A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining

Abdul Hamid M. Ragab
Faculty of Computing and Information Technology,
King Abdulaziz University
Jeddah, Saudi Arabia
aragab@kau.edu.sa

Amin Y. Noaman
Faculty of Computing and Information Technology,
King Abdulaziz University
Jeddah, Saudi Arabia
anoaman@kau.edu.sa

Abdullah S. AL-Ghamdi
Faculty of Computing and Information Technology,
King Abdulaziz University
Jeddah, Saudi Arabia
aalmalaise@kau.edu.sa

Ayman I. Madbouly
Research and Consultancy Dept., Deanship of
Admission and Registration,
King Abdulaziz University
amadbouly@kau.edu.sa

ABSTRACT

In big data universities, there may be several problems related to college admission and enrollment due to the increasing volume of students' data applying for higher education. So that there is a need to apply efficient data mining algorithms for better decision making for students' data classification. In this paper, nine classification algorithms are comparatively tested to find the optimum algorithm for students' dataset classification. The KAU-ODUS+ data base for the preparatory year students are used as an approval dataset for the experimental purposes. The Weka-knowledge analysis tool which is open source data mining workbench software is used for simulation of practical measurements. The classification technique that has the potential to significantly improve the performance is suggested for use in colleges' admission and enrollment applications. Impact of students GPA and their grades qualified materials with respect to their colleges' admission desire are visually analyzed. Results show that C4.5, PART and Random Forest algorithms give the highest performance and accuracy with lowest errors while IBK-E and IBK-M algorithms give high errors and low accuracy.

Keywords

College admission and enrollment, Classification, Data Mining, Machine Learning, WEKA.

1. INTRODUCTION

University colleges require rules for making decision to classify students to be enrolled to suitable colleges. These include students desires, their GPA, and colleges criteria requirements. The objective of this paper is to investigate the performance of different classification algorithms for colleges' admission. A major problem in colleges' admission analysis is to build an ultimate model that yields fruitful results on certain given information. Therefore, different classification models must be evaluated to attain the ultimate model. This kind of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial intelligent algorithms which perform the role as classifiers. Classification algorithms always find a set of rules to represent data into classes.

There are several data mining (DM) tools available [1, 2]. In this paper we use Weka DM workbench, which is an open source tool developed using JAVA [3]. It contains tools for data preprocessing, classification, regression, clustering, association rule and visualization. It not only supports DM algorithms, but also data preparation and Meta learners like bagging and boosting. Also, for the applicability issue, Weka toolkit has achieved the highest applicability followed by Orange, Tanagra, and KNIME, respectively [4]. While using Weka for classification, performance can be better improved by applying cross validation test mode instead of using percentage split test mode [4]. The popular and well accepted algorithms supported using Weak, for classification task are decision trees such as ID3, C4.5, and PART [5, 6]. Another algorithm which is based on probability theory is Naïve Bayes algorithms [7, 8]. Other two algorithms that based on Artificial Neural Networks (ANN) algorithms are MPL and Multilayer Perceptron [9, 10], and Support Vector Machine (SVM) [11]. In this paper, nine classification algorithms are investigated for their performance, as explained in section 1. In sections 2 and 3 we discuss some of the related work. In sections 4 and 5 we investigate types of classification algorithms including their performance measures. Results and comparative analysis, and conclusions are explained in sections 6 and 7 respectively.

2. RELATED WORK

DM is the process of extracting knowledge form huge dataset. Their techniques apply advanced computation methods to discover unknown relations and sum up results by analyzing the observed dataset to make these relations clear and understandable. Hu and et.al [12] conducted experimental comparison of LibSVMs, C4.5, Bagging C4.5, AdaBoosting C4.5, and Random Forest on seven Microarray cancer data sets. They concluded that C4.5 was better among all algorithms and also found that data preprocessing and cleaning improves the efficiency of classification algorithms. Shin in [13] conducted comparison between C4.5 and Naïve Bayes and concluded that C4.5 is out performing algorithm then Naïve Bayes. Sharma and Sahni [14] conducted experiment in the Weka environment by using four algorithms namely ID3, J48, Simple CART and Alternating Decision Tree (ADTree). They compared these four algorithms for spam email dataset in terms of classification accuracy. According to their simulation results, the J48 classifier outperforms the ID3, CART and ADTree in terms of classification accuracy [15].

Table-1 shows a summary for some recent work related to classification algorithms performance and type of the applications area for the experimental datasets used. It illustrates several DM algorithms that can be applied into different application area compared with the algorithms investigated in this paper.

Table-1 a summary of related DM-Algorithms and the application datasets used.

