OPTIMAL ALGORITHM FOR PREDICTING STUDENTS' ACADEMIC PERFORMANCE

[1]Kuyoro 'Shade O., [2]Prof. Nicolae Goga, [1]Dr. Oludele Awodele and [1]Dr. Samuel Okolie
[1]Department of Computer Science, Babcock University, Nigeria
[2]University of Groningen, The Netherlands or Politenica Bucharest

ABSTRACT

Machine learning has been successfully applied to numerous domains such as pattern recognition, image recognition, fraud detection, medical diagnosis, banking, bioinformatics, commodity trading, computer games and various control applications. Recently, this paradigm is been employed to enhance and evaluate higher education tasks. The focus of this work is on identifying the optimal algorithm suitable for predicting first-year tertiary students academic performance based on their family background factors and previous academic achievement. One thousand five hundred (1,500) enrolment records of students admitted into computer science programme Babcock University, Nigeria between 2001 and 2010 was used. The students' first year academic performance was measured by Cumulative Grade Point Average (CGPA) at the end of the first session and the previous academic achievement was measured by SSCE grade score and UME score. Waikato Environment for Knowledge Analysis (WEKA) was used to generate 10 classification models(five decision tree algorithms -Random forest, Random tree, J48, Decision stump and REPTree and five rule induction algorithms –JRip, OneR, ZeroR, PART, and Decision table) and a multilayer perceptron, an artificial neural network function. These algorithms were compared using 10-fold cross validation and hold-out method considering accuracy level, confusion matrices and CPU time to determine the optimal model. This work will be taken further by designing a framework of predictive system based on the rules generated from the optimal model.

Keywords: decision trees, neural networks, family background, machine learning, predictive system.

INTRODUCTION

Machine learning has proven to be of great value in various application domains. It is especially useful in data mining problems where large databases contain valuable implicit regularities that can only be discovered automatically; in poorly understood domains where humans might not have the knowledge needed to develop effective algorithms such as face recognition from images; and in domains where the program must dynamically adapt to changing conditions. (Schaffer, 1994) Machine Learning (ML) techniques embody some of the facets of the human mind that allow us solve complex problems at speeds which outperform even the fastest computers (Schank, 1982). ML techniques have been used successfully in solving many difficult problems such as speech recognition from text (Sejnowski and Rosenberg, 1987), adaptive control (Narendra and Parthasarathy, 1987) and markup estimation in the construction industry (Hegazy and Moselhi, 1994). Designing machine learning approach to solve problems involves a number of choices such as choosing the type of training experience, the target function to be learned, a representation for this target function, and an algorithm for learning the target function from training examples. The most commonly used machine learning algorithms are Artificial Neural Network, Decision trees, Genetic Algorithms, Rule Induction, Regression Methods, and so on. In recent years, machine learning is finding larger and wider applications in higher education learning. It is showing an increasing trend in institutional research. This has to do with the growing interest in knowledge management and in moving from data to information and finally to knowledge discovery.

Higher learning institutions encounter many problems which keep them away from achieving their quality objectives. Some of these problems stem from knowledge gap. Knowledge gap is the lack of significant knowledge at the educational main processes such as counseling, planning, registration, evaluation and marketing. For instance, many learning institutions do not have access to the necessary information to counsel students. Therefore they are not able to give suitable recommendation to the students. The hidden patterns, associations, and anomalies that are discovered by machine learning techniques can help bridge this knowledge gap in higher learning institutions. This knowledge would enable the higher learning institutions in making better decisions, having more advanced planning in directing students, predicting individual behaviors with higher accuracy, and enabling the institution to allocate resources and staff more effectively. It results in improving the effectiveness and efficiency of the processes.

Machine learning is considered the most suitable technology in giving additional insight into educational entities such as; student, lecturer, staff, alumni and managerial behavior. It acts as an active automated assistant in helping them make better decision on their educational activities. A series of recent application of machine learning algorithms to education policy questions including forecasting educational spending and analyzing educational productivity has been carried out revealing the complexities or simplicities of our educational system. (Lemke,1997; Golding and Donalson, 2006; Ventura and Romero, 2011) Given the success of the use of machine learning algorithms in many fields and applications, it seems reasonable that these methodologies should be able to provide us with some new insights into the types of patterns that exist in educational data. The implementation of machine learning algorithms as a tool for determining educational achievement and assessment at all levels continues to increase.

The differential students' performance in tertiary institutions is a source of great concern and research interest to the higher education managements, government, parents and other stakeholders because of the importance of education to national development. Academic institutions are increasingly required to monitor both their performance and that of their students. This gives rise to a need to extract useful information from the available students' large datasets to inform academic policies on how best to improve student retention rates, allocate teaching and support resources, or create intervention strategies to mitigate factors that affect student performance adversely. Maximizing the potential of students, providing evidence of delivering value for money to the bodies that fund them, and performing up to expectation is very crucial to tertiary institutions. Most institutions are often judged by the quality of the awards they provide; for instance, the more honours level graduates a course provides, the better the course is perceived to be. This provides additional quest for institutions to take proactive steps to investigate students' data with a view of finding useful information that can aid planning activities, decision making and students' intervention strategies. It is necessary to carefully measure student outcomes or expected outcomes that may provide evidence as to whether student potential is being realized against some benchmarks.

From diverse literature, the observed poor performance of students in tertiary institutions has been partly traced to poor academic background and wide range of other predictors, including personality factors, intelligence, gender and aptitude tests, academic achievement, previous college achievements, and demographic data. Many researchers have come to some interesting conclusion as to which of these predictors has impacted students' academic performance in tertiary institutions. There is a growing interest and concern in many countries about the problem of school failure and the determination of its main contributing factors. This problem has been referred to as "the one hundred factors problem".(Ventura and Romero, 2011) Different predictors including gender, personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data have been identified in literature as contributors to students' academic performance. The objective of this work is to identify the optimal algorithm suitable for predicting first-year tertiary students academic performance based on their family background factors and previous academic achievement. The dataset used for this work comprise one thousand five hundred (1,500) records of students admitted between 2001 and 2010 into computer science programme Babcock University, Nigeria. The students' first year academic performance was measured by Cumulative Grade Point Average (CGPA) at the end of the first session and the previous academic achievement was measured by SSCE grade score and UME score.

RELATED WORKS

The literature is full of works relating machine learning algorithms or data mining techniques to university admission, student performance, and related problems. Recently, the focus of literature is on application of machine learning to educational datasets to proffer solutions to education challenges especially in relation to predicting students' academic performance.

