More, Less, Average}
22	BP	Back Papers	{Y, N}

B. Selection of Attributes

In this step we are going to find out most correlated attributes based on CorrelationAttributeEval in WEKA, which is used to evaluate the correlation between the class and other attributes. In this step we are able to know how much these correlated attributes going to affect the final class. Here we are determining average correlation of the attributes to the final class. In turn this step will help us to find out the attributes with less correlation value and the attributes with high correlation value. After getting the correlation values, attributes with less correlation values will be removed to maintain the accuracy.

Following Table no 2 shows the average correlation of different attributes

Table no 2: Correlation values of attributes**.**

Seqence	Attribute name	Correlation values
1	HSP	0.6354
2	BP	0.3417
3	HP	0.2931
4	IA	0.2685
5	ER	0.1668
6	Atd	0.1043
7	FaO	0.0935
8	MT	0.0907
9	SH	0.0884
10	FaQ	0.0857
11	MoQ	0.0781
12	PMS	0.0664
13	NoF	0.0642
14	HT	0.0633
15	LOC	0.0551
16	Caste	0.0491
17	FamI	0.0488
18	Gen	0.0412
19	MoO	0.0392
20	Trans	0.0356
21	MS	0.0301

IV. Experiments and Results

The WEKA tool provides built in classification algorithms to get results in a flexible way. In this paper J48, Random Forest, Rap Tree, LMT, Naïve Bayes, BayesNet and PART classification algorithms available in WEKA have been used.There are 141 records from Computer Science and Engineering Department, Assam Kaziranga University, Jorhat, Assam with 12 selected attributes. Below tables show the performance of different algorithms based on different parameters.

Table no 3: Comparative study of different algorithms based on accuracy

Algorithms	Accuracy	Correctly classified instances	Incorrectly classified instances
J48	85.81%	121	20
Random Forest	100%	141	0
Rap Tree	68%	97	44
LMT	80%	113	28
Naïve Bayes	78%	110	31
BayesNet	70%	111	30
PART	90%	127	14

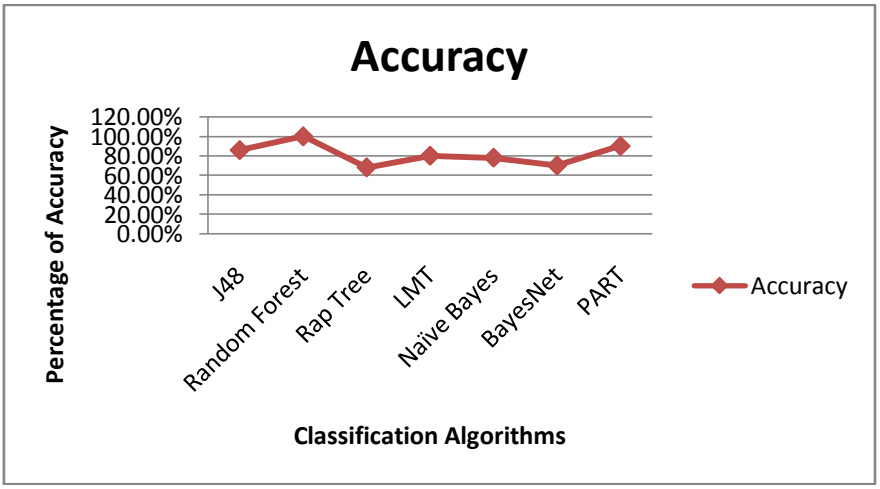

Figure no 1: Algorithms vs. Accuracy Graph

Observation from Table no 3 andFigure no 1: From the comparative study of the seven algorithms, the Random Forest algorithm shows the maximum rate of accuracy with zero incorrectly classified instances.

