Indian Journal of Science and Technology, Vol 11(47), DOI: 10.17485/ijst/2018/v11i47/130870, December 2018
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Predicting Student Academic Performancein Computer Organization Course: Using J48 Algorithm

Devine Grace D. Funcion*

Leyte Normal University, Tacloban City, Philippines; dgd_35@lnu.edu.ph

Abstract

Objective: Education acts as a significant role in student's life were low scholastic performance create a vast impact on the final level of the scholars. However, no model will guide the student and teachers in predicting academic performance and subsequently help improve the ranks of the students. **Methods:** Data mining technique explicitly utilized the J48 algorithm to predict the academic performance of the students. The 10-Folds Cross-validation and Receiving Operating Characteristics Curve (ROC) was deployed to create a model and test the result based on the attributes. The collected datasets of this study are from the previous grades of the 2nd year BSIT students enrolled in the Computer Organization Course from S.Y. 2016-2017 and 2017-2018. **Findings:** The result generated in the decision tree model and decision rule classification, Confusion matrix, ROC and AOC show that Lab exercise/Project is the most critical attribute that profoundly affects the students' academic performance followed by quizzes, finals, recitation and midterm attribute in the Computer Organization class. Additionally, from the result, the model was able to identify students who will pass at 89.0% accuracy, failed at 92.60% accuracy and conditional at 74.90%. Finally, the model has high acceptability and accuracy rate in predicting the Student Academic Performance in Computer Organization. **Application/Improvements:** This study can be used to develop or create a model that will predict the academic performance of the students in Computer Organization. For more improvement of the subject area, it recommended other data mining technique would be to predict academic performance with additional parameters to test the accuracy of the algorithm.

Keywords: Academic Performance, Computer Organization, Data Mining, Decision Tree, Information Technology

1. Introduction

Strengthening the academic performance of students is necessary since they are the next generation of professionals. In this generation the global competition of hiring competent workers is stiff. Companies hire workers who have excellent academic qualification and training. Nonetheless, the education sector like Universities in the Philippines should equip their students with a comprehensive education to be globally competitive in term of educational qualification and to be at par with the graduates aboard. Moreover, the Commission on Higher Education (CHED) issued a memorandum no. 46 s. 2012 paradigm shifts from traditional teaching to outcomes-based education were graduates imbued to have analytical and problem-solving skills, lifelong learning, teamwork attitude and communication skills. However, with the revision of the curriculum, educators are encouraged to utilize a different strategy and teaching methods to raise the academic performance of the scholars. Unfortunately, some students show poor academic performance. Poor academic performance defines as a public presentation that is judged by the analyst as falling below an average standard[1]. It additionally pointed out that academic failures are not only frustrating to the students and the guardians but as well as to the teachers. Surveys have indicated that many factors influence poor academic performance these are poor study habits, lack of involvement in a school program, low retention, association with wrong peers, low achievement motivation and emotional

*Author for correspondence

problems affect the academic operation of the students[2,3]. Having this situation where the student shows poor academic performance, no concrete model will guide the teacher to predict academic performance.

The application of data mining technique to predict academic performance will guide the teachersand student reach the better academic achievement. The principal components of data mining are applying different techniques and calculations to extract and identify trends from massive amounts of data[4]. Hence, literature has identified factors that affect student academic performance, such as personal, social and environmental problems. Moreover, several studies conducted on the use of data mining technique to analyze large data and create models to predict student performance are available in the literature, and a few specific studies is listed below for reference.

A study of 50 students of the VBS Purvanchal University utilized the data classification, the decision tree method to predict the students' division by the previous database[5]. Information like Attendance, Class test, Seminar and Assignment was collected from the last student record to determine the academic performance of the students at the end of the semester. Based on the study the algorithm was able to identify those students who needed special attention and reduce the number of students who failed the subject[6].