Varapron et al 2003 used Rough Set theory as a classification approach to analyze student data where the Rosetta toolkit was used to evaluate the student data to describe different dependencies between the attributes and the student status. The dataset used in their experiments is the student data of Suranaree University of Technology (SUT) during 2001-2002 academic year. Delavari and Beikzadeh 2004 proposed a model to represent how data mining can be used in a higher educational system to improve the efficiency and effectiveness of the traditional processes. The model is presented as a guideline for higher educational system to improve the decision-making processes. Mierle et al 2005 describes the results of analyzing data from a large collection of the so-called concurrent version system (CVS) created by many students working on a small set of identical projects (course assignments) in the 2nd year undergraduate computer science course. The proposed model is used to extract all information of student behavior in writing the code of assignments and to find some statistical patterns or predicators that can be used to enhance students' performance in writing code. The result suggests that aspect such as student work habits, even code quality, have little bearing on the student's performance. Kalles and Pierrakeas 2004 discussed different machine learning techniques (decision trees, neural networks, Naive Bayes, instance-based learning, logistic regression and support vector machines) and compared them with genetic algorithm based induction of decision trees. They discussed why the approach has a potential of developing into an alert tool. They embarked in an effort to analyze students' academic performance through the academic years, as measured by the students home work assignments, attempted to derive short rules that explain and predict success or failure in the final exams. Delavari et al 2005 enhance the proposed analysis model of Delavari and Beikzadeh

2004; which they used as a roadmap for the application of data mining in higher educational system. The enhanced model is called Data Mining for Higher Education (DM_EDU). The model allows decision makers to better predict which students are less likely to perform well in that specific course.

Al-Radaideh et al, 2006 use data mining processes, classification tasks –decision trees (ID3,C4.5) and Naïve Bayes, to enhance the quality of higher educational system by evaluating students' data and studying the main attributes that affect the student performance in courses. The data of students who took C++ in 2005 was collected from Yarmouk University. The result shows that the classification accuracy for the 3 algorithms used is rather low which indicates that the collected samples and attributes were not sufficient to generate a classification model of high quality.
Golding and Donalson 2006 stated that the use of performance in first year computer science course is a possible factor which may determine academic performance showing that; gender and age have no significant correlation as predictive factors. Hamalainen and Vinni 2006 compared machine learning algorithms for intelligent tutoring system tackling problems where educational datasets are so small that ML methods cannot be applied directly. They recommended variation of naïve Bayesian classifiers which are robust. Hijazi and Naqvi 2006 used linear regression to determine factors influencing students' academic performance. It was found out that mother's education, family income were high determinants of student academic performance.

Superby et al 2006 & Vandamme et al 2007 studied correlations of various parameters such as attendance, estimated success, previous academic experience and study skill. It was discovered that changing process factors during students stay in the university plays a large part in academic performance. The rate of prediction obtained from the techniques used was not particularly good due to the difficulty in classifying students into 3 groups: high, medium and low risk before the first university exam. Nguyen et al, 2007 compares the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two different academic institutes. The suitability of using data mining techniques for prediction of academic performance was investigated using 2 case studies. In Can Tho University (CTU), Viet Nam. 20492 students' records were used selecting; students' records and GPA at the end of the 2nd year to predict performance in the 3rd year. For Asian Institute of Technology (AIT), Thailand admission information such as academic institute and GPA was used to predict GPA at the end of first year using 936 records. The result shows that decision tree was significantly more accurate than Bayesian networks algorithm for predicting student performance; and that prediction accuracies for minority classes are consistently lower for both data sets and for all classes. To correct this problem re-sampling function was used to oversample the minority classes and under-sample the majority classes thereby achieving more balanced distribution. The research was compared with Bekele and Menzel 2005 and Minaei-Bidgoli et al 2003. The overall result was slightly better than Minaei-Bidgoli et al 2003's. This comparison is to appreciate the use of different approaches in predicting student performance. The overall prediction of the analysis was high showing that the system is reliable for identifying excellent students.

Quadri and Kalyankar, 2010 make use of decision tree analysis to analyze the problem of drop outs in any higher educational institution. Decision trees are used to make important design decisions and explain the interdependencies among the properties of drop out students; providing an instance machine learning technique that can be used to improve the effectiveness and efficiency of modeling process. The study address the capabilities and strengths of decision tree algorithm in identifying drop out students to guide the teachers in concentrating on appropriate features associated with counseling students or arranging financial aid to them. The study is an extension of the educational model developed by Shyamala & Rajagopalan, 2006 . Paris et al, 2010, evaluate the performance of different prediction techniques for prediction of students' CGPA class targeting weak students of second class lower and third class CGPA. Decision trees and Bayesian methods that have comprehensive visual representation were used. The proposed voting technique accuracy was compared with C4.5 NBTree, BayesNet, naïve Bayes, hidden naïve Bayes (HNB) and voting technique on 3 weak classifiers (naïve Bayes, OneR and Decision stump). The result shows that HNB performed well on most classes except for high distribution class which decision stump classifier compliment. Affendey et al, 2010 used attribute importance analysis to rank influencing factors (courses) that contributes to the prediction of students' academic performance. It was determined whether a first year student will graduate higher or lower than a second class upper. 2427 complete records of bachelor of computer science students admitted from 2000 to 2006 were collected. The prediction results using CfS as attribute selection technique shows that Naïve Bayes, AODE and RBFNetwork performed best on the data sets with 95.29% accuracy, on the other hand AODE score best with CoE showing 95.29% accuracy. The result agrees with Golding and Donaldson 2006's findings that first year courses are possible factors determining academic performance.

Bhardwaj and Pal, 2011 justifies the capabilities of data mining techniques in context of higher education. Decision tree is used to evaluate students' performance at the end of semester. Variables considered are previous semester marks, class test grade, seminar performance, assignment, general proficiency, attendance, lab work, and end semester marks. The classification task used is able to predict the student division on the basis of the previous database. This helps to reduce failure ratio because early identification will enable appropriate action. In another study of these authors (Bhardwaj and Pal, 2011), they focus on using Bayesian classification algorithm to predict students' performance in BCA dept of Indian Universities. Variables considered are Sex, Category, medium of teaching, student food habit, other habit, living condition, accommodation, family size, family status, family annual income, grade in senior secondary school, students' college type, father's qualification, mother's qualification, father's occupation, mother's occupation and grade obtained in BCA. Naïve Bayes classification algorithm was used as a technique to design the student performance prediction model. It is found that grade in senior secondary school, living condition, medium of teaching, mother's qualification, student other habit, family income and family status were high potential variable for student performance. The investigation shows that other factors outside students' effort have significant influence over students' performance.