Year	Authors	DM-Algorithms	Application Area Datasets
2011	Aman Kumar Sharma [16]	ID3, J48, Simple CART and Alternating Decision Tree (ADTree)	Spam Email Data
2012	Rohit Arora, Suman [17]	C4.5, MLP	Diabetes and Glass
2012	Abdullah H. Wahbeh, Mohammed Al-Kabi [18]	C4.5,SVM, Naïve Bayes	Arabic Text
2013	Murat Koklu and Yavuz Unal [19]	MLP, J48, and Naïve Bayes	Diabetic Patients
2013	S. Vijayarani, M. Muthulakshmi [20]	Attribute Selected Classifier, Filtered Classifier, LogitBoost	Classifying computer files
2014	Devendra Kumar Tiwary [21]	Decision Tree (DT), Naïve Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machine (SVM).	Credit Card
This paper	Abdul Hamid M. Ragab, and et.al.	C4.5, Random Forest, IBK-E, IBK-M, LIBSVM, MLP, Multilayer Perceptron, Naïve Bayes, PART	Students Colleges Enrollments

3. EXPERIMENTAL DESIGN METHODOLOGY

Fig.1 shows a block diagram for the components of the practical experiment used. Nine classification algorithms are used for measuring their performance. We used Weka DM workbench in the investigation. The training and testing dataset used in the experiment was extracted from the KAU-ODUS+ database. It contains information related to the preparatory year students' enrollment process.

3.1 Classification

Classification is one of the DM techniques that is mainly used to analyze a given dataset and takes each instance of it and assigns this instance to a particular class with the aim of achieving least classification error. It is used to extract models that correctly define important data classes within the given dataset. It is a two-step process. In first step the model is created by applying classification algorithm on training data set. Then in second step, the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy. So, classification is the process to assign class label for this dataset whose class label is unknown. Versatile list of techniques are available for classification like decision tree induction, Bayesian classification, and Bayesian network. Table-2 defines the nine classifications algorithms to be investigated in this paper.

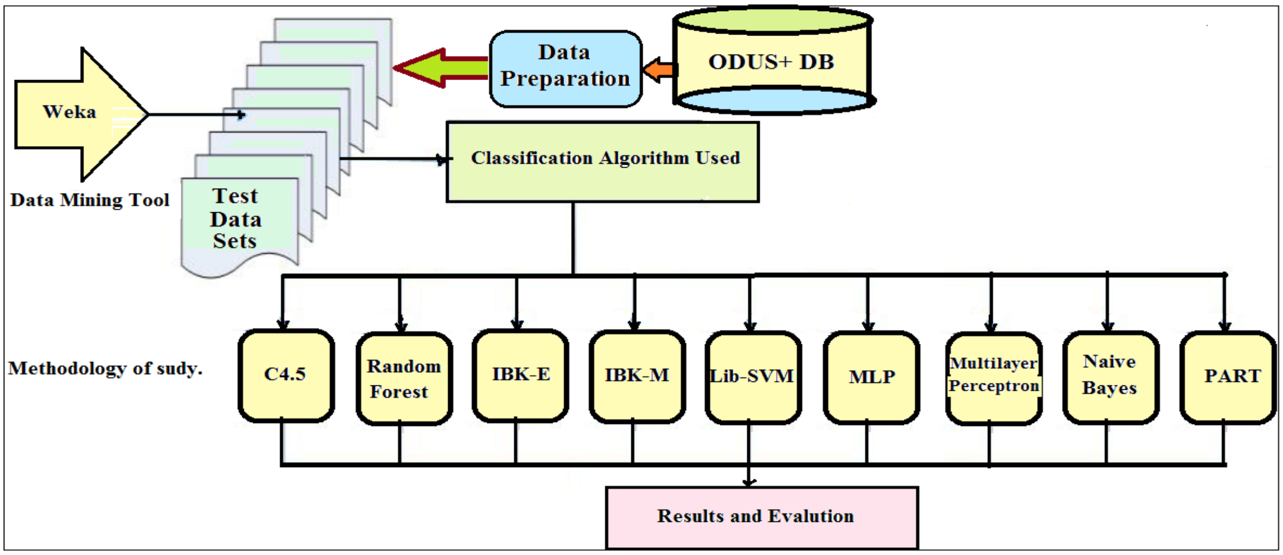


Fig.1 block diagram of experiment components used for measuring classifiers performance.

Table-2 the nine Weka classification algorithms to be experimentally investigated.