Also, the J48 decision tree algorithm was used to identify the most critical course in the students' study plan based on their final grades in the mandatory classes[7]. Hence, according to the study J48 gives 96.73% accuracy rate on predicting soil fertility based on the attributes like pH value of soil, Organic Carbon %, Electrical conductivity, decision per meter[8]. The J48 algorithm achieves accuracy up to 99.87%[9]. The researcher utilized the J48 algorithm, because of high accuracy rate on predicting academic performance of the students in Computer Organization course[10]. Furthermore, the researcher used the previous records of the students like quizzes, recitation, projects and significant examinations (Midterms and Final) to develop a model using a decision tree and to predict what factorscontribute to student low academic performances.

Hence, several studies apply classification methods like decision trees and the Bayesian network to the educational data for predicting the student's performance in examinations. The prediction will help to identify the weak students and help them to score better marks[11]. The C4.5, ID3 and CART decision tree algorithms are applied to engineering student's data to predict their performance in the final exam[12]. Moreover, a study on data mining was conducted to predict student dropout in the management of Engineering program. In this study, the model can produce an accurate prediction of students who tend to dropoutof the program. The results show that the machine-learning algorithm can establish a useful predictive model from the current student dropout data[13].

Moreover, this study is similar to the above literature, which will apply data mining, specifically the J48 algorithm to predict student academic performance in Computer Organization course. The C4.5 and decision tree is used to create a model and utilized the 10-fold cross-validation of the result using Receiving Operating Characteristics Curve (ROC) Area under ROC Curve is a graph that is used to visualize the result of the algorithm. Likewise, the study will help the students and teachers to improve the academic level of the student. Hence, the ability to predict student academic performance is essential for the teachers as well as to students who need particular attention to prevent in getting a failing grade in the final semester.

2. Methodology

The study utilized Knowledge Discovery in Database (KDD) process. The process starts with data collection and data preprocessing followed by classification model construction and ends with model evaluation and interpretations[6].

2.1 Data Mining Process

2.1.1 Data Collection

The collected datasets of this study are from the previous grades of the 2nd year BSIT students enrolled in the Computer Organization Course from S.Y. 2016-2017 and 2017-2018. There are 151 datasets with six variables (Quizzes, Recitation, Projects, Midterm, Final Exam, Remarks). Table 1 shows the attribute description and their possible values.

The variables in Table 1 are defined to understand the process:

Quizzes – There is no definite time or schedule to conduct the assessment to evaluate the learning process of the students. Computation for examinations is Raw score divide by the total number of Items multiply to thirty-five

(35) plus sixty to get the highest equivalent grade of 95 and lowest equivalent grade of 60.

Recitation – An equal opportunity is given to the student to recite during the class. Wherein, recitation is being conducted by row or by columns. The point has given **5** – Concepts shared is broad and accurate, **4** - Concepts shared is somewhat full and precise, but limited, **3** – Concepts shared is definite but limited, **2** - Concept shared is not accurate and limited, **1** – Attempted to share knowledge, **0** – did not try to share new concepts or ideas.

Project – One project is being given to the students to encapsulate and apply the learned skills and knowledge.

Table 1. Student related variables

Variables	Description	Possible Values
Q	Quizzes	3 grade
R	Recitation	0-5
P	Project	1-10
M	Midterm Examination	60-95 grade
F	Final Examination	60-95 grade

2.1.2 Software Used

To apply the classification algorithm, in predicting data the University of Waikato in New Zealand developed the widely used software WEKA toolkit. This toolkit provides a wide range of different data mining algorithms implemented in JAVA. It has been commonly used in educational data mining researches and for teaching purposes[8].

2.1.3 Data Preparation and Processing

During this phase, a pre-processing of collected and prepared the data for the mining techniques. At first, we eliminated some irrelevant attributes, e.g., student name, student number, teacher's name and schedule. Then, each student will have the following characteristicsas shown in Table 2.

Second step data preparation, student grades are stored in MS Excel and later converted to Microsoft Excel Comma Separated Values File (.csv). The .csv file was then loaded to Notepad++. In this stage, data cleaning is done to eliminate unwanted symbols (e.i., comma, colon, spaces).