Yadav et. al, 2012 focus on generating predictive models for student retention management using decision tree algorithms (ID3, C4.5 and ADT) in WEKA. Study shows that intervention programs can have significant effects on retention, especially for the first year. Machine learning algorithms were applied to analyze and extract information from existing student data to establish predictive models. The predictive models are then used to identify among new incoming first year students those who are most likely to benefit from the support of the student retention program. The empirical results show that short but accurate prediction list for the student retention purpose can be produced by applying the predictive models to the records of incoming new students. The study identifies students which needed special attention to reduce drop-out rate.

Other studies, outside those reviewed here, tried to identify the significant factors that influence students' academic performance in more detailed way revealing wide range of potential predictors, including personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data, as contributors. Some of these factors seemed to be stronger than others; but there is no consistent agreement among different studies. However, all studies show that academic success is dependent on one factor or the other. Grades and achievements, personality and expectations, as well as sociological background all play significant roles in determining the students' academic performance. In summary, the studies show that various predictors at various time and different location contribute to the outcome of students, and that various techniques have been employed to determine these predictors. This study focus however is on identifying optimal machine learning algorithm suitable for predicting first-year tertiary students academic performance based on their family background factors and previous academic achievement.

DESIGN OF EXPERIMENT

This work focus on comparing the performance of machine learning algorithms on data relating to students family background factors and previous academic achievement with the aim of identifying the optimal model for predicting students performance. One thousand five hundred (1,500) records of students admitted between 2001 and 2010 into Computer Science programme of Babcock University, Nigeria were used. The students' first year academic performance is measured by Cumulative Grade Point Average (CGPA) at the end of the first session and the previous academic achievement is measured by Senior Secondary Certificate Examination (SSCE) grade score and University Matriculation Examination (UME) score. In the design of experiment, data relating to students' academic performance was collected from the students' record; data relating to students' family background and previous academic achievement was extracted from the enrolment form in the students' files; and data repositories that interface with WEKA computing environment was created. 66% of the data was used to train the models, while the remaining was used to test. WEKA computing tools was used to generate 10 classifiers and multilayer perceptron (artificial neural networks) machine learning algorithms. The machine learning algorithms generated from the students' data was compared using 10-fold cross-validation and hold-out methods. Accuracy level and confusion matrices benchmarks are used to determine the optimal predictive model. The methodology is detailed as follows:

a. Data Collection and Preparation

The dataset for the purpose of this study comprise of one thousand, five hundred records of students admitted into computer science programme, Babcock University, Nigeria obtained from the Students Record Systems. The real-world dataset from the Students Record did not contain sufficient students' family background information; therefore some background information was extracted from the enrolment forms that are given to students to fill as part of entrance registration requirements. Other variables that was extracted from these forms include SSCE grade in English, Mathematics, Physics, Chemistry, Biology and one other relevant subject, UME score, mother's educational qualification, father's educational qualification, sponsor, family size, student's position in the family, mother's occupation, father's occupation, marital status of parents, and average family income.

After the data collection, incomplete data was eliminated and the data was cleaned by smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. The SSCE grade was ranked to generate total SSCE score for each student; also the first year cumulative grade point average was grouped into different classes for easy identification. All other variables were grouped appropriately as shown in Table 1. Data repository that interfaces with WEKA was created for the data collected.

b. Variables Selection and Transformation

Information for the variables selected was extracted from the data repository that was created for the purpose of this study. Predictor and response variables derived from the data repository are given in Table 1.

Table 1: Data Format

S/N	Variable Name	Variable format	Variable Type
1.	Gender	Male, Female	Categorical
2.	Average Family Income		Continuous
3.	Mother’s educational qualification	No formal education, Primary, SSCE, 1st degree, 2nd degree, PhD	Categorical
4.	Father’s educational qualification	No formal education, Primary, SSCE, 1st degree, 2nd degree, PhD	Categorical
5.	Marital status of parents	Married, Divorced, Separated, Widowed	Categorical
6.	Mother’s occupation	Unemployed, Government worker, Private, Self employed	Categorical
7.	Father’s occupation	Unemployed, Government worker, Private, Self employed	Categorical
8.	Family size		Continuous
9.	Ethnicity	Yoruba, Hausa, Igbo, Others	Categorical
10.	Religion	Christianity, Islam, Traditional, Others	Categorical
11.	Sponsor	Parents, Scholarship, Self, Others	Categorical
12.	SSCE Grade Score	A1-8, B2-7, B3-6,C4-5,C5-4,C6-3,D7-2,D8-1,F9-0	Continuous
13.	UME Score		Continuous
14.	Age on entry		Continuous
15.	Current CGPA	A: 4.5-5.0, B+:4.0-4.49, B: 3.5-3.99, C+: 3.0-3.49, C: 2.5-2.99, D:2.0-2.49, E: 1.0-1.99, F:<1.0	Categorical

c. Model Building

Waikato Environment for Knowledge Analysis (WEKA) was used to build software tool for all experiments. WEKA is a collection of machine learning algorithms tools for data pre-processing, classification, regression, clustering, association rules and visualization. There are many machine learning algorithms implemented in WEKA including Bayesian classifiers, Decision Trees, Rules, Functions, Lazy classifiers and miscellaneous classifiers.. WEKA was used to generate 10 classification models(five decision tree algorithms -Random forest, Random tree, J48, Decision stump and REPTree and five rule induction algorithms –JRip, OneR, ZeroR, PART, and Decision table) and a multilayer perceptron, an artificial neural network function. These algorithms were compared using 10-fold cross validation and hold-out method considering accuracy level, confusion matrices and CPU time to determine the optimal model. The ten classification algorithms have been selected because they are considered as “white box” classification model, that is, they provide explanation for the classification and can be used directly for decision making. Each classifier belongs to a different family of classifiers implemented in Weka: Random forest, Random tree, J48, Decision stump and REPTree related to Decision Trees, JRip, OneR, ZeroR, PART, and Decision table belong to Rules, and multilayer perceptron belong to neural networks functions. Attribute importance analysis was carried out to rank the attributes by significance using Information gain and gain ratio attribute evaluators. Ranker’s Search method was used to achieve this. The models built from the supervised algorithms (decision trees and neural networks) were trained with 66% of the data (hold-out method) and 10-fold cross-validation was used to compute confusion matrices and accuracy level to compare the models.

