IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. III (Jan. 2014), PP 63-69
www.iosrjournals.org

An Analysis of students' performance using classification algorithms

Mrs. M.S. Mythili[1], Dr. A.R.Mohamed Shanavas[2]
[1]*Ph.D Research Scholar, Bharathidasan University & Assistant Professor,*
Department of Computer Applications, Bishop Heber College, Tiruchirappalli 620 017, TamilNadu, India.
[2]*Associate Professor, Dept. of Computer Science, Jamal Mohamed College,*
Tiruchirappalli 620 020,TamilNadu, India.

Abstract: *In recent years, the analysis and evaluation of students' performance and retaining the standard of education is a very important problem in all the educational institutions. The most important goal of the paper is to analyze and evaluate the school students' performance by applying data mining classification algorithms in weka tool. The data mining tool has been generally accepted as a decision making tool to facilitate better resource utilization in terms of students' performance. The various classification algorithms could be specifically mentioned as J48, Random Forest, Multilayer Perceptron, IB1 and Decision Table are used. The results of such classification model deals with accuracy level, confusion matrices and also the execution time. Therefore conclusion could be reached that the Random Forest performance is better than that of different algorithms.*

Keywords*: Decision Table, IB1, J48, Multilayer Perceptron, Random Forest*

I. Introduction

Data Mining could be a promising and flourishing frontier in analysis of data and additionally the result of analysis has many applications. Data Mining can also be referred as Knowledge Discovery from Data (KDD).This system functions as the machine-driven or convenient extraction of patterns representing knowledge implicitly keep or captured in huge databases, data warehouses, the Web, data repositories, and information streams. Data Mining is a multidisciplinary field, encompassing areas like information technology, machine learning, statistics, pattern recognition, data retrieval, neural networks, information based systems, artificial intelligence and data visualization.

The application of data mining is widely prevalent in education system. Educational data mining is an emerging field which can be effectively applied in the field of education. The educational data mining uses several ideas and concepts such as Association rule mining, classification and clustering. The knowledge that emerges can be used to better understand students' promotion rate, students' retention rate, students' transition rate and the students' success. The data mining system is pivotal and crucial to measure the students' performance improvement. The classification algorithms can be used to classify and analyze the students' data set in accurate manner. The students' academic performance is influenced by various factors like parents' education, locality, economic status, attendance, gender and result.

The main objective of the paper is to use data mining methodologies to study and analyze the school students' performance. Data mining provides many tasks that could be used to study the students' performance. In this paper, the classification task is employed to gauge students' performance and deals with the accuracy, confusion matrices and the execution time taken by the various classification data mining algorithms.

This paper is catalogued as follows. Section 2 enumerates a related work. Section 3 presents the idea of Classification and discusses the aspects of classification algorithm. Section 4 elaborates a Data Preprocessing. Section 5 explains the Implementation of model construction. Section 6 describes the results and discussions. Section 7 provides the conclusion.

II. Related Work

Alaael-Halees 2009 suggested that Data Mining is an emerging methodology used in educational field to enhance the understanding of learning process. The application of Data mining is widely spread in Higher Education system. In Education domain many researchers and authors have explored and discussed various applications of data mining in higher education. The authors had gone through the survey of the literature to understand the importance of data mining applications. In the year 2001 Luan al. suggested a powerful decision support tool called data mining. Data Mining is a powerful tool for academic purposes Alumni, Institutional effectiveness, marketing and enrollment can benefit from the use of data mining Data Mining is the most suited technology that can be used by lecturer, student, alumnus, manager and other educational staff and is a useful tool for decision making on their educational activities

Delmater et.al. Placed stress on underlying predictive modeling which is a mixture of mathematics, computer science and domain expertise Qasem et.al. Started an attempt to use data mining functions to analyze and evaluate student academic data and to enhance the quality of the higher educational system. The higher managements can use such classification model to enhance the courses outcome according to the extracted knowledge. Such knowledge can be used to give a deeper understanding of student's enrollment pattern in the course under study, and the faculty and managerial decision maker in order to utilize the necessary actions needed to provide extra basic course skill classes and academic counseling. On the other hand, using such knowledge the management system can improve their policies, enhance their strategy, and improve the quality of management system.

III. Classification

This process is employed to classify data into predefined categorical class labels. Classification can be a two step process consisting of training and testing. In the first step, a model is constructed by analyzing the data tuples from training data having a collection of attributes. For every tuple in the training data, the worth of class label attribute is understood. Classification rule is applied on training data to form the model. In the second step of classification, test data is employed to examine the accuracy of the model. If the accuracy of the model is appropriate then the model can be used to classify the unknown data tuples. The fundamental techniques for classification are decision tree classifier, neural networks, rule based classifier and Lazy based classifier.