Additionally,in the notepad application declaration of syntaxes like @Relation, @Attribute and @Data is included as a requirement in the WEKA application. The converted text file to Attribute-Relation File Format (ARFF) in notepad describes the list of instances sharing a set of attributes and the accepted file format for WEKA application. Next, information is uploaded to the WEKA Application and conducted the pre-processing of raw data to a more understandable file format. The third step, data modeling, WEKA used to predict Student Academic Performance of the BSIT students. This stage consists of five phases of training, pattern, testing, result evaluation and knowledge representation. In this stage, it divides the cleaned data into two stages; the training and testing stage. In the training stage, the J48 algorithm is used to build a model. The J48 algorithm is C4.5 decision tree approach that is useful in the classification of the problem which creates a binary tree model[12].

Additionally, on the testing stage, the k-fold cross validation using 10-fold cross-validation was observed. The cross-validation the techniqueis about dividing the data into k number of equally sized folds. Also in the testing stage Receiving Operating Characteristics Curve (ROC) Area under ROC Curve is a graph that is used to visualize the selected classifier based on the passed, failed and conditional remarks[14].

Table 2. Sample datasets of student records from the previous semester

Lab Exercises/ Project	Quizzes	Midterm	Finals	Remarks
90.33	82.70	75.75	86.47	Pass
76.67	84.23	71.67	82.74	Pass
88.00	92.69	79.25	89.80	Pass
95.67	85.76	84.20	83.90	Pass

2.1.4 Data Visualization

After loading the data to WEKA, we set out some primary useful knowledge about the attributes before applying any data mining method by using the visualizing technique in the software. For example, we found that in the Remarks attribute there are eighteen (18) students under passed status, fourteen (14) students under conditional status and six (6) students under the failed state. Nevertheless, the problem is on the student in the conditional remarks. Under the policy of the University, students who get a limited remark is given a one (1) year to comply for the missing requirements. However, conditional statements mean that students will get either passed or 3.0-grade equivalent or failed or 5.0-grade equivalent. Likewise, having a model that will predict academic performance,

which is helpful for the fourteen (14) students. This will serve as a guide to improve and enhance more their chances of passing the course after completion.

2.1.5 Decision Trees

Decision trees are standard supervised learning algorithms, easy to understand and easy to use. Decision trees are trees that separate instances by sorting them based on feature values[15]. Decision trees are the most popular classification technique in data mining[16]. They represent the group of classification rules in a tree form, and they have several advantages over other methods as stated in[17]:

- The simplicity of its presentation makes them easy to understand.
- They can work for different types of attributes, nominal or numerical.
- They can classify new examples fast.

The tree has as its root a hub determining the test and, for every result thus, the comparing subtree is gotten by applying a similar methodology to the subset of occurrences with that result. A tree is either a leaf hub named in a class or a structure comprising of a test hub connected to at least two subtrees. A test hub Figure 1 some result in light of the trait estimations of a case, where every likely resultis related to one of the subtrees. On the off chance that this hub is a test, is resolved andoutcome for the matterand the procedure keeps utilizing the proper subtree[18].

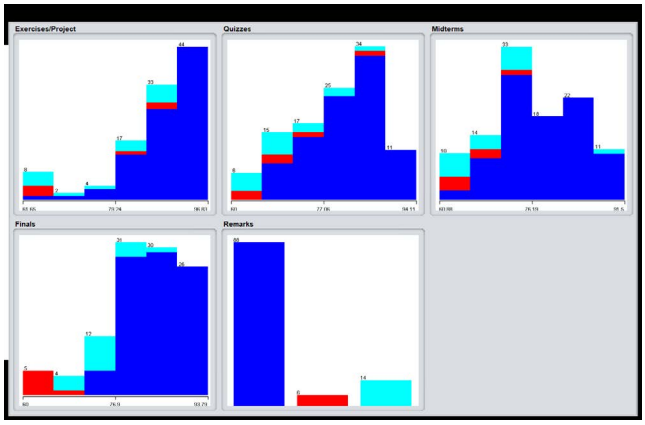


Figure 1. Primary visualization of the different attributes.