EXPERIMENTAL RESULTS

This section presents the experimental result generated from the study. The attributes relating to students’ family background factors and previous academic achievement were considered. Figure 1 presents the visualization of all the attributes used in this study. The attributes were ranked in order of importance using information gain and gain ratio measures. The outcome is presented in table 2 and figure 2. The ranking of both attribute evaluators was done using ranker search method. Among the fourteen attributes used in this study, it was discovered that students JAMB Score, Age on entry, Father's occupation, Mother's occupation are the best five attributes. The outcome of both evaluators is similar as shown in figure 2. The accuracy level and CPU time taken to build the ten classification models and multilayer perceptron, an artificial neural network function using WEKA intelligent tool are presented in tables 2, table 3, figure 3 and figure 4.

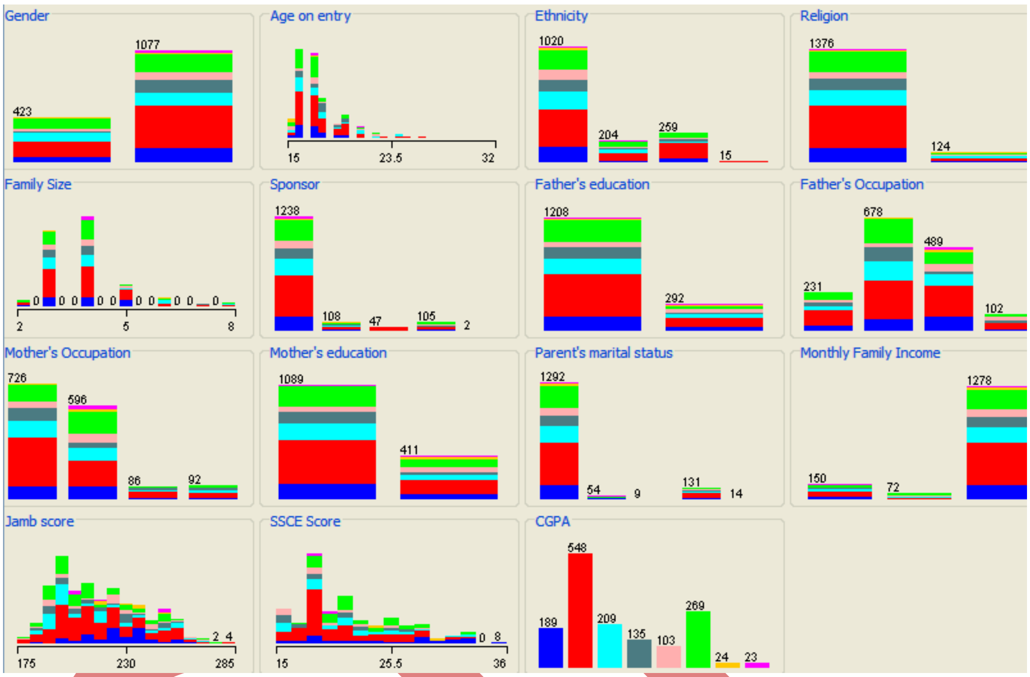

Figure 1: Attributes visualization

Table 2 Attributes Ranking using information gain and gain ratio

S/N	Attribute	Information Gain		Gain Ratio	
		Value	Rank	Value	Rank
1	Gender	0.0389	10	0.0453	11
2	Age on entry	0.1689	5	0.0951	5
3	Ethnicity	0.0609	8	0.0478	10
4	Religion	0.0277	14	0.0673	9
5	Family Size	0.1465	6	0.0745	7
6	Sponsor	0.064	7	0.0681	8
7	Father's education	0.0313	12	0.044	12
8	Mother's education	0.0359	11	0.0424	13
9	Father's Occupation	0.4343	2	0.1088	2
10	Mother's Occupation	0.374	3	0.1063	4
11	Parent's marital status	0.0588	9	0.0761	6
12	Monthly Family Income	0.0293	13	0.0397	14
13	Jamb score	0.8013	1	0.1658	1
14	SSCE Score	0.2164	4	0.1072	3

Attribute ranking (with respect to the class attribute) according to information gain and gain ratio criteria show that students JAMB Score, Age on entry, Father's occupation, Mother's occupation are the best attributes.

These attributes outperform other attributes in their contribution to the outcome of students' first year performance in tertiary institution as shown in figure 2.

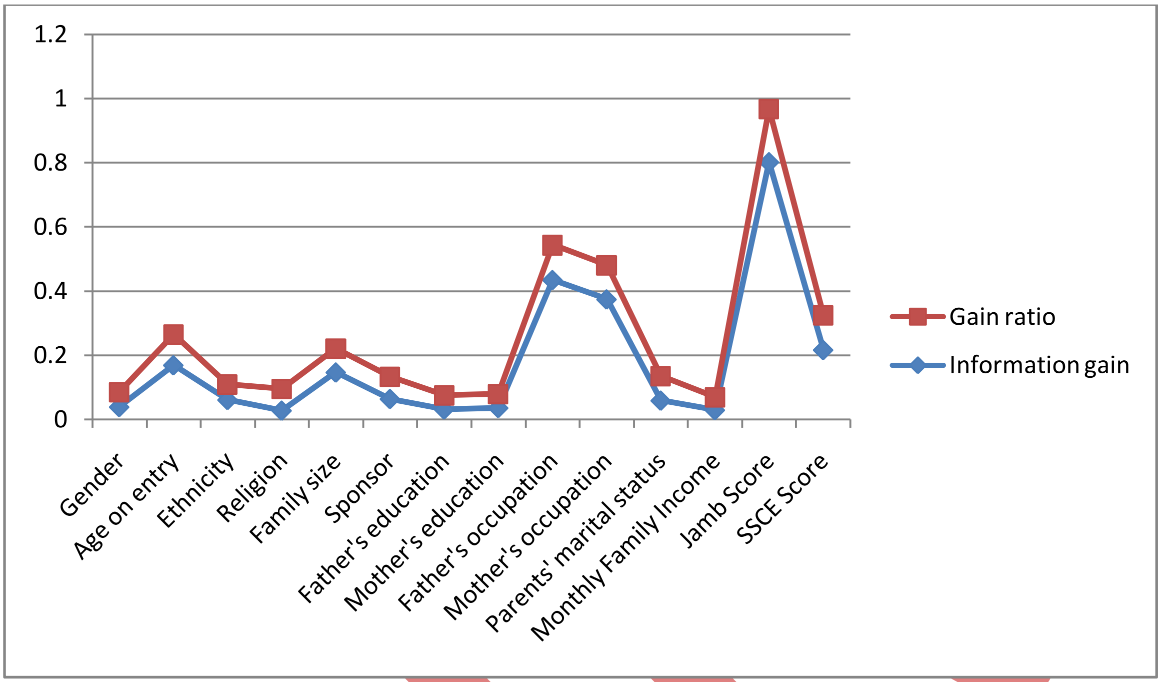

Figure 2: Information gain and gain ratio of the attributes

Table 3: Classification Accuracy on 10-fold crossvalidation and hold out methods

	Random forest	**Random tree**	**Reptree**	**C4.5(J48)**	**Decision stump**	**JRip**	**OneR**	**PART**	**Decision table**	**ZeroR**	**MLP**
10-fold	**95.87**	**96.07**	**74.87**	**80.6**	**39.2**	**88.87**	**45.33**	**80.4**	**63.6**	**36.53**	**87.13**
Hold out	**85.29**	**85.69**	**69.22**	**77.03**	**34.90**	**72.94**	**44.9**	**76.67**	**63.92**	**34.9**	**77.25**

