Predicting Students Final GPA Using Decision Trees: A Case Study

Mashael A. Al-Barrak and Muna Al-Razgan

***Abstract*—Educational data mining is the process of applying data mining tools and techniques to analyze data at educational institutions. In this paper, we used educational data mining to predict students' final GPA based on their grades in previous courses. In our case study, we collected students' transcript data that included their final GPA and their grades in all courses. After pre-processing the data, we applied the J48 decision tree algorithm to discover classification rules. We extracted useful knowledge for final GPA, and identify the most important courses in the students' study plan based on their grades in the mandatory courses.**

***Index Terms*—Educational data mining, classification, decision tree, analysis.**

I. Introduction

The availability of educational data has been growing rapidly, and there is a need to analyze huge amounts of data generated from this educational ecosystem, Educational Data Mining (EDM) field that has emerged. Educational data mining is the process of applying data mining tools and techniques to analyze the data at educational institutions [1]. Recently, educational data mining is evolving and helping the educational sector to adapt new teaching techniques for the learning process and learners. There is a large number of published papers in the area of educational data mining. This area of research is gaining popularity due to the potential benefits to the educational field [1]. Educational institutions used educational data mining (EDM) to gain deep and thorough knowledge to enhance its assessment, evaluation, planning, and decision-making in its educational programs. EDM will help academic programs identify and discovered hidden patterns in the data. These extracted patterns can be used to predict student performance and behaviors easily. Knowing this educational information, will help administrations to allocate facilities and resources more effectively [2].

Universities have been using many data mining techniques to analyze educational report stored in the educational institute such as enrollment data, students' performance, teachers' evaluations, gender differences, and many others. Data mining techniques may, for example, give a university the needed information to better plan a number of students' enrollment, students drop out, early identification of weak students, and to efficiently allocate resources with a precise approximation.

Manuscript received November 9, 2014; revised February 4, 2015.

The authors are with King Saud University, Saudi Arabia (e-mail: mbarrak@ksu.edu.sa).

There are many techniques in data mining that can be applied to educational data, such as classification, clustering, and association rules to name a few. These techniques will help extract hidden knowledge and useful information.

Classification is one of the supervised learning techniques that build a model to classify a data item into a predefined class label. The aim of classification is to predict the future output based on the available data. Hence, educational institute is looking to predict the future output of their enrolled students based on their available previous and current students data, which make classification one of the techniques better suited for educational analysis.

Most of the previous studies focus on the use of classification for predictions based on enrollment data, performance of students in certain course, grade inflation, anticipated percentage of failing students, and assist in grading system. Up to our knowledge, there are no studies that use classification to predict a student final outcome based on his/her grades in a program study plan. Analyzing all the courses that are required in the study plan will identify the list of courses that have a huge impact on final GPAs.

Our contribution in this paper is to utilize classification to help predict students final GPA based on their grades in all courses. As a result of this insight, it will produce useful and hidden knowledge to student, teachers and universities management to take an appropriate action to improve students' results, and contribute to a better quality education. In our research of this paper, we will be analyzing the obtained data for the information technology department, at King Saud University using decision trees. In this case study, we will use the J48 decision tree classification algorithm several times to answer the following questions:

1) How can we predict students final GPA given the grades of all the mandatory courses?
2) How can we predict students final GPA earlier given the grades of Java1 and Java2 courses? [Since these two courses are the first pure programming courses taken in early semesters]
3) How can we predict the students final GPA given the grades of Java1, Java2 and data structure courses. [These are the programming courses taken in levels 3, 4 and 5 respectively during the program plan].
4) How can we predict the students final GPA after each semester, given the grades of the program mandatory courses in each semester? Also to identify the important course in each level.

The information technology department management would like to know which courses in the available data are the strongest predictors of student final GPA.

To answer these questions, our paper is organized into the

DOI: 10.7763/IJIET.2016.V6.745

following: Section II illustrates the research work that has been conducted in EDM. In Section III, understanding the domain of study will be defined. In Section IV consists of the description of the process of building a model that includes data collection, used tools, pre-processing, data visualization, building a classification model, and then analyzing the results. We conclude the paper in Section V.

II. Related Work

In the arena of educational data mining, there has been a recent surge in research paper and publishing. For example, the authors in [3] built a classification model for predicting the suitable study track for school students. The data consists of 248 instances that were collected from six basic schools in Mafraq city in Jordan. The decision tree reached an overall accuracy of 87%.