2.1.6 J48 Algorithm

The J48 is implemented Java using C4.5 algorithms. A decision tree is created based on a set of labeled input data and process in the C4.5 is a program, which Ross Quinlan developed the algorithm[19]. J48 uses a divide-and-conquer algorithm to break open a root node into a subset of two partitions until the leaf node (object node) occurs in the tree. The creation of a decision tree it follows the below process:

Step 1: If all T belongs to the same group of instances, class and T is having fewer cases, than the tree is leaf labeled with the most frequent type in T.

Step 2: If step 1 occurs an error, possible selection of outcome in single or more test-based attributed then consider, this test as a root node of the tree with one branch of each issue of the trial, partition T into the corresponding T1, T2, T3 ....., according to the result for each respective cases, and the same applied a recursive way to each subnode.

Step 3: Data gain and default gain proportion are positioned utilizing two heuristic criteria by calculation J48[20].

2.1.7 Cross-Validation

Cross-validation (CV) method used to validate the predicted model. CV test divides the training data into some partitions or folds. The classifier is assessed by exactness on one stage after gained from another. In the process, all used the repeated separation until the end of evaluation[16]. The most common types are 10-fold and the bootstrap result obtained into a single estimation[20].

3. Result and Discussion

3.1 The Model

Figure 2 illustrates the graphical presentation of the pruned decision tree of Student Academic Performance. Wherein, Project as the highest instances and become the first split between the (Quizzes < = 60) and (Quizzes > 60) in predicting student academic performance. Moreover, in Figure 3, shows the academic performance decision rule that the project has the highest factor to receive pass, failed or conditional remarks.

The decision tree has correctly classified 13 instances as shown in Table 3 the confusion matrix that correctly classified instances and misclassification of Student Academic Performance. Moreover, is interpreted as:

- The decision tree has correctly classified 106 instances as PASS and thirteen cases as conditional leading to Misclassification.

- The decision tree has correctly classified 13 instances as FAILED and three instances as conditional leading to Misclassification.
- The decision tree has incorrectly classified 6 PASS and five failed to lead to Misclassification and correctly classified five instances as conditional.

Table 4 shows the cross-validation summary wherein there are 82.12% correctly classified instances and 17.88% incorrectly classified instances this is supported by Table 5 that shows the detailed accuracy by class wherein the Precision Weight Average of the academic performance of the students is 82.10%. Additionally, the study utilized Receiving Operating Characteristics Curve (ROC) and the Area under ROC Curve (AUC) for model accuracy a shown in Figure 4 about ROC curve and AUC curve. Also, results revealed that the attribute Pass 89.0% accuracy, failed has 92.60% accuracy and Conditional has 74.90%. Finally, the model has high acceptability and accuracy in predicting the Student Academic Performance in Computer Organization.

Table 3. Confusion matrix

Pass	Failed	Conditional	<-- classified as
106	0	6	| a = Pass
0	13	5	| b = Failed
13	3	5	| c = Conditional

3.2 Student Academic Performance

Predicting student academic performance in higher education is essential to reduce the number of students who will fail the course. It is additionally useful for the teacher to appropriately lead the students towards learning and have the capacity to comprehend those students who are not performing well in the subject. Likewise, the prediction will serve as a field guide to improve their performance in the classroom.

As shown in the decision tree model the J48 algorithm was able to predict 89% accuracy for passed, failed has 92.60% accuracy and conditional has 74.90% accuracy on computer organization course based on Table 4 result. Hence, Exercise/Project attributes revealed the highest indicator that can affect the academic perfor-

Table 4. Cross-validation summary

Correctly Classified Instances	124			82.1192		%
Incorrectly Classified Instances	27			17.8808		%
Kappa Statistic	0.5393					
K&B Relative Info Score	7935.1148		%			
K&B Information Score	87.3548		Bits	0.5785		bits/instance
Class complexity | order 0	163.5452		Bits	1.0831		bits/instance
Class complexity | scheme	8655.735		Bits	57.3227		bits/instance
Complexity improvement (Sf)	-8492.19		Bits	-56.2397		bits/instance
mean absolute error	0.1277					
Root mean squared error	0.3094					
Relative absolute error	45.3953		%			
Root relative squared error	83.0006		%			
Total Number of Instances	151					