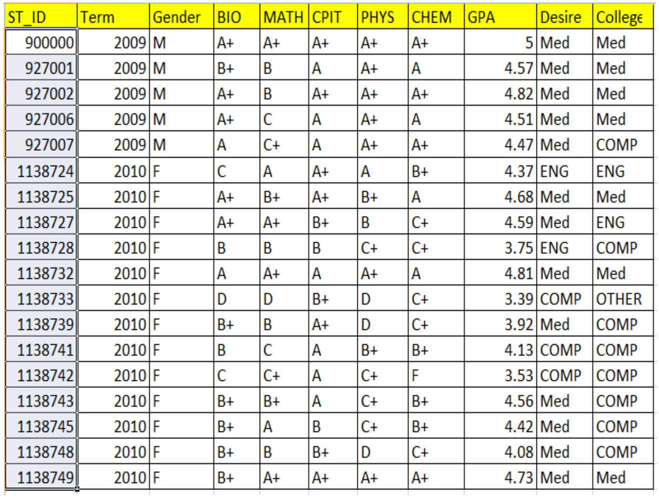
Classifier	Description
C4.5	It is a decision tree algorithm developed by Ross Quinlan [5]. C4.5 is an extension of Quinlan's earlier ID3 algorithm. Unlike ID3, C4.5 can generate pruned and unpruned decision trees. We can use the decision trees generated by C4.5 for classification problems; C4.5 is often referred to as a statistical classifier [6].
Random Forest	The Random Forest algorithm combined multiple random trees that votes on a particular outcome. In the random forest algorithm each vote is given equal weight. The forest chooses the classification that contains the most votes [22, 23].
IBK, IBK-E, IBK-M	K-nearest neighbor’s classifier can select appropriate value of K based on cross-validation [7, 24]. Can also do distance weighting. IBK-E: K-nearest neighbor’s classifier- with Euclidean Distance Measure. IBK-M: K-nearest neighbor’s classifier- with Minkowski Distance Measure. Can select appropriate value of K based on cross-validation. Can also do distance weighting.
LIBSVM	A Library for Support Vector Machines. Support vector machines (SVMs) [11] are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.
MLP	MLP (Multilayer Perceptron) - is a feed forward neural network in which the neurons are organized in layers. There is no feedback of the output of a neuron as input to another neuron in which it depends [8].Trains a multilayer Perceptron with one hidden layer using WEKA's Optimization class by minimizing the squared error plus a quadratic penalty with the Broyden–Fletcher–Goldfarb–Shannon (BFGS) method [9]. In numerical optimization, the BFGS algorithm is an iterative method for solving unconstrained nonlinear optimization problems.
Multilayer Perceptron	The Multilayer Perceptron classifier uses back-propagation to classify instances. The network can be built by hand, created by an algorithm or both [10]. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid, except when the class is numeric in which case the output nodes become un-threshold linear units.
PART	PART (Partial Decision Trees) adopts the divide-and-conquer strategy of RIPPER and combines it with the decision tree approach of C4.5 [25]. To generate a single rule, PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest coverage as the new rule. Afterwards, the partial decision tree is discarded which avoids early generalization. Rule sets can be learned one rule at a time without any need for global optimization.

4. THE TEST DATASETS

The datasets used in this investigation is students’ approval data file extracted from KAU-ODUS+ data base repository. It has 5260 instances with 8 attributes, and one class as shown in the sample dataset in Table-3. The dataset contains good mix of attributes - continuous, nominal with small numbers of values, and nominal with larger numbers of values. The tens cross-validation method is used for testing the accuracy of the classification of the selected classification methods.

TEN-FOLD CROSS-VALIDATION: Tenfold cross-validation is used in this experiment. In ten folds cross-validation, a data set is equally divided into 10 folds (partitions) with the same distribution. In each test 9 folds of data are used for training and one fold is for testing (unseen data set). The test procedure is repeated 10 times. The final accuracy of an algorithm will be the average of the 10 trials.

Table-3 Sample students dataset used for Students’ Colleges enrollment



ST_ID	Term	Gender	BIO	MATH	CPIT	PHYS	CHEM	GPA	Desire	College
900000	2009	M	A+	A+	A+	A+	A+	5	Med	Med
927001	2009	M	B+	B	A	A+	A	4.57	Med	Med
927002	2009	M	A+	B	A+	A+	A+	4.82	Med	Med
927006	2009	M	A+	C	A	A+	A	4.51	Med	Med
927007	2009	M	A	C+	A	A+	A+	4.47	Med	COMP
1138724	2010	F	C	A	A+	A	B+	4.37	ENG	ENG
1138725	2010	F	A+	B+	A+	B+	A	4.68	Med	Med
1138727	2010	F	A+	A+	B+	B	C+	4.59	Med	ENG
1138728	2010	F	B	B	B	C+	C+	3.75	ENG	COMP
1138732	2010	F	A	A+	A	A+	A	4.81	Med	Med
1138733	2010	F	D	D	B+	D	C+	3.39	COMP	OTHER
1138739	2010	F	B+	B	A+	D	C+	3.92	Med	COMP
1138741	2010	F	B	C	A	B+	B+	4.13	COMP	COMP
1138742	2010	F	C	C+	A	C+	F	3.53	COMP	COMP
1138743	2010	F	B+	B+	A	C+	B+	4.56	Med	COMP
1138745	2010	F	B+	A	B	C+	B+	4.42	Med	COMP
1138748	2010	F	B+	B	B+	D	C+	4.08	Med	COMP
1138749	2010	F	B+	A+	A+	A+	A+	4.73	Med	Med

5. CLASSIFIER PERFORMANCE MEASURES

For classification tasks, the terms *true positives (TP)* and *false positives(FP* compare the results of the classifier under test with trusted external judgments. The terms *true* and *false* refer to whether that prediction corresponds to the external judgment, sometimes known as the *observation*. Let us define an experiment from *P positive instances and N negative instances* for some condition. Then, true positive rate (TPR) and false positive rate (FPR) is computed as shown in equation (1):

$$\boldsymbol{TPR = TP / P = TP / (TP + FN);}$$

$$\boldsymbol{FPR = FP / N = FP / (FP + TN) \qquad ....(1)}$$

Precision: is the probability that a (randomly selected) retrieved document is relevant. *Recall:* is the probability that a (randomly selected) relevant document is retrieved in a search. Precision and recall are then defined as in equation (2):

Precision = TP / (TP + FP);

Recall = TP / (TP + FN) …… (2)

F-measure: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score is defined as in equation (3):

F = 2 x (*Precision x Recall)* / (*Precision + Recall) … (3)*

The *Matthews correlation coefficient (MCC)* is used in machine learning as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The *MCC* is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and−1 indicates total disagreement between prediction and observation.