Table no 4: Comparative study of different algorithms based on Kappa statistics

Algorithms	Kappa statistics
J48	0.7862
Random Forest	1
Rap Tree	0.5097
LMT	0.7034
Naïve Bayes	0.6771
BayesNet	0.6866
PART	0.8517

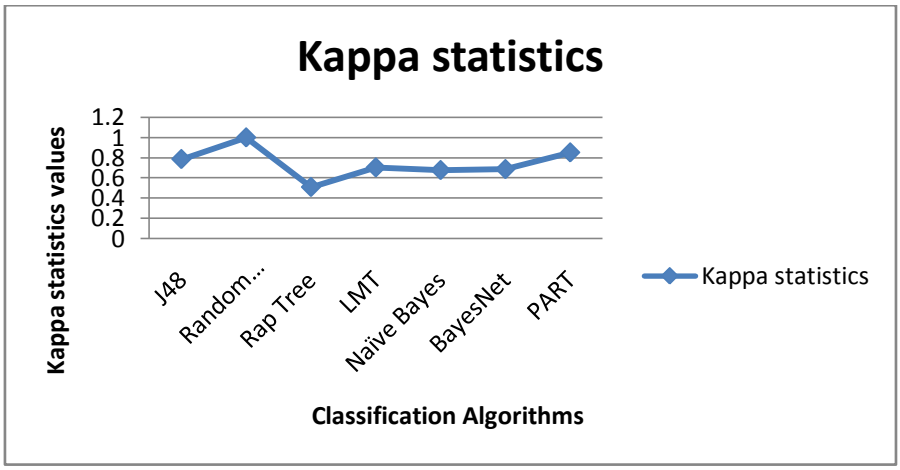

Figure no 2: Algorithms vs. Kappa statistics values

Observation from Table no 4 and Figure no 2: In this comparison also Random Forest algorithm is showing the Kappa statistics value 1, which is greater than all the other algorithms. This means Random Forest algorithm is highly significant based on kappa statistics.

Table no 5 : Comparative study of different algorithms based on MAE(Mean Absolute Error)

Algorithms	MAE(Mean Absolute Error)
J48	0.0888
Random Forest	0.0702
Rap Tree	0.185
LMT	0.1333
Naïve Bayes	0.1015
BayesNet	0.0975
PART	0.0715

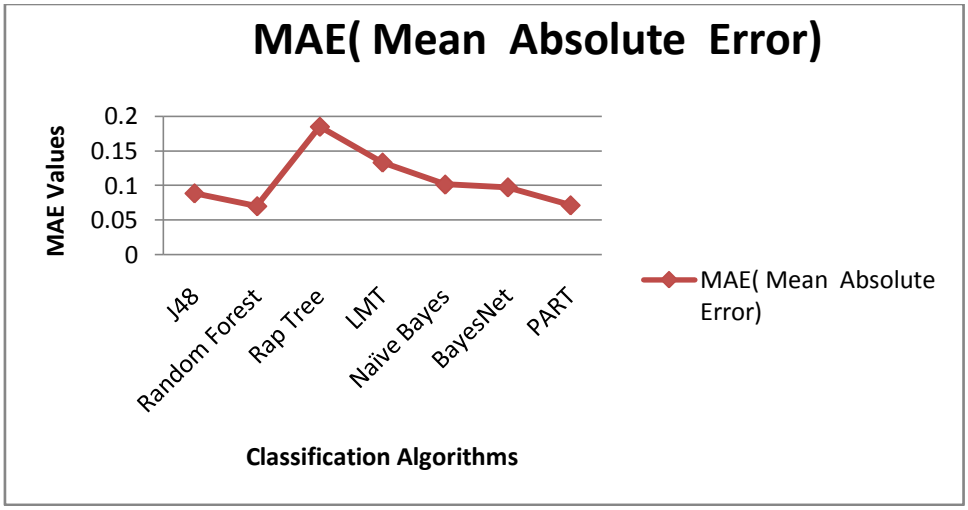

Figure no 3: Algorithms vs. MAE values

Observation from Table no 5 and Figure no 3: Here Random Forest algorithm is showing low rate of Mean Absolute Error (MAE), i.e. this algorithm is efficient than all the other algorithms in terms of MAE.