Table 5. Detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.946	0.333	0.891	0.946	0.918	0.657	0.89	0.927	Pass
	0.722	0.023	0.813	0.722	0.765	0.737	0.926	0.771	Failed
	0.238	0.085	0.313	0.238	0.270	0.173	0.749	0.342	Conditional
Weighted Avg.	0.821	0.262	0.801	0.821	0.809	0.599	0.875	0.827	

mance followed by the attribute quizzes, which the cause of the second split in the decision tree model is based on Figure 3 for the students to passed or failed the Computer Organization course. The computer organization course by nature is more on skills demonstration that the student should be able to demonstrate their knowledge in binary conversion, boolean algebra, logic gates, microprogramming and memory management. Hence, for the student to demonstrate their skills and learnings are giving lab exercises and projects will help the student use their cognitive abilities in the actual application or demonstration. A project-based learning approach encourages the student to reflect on their knowledge and work collaboratively on research projects[21]. For example, an individual project plan is presented to the students to develop simulations in making a traffic light out of logic gates. The students should be able to use the correct logic gates for the LED lights to reverse on. Additional Digital clock signal or seven-segment registry is another requirement for the project that will change the LED light accordingly. Moreover, the student should be able to demonstrate and create the traffic light based on the acquired skills and knowledge in binary and logic gates.

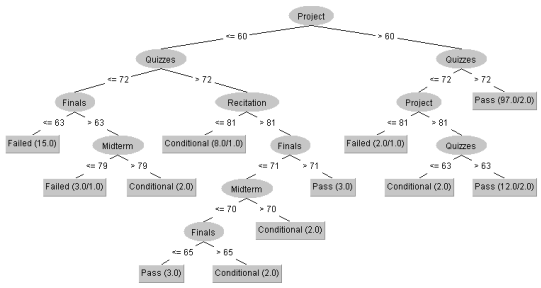
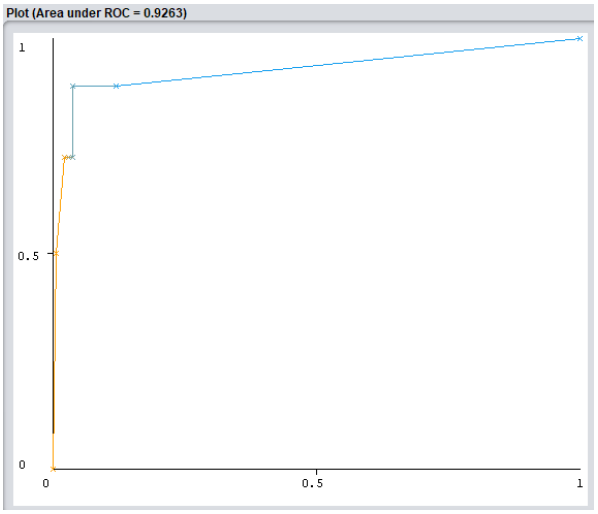


Figure 2. Decision tree model on student academic performance.

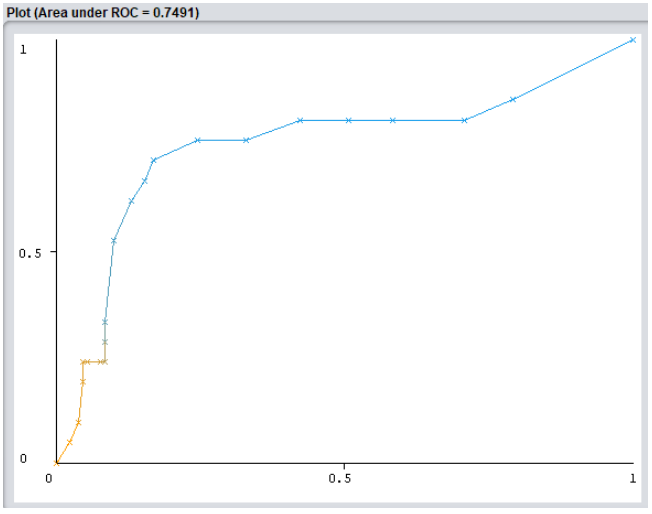
Likewise, students are given situational problems to analyze and use their skills in Boolean algebra, microprogramming and construct a state transition diagram and Karnaugh Mapping. Problem-based learning can have a positive effect on students' knowledge primarily on critical thinking skills[22]. Additionally, Quizzes as the second attribute that affects student academic performance is essential to evaluate the retention skill and assess those students who display low academic achievement, possible re-teaching of the topics is needed for the students to understand the lessons in Computer Organization thoroughly.