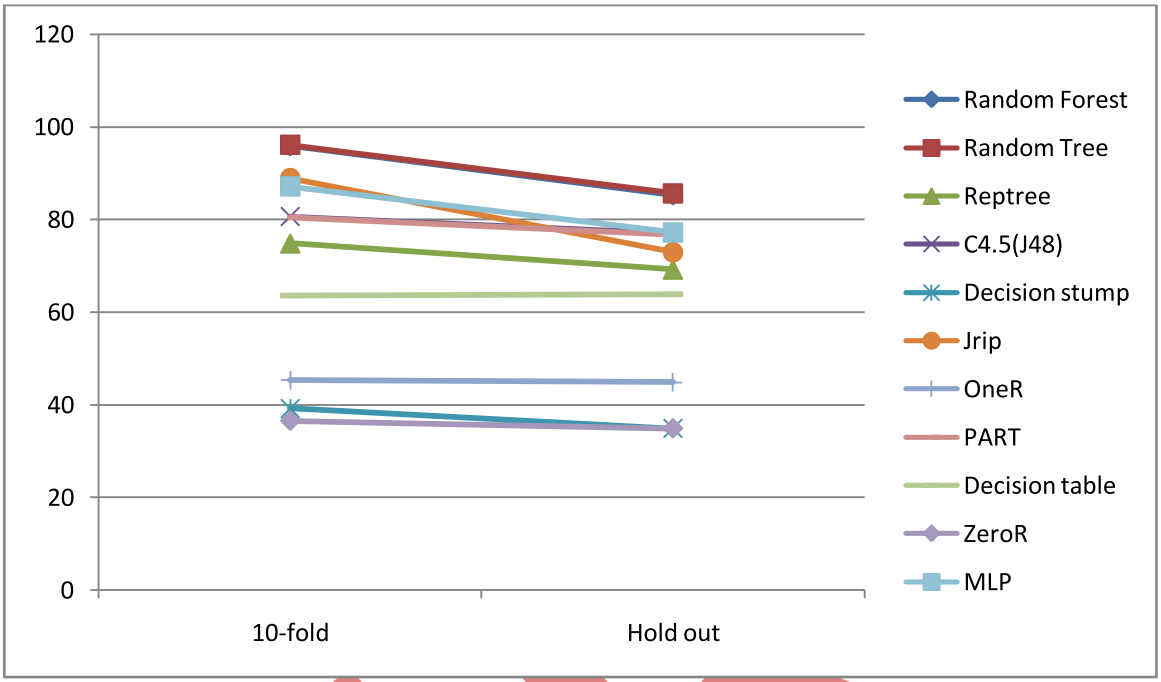

Figure 3: Prediction accuracy for classifiers for 10-fold cross validation and holdout

The outcome of both 10-fold crossvalidation and hold-out method is similar for all the classifiers. Random tree outperforms all other classifiers on both counts. Random forest, Reptree, J48, JRIp, PART, Decision table and multilayer perceptron perform well with the lowest accuracy for both hold-out and 10-fold crossvalidation being 63.6%. Decision stump, OneR and ZeroR slightly fall behind in accuracy. But overall random tree gives accuracy of 96.07% for 10-fold cross validation and 85.69% for holdout method which outperform all other classifiers used in this study.

Table 4: Time taken (seconds) to build the algorithms

	Random forest	Random tree	Reptree	C4.5(J48)	Decision stump	JRip	OneR	PART	Decision table	ZeroR	MLP
10-fold	**0.89**	**0.05**	**0.56**	**0.27**	**0.05**	**4.42**	**0.03**	**1.09**	**1.25**	**0.01**	**337.08**
Hold out	**0.22**	**0.01**	**0.08**	**0.11**	**0.02**	**2.64**	**0.03**	**1.87**	**1.51**	**0.01**	**279.5**

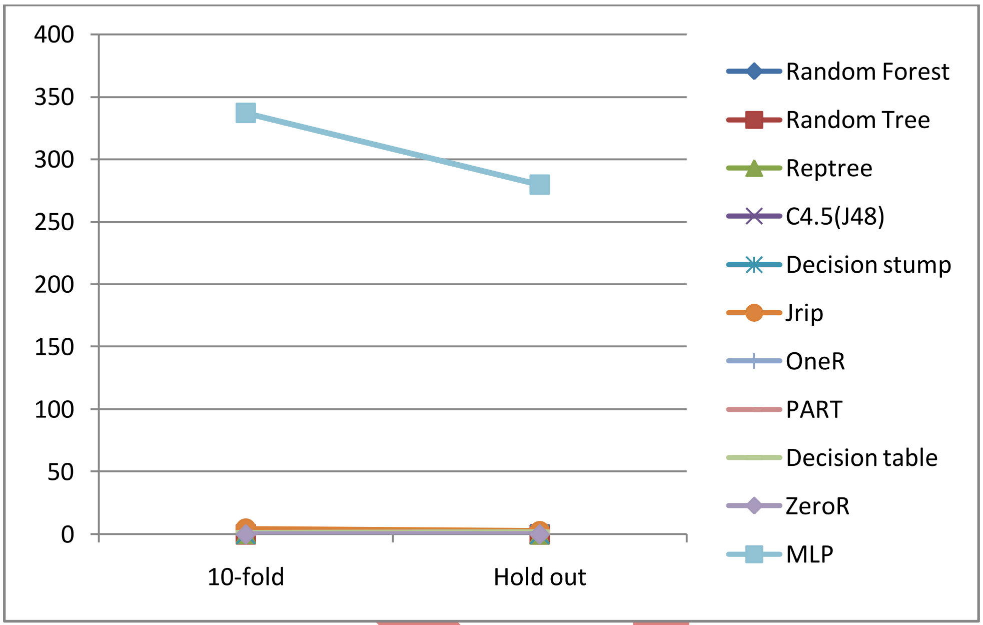

Figure 4: Time taken (seconds) to build the classifier algorithms

The disparity between time taken to build multilayer perceptron and other classification algorithms is very wide as shown in figure 4. Multilayer perceptron consumes much computer resources. Other classifiers took considerable time to execute and consume less system resources. Considering the time taken for building the models in relation to the accuracy level and the performance of the model, it can be established that random tree takes very short time and outperform all other classifiers in this study. Therefore, it can be deduced that random tree according to the outcome of this study is a very good classifier for predicting student first year academic performance in relation to other algorithms used in this study.