3.1. Classification Algorithms

This research paper contains a rule based classifier (Decision Table), a common decision tree classifier C4.5 (J48), Random Forest, a neural network (Multilayer Perceptron) and a Lazy based classifier (IB1).The classifiers are mentioned in brief.

3.2. Decision tree classifiers

A decision tree can be a flow chart resembling a tree structure, where every internal node is denoted by rectangles and the leaf nodes are denoted by ovals. This is often used algorithm because of easy implementation and easier to understand compared to different classification algorithms. Decision tree starts with a root node that helps the users to take required actions. From this node, users split every node recursively according to decision tree learning algorithm. The ultimate result is a decision tree in which each branch represents an outcome.

3.2.1. C4.5 (J48)

This algorithm can be a successor to ID3 developed by Quinlan Ross. It is additionally supported the Hunt's algorithm.C4.5 handles each categorical and continuous attributes to create a decision tree, so as to handle continuous attributes. C4.5 splits the attribute values into 2 partitions based on the chosen threshold. It additionally handles missing attribute values. C4.5 has the concept of Gain Ratio as an attribute selection measure to create a decision tree. It prunes the biasness of information gain once there are many outcome values of an attribute. At first, calculate the gain ratio of every attribute. The root nodes are the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to get rid of unessential branches with in the decision tree to enhance the accuracy of classification.

3.2.2. Random Forest

Random Forests is a bagging tool that leverages the ability of multiple varied analyses, organization strategies, and ensemble learning to supply correct models, perceptive variable importance ranking, and laser-sharp coverage on a record-by-record basis for deep data understanding. Its strengths are recognizing outliers and anomalies in knowledgeable data, displaying proximity clusters, predicting future outcomes, characteristic necessary predictions, discovering data patterns, exchange missing values with imputations, and providing perceptive graphics.

3.3. Neural Network

Multilayer Perceptron (MLP) algorithm is one of the most widely used and common neural networks. Multilayer Perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a collection of acceptable output. An MLP consists of multiple layers of nodes in an exceedingly directed graph, with every layer totally connected to the consequent one. Their current output depends solely on the present input instance. It trains victimization back propagation.

3.4. IB1

IBI is nearest neighbor classifier. It uses normalized Euclidean distance to search out the training instance nearest to the given test instance, and predicts the identical category as this training instance. If many instances have the smallest distance to the test instance, the primary one obtained is employed. Nearest neighbor methodology is one of the effortless and uncomplicated learning/classification algorithms, and has been effectively applied to a broad variety of issues.

3.5. Decision Table

Decision Tables are classification models elicited by machine learning algorithms and are used for creating predictions. A decision table consists of a hierarchical table within which entry in a higher level table gets broken down by the values of a pair of additional attributes to make another table.

IV. Data Preprocessing

Datasets utilized within the classification algorithm ought to be clear and can be preprocessed for handling missing or redundant attributes. The data are to be handled with efficiency to induce the best outcome from the Data Mining process.

4.1. Attribute Identification

Dataset collected from student database consists of

Attributes	Description	Possible values
Gender	Gender (male, female)	M,f
Locality	Living locality	Urban, Rural
paredu	Parental education	Edu,unedu
Eco	Economic status	High, Low
Attendance	Class attendance	High, Low
Result	Students' result	First,Second,Third,Fail

V. Implementation of Model Construction

Weka is open source software system that implements a large collection of machine learning algorithms and is widely utilized in data mining applications. From the above data, student.arff file was created. This file was loaded into a WEKA explorer. The students' academic performance is influenced by various factors like parents' education, locality, economic status, attendance, gender and result from the different school students.260 samples were taken for the implementation. The classify panel permits the user to use classification algorithms to the dataset, to estimate the accuracy of the resulting predictive model, and to visualize the model.

The decision tree classifier C4.5 (J48), Random Forest, Neural Network (Multilayer Perceptron) and Lazy based classifier (IB1) Rule based classifier (Decision Table) were enforced in weka. Under the "Test options", the 10 fold cross validation is chosen.

VI. Results and Discussion

The analysis and interpretation of classification is time consuming process that needs a deep understanding of statistics. The process needs a large amount of time to finish and expert analysis to look at the classification and relationships within the data.