R1: IF (Project<=60) AND (Quizzes <=72) AND (Finals <= 63) THEN
Performance = "FAILED"
R2: IF (Project <=60) AND (Quizzes <=72) AND (Finals > 63) AND (Midterm <= 79) THEN
Performance = "FAILED"
R3: IF (Project <=60) AND (Quizzes <=72) AND (Finals > 63) AND (Midterm > 79) THEN
Performance = "CONDITIONAL"
R4: IF (Project <=60) AND (Quizzes >72) AND (Recitation <= 81) THEN
Performance = "CONDITIONAL"
R5: IF (Project <=60) AND (Quizzes >72) AND (Recitation > 81) AND (FINALS <= 71)
AND (Midterm <= 70) AND (Finals <=65) THEN
Performance = "PASS"
R6: IF (Project <=60) AND (Quizzes >72) AND (Recitation > 81) AND (FINALS <= 71)
AND (Midterm <= 70) AND (Finals >65) THEN
Performance = "CONDITIONAL"
R7: IF (Project <=60) AND (Quizzes >72) AND (Recitation > 81) AND (FINALS <= 71)
AND (Midterm > 70) THEN
Performance = "CONDITIONAL"
R8: IF (Project <=60) AND (Quizzes >72) AND (Recitation > 81) AND (FINALS > 71)
Performance = "PASS"
R9: IF (Project >60) AND (Quizzes <=72) AND (Project <= 81) THEN
Performance = "FAILED"
R10: IF (Project >60) AND (Quizzes <=72) AND (Project <= 81) AND
(QUIZZES <= 63) THEN
Performance = "CONDITIONAL"
R11: IF (Project >60) AND (Quizzes <=72) AND (Project > 81) AND
(QUIZZES > 63) THEN
Performance = "PASS"
R12: IF (Project >60) AND (Quizzes >72) THEN
Performance = "PASS"

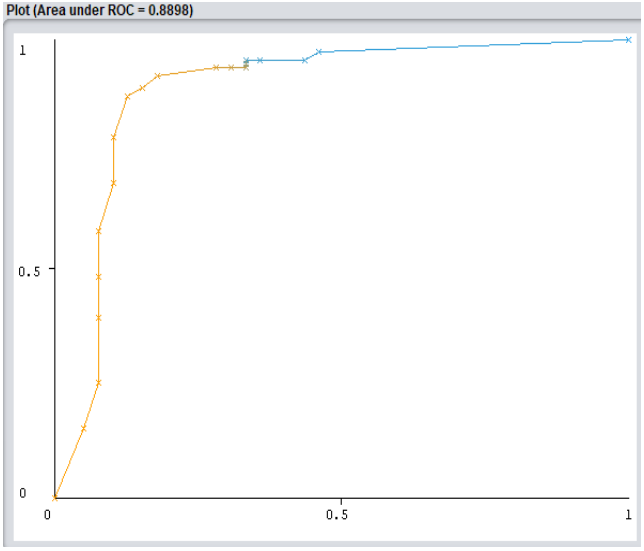
Figure 3. Student academic performance decision rule.



ROC Curve for Failed



ROC Curve for Conditional



ROC Curve for Pass

Figure 4. The area under the ROC curve.