Confusion Matrix: also known as a contingency table or an error matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). The MCC can be calculated directly using the formula (4):

MCC=(TPxTN–FPxFN) / √ (TP+FP)(TP+FN)(TN+FP)(TN+FN) ..(4)

If any of the four sums in the denominator is zero, the denominator can be arbitrarily set to one; this results in a Matthews correlation coefficient of zero, which can be shown to be the correct limiting value. In equation (5), we define the classifiers accuracy which is the percentage of predictions that is correct, as follows: In equation (6), we define the mean absolute error. It is the prediction probability of the correct class, divided by the actual probability of the class and N is the number of classes.

Accuracy = (TP + TN) / (TP + TN + FP + FN) … (5)
Error Rate = (FP + FN) / (TP + TN + FP + FN) … (6)

***Receiver Operating Characteristic* (ROC):** In signal detection theory, ROC curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings. The ROC is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (TPR and FPR) as the criterion changes.

Precision-Recall Curve (PRC): is a two-dimensional graph where x-axis represents the precision which measures the fraction of instances classified as positive that are truly positive, and y-axis represents the recall which is the same as true positive rate. Precision-recall curves are important to visualize classifier performances. The goal is to observe whether the precision-recall curve is towards the upper right corner of the chart.
In the next sections we discussed results obtained for comparing these performance measures for classifiers algorithms used which is shown in Table-2 of section 3.

6. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

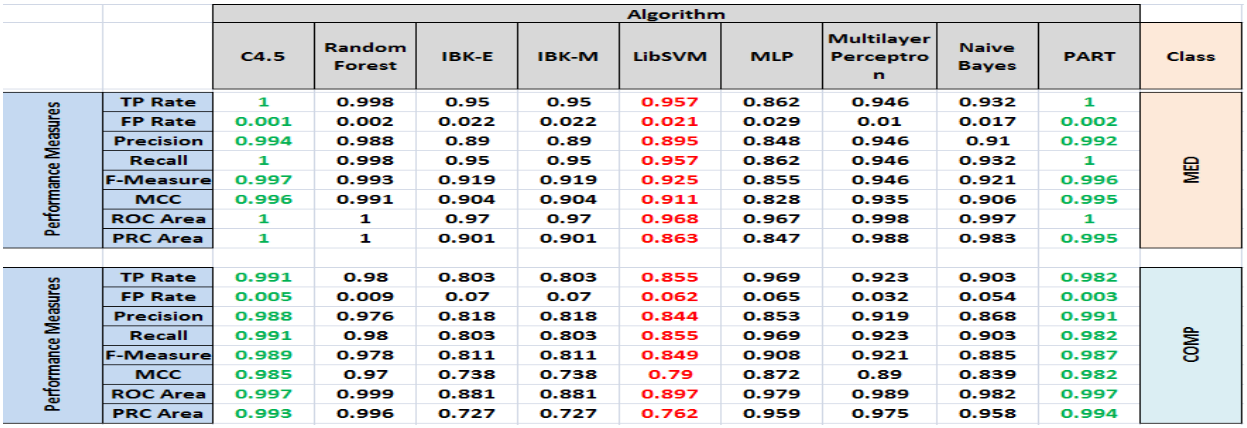
We investigate the performance of selected classification algorithms described in Table-2. The classifications have been done using 10-fold cross-validation. In WEKA, all data is considered as instances and features in the data are known as attributes. The simulation results are partitioned into several sub items for easier analysis and evaluation.

6.1 CLASSIFIERS ALGORITHMS PERFORMANCE

The results of the simulation are shown in Tables 4 and 5 below. Table 4 mainly summarizes the result of nine classifiers algorithms used and their corresponding performance measures for medical and computer colleges (class). Meanwhile, Table 5 shows the results of the nine classifier algorithms and their performance measures for engineering and others colleges (class).

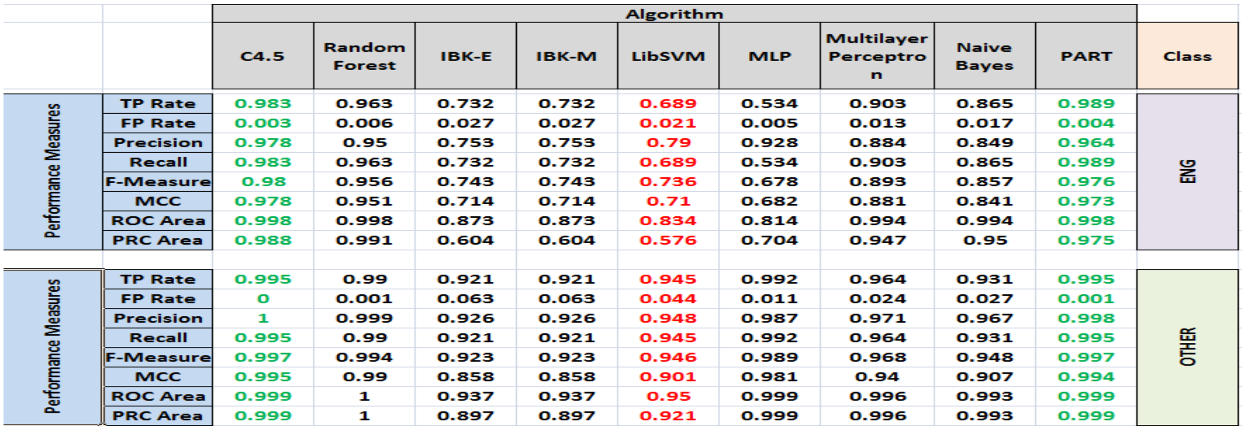
Results show that C4.5 gives the best performance, then PART, Random Forest, Multilayer Perceptron, MLP and Naïve Bayes, respectively. The LibSVM algorithm come in the seven’s order. IBK-E and IBK-M gives equal performance and they come in the eights order.