Table 5: Detailed Accuracy of Classifiers using 10-fold crossvalidation

Classifiers	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC	RFC
Random Forest	0.959	0.013	0.959	0.959	0.958	0.947	0.993	0.985
Random tree	0.961	0.013	0.961	0.961	0.961	0.949	0.979	0.943
Reptree	0.749	0.086	0.751	0.749	0.746	0.674	0.936	0.818
C4.5	0.806	0.056	0.807	0.806	0.806	0.753	0.965	0.897
Decision stump	0.392	0.332	0.2	0.392	0.247	0.09	0.535	0.236
JRIp	0.789	0.087	0.8	0.789	0.789	0.722	0.911	0.775
OneR	0.453	0.2	0.439	0.453	0.432	0.272	0.627	0.303
PART	0.804	0.055	0.803	0.804	0.803	0.752	0.969	0.899
Decision table	0.636	0.071	0.818	0.636	0.678	0.626	0.915	0.772
ZeroR	0.365	0.365	0.133	0.365	0.196	0	0.495	0.213

TP Rate- True Positive Rate (proportion of cases correctly classified to the actual class)
FP Rate- False Positive Rate (proportion of cases belonging to another class misclassified as other class)
Precision- (Positive predictive value determined by {TP/[TP+FP]})
Recall- same as TP Rate ROC- Receiver Operating Characteristic (graphical display of TPR vs FPR)

Table 6: Detailed Accuracy of Classifiers using hold-out method

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC	RFC
Random Forest	0.853	0.046	0.857	0.853	0.853	0.811	0.96	0.919
Random tree	0.857	0.043	0.86	0.857	0.857	0.817	0.91	0.775
Reptree	0.692	0.086	0.696	0.692	0.689	0.611	0.884	0.69
C4.5	0.771	0.07	0.773	0.771	0.769	0.709	0.904	0.738
Decision stump	0.349	0.349	0.122	0.349	0.181	0	0.57	0.228
JRIp	0.729	0.098	0.742	0.729	0.729	0.651	0.859	0.646
OneR	0.449	0.187	0.428	0.449	0.424	0.275	0.631	0.295
PART	0.767	0.067	0.768	0.767	0.766	0.707	0.903	0.75
Decision table	0.639	0.098	0.774	0.639	0.665	0.596	0.883	0.716
ZeroR	0.349	0.349	0.122	0.349	0.181	0	0.5	0.205

TP Rate- True Positive Rate (proportion of cases correctly classified to the actual class)
FP Rate- False Positive Rate (proportion of cases belonging to another class misclassified as other class)
Precision- (Positive predictive value determined by {TP/[TP+FP]})
Recall- same as TP Rate **ROC**- Receiver Operating Characteristic (graphical display of TPR vs FPR)

Table 6 and 7 show the detailed accuracy level achieved by the ten classification algorithms. This is to further reveal the performance of each algorithm based on the true positive rate (TP rate), false positive rate (FP rate), precision, recall and other measures. The True Positive (TP) rate is the proportion of cases which were classified as the actual class, that is, how much part of the class was captured. It is equivalent to Recall. The False Positive (FP) rate is the proportion of cases which were classified as the one class, but belong to a different class. The Precision is the proportion of the cases which truly have the actual class among all those which were classified as the class. The F-Measure is simply 2*Precision*Recall/(Precision+Recall), a combined measure for precision and recall. The Receiver Operating Characteristic (ROC) is the graphical display of TPR versus FPR. These measures are useful for comparing classifiers based on the accuracy. As previously established using prediction accuracy, random tree outperform all other algorithms used in this work. The confusion matrices showing the numbers misclassified and the correctly classified for all algorithms based on classes using both 10-fold cross validation and hold-out method are presented in table 7 and 8 in the Appendix.

CONCLUSION

This work explores the efficiency of several machine learning algorithms in determining the influence of family background factors and previous academic achievement on first year tertiary student academic performance in order to identify the optimal model. It is discovered that random tree performance is better than that of other algorithms used in this study. Although, the application of machine learning algorithms to education datasets is not entirely new, this work has been able to identify random tree as a good classifier in predicting first-year tertiary students academic performance considering the family background factors and previous academic achievement. The outcome agrees with Al-Radaideh et al, 2006, Nguyen et al, 2007, Quadri and Kalyankar, 2010 ,Yadav et. al, 2012 whose outcomes reveal that classes of decision trees are the best algorithms for predicting students academic performance. This work will be further improved by designing a predictive/recommender system based on the findings of this work.

REFERENCES

[1] Affendey, L.S., I.H.M. Paris, N. Mustapha, M.N. Sulaiman and Z. Muda, 2010. Ranking of influencing factors in predicting students academic performance. Inform. Technol. J., 9: 832-837.

[2] Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). Mining student data using decision trees. In the Proceedings of the 2006 International Arab Conference on Information Technology (ACIT'2006).

[3] Bhardwaj B. K. and Pal S. Data Mining: A prediction for performance improvement using classification, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011

[4] Bhardwaj B. K. and Pal S. Mining Educational Data to Analyze Students‟ Performance (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

[5] Delavari N, Beikzadeh M. R, Amnuaisuk S. 2005. Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System. 6th Annual International Conference: ITEHT Juan Dolio, Dominican Republic.

[6] Delavari N, Beikzadeh M. R. 2004 A New Model for Using Data Mining in Higher Educational System, 5th International Conference on Information Technology based Higher Education and Training: ITEHT '04, Istanbul, Turkey.

[7] Golding, P. and O. Donaldson, 2006. Predicting academic performance. Proceedings of the 36th

ASEE/IEEE Frontiers in Education Conference T1D-21, San Diego, CA.,1-6.

[8] Hämäläinen, W. and M. Vinni, 2006. Comparison of machine learning methods for intelligent tutoring systems. Proceedings of the 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan, 525-534.

[9] Hegazy, T. and Moselhi, O. (1994). Analogy Based Solution To Markup Estimation Problem. Journal of Computing in Civil Engineering, vol 8(1), pp72-87.

[10] Herrera, O. L. (2006). Investigation of the role of pre- and post-admission variables in undergraduate institutional persistence, using a Markov student flow model. PhD Dissertation, North Carolina State University, USA.