TABLE 1: Attributes Ranking using information gain and gain ratio

S.No	Attribute	Information Gain		Gain Ratio	
		Value	Rank	Value	Rank
1.	Gender	0.0286	5	0.035	5
2.	Locality	0.0544	4	0.0544	4
3.	P.ed	0.1016	3	0.1193	3
4.	Attendance	0.6429	1	0.6592	1
5.	Eco	0.579	2	0.582	2

This section presents the results generated from the study. The attributes were ranked in order of its importance using information gain and gain ratio measures. The ranking of each Attribute evaluators was done using ranker search method.

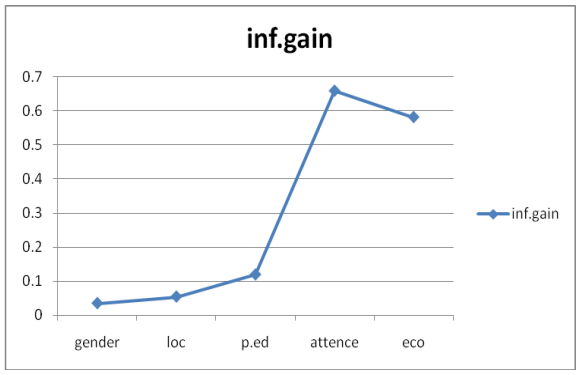

Figure 1: information gain of the attributes

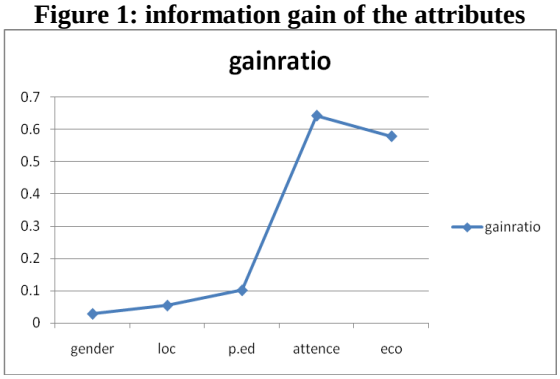

Figure 2: gain ratio of the attributes

The Figure 1 and Figure 2 clearly identify the Attribute ranking according to information gain and gain ratio of the attributes.

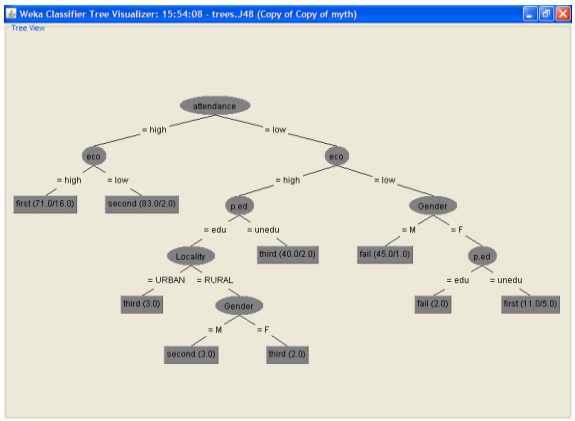

Figure 3: Visual image of generated decision trees

The attendance of the students' is taken as root node from that economical status and parents' education support taken as branch node and so on. The knowledge represented by decision tree can be extracted within the form of IF-THEN rules.

1. IF attendance="high" AND eco="high" THEN result="first"
2. IF attendance="high" AND eco="low" THEN result="second"
3. IF attendance="low" AND eco="high" AND p.ed=edu AND locality=urban THEN result="third"
4. IF attendance="low" AND eco="high" AND p.ed=edu AND locality=rural AND gender="m" THEN result="second"
5. IF attendance="low" AND eco="high" AND p.ed=edu AND locality=rural AND gender="f" THEN result="third"
6. IF attendance="low" AND eco="high" AND p.ed=unedu THEN result="third"
7. IF attendance="low" AND eco="low" AND gender=m THEN result="fail"
8. IF attendance="low" AND eco="low" AND gender=f AND p.ed=edu THEN result="fail"
9. IF attendance="low" AND eco="low" AND gender=f AND p.ed=unedu THEN result="first"

From the above set of rules an inescapable conclusion emerges the attendance is considerably related with student performance. From the rule set it was found that parent education, locality, gender, Economic Status, and different factors are of high potential variable that have an effect on students' performance for getting good performance in examination result.

Table 2 Performance result for Classifiers

Evaluation Criteria	Classifiers				
	J48	**Random Forest**	**Multilayer Perceptron**	**IB1**	**Decision Table**
Timing to build the model(sec)	0.03	0	1.11	0.01	0.02
Correctly classified instances	224	232	227	209	224
Incorrectly classified instances	36	28	33	51	36
Accuracy (%)	86.15%	89.23%	87.30%	80.38%	86.15%

In Table 2 the time build by the Random Forest is less than the remaining classifier is shown and therefore the percentage of correctly classified instances is usually referred as accuracy of the model. Hence Random Forest classifier can be termed as more accurate than other classifiers.