4. Conclusion

An early prediction of students at risk of poor academic performance helps the instructor primarily to give proper guidance to improve their performance through individual tutoring and counseling. This study focused on identifying the attributes that can influence students' academic performance. Moreover, models are used to determine the academic performance of the students with highly acceptable results as shown in the decision tree, confusion matrix, Receiving Operating Characteristics Curve (ROC) and the Area under the ROC Curve (AUC). The rules derived from the model emphasize lab exercises/project is the essential attributes that the student will likely to pass orfail the course. Those students identified for poor performance can be considered for proper guidance to improve further.

5. References

1. Aremu OA, Sokan BO. A multi-causal evaluation of academic performance of Nigerian learners: Issues and implications for national development. Department of Guidance and Counselling, University of Ibadan, Ibadan; 2003. p. 365–75.
2. Mekonnen S. Problems are challenging the academic performance of physics students in higher governmental institutions in the case of Arbaminch, Wolayita Sodo, Hawassa and Dilla Universities. Natural Science. 2014; 6(5):362–75. https://doi.org/10.4236/ns.2014.65037
3. Ramaswami M, Bhaskaran RA. CHAID based performance prediction model in educational data mining. International of Computer Science. 2010; 7(1):1–9.
4. Hand DJ. Principles of data mining. Drug Safety. 2007; 30(7):621–2. PMid: 17604416. https:// doi.org/ 10.2165/ 00002018-200730070-00010
5. Yadav SK, Pal S. Data mining: A prediction for performance improvement of engineering students using classification. Journal of Information Technology. 2012; 2(2):51–6.
6. A review of research on project-based learning. 2000. http://www.bie.org/images/uploads/general/9d06758fd346969cb63653d00dca55c0.pdf
7. Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications. 2013; 6(2):256–61.
8. Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications. 2014; 98(22):1–5.
9. Al-Barrak MA, Al-Razgan M. Predicting student's final GPA using decision trees: A case study. International Journal of Information and Education Technology. 2016; 6(7):1–6. https://doi.org/10.7763/IJIET.2016.V6.745
10. Data mining view on classroom teaching language. 2011. https://arxiv.org/abs/1104.4164
11. Sundar PP. A comparative study for predicting student's academic performance using Bayesian network classifiers. IOSR Journal of Engineering (IOSRJEN). 2013; 3(2):1–6. https://doi.org/10.9790/3021-03213742
12. Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications. 2013; 6(2):256–61.
13. Pal S. Mining educational data to reduce dropout rates of engineering students. International Journal of Information Engineering and Electronic Business. 2012; 4(2):1–7. https://doi.org/10.5815/ijieeb.2012.02.01
14. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27(8):861–74. https://doi.org/10.1016/j.patrec.2005.10.010
15. Las Johansen BC. Predicting teachers' ICT competence in a Philippine University using J48 algorithm. Indian Journal of Science and Technology. 2018; 11(7):1–7. https://doi.org/10.17485/ijst/2018/v11i7/119062
16. Baradwaj BK, Pal S. Mining educational data to analyze students' performance. International Journal of Advanced Computer Science and Applications. 2012; 2(6):1–7.
17. Quinlan JR. Learning decision tree classifiers. ACM Computing Surveys (CSUR). 1996; 28(1):71–2. https://doi.org/10.1145/234313.234346
18. Gholap J. Performance tuning of J48 algorithm for prediction of soil fertility. International Journal of Computer Science and Information Technology. 2012; 2(8):1–5.

19. Gupta DL, Malviya AK, Singh S. Performance analysis of classification tree learning algorithms. International Journal of Computer Applications, 2012; 55(6):1–6. https://doi.org/10.4018/jaec.2012100101
20. Lavesson N, Davidson P. Multi-dimensional measures function for classifier performance. 2nd IEEE International Conference on Intelligent System; 2004. p. 508–13. https://doi.org/10.1109/IS.2004.1344802
21. Romero C, Ventura S, Espejo PG, Hervas C. Data mining algorithms to classify students. Educational Data Mining; 2008. p. 1–10.
22. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: An update. Presented at the Special Interest Group on Knowledge Discovery and Data Mining, SIGKDD. 2009; 11(1):10–8.