Table-4 Classifiers Algorithms results for Medical and Computer Colleges.



		Algorithm									
		C4.5	Random Forest	IBK-E	IBK-M	LibSVM	MLP	Multilayer Perceptron	Naive Bayes	PART	Class
Performance Measures	TP Rate	1	0.998	0.95	0.95	0.957	0.862	0.946	0.932	1	MED
	FP Rate	0.001	0.002	0.022	0.022	0.021	0.029	0.01	0.017	0.002	
	Precision	0.994	0.988	0.89	0.89	0.895	0.848	0.946	0.91	0.992	
	Recall	1	0.998	0.95	0.95	0.957	0.862	0.946	0.932	1	
	F-Measure	0.997	0.993	0.919	0.919	0.925	0.855	0.946	0.921	0.996	
	MCC	0.996	0.991	0.904	0.904	0.911	0.828	0.935	0.906	0.995	
	ROC Area	1	1	0.97	0.97	0.968	0.967	0.998	0.997	1	
	PRC Area	1	1	0.901	0.901	0.863	0.847	0.988	0.983	0.995	
Performance Measures	TP Rate	0.991	0.98	0.803	0.803	0.855	0.969	0.923	0.903	0.982	COMP
	FP Rate	0.005	0.009	0.07	0.07	0.062	0.065	0.032	0.054	0.003	
	Precision	0.988	0.976	0.818	0.818	0.844	0.853	0.919	0.868	0.991	
	Recall	0.991	0.98	0.803	0.803	0.855	0.969	0.923	0.903	0.982	
	F-Measure	0.989	0.978	0.811	0.811	0.849	0.908	0.921	0.885	0.987	
	MCC	0.985	0.97	0.738	0.738	0.79	0.872	0.89	0.839	0.982	
	ROC Area	0.997	0.999	0.881	0.881	0.897	0.979	0.989	0.982	0.997	
	PRC Area	0.993	0.996	0.727	0.727	0.762	0.959	0.975	0.958	0.994	

Table-5 Classifiers Algorithms results for Engineering and Other Colleges.



		Algorithm									
		C4.5	Random Forest	IBK-E	IBK-M	LibSVM	MLP	Multilayer Perceptron	Naive Bayes	PART	Class
Performance Measures	TP Rate	0.983	0.963	0.732	0.732	0.689	0.534	0.903	0.865	0.989	ENG
	FP Rate	0.003	0.006	0.027	0.027	0.021	0.005	0.013	0.017	0.004	
	Precision	0.978	0.95	0.753	0.753	0.79	0.928	0.884	0.849	0.964	
	Recall	0.983	0.963	0.732	0.732	0.689	0.534	0.903	0.865	0.989	
	F-Measure	0.98	0.956	0.743	0.743	0.736	0.678	0.893	0.857	0.976	
	MCC	0.978	0.951	0.714	0.714	0.71	0.682	0.881	0.841	0.973	
	ROC Area	0.998	0.998	0.873	0.873	0.834	0.814	0.994	0.994	0.998	
	PRC Area	0.988	0.991	0.604	0.604	0.576	0.704	0.947	0.95	0.975	
Performance Measures	TP Rate	0.995	0.99	0.921	0.921	0.945	0.992	0.964	0.931	0.995	OTHER
	FP Rate	0	0.001	0.063	0.063	0.044	0.011	0.024	0.027	0.001	
	Precision	1	0.999	0.926	0.926	0.948	0.987	0.971	0.967	0.998	
	Recall	0.995	0.99	0.921	0.921	0.945	0.992	0.964	0.931	0.995	
	F-Measure	0.997	0.994	0.923	0.923	0.946	0.989	0.968	0.948	0.997	
	MCC	0.995	0.99	0.858	0.858	0.901	0.981	0.94	0.907	0.994	
	ROC Area	0.999	1	0.937	0.937	0.95	0.999	0.996	0.993	0.999	
	PRC Area	0.999	1	0.897	0.897	0.921	0.999	0.996	0.993	0.999	