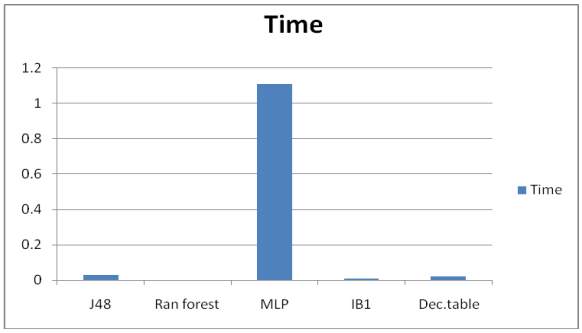

Figure 4: Time taken to build the classifier algorithm

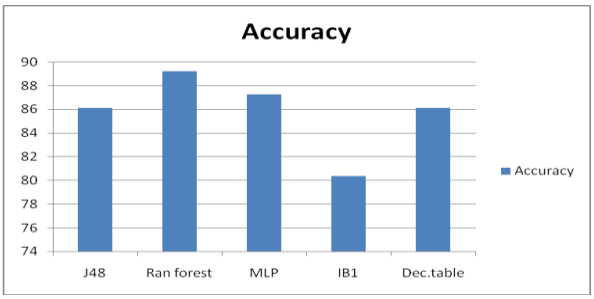

Figure 5: Accuracy of the classifier algorithm

The Figure 4 and Figure 5 shows that the graphical representation of time and accuracy results of school students' performance analysis based on students' dataset. It clearly reveals that Random Forest is a very best classifier for analyzing the school students' performance result consuming less time coupled with good accuracy.

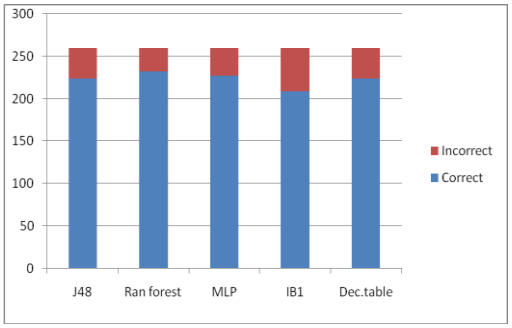

Figure 6: Efficiency of different classifiers

The Figure 6 explains the graphical representation of correctly classified instances of results of school students' performance analysis based mostly on student dataset. The highest percentage of correctly classified instances is the Random Forest classifier.

Table 3: Error measurement for classifiers

Evaluation Criteria	Classifiers				
	J48	Random Forest	Multilayer Perceptron	IB1	Decision Table
Kappa statistic	0.8116	0.8531	0.8278	0.7306	0.8118
Mean absolute error	0.0984	0.0836	0.0836	0.0981	0.1328
Root mean squared error(RMSE)	0.2336	0.2128	0.23	0.3132	0.2384
Relative absolute error(RAE)	26.9378%	22.8844%	22.8995%	26.8611%	36.3805%
Root relative squared error(RRSE)	54.6895%	49.8177%	53.8346%	73.3071%	55.8143%

In Table 3 it explains the time build by the Random Forest is less than the remaining classifier. Kappa statistics is a measure of the degree of non random agreement between observers and measurement of a particular categorical variable. The root mean square error and Mean absolute error of Random Forest are minimum when compared to other classifiers. Therefore the Random Forest is that the efficient classification technique among remaining classification technique.

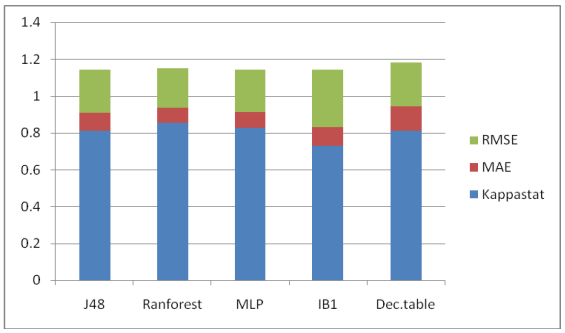

Figure 5: Error rate of different classifiers

The Figure 5 compares errors among completely different classifiers (root mean square error and Mean absolute error) Random Forest has lower error rate compared to different classifiers. Therefore the Random Forest is the efficient classification technique among remaining classifiers.