[11] Hijazi S. T., and Naqvi R. S. M. M.(2006), —Factors affecting student's performance: A Case of Private College, Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[12] Ikmal Hisyam Mohamad Paris, Lilly Suriani Affendey, and Norwati Mustapha Improving Academic Performnce Prediction using Voting Technique in Data Mining World Academy of Science, Engineering and Technology 62 2010 820-823

[13] Kalles D., Pierrakeas C. 2004, Analyzing student performance in distance learning with genetic algorithms and decision trees, Hellenic Open University, Patras, Greece.

[14] Kember, D. (1995). Open learning courses for adults: A model of student progress. Englewood Cliffs, NJ: Education Technology.

[15] Kuyoro S. O. 2010, Investigating the Effect of Students Socio-Economic/Family Background on Students Academic Performance in Tertiary Institutions Using Decision Tree Algorithms. Department of Computer Science, University of Ibadan. Unpublished MSc Thesis

[16] Marquez-Vera C., Romero C. and Ventura S. Predicting School Failure Using Data Mining. Proceedings of the 4th International Conference on Educational Data Mining, pp 271-276, 6 Jul 2011

[17] Michie D., Spiegelhalter D.J., Taylor C.C. 1994 Machine Learning, Neural and Statistical Classification. Prentice Hall Inc. Available at: http://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf Retrieved 17-09-2012

[18] Nandeshwar, A., & Chaudhari, S. (2009). Enrollment prediction models using data mining. Retrieved January 10, 2010, from http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf

[19] Narendra, K.S. and Parthasarathy, K. (1987). Identification and control of dynamical systems using neural networks. IEEE Transactions on Neural Networks, vol. 1. pp4-27.

[20] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy 2007 A comparative analysis of techniques for predicting academic performance 37th ASEE/IEEE Frontiers in Education Conference, Milwaukee, WI T2G-7 to T2G-12

[21] Quadri M. N. and Kalyankar N.V. 2010 Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques, Global Journal of Computer Science and Technology Vol. 10 Issue 2, pp2-5

[22] Quinlan J.R. 1996. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 4:77-90.

[23] Schaffer, C. 1994. A conservation law for generalization performance. In Proceedings of the Eleventh International Conference on Machine Learning, Pages 153_178, New Brunswick, USA, July 10-13.

[24] Schank, R. (1982). Dynamic Memory: A theory of reminding and learning in computers and people. Cambridge University Press.

[25] Sejnowski, T.J. and Rosenberg, C.R. (1987). Parallel networks that learn to pronounce English text. Complex Systems, vol 1pp145-168.

[26] Sembiring S., Zarlis M., Hartama D., Ramliana S, and Elvi W. Prediction of student academic performance by an application of data mining techniques in the proceeding 2011 International Conference on Management and Artificial Intelligence IPEDR vol.6 (2011) © (2011) IACSIT Press, Bali, Indonesia

[27] Superby, J.F., J.P. Vandamme and N. Meskens, 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. Proceedings of the 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop, (ITS'06), Jhongali, Taiwan, 37-44.

[28] Vandamme, J.P., N. Meskens and J.F. Superby, 2007. Predicting academic performance by data mining methods. Educ. Econ., 15: 405-419.

[29] Vialardi C., Bravo J., Shafti L. and Ortigosa A. (2009). Recommendation in Higher Education Using Data Mining Techniques, Educational Data Mining, pp190-199, 2009

[30] Yadav S. K., Bhardwaj B. K. and Pal S. (2012) Mining Education Data to Predict Student's Retention: A comparative Study, International Journal of Computer Science and Information Security (IJCSIS), Vol. 10, No. 2, 113-117

[31] Yu C. H., DiGangi S., Jannasch-Pennell A. and Kaprolet C. A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. Journal of Data Science 8(2010), 307-325

APPENDIX

Table 7: Confusion matrix of classifiers using 10-fold cross validation

Random Forest	**a**	**b**	**c**	**D**	**e**	**f**	**g**	**H**	**Rando m Tree**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**181**	5	0	2	1	0	0	0	**a**	**181**	4	0	2	0	0	0	2
b	2	**538**	0	2	0	6	0	0	**b**	0	**536**	2	0	2	8	0	0
c	2	3	**199**	1	0	4	0	0	**c**	0	4	**202**	0	2	1	0	0
d	0	1	1	**133**	0	0	0	0	**d**	1	1	0	**132**	1	0	0	0
e	0	3	2	2	**96**	0	0	0	**e**	0	6	0	0	**97**	0	0	0
f	2	12	6	0	5	**244**	0	0	**f**	1	12	4	0	**6**	**246**	0	0
g	0	0	0	0	0	0	**24**	0	**g**	0	0	0	0	0	0	**24**	0
h	0	0	0	0	0	0	0	**23**	**h**	0	0	0	2	0	0	0	**23**
RepTree	**a**	**b**	**c**	**D**	**e**	**f**	**g**	**h**	**C4.5**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**139**	23	2	5	5	15	0	0	**a**	**149**	21	4	3	3	9	0	0
b	10	**466**	17	11	12	27	3	2	**b**	10	**470**	22	6	17	20	3	0
c	11	34	**147**	5	2	10	0	0	**c**	1	18	**165**	7	7	10	1	0
d	9	19	2	**99**	5	1	0	0	**d**	8	10	6	**103**	2	6	0	0
e	2	31	1	3	61	3	**0**	2	**e**	1	16	2	3	**75**	**5**	1	0
f	15	63	9	3	7	172	0	0	**f**	2	29	14	8	**8**	**207**	0	1
g	0	1	1	0	0	2	**20**	0	**g**	1	2	1	0	1	1	**18**	0
h	0	1	0	2	0	1	0	19	**h**	1	0	0	0	0	0	0	**22**
Decision stump	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**JRIp**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**56**	133	0	0	0	0	0	0	**a**	**144**	38	0	0	2	5	0	0
b	16	**532**	0	0	0	0	0	0	**b**	16	**478**	11	7	5	31	0	0
c	15	194	0	0	0	0	0	0	**c**	3	32	**161**	4	1	8	0	0
d	3	132	0	0	0	0	0	0	**d**	0	29	1	**101**	2	2	0	0
e	0	103	0	0	0	0	0	0	**e**	2	35	1	2	**63**	0	0	0
f	17	252	0	0	0	0	0	0	**f**	11	52	5	1	3	**197**	0	0
g	11	13	0	0	0	0	0	0	**g**	0	2	1	0	0	1	**20**	0
h	0	23	0	0	0	0	0	0	**h**	1	2	0	0	0	1	0	**19**
OneR	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**PART**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**60**	80	4	4	22	16	0	3	**a**	**147**	19	10	4	3	5	0	1
b	32	**391**	37	16	22	40	0	10	**b**	17	**476**	10	13	6	22	1	3
c	9	105	**76**	2	2	12	3	0	**c**	6	16	**162**	5	4	13	2	1
d	2	39	11	**36**	1	46	0	0	**d**	11	13	5	**98**	6	2	0	0
e	9	54	7	8	**22**	1	0	2	**e**	3	10	5	7	**73**	**4**	1	0
f	9	127	7	25	14	**85**	0	2	**f**	6	31	9	**6**	**5**	212	0	0
g	10	3	9	0	0	0	**0**	2	**g**	0	1	4	0	0	0	**19**	0
h	0	11	2	0	0	0	0	**10**	**h**	1	2	0	0	1	0	0	**19**
Decision table	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**ZeroR**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**173**	11	1	0	1	3	0	0	**a**	0	189	0	0	0	0	0	**0**
b	204	**333**	3	0	3	4	0	1	**b**	0	548	0	0	0	0	0	**0**
c	68	15	**123**	0	0	3	0	0	**c**	0	209	0	0	0	0	0	**0**
d	43	4	0	**85**	0	3	0	0	**d**	0	135	0	0	0	0	0	**0**
e	39	9	0	1	**54**	0	0	0	**e**	0	103	0	0	0	0	0	**0**
f	89	26	3	0	**21**	**49**	0	0	**f**	0	269	0	0	0	0	0	**0**
g	5	0	0	0	1	0	**18**	0	**g**	0	24	0	0	0	0	0	**0**
h	4	0	0	0	0	0	0	**19**	**h**	0	23	0	0	0	0	0	**0**