In addition, the authors in [4] analyzed students performance data using classification algorithm named ID3 to predict students marks at the end of the semester. This was applied for master of computer applications course from 2007 to 2010 in VBS Purvanchal University, Jaunpur. Their study aimed to help students and teachers find ways to improve students' performance. Data was collected from 50 students, and then a set of rules was extracted for their analysis.

Another study that focused on the behavior to improve students' performance using data mining techniques is illustrated in [5]. The data consisted of 151 instances from a data base management system course held at the Islamic University of Gaza. The data was collected from personal records and academic records of students. The author performed the data mining techniques, namely: association rules, classification, clustering and outlier detection. The results revealed useful information from association rules and classification models. Furthermore, the study had clustered students' data to identify their characteristic and detect all outliers. The knowledge obtained was useful to improve students' performance in the database course.

Another outline of the research was comparing the performance of various classifiers using educational data mining as in the following studies: The study in [6] aimed to predict students enrollment using admissions data. The researchers used applicants' data from West Virginia University that consists of 112,390 instances. Various classification learners' models had been built. They compared the result of the different learners and identified that the rules from J48 and Rido to be the best.

Furthermore, in [7], the enrollment and personal data of the University of National and World Economy (UNWE) in Bulgaria was analyzed. The goal was to predict the student performance based on the pre-university characteristics. The researcher applied several classification algorithms and the result showed that J48 decision tree had the highest overall accuracy followed by the rule learner (JRip) and the k-NN classifier.

Moreover, in [8] the authors attempted to apply different classification techniques to an educational data set to compare their performance and choose the best algorithms to be integrated in their (E-learning Web Miner) tool. This tool aimed to help teachers discover their students' performance. They used the data from the course named "Introduction to multimedia methods", offered in three academic years from 2007 to 2010 at the University of Cantabria. They had used different classification methods and they found that the performance and the accuracy of the techniques rely on the type of the attributes and the size of the dataset. Among the findings, J48 was found suitable for datasets with more than 100 instances and nominal attributes with missing data.

Another comparison of different data mining algorithms was carried out by Romero and others in [9]. Their research aimed to classify students with the same final marks into different groups depending on the activities done in a web-based course. These activities includes: the number of assignments done, the number of quizzes taken, the number of quizzes failed, and the total time used on quizzes, etc. Their dataset consist of 438 Cordoba University students in 7 Moodle courses. They evaluated the performance of the algorithms based on the type of the attributes: numerical, categorical and numerical rebalanced data. They found that two decision tree algorithms CART and C4.5 were the best algorithms for categorical data.

In [10] the author used classification to predict student success based on socio-demographic variables(age, gender, ethnicity, education, work status, and disability) and study environment (course program and course block). The dataset contains students' data in Information Systems course at the Open Polytechnic of New Zealand. Four classification trees, namely: CHAID, exhaustive CHAID, QUEST and CART were used in this study and CART was the best-performed tree with an overall correct classification percentage of 60.5%.

Most of the previous studies focused on the use of classification for prediction based on enrollment data and performance of students in a certain course. Along the same line of focus, in this research, we will be building decision tree classification models to predict students final GPA based on their grades in their study plan.

III. Understanding the Domain of Study

To perform our experiment, we first need to understand the domain from which we are going to collect the data from. We obtained the study plan and courses taught in the Information Technology department (IT) for females at Computer Sciences College, King Saud University, Riyadh. This program has a four-year study plan, with two semesters. The first year is the preparatory year, where students have to take general course in mathematics, English, religion and communication. Starting from the second year (third semester) students' will start taking computer specialized courses along with another general education courses. Students are required to take 16 mandatory specialized courses and 7 elective specialized courses from their choices in order to graduate. In this study, we will focus on the mandatory courses offered by the program because of their domination in the study plan, which in turn have a greater effect on the final graduation grade. We also want to produce useful knowledge about these courses to the department management about the importance

of each course that (IT) department requires from the students.

IV. Building the Model

A. Data Collection

Transcripts data for female students who graduated from Computer Sciences College at King Saud University in the year 2012 were collected from the database management system and the total number of students was236 students. The collected data was organized in Microsoft Excel sheet. Each student record had the following attributes: student name, student ID, final GPA, semester of graduation, major, nationality, campus, and all the courses taken by the student including the course' grade.

B. Tools Used

To apply the classification algorithm, we used WEKA toolkit [11], a widely used software for data mining that was developed at the University of Waikato in New Zealand. This toolkit provides a wide range of different data mining algorithms implemented in JAVA. It has been widely used in educational data mining researches and for teaching purposes.