Table 4: Class label accuracy for classifiers

classifier	TP	FP	Precision	Recall	Class
J48	0.919	0.111	0.722	0.809	First
	0.88	0.029	0.88	0.88	Fail
	0.844	0.018	0.964	0.9	Second
	0.808	0.024	0.894	0.848	Third
Random Forest	0.935	0.101	0.744	0.935	First
	0.94	0.014	0.94	0.94	Fail
	0.875	0.018	0.966	0.875	Second
	0.827	0.01	0.956	0.827	Third
MLP	0.887	0.086	0.764	0.887	First
	0.96	0.038	0.857	0.96	Fail
	0.844	0.012	0.976	0.844	Second
	0.827	0.029	0.878	0.827	Third
IB1	0.645	0.086	0.702	0.645	First
	0.94	0.029	0.887	0.94	Fail
	0.865	0.104	0.83	0.865	Second
	0.75	0.053	0.78	0.75	Third
Decision Table	0.887	0.101	0.733	0.887	First
	0.92	0.038	0.852	0.92	Fail
	0.844	0.018	0.964	0.844	Second
	0.808	0.024	0.894	0.808	Third

The Table 4 clearly shows the performance of every classifier based on the true positive rate (TP rate) and false positive rate (FP rate), precision, recall and different measures. These measures are very helpful for comparing the classifiers based on the accuracy. The Random Forest outperforms all different classifiers within the students' dataset.

Table 5: Confusion Matrix

Classifier	First	FAIL	Second	Third	Class
J48	57	4	1	0	First
	6	44	0	0	Fail
	10	0	81	5	Second
	6	2	2	42	Third
Random Forest	58	2	2	0	First
	3	47	0	0	Fail
	10	0	84	2	Second
	7	1	1	43	Third
MLP	55	6	1	0	First
	1	48	0	1	Fail
	10	0	81	5	Second
	6	2	1	43	Third
IB1	40	4	10	8	First
	2	47	1	0	Fail
	10	0	83	3	Second
	5	2	6	39	Third
Decision Table	55	6	1	0	First
	4	46	0	0	Fail
	10	0	81	5	Second
	6	2	2	42	Third

The Table 5 reveals that the confusion matrices are very helpful for analyzing the classifiers.

VII. Conclusion

The work explores the potency of machine learning algorithms in deciding the influence of result, parental education, gender, economy and the locality within the study and analyze of school students'performance.It is discovered that Random Forest performance is best than that of different algorithms employed in the study. This study is going to be terribly useful for the educational institutions. In future, it is doable to increase the analysis by using different clustering techniques and association rule mining for the students' dataset.

References

[1]. Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006) *'Mining Student Data Using Decision Trees'*, The 2006 International Arab Conference on Information Technology (ACIT'2006) – Conference Proceedings.

[2]. Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) *'Data Mining Model for Higher Education System',* European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

[3]. Baradwaj, B. and Pal, S. (2011) *'Mining Educational Data to Analyze Student s' Performance'*, International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

[4]. Chandra, E. and Nandhini, K. (2010) *'Knowledge Mining from Student Data'*, European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

[5]. El-Halees, A. (2008) *'Mining Students Data to Analyze Learning Behavior: A Case Study'*, The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18.

[6]. Han, J. and Kamber, M. (2006) *Data Mining: Concepts and Techniques*, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

[7]. Kumar, V. and Chadha, A. (2011) *'An Empirical Study of the Applications of Data Mining Techniques in Higher Education',* International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84.

[8]. Mansur, M. O. , Sap, M. and Noor , M. (2005) *'Outlier Detection Technique in Data Mining: A Research Perspective',* In Postgraduate Annual Research Seminar.

[9]. Romero, C. and Ventura, S. (2007) *'Educational data Mining: A Survey from 1995 to 2005'*, Expert Systems with Applications (33), pp. 135-146

[10]. Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, *"Mining student data using decision trees"*, International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[11]. U. K. Pandey, and S. Pal, *"A Data mining view on class room teaching language",* (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[12]. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, *"Data mining model for higher education system",* European Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.

[13]. M. Bray, *The shadow education system: private tutoring and its implications for planners*, (2nd ed.), UNESCO, PARIS, France, 2007.

[14]. B.K. Bharadwaj and S. Pal. *"Data Mining: A prediction for performance improvement using classification",* International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[15]. J. R. Quinlan, *"Introduction of decision tree: Machine learn",* 1: pp. 86-106, 1986.

[16]. Vashishta, S. (2011). Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm. *IJACSA - International Journal of Advanced Computer Science and Applications, 2*(4), 77-80.