a = B+ b = C c = C+ d = D e = E f = B g = A h = F

Table 8: Confusion matrix of classifiers using hold-out method

Random Forest	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**Random Tree**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**61**	9	0	0	0	4	0	0	**a**	**59**	7	2	0	0	4	0	2
b	6	**156**	4	2	8	2	0	0	**b**	2	**158**	10	0	2	6	0	0
c	3	6	**57**	0	0	0	0	0	**c**	0	6	**57**	0	0	2	0	1
d	0	5	1	**43**	2	0	0	0	**d**	0	4	1	**43**	0	3	0	0
e	0	4	0	0	**30**	0	0	0	**e**	0	2	0	0	**32**	**0**	0	0
f	2	9	4	2	0	**71**	0	0	**f**	2	9	2	4	**0**	**71**	0	0
g	0	0	0	0	0	0	**7**	0	**g**	0	0	0	0	0	0	**7**	0
h	0	0	0	2	0	0	0	**12**	**h**	0	0	0	2	0	0	0	**10**
RepTree	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**C4.5**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**49**	11	2	7	0	5	0	0	**a**	**53**	13	4	3	0	1	0	0
b	11	**140**	7	3	8	7	0	2	**b**	9	**154**	6	2	2	5	0	0
c	1	12	**49**	2	1	1	0	0	**c**	2	5	**53**	3	0	2	0	1
d	1	7	5	**35**	0	3	0	0	**d**	3	7	2	**37**	0	1	0	1
e	5	4	2	3	**18**	**1**	0	1	**e**	0	5	0	1	**24**	4	0	0
f	7	20	9	1	**1**	**50**	0	0	**f**	3	15	6	3	4	**57**	0	0
g	0	0	4	0	0	0	**3**	0	**g**	0	0	1	0	0	0	**6**	0
h	0	0	0	2	0	1	0	**9**	**h**	1	2	0	0	0	0	0	**9**
Decision stump	**a**	**b**	**c**	**d**	**e**	**F**	**g**	**h**	**JRIp**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**0**	74	0	0	0	0	0	0	**a**	**49**	18	2	0	2	3	0	0
b	0	**178**	0	0	0	0	0	0	**b**	11	**147**	5	2	5	8	0	0
c	0	66	**0**	0	0	0	0	0	**c**	0	9	**52**	1	2	2	0	0
d	0	51	0	**0**	0	0	0	0	**d**	2	9	1	**35**	0	4	0	0
e	0	34	0	0	**0**	0	0	0	**e**	0	10	1	1	**19**	3	0	0
f	0	88	0	0	0	**0**	0	0	**f**	2	25	3	0	3	**55**	0	0
g	0	7	0	0	0	0	**0**	0	**g**	0	1	0	0	0	0	**6**	0
h	0	12	0	0	0	0	0	**0**	**h**	0	3	0	0	0	0	0	**9**
OneR	**a**	**b**	**c**	**d**	**e**	**F**	**g**	**h**	**PART**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**24**	34	4	0	8	4	0	0	**a**	**53**	7	1	5	2	6	0	0
b	16	**125**	22	4	3	8	0	0	**b**	5	**155**	7	0	5	6	0	0
c	7	26	**30**	0	0	3	0	0	**c**	0	6	**49**	1	1	8	1	0
d	1	15	4	**13**	0	18	0	0	**d**	3	8	0	**35**	2	3	0	0
e	3	13	6	2	**8**	2	0	0	**e**	0	7	0	2	**21**	4	0	0
f	1	41	7	6	4	**29**	0	0	**f**	4	12	6	1	3	**62**	0	0
g	4	1	2	0	0	0	**0**	0	**g**	0	0	1	0	0	0	**6**	0
h	0	5	2	0	0	5	0	**0**	**h**	0	2	0	0	0	0	0	**10**
Decision table	**a**	**b**	**c**	**d**	**e**	**F**	**g**	**h**	**ZeroR**	**a**	**b**	**c**	**d**	**E**	**f**	**g**	**h**
a	**61**	13	0	0	0	0	0	0	**a**	**0**	74	0	0	0	0	0	0
b	57	**118**	0	0	1	2	0	0	**b**	0	**178**	0	0	0	0	0	0
c	16	10	**40**	0	0	0	0	0	**c**	0	66	**0**	0	0	0	0	0
d	10	6	0	**32**	2	1	0	0	**d**	0	51	0	**0**	0	0	0	0
e	13	4	0	0	**17**	0	0	0	**e**	0	34	0	0	**0**	0	0	0
f	28	17	0	0	**0**	**43**	0	0	**f**	0	88	0	0	0	**0**	0	0
g	1	0	0	0	0	0	**6**	0	**g**	0	7	0	0	0	0	**0**	0
h	1	2	0	0	0	0	0	**9**	**h**	0	12	0	0	0	0	0	**0**

a = B+ b = C c = C+ d = D e = E f = B g = A h = F