C. Data Preparation and Pre-Processing

During this phase, we applied some pre-processing for the collected data to prepare it for the mining techniques. At first, we eliminated some irrelevant attributes, e.g. student name, nationality, and campus. We also removed all the data related to the general and elective courses to focus only on the program mandatory courses. Then, we re-arranged the Table I so that each student has the following attributes: ID, final GPA, and the course grades student took during a four-year study program. In the final step, we discretized the numerical attributes to categorical ones. For example, we grouped the final GPA into five groups: excellent, very good, good, average and poor. In the same way, we discretized the students' grade in each course into: A+, A, B+, B, C+, C, D, D+ and F. The following table demonstrates a sample of the data that we worked on.

TABLE I: Sample of the Dataset

ID	Final GPA	JAVA1	JAVA2	.....
1	Excellent	A+	A	.....
2	Good	C	B+	.....

D. Data Visualization

After loading the data to WEKA, we got some primary useful knowledge about the attributes before applying any data mining method by using the visualizing technique in the software. For instance, we discovered that most of the graduated students during that year had an excellent final GPA and none of the students graduated with a poor GPA as shows in the Fig. 1.

This information indicates that there is a high level of grade Inflation. This problem exists due to the tough requirements that students have to meet before getting accepted in this college. Therefore, only high-level students with excellent high school grades and a very good academic background in many science subjects have the opportunity to study in the Computer Sciences College at King Saud University (The oldest university in Saudi Arabia).Another reason for this problem is the material that is being taught to the students and how the students are really evaluated. The authors in [12] suggests that the solution to this problem should be centered around the learning process rather than the grades. In order to overcome the grade inflation problem, some solution were suggested by the [12] such as: discussing the evaluation methods used after the end of each year, offering faculty development programs regularly and educating students regardless of the grading practices.

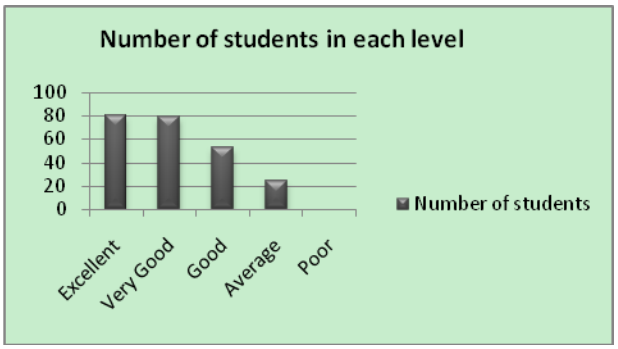

Fig. 1. Number of Graduated students in each level.

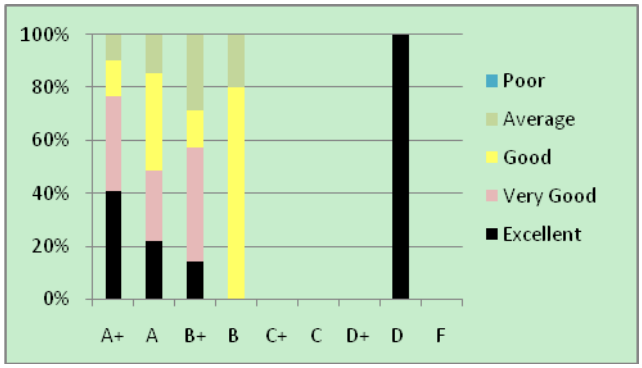

Fig. 2. Visualizing students grades in their final project with their final GPA.

More interesting information from visualizing the data was the relationship between students final GPA and their grades in each course. For example, Fig. 2 demonstrates the students' grade in the final project course. As it shows, the majority of the students that year received an A+ in that course and most of this group graduated with an "Excellent" GPA. On the other hand, most of the students earned an A in this course and graduated with a "Good" GPA. From this information, we can conclude that a student can have an A or even an A+ plus in her final project regardless of her academic performance in the previous courses.

E. Classification

In our research, we aim to predict students' final GPA based on their grades on mandatory courses. This will give us an insight on how much specific courses affect the students' graduation grades. We chose to use classification because the objective of classification techniques in educational data mining is to identify what are the important factors that contribute to categorizing students' final grades. Decision trees are the most popular classification technique in data mining [13]. They represent the group of classification rules in a tree form, and they have several advantages over other

techniques as stated in [13]:

- 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.