Table no 6: Comparative study of different algorithms based on RMSE(Root Mean Square Error)

Algorithms	RMSE(Root Mean Square Error)
J48	0.2107
Random Forest	0.1095
Rap Tree	0.3041
LMT	0.2452
Naïve Bayes	0.2409
BayesNet	0.2383
PART	0.1776

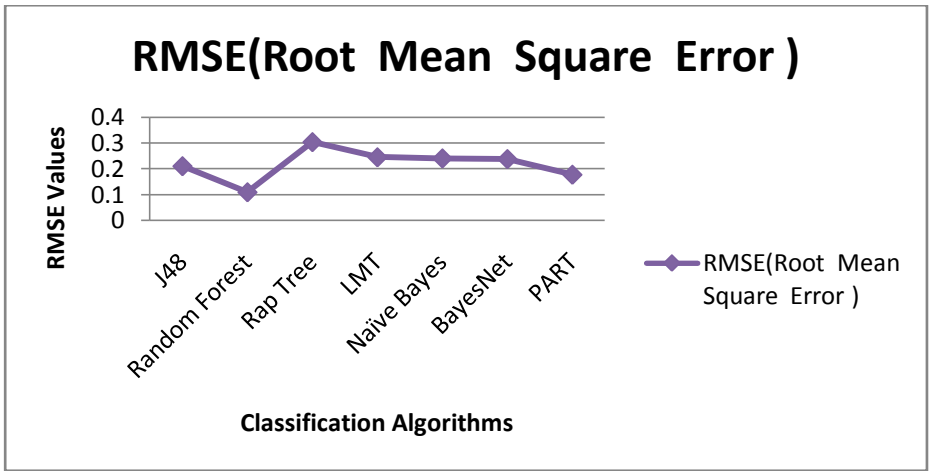

Figure no 4: Algorithms vs. RMSE values

Observation from Table no 6 and Figure no 4:Here also Random Forest algorithm is showing low rate of Root Mean Square Error (RMSE).

Table no 7: Comparative study of different algorithms based on RAE(Relative Absolute Error)

Algorithms	RAE(Relative Absolute Error)
J48	32.72
Random Forest	25.867
Rap Tree	68
LMT	49
Naïve Bayes	37.41
BayesNet	35.92
PART	29.42

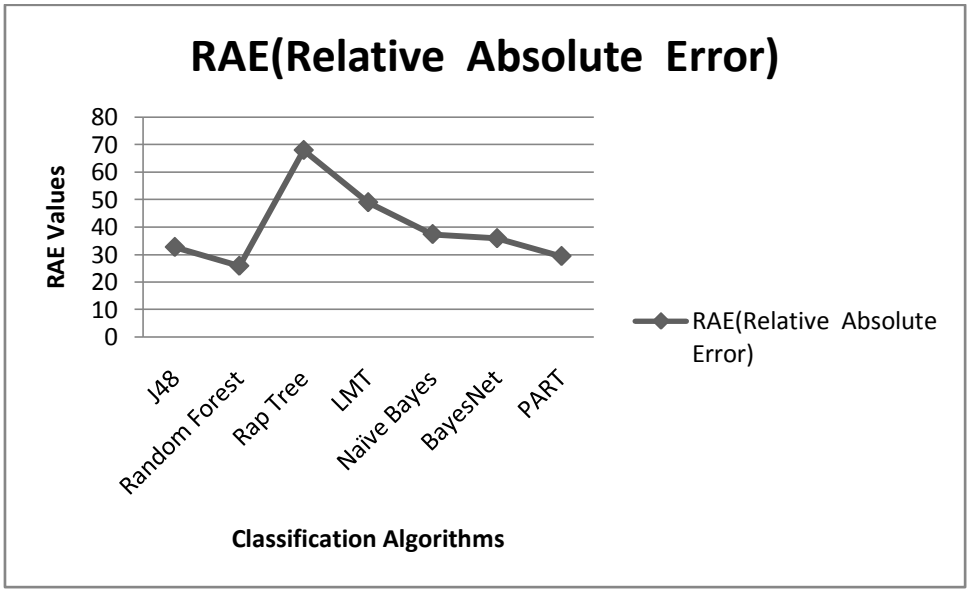

Figure no 5: Algorithms vs. RAE values

Observation from Table no 7 and Figure no 5:Here Random Forest algorithm is showing low rate of Relative Absolute Error (RAE), i.e. this algorithm is efficient than all the other algorithms in terms of RAE.