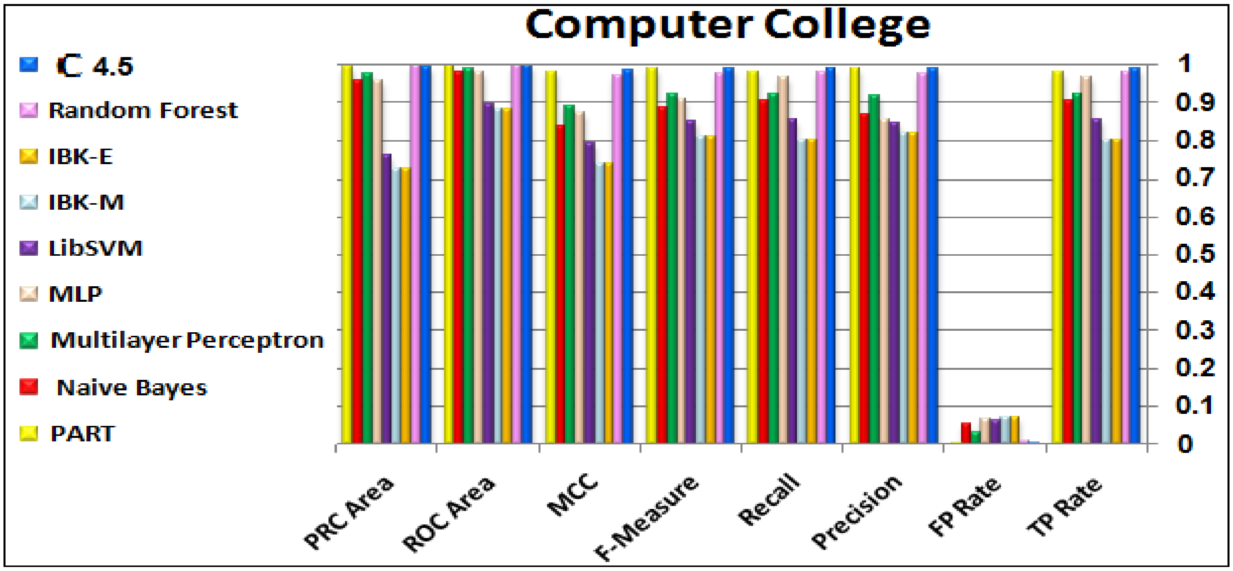


Fig.2 Classifiers Algorithms Performance results for Computer College.

6.2 COLLEGES ADMISSION AND ENROLLMENTS

Fig.2 shows graphical representation of the simulation results for Students who enrolled into Computer College. It shows that C4.5 and Random Forest give the best performance among the others algorithms investigated. Similar results are obtained for Medical, Engineering and Other Colleges.

Fig.3 shows graphical representation of the simulation results for classifiers accuracy. It shows that C4.5 and PART give the best accuracy among the other algorithms investigated. IBK-M and IBK-E algorithms give the lowest accuracy. Fig.4 shows graphical representation of the simulation results for classifiers algorithms mean absolute errors. It shows that C4.5 and PART give the lowest error among the others algorithms investigated. MLP algorithm gives the highest error.

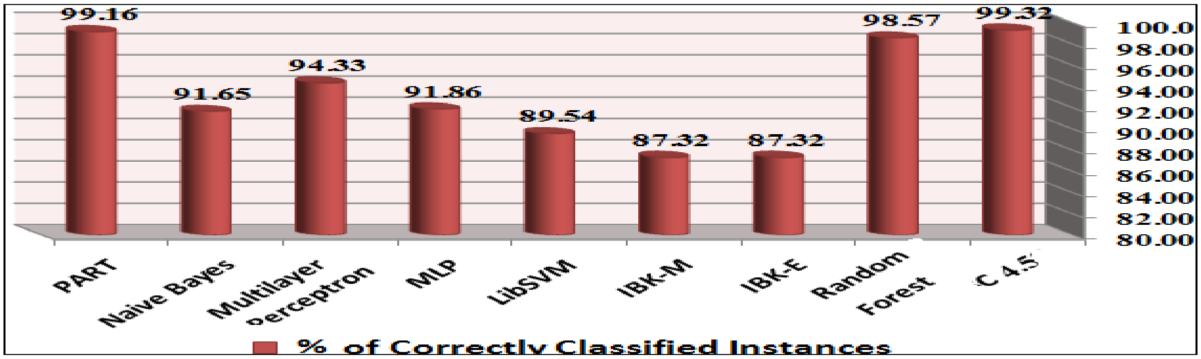


Fig.3 Classifiers Algorithms Performance results for classifiers accuracy.