One of the earliest decision tree algorithms is the C4.5 tree developed by Ross Quinlan [14]. The basic idea of this tree is to built trees from a group of training data using the concept of information entropy [7]. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA. We chose this algorithm after proving its capabilities to handle educational dataset and provide a high accuracy results as mentioned in[3], [6]-[8].

F. Results and Discussion

To answer the first research questions and predict the final GPA, given the grades of all the mandatory courses, we applied the J48 tree on the data containing the grades of all the mandatory courses and the final GPA for all students in WEKA. Interestingly we found the root node to be the "Software Engineering-1" course taken in level 5 doing in the middle of the specialization year. This means that this course is the most important course and it is closely related to students' final GPA. Moreover the tree showed that if a student received an A+ in this course, she would graduate with an Excellent GPA. The represented tree is large, therefore some of the strongest rules in the tree are:

IF SoftwareEngineering-1=A+ THEN final GPA=excellent
IF SoftwareEngineering-1=A and(java1=A+ OR java1=A OR java1=B+)THEN final GPA= excellent
IF SoftwareEngineering-1=B and Assembly Language=B+ THEN final GPA= very good
IF SoftwareEngineering-1=B and Assembly Language =C+ and Project2=A+THEN final GPA= very good
IF SoftwareEngineering-1=C and Networks-1=D THEN final GPA= Average

---

This information is very helpful for the IT department administration to allow them to focus more on this particular course in order to help students pass it with the best grades and later graduate with a high GPA.

To answer the second research questions and predict students' final GPA on their grades in the first two programming courses (Java1 and Java2), we ran the J48 algorithm on this data. The resulted tree in Fig. 3 showed only Java2 course as the root node and no representation of Java1. This means that Java2 course is closely related to the final GPA more than Java1, unlike what the majority of the students thought.

This model can be useful to predict the students final GPA earlier in the first year of their major. Providing the students, faculty and administration enough time to work together and help weak students upgrade their performance in the future and have better results.

Interestingly, when we added the students' grades on the Data Structure course to Java1 and Java2, and ran the algorithm again to answer the third research questions, the resulted tree was exactly the same as the previous tree in Fig. 3 with only Java2 course as the root node. This information confirms our previous result which emphasizes the importance of Java2 course and its influence on the final GPA regardless of the data structure course, which means that JAVA 2 is the most important programming course taken by the students.

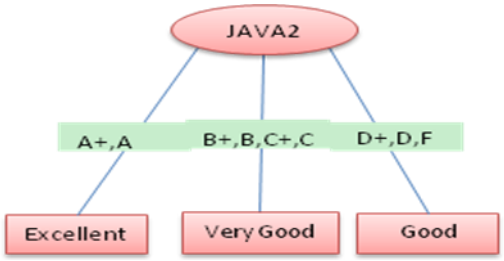

Fig. 3. The resulted J48 tree for predicting students final GPA based on Java2 course grades.

Students can benefit from this information by focusing more on the Java2 course and making an effort to pass it with the best grade they can in order to increase their chances of graduating at substantial level. The IT department management can also find this information useful to improve the content of this course and assign it to the specialized faculties to support the students.

In order to answer the fourth research questions and attempt to predict the students final GA after each semester of their specializes years, we ran the J48 algorithm 6 times, one for each semester. The results were as follows:

- For the third semester, the study plan contained only Java1 course as the specialized course, so we ran the algorithm based only on the students grades in Java1 and their final GPA. The resulted tree is in Fig. 4.

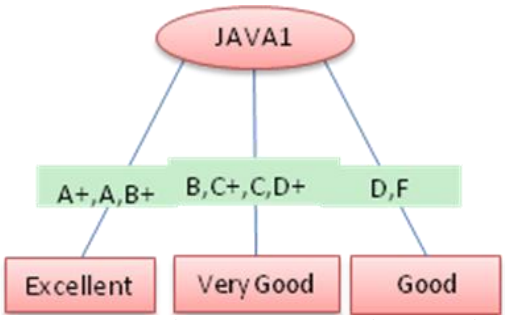

Fig. 4. The resulted J48 tree for predicting students final GPA after the third semester.

The knowledge gained from this tree may be questionable and not that reliable since it is too early to predict students final GPA based on their grades of their first specialized course. However, it can give some idea of what to expect and plans the future for improvement.

- The classification tree that predicts students final GPA after the fourth semester is shown in Fig. 5. In this semester students take 3 mandatory courses: Database principles, Assembly language and Java2 courses.