Table no 8 : Comparative study of different algorithms based on RRSE(Root Relative Squared Error)

Algorithms	RRSE(Root Relative Squared Error)
J48	57.38
Random Forest	29.82
Rap Tree	82
LMT	66

Naïve Bayes	65.6
BayesNet	64.89
PART	45

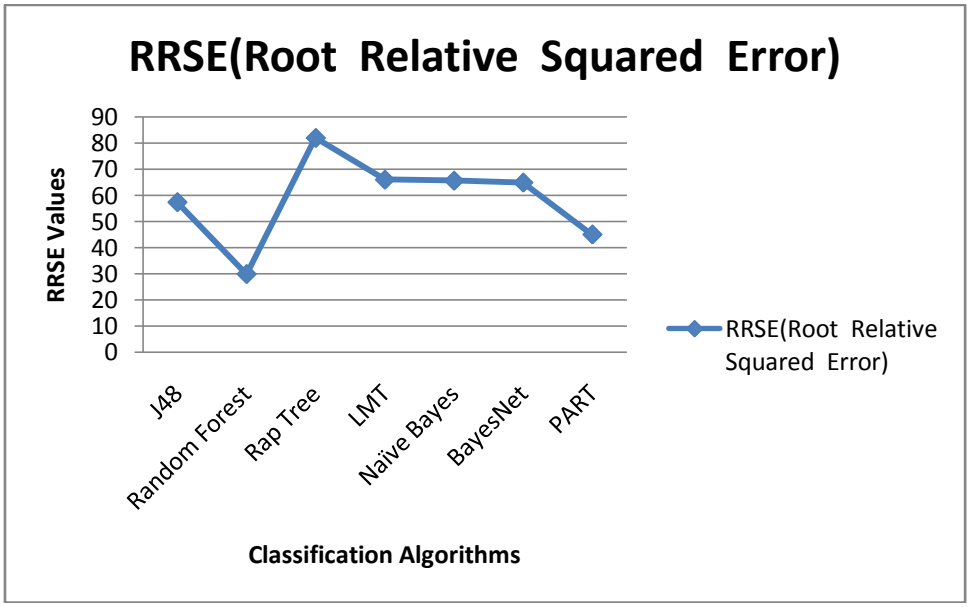

Figure no 6: Algorithms vs. RRSE values

Observation from Table no 8 and Figure no 6:In this comparison also Random Forest algorithm is showing low rate of Root Relative Square Error (RRSE).