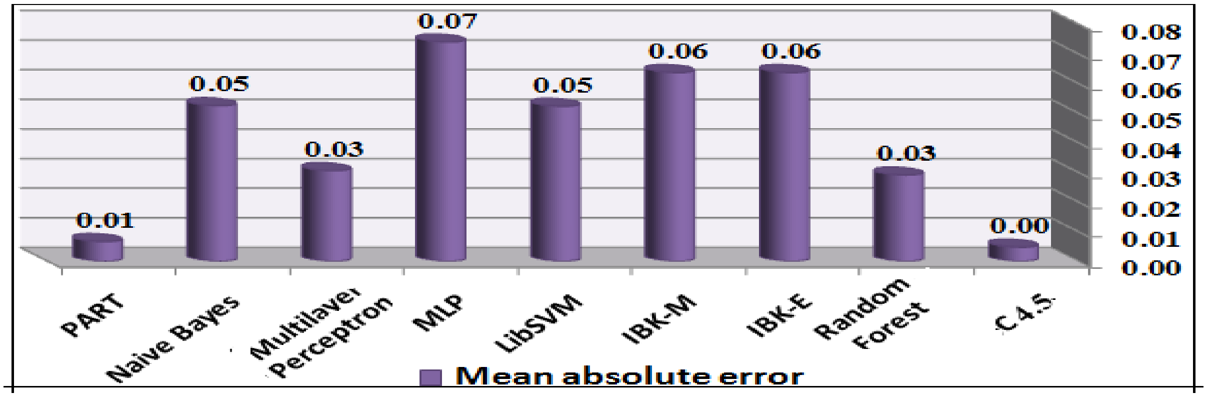


Fig.4 Classifiers Algorithms Mean absolute errors.

6.3 IMAPCT OF GPA AND GRADES QUALIFIED MATERIALS ON ENROLLMENT

Weka Toolkit provides histogram for visualization analysis of the attributes. We can trace the distribution of students within each college according to their GPA, and their Desire when achieving Colleges criteria, which include *Grades Qualified Materials* (GQM) shown in Table 6. Results in Fig.5 show that 828 students enrolled into Medical College (379 Male & 449 Female), 1479 students enrolled into Computer College (1077 Male & 402 Female), 534 students enrolled into Engineering College (217 Male & 317 Female), and 2419 students enrolled into Other College (1992 Male & 427 Female).

Fig.6 shows a histogram distribution of the GPA of all students and their actual enrollment. Students who have GPA >= 4.5 are enrolled into Medical College while some of them preferred to join the engineering college and others preferred to join the Computer College. While almost all students that have GPA >= 4 joined the Engineering College. There are some of them could not join that College and they actually joined Other Colleges based on other attributes of classification (Grades qualified materials). Also, it is clear that students who joined Computer College have GPA >= 3.5, also not all of these students could join the Computer College and they actually joined Other Colleges based on other attributes classification *(Grades qualified materials).* Finally, all students with GPA < 3.5 joined Other Colleges.

Table-6 Grades qualified materials (GQM) for all Colleges.

College	GQM	Grade
Medical	CHEM, PHYS, BIO	>= "B+"
Engineering	PHYS, MATH	>= "B"
Computer	CPIT	>= "B"
Other	NA(Not Available)	NA

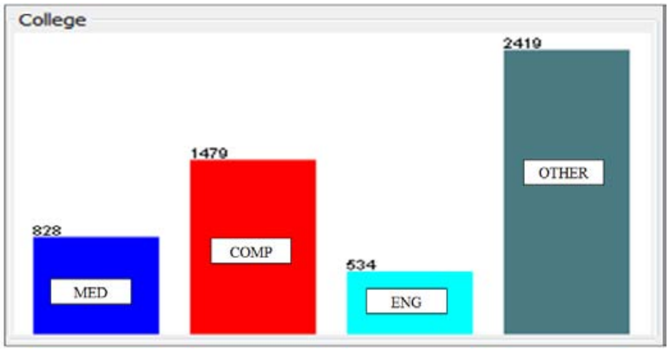


Fig.5 students' enrollment distribution Histogram into Colleges.

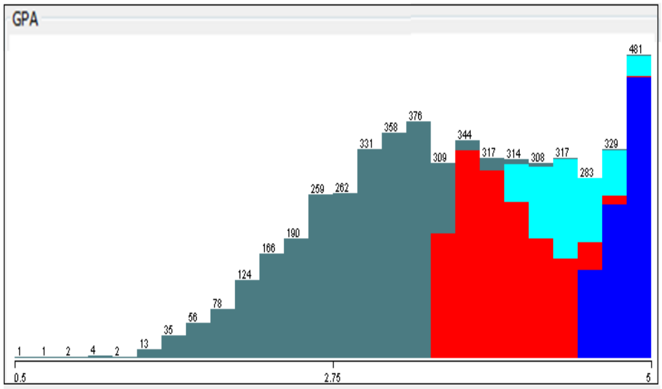


Fig.6 Histogram for distribution of the GPA of all students and their actual enrollment.

7. CONCLUSION

In this paper, several university colleges have been considered on this study including Medical, Engineering, Computing and other colleges. Impact of students GPA and their Grades Qualified Materials with respect to their Colleges admission desire are visually analyzed. This helped us for determining the actual numbers of qualified students to be enrolled into their desired college as long as they achieved college criteria rules.

Nine classification algorithms are investigated. They included C4.5, Random Forest, IBK-E, IBK-M, LibSVM, MLP, Multilayer Perceptron, Naïve Bayes, and PART. Comparative study and analysis related to classification measures included *Recall, Precision, F-Measure, Matthews's correlation coefficient* (MCC), *Precision-Recall Curve (PRC), ROC curve, FP-rate, and TP-rate* have been computed by simulation using Weka Toolkit.