As it shown in the tree, database principles course is the root node, which makes it the most important course in this semester and a better predictor for the final GPA. Students are predicted to graduate with an excellent grade if they earned an A+, A or B+ in that course regardless of the other two courses taken in that semester.

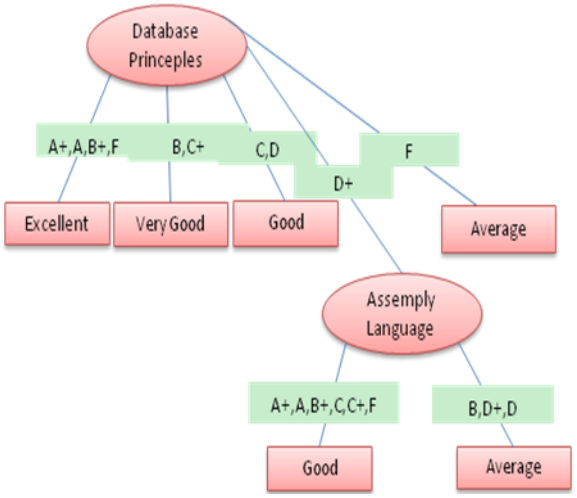

Fig. 5. The resulted J48 tree for predicting students final GPA after the fourth semester.

After the database course in the tree is the Assembly Language course. When students earned a D+ in database course, Assembly Language will improve the students final GPA to be good or average. There is no appearance of Java2 course in this tree, which indicates that Java2, compared to the database and assembly language courses, has no direct relationship with the final GPA.

- Students in the fifth level of their study plan take five mandatory courses: Data structure, HCI and Visualize Programming, Web Applications, Software Engineering-1 and Networks-1. The resulted tree after running the algorithm on these courses showed the Software Engineering-1 course to be the root node, which emphasizes our previous result regarding the importance of this course. The decision tree is too large to represent, however, below are some of the strong rules:

IF Software Enginerring-1=A+ OR A THEN final GPA=Excellent

IF Software Enginerring-1=B+ and (Networks1=A+ OR Networks1=A) THEN final GPA= Excellent

IF Software Enginerring-1=B THEN final GPA= Very Good

IF Software Enginerring-1=B and Networks1=B THEN final GPA= very good

---

- Mandatory courses taken by students in the six semesters are: Computer Architecture, Software Engineering-2, and Information Security. After running the algorithm on these courses, information security course becomes at the decision tree root. Furthermore, after the information security course is Computer Architecture course. For example, when the student received a C in information security, and an A+ in computer architecture student will earn very good as their final GPA. Fig. 6 displays the classification result.

The tree shows no indication of the Software Engineering-2 course, which implies that this course is the least related course to the final GPA.

- In the seventh semester, students are required to take 3 courses: Networks-2, Project-1, and Computer Ethics. Classifying students final GPA based on these courses, the resulted tree shows only Computer Ethics and Project-1 courses where the Computer Ethics as the root node. This indicates that in this semester these two courses are related to the final GPA and can be used to predict the students' final GPA regardless of their grades in Network-2. The resulted tree is too large to represent, however, below are some of the strong rules produced from the tree:

IF Computer Ethics=A+ THEN final GPA=Excellent

IF Computer Ethics=A and Project-1=A+ THEN final GPA= Excellent

IF Computer Ethics=A and Project-1=A THEN final GPA= very good

IF Computer Ethics=B+ THEN final GPA= Very Good

IF Computer Ethics =B THEN final GPA= Good

IF Computer Ethics=C+ THEN final GPA= Average

- In the final semester, students are required to take only one mandatory course, Project-2 and two elective courses, which was eliminated in this study. The resulted tree in Fig. 7 confirms the visualization result we mentioned above in section D Fig. 2.It shows students can have an Excellent final GPA if they earn a A+, a C+, a C or even a D in their Project-2. This is because Project-2 is their final course in their last semester and the students final GPA is a cumulative calculation of their performance in all courses in their study plan.