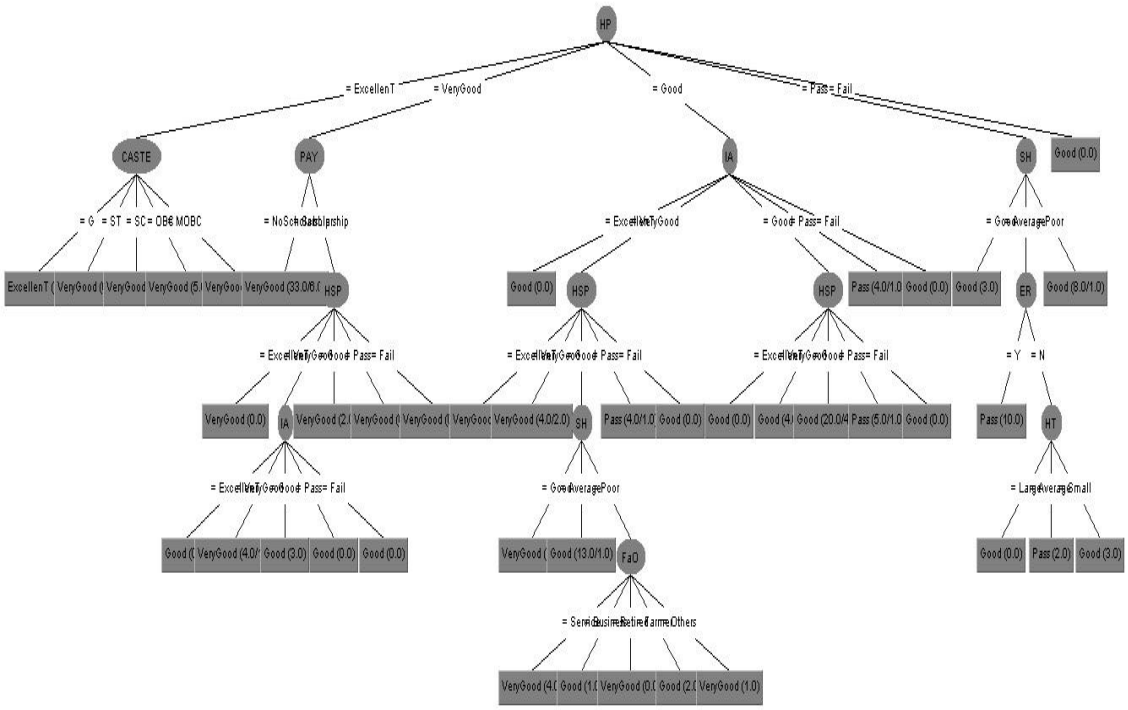

Figure no 7: J48 Tree Visualization

Figure no 7shows the result tree of the J48 algorithm, which is the best approach to show the most correlated attributes to the final class ESM. Each node in the tree is an attribute, and its branches are drawn on the basis of the responses. Each node can be considered as a decision. The tree can be used also for predicting end semester result by giving responses to the nodes of the tree.

V. Conclusion

This study aims to explore and evaluate the academic performance of students based on data collected from Assam Kaziranga University, Jorhat, Assam. Total 152 responses were collected with 22 questionnaires/ attributes. Based on attribute selection process, 12 highly correlated attributes are selected. A total of 10 rows are removed having more than one empty column. After removing these rows, we obtain a total of 141 responses. After Pre-processing and selection of attributes, seven different classification algorithms are applied on the dataset like J48, Random Forest, Rap Tree, LMT, Naïve Bayes, BayesNet and PART. After doing the analysis based on parameters Accuracy, Kappa statistics, MAE(Mean Absolute Error), RMSE(Root Mean Square Error), RAE(Relative Absolute Error), RRSE(Root Relative Squared Error), it can be concluded that Random Forest algorithm is highly efficient than all other algorithms used for the purpose of analysis

References

[1]. BakerRSJd, Yacef K.The state of educational datamining in 2009: A review and future visions. J EduData Min **2009.**
[2]. Khasanah, A.U. and Harwati, A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques. IOP Conf. Series: Materials Science and Engineering, **2017**. 215(012036): p. 7.
[3]. NikitabenShelke and ShriniwasGadage, "A survey of data mining approaches in performance analysis and evaluation", International Journal of Advanced Research in Computer Science and Software Engineering ,vol 5, iss 4, **2015**.
[4]. Nichat, A.A. and D.A.B. Raut, Analysis of Student Performance Using Data Mining Technique. International Journal of Innovative Research in Computer and Communication Engineering, **2017**. 2007(An ISO 3297): p. 5.
[5]. V. Hegde, "Dimensionality Reduction Technique for Developing Undergraduate Student Dropout Model using Principal Component Analysis through R Package," pp. 1–6, **2016**.
[6]. Vanaja, S., and Ramesh Kumar, K. Analysis of Feature Selection Algorithms on Classification: A Survey. International Journal of Computer Applications, 96(17)**2014**.
[7]. Bharti Thakur, Data Mining With Big Data Using C4.5 and Bayesian Classifier, , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 8, August **2014**

Er. AmlanJyoti Baruah, etal. "A Comparative Analysis of Different Classification Algorithms based on Students' Academic Performance Using WEKA." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22.1 (2020), pp. 49-56.