Experimental Results show that C4.5 gives the best performance and accuracy and lowest absolute errors, then PART, Random Forest, Multilayer Perceptron, and Naïve Bayes, respectively. The LibSVM algorithm come in the seven's order. IBK-E and IBK-M gives low equal accuracy and equal high errors and they come in the eights order. MPL gives highest error.

Hence, these results can help higher university authorities to select the optimal classification algorithms suitable for the datasets related to students' admission and enrollments for KAU University colleges.

8. ACKNOWLEDGMENTS

This study has been supported by Research Project of DSR at King Abdulaziz University, Jeddah, Saudi Arabia.

9. REFERENCES

[1] Karina Giberta, Miquel Sànchez-Marrèa, Beatriz Sevilla," Tools for Environmental Data Mining and Intelligent Decision Support", International Congress on Environmental Modelling and Software, 2012, http://www.iemss.org/society/index.php/iemss-2012-proceedings, May 2014.

[2] C. Giraud-Carrier and O. Povel," Characterising Data Mining software", Intelligent Data Analysis, PP 181–192, 2003.

[3] www.cs.waikato.ac.nz/~ml/weka, May 2014.

[4] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa," A Comparison Study between Data Mining Tools over some Classification Methods", (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence,pp18-26, 2012.

[5] Quinlan, J. R. C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[6] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao," Predicting Students' Performance using ID3 and C4.5 Classification Algorithms", Int. Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, Sep.2013

[7] http://www.knime.org/files/nodedetails/weka_classifiers_lazy_IBk.html, May 2014.

[8] Md. Faisal Kabir, etal ,"Enhanced Classification Accuracy on Naive Bayes Data Mining Models" , International Journal of Computer Applications IJCA, No.3 - Article 2, 2011. http://jsalatas.ictpro.gr/implementation-of-elman-recurrent-neural-network-in-weka/, May 2014.

[9] Galen Andrew, "Overview of Quasi-Newton optimization methods", Informal tutorial at Microsoft Research, on Jan. 18, 2008. https://homes.cs.washington.edu/ ~galen/files/quasi-newton-notes.pdf, May 2014.

[10] Martin Riedmiller, Machine Learning: Multi Layer Perceptrons, Albert Ludwigs University Freiburg AG Maschinelles Lernen, http://ml.informatik.uni-freiburg.de/_media/documents/teaching/ss09/ml/mlps.pdf, May 2014.

[11] Dustin Boswell," Introduction to Support Vector Machines", August 6, 2002. www.work.caltech.edu/ ~boswell/IntroToSVM.pdf , May 2014.

[12] S. Aruna, S.P. Rajagopalan and L.V. Nandakishore, "An Empirical Comparison of Supervised Learning Algorithms in Disease Detection," International Journal of Information Technology Convergence and Services, vol. 1, no. 4, pp. 81-92, 2011.

[13] Hong Hu, Jiuyong Li, Ashley Plank, Hua Wang, Grant Daggard, "A Comparative Study of Classification Methods for Microarray Data Analysis", in Australasian Data Mining Conference, Sydney, 2006.

[14] My Chau Tu, Dongil Shin, Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms", in Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.

[15] Aman Kumar Sharma, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis,"

International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 5, pp 1890- 1895, May 2011.

[16] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar,” Comparative Study on Email Spam Classifier using Data Mining Techniques”, Proceedings of Int. MultConf. of Engineers and Computer Scientists, Hong Kong, Vol. I, March 14-16, 2012.

[17] Rohit Arora, Suman,” Comparative Analysis of Classification Algorithms on Different Datasets using WEKA”, International Journal of Computer Applications. Vol. 54 No.13, pp 21-25, Sep. 2012.

[18] Abdullah H. Wahbeh and Mohamed Al-Kabi,”Comparative Assessment of the Performance of Three”, ABHATH AL-YARMOUK: "Basic Sci. & Eng." Vol. 21, No. 1, pp. 15- 28, 2012.

[19] Murat Koklu and Yavuz Unal,” Analysis of a Population of Diabetic Patients Databases with Classifiers”, World Academy of Science, Engineering and Technology, International Journal of Medical, Pharmaceutical Science and Engineering Vol.7 No.8, pp 176- 178, 2013.

[20] S. Vijayarani1 Mrs. M. Muthulakshmi,” Comparative Study on Classification Meta Algorithms”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 8, pp 1768- 1774, Oct. 2013.

[21] Devendra Kumar Tiwary,” A Comparative Study of Classification Algorithms for Credit Card Approval using WEKA” GALAXY International Interdisciplinary Research Journal, GIIRJ, Vol.2 No.3 ,pp 165 – 174, Mar. 2014.

[22] Andy Liaw and Matthew Wiener,”Classification and Regression by randomForest”, R News, Vol. 2/3, Dec. 2002.

[23] Frederick Livingston, “Implementing Breiman’s Random Forest Algorithm into Weka”, ECE591Q Machine Learning Conference Papers, Nov. 27, 2005.

[24] D. Aha, D. Kibler , M. Albert,” Instance-based learning algorithms”, Machine Learning. Vo.6, PP 37-66, 1991.

[25] S.Vijayarani, M.Divya,” An Efficient Algorithm for Generating Classification Rules“, IJCST Vol. 2, Iss ue 4, Oct. - Dec. 2011.