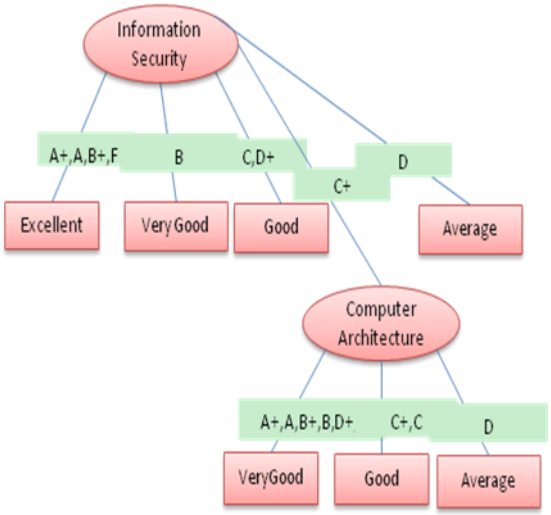

Fig. 6. The resulted J48 tree for predicting students final GPA after the sixth semester.

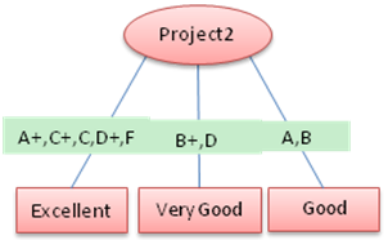

Fig. 7. The resulted J48 tree for predicting students final GPA after the final semester.

To summarize the results of the fourth research question, the following Table II shows the most important course in

each semester:

TABLE II: A SUMMARY OF THE MOST IMPORTANT COURSE IN EACH SEMESTER

Semester	Course Name
3	Java1 (the only specialized course)
4	Database Principles
5	Software Engineering 1
6	Information security
7	Computer Ethics
8	Project 2

V. CONCLUSION

In this paper, we presented a case study in educational data mining. It shows the potential of data mining in higher education. It was especially used to improve students' performance and detect early predictor of their final GPA. We utilized the classification technique, decision tree in particular, to predict students' final GPA based on their grades on previous courses. We discovered classification rules to predict students final GPA based on their grades in mandatory courses. We also evaluate the most important courses in the study plan that have a big impact on the students' final GPA. For future work, we will generalize the study and add the elective and general courses to get more accurate results. We will extend the experiment using other data mining techniques, such as neural network and clustering.

REFERENCES

[1] M. Al-Razgan, A. S. Al-Khalifa, and H. S. Al-Khalifa, "Educational data mining: A systematic review of the published literature 2006-2013," in *Proc. the 1st International Conference on Advanced Data and Information Engineering*, 2013, pp. 711-719.

[2] F. Siraj and M. A. Abdoulha, "Mining enrolment data using predictive and descriptive approaches," *Knowledge-Oriented Applications in Data Mining*, pp. 53-72, 2007.

[3] Q. A. Al-Radaideh, A. A. Ananbeh, and E. M. Al-Shawakfa, "A classification model for predicting the suitable study track for school students " *International Journal of Research and Reviews in Applied Sciences*, vol. 8, 2001.

[4] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance " *International Journal of Advanced Computer Science and Applications*, vol. 2, 2011.

[5] A. El-Halees, "Mining student data to analyze learning behavior: A case study," presented at the International Arab Conference of Information Technology (ACIT2008), Tunisia, 2008.

[6] A. Nandeshwar and S. Chaudhari. (2009). Enrollment Prediction Models Using Data Mining. [Online]. Available: http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf

[7] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybernetics and Information Technologies*, vol. 13, 2013.

[8] D. Garc´ıa-Saiz and M. Zorrilla, "Comparing classification methods for predicting distance students' performance," *The Journal of Machine Learning Research*, 2011.

[9] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "Data mining algorithms to classify students," presented at the the 1st International Conference on Educational Data Mining, 2008.

[10] Z. J. Kovačić, "Early prediction of student success: Mining students enrolment data," presented at the Informing Science & IT Education Conference, 2010.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten," The weka data mining software: an update," presented at the Special Interest Group on Knowledge Discovery and Data Mining, SIGKDD, 2009.

[12] E. Boretz, "Grade inflation and the myth of student consumerism," *College Teaching*, vol. 52, 2004.

[13] W. H¨am¨al¨ainen and M. Vinni, "Classifiers for educational data mining," *Handbook of Educational Data Mining*, 2010.

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

Mashael A. Al-Barrak is a lecturer in Natural Sciences and Engineering Department, at the College of Applied Studies and Community Service, King Saud University, Riyadh, Saudi Arabia. She received her master degree in information systems from King Saud University in 2013. Her research focuses are in data mining, educational data mining.

Muna Al-Razgan is an assistant professor in Information Technology Department, at the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. She received her PhD degree in information technology from George Mason University, VA, USA in 2008. Her research focuses are in data mining, educational data mining, and assistive technologies.