Journal of
Information Technology & Software Engineering

Research Article Open Access

Predicting Student Academic Performance in KSA using Data Mining Techniques

Nawal Ali Yassein[1], Rasha Gaffer M Helali[2*] and Somia B Mohomad[1]

[1]Community College, Najran University, King Abdulaziz Rd, Najran, Saudi Arabia
[2]College of Sciences and Home Economics, University of Bisha, Bishah, Saudi Arabia

Abstract

The main objective of higher education institutions is to provide quality education to its students. One way to achieve highest level of quality is to identify factors affecting academic performance and then trying to resolve weakness of these factors. The specific objective of the proposed research work is to find out if there are any patterns in the available data (student and courses records) that could be useful for predicting students' performance. The study involved a sample of 150 students collected from Najran University students in Saudi Arabia. The data was captured and arranged with the use of statistical package for social sciences (SPSS) and data mining tool (clementine). Developing an accurate student's performance prediction model is challenging task. Data mining based model were used to identify which of the known factors can give an early indicator of expected performance. This paper employs both feature reduction and classification technique to reduce error rate. The experimental results reveal significant relationships between including both practical work and assignments in course and its success rate. But, on the other hand the number of given assignment has a negative impact on course academic performance. In context of factors affect student academic performance, the most affecting factor is student attendance in class in addition to final exam and mid exam grades.

Keywords: Academic performance; Prediction; Decision tree; Educational data mining

Introduction

Student's academic performance affected by many factors, like personal, socio-economic and other environmental variable [1]. Knowledge about these factors and their effect on student performance can help managing their effect. Recently, much attention has been paid to educational mining research. Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people learning activities in educational environment [2]. Predicting student's performance becomes more challenging due to the large volume of data in educational databases [3]. The topic of explanation and prediction of academic performance is widely researched. The ability to predict student performance is very important in educational environments. Increasing student success is a long term goal in all academic institutions. If educational institutions can predict students' academic performance early before their final examination, then extra effort can be taken to arrange proper support for the low performing students to improve their studies and help them to success [3]. On the other hand, identifying attributes that affect course success rate can assist in courses improvement. Newly developed web-based educational technologies and the application of quality standard offer researchers' unique opportunities to study how students learn and what approaches to learning lead to success. The main objective of the paper is to identify both factors that affect courses success rate and student success rate then using these factors as early predictor to expected success rate and handling their weakness.

The subsequence sections are organized as follows: section II contains what had been done in the area of educational mining. Section III describes the proposed predictive model and used data set. Then the following sections highlight analysis and results. Finally, conclusion is presented.

Data Mining Techniques

Data mining is a computational method of processing data which is successfully applied in many areas that aim to obtain useful knowledge from the data [4]. Data mining techniques are used to build a model to identify new knowledge information [5]. There are several major data mining techniques have been developed and used including association, classification, clustering, prediction, sequential patterns and decision tree. The following is the description of main techniques used in the area.

Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into separate groups [6].

Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster [6].

***Corresponding author:** Rasha Gaffer M Helali, College of Sciences and Home Economics, University of Bisha, Bishah 67714, Saudi Arabia, Tel: + 00966583355160; E-mail: rasha_800@hotmail.com

Received November 04, 2017; **Accepted** November 13, 2017; **Published** November 20, 2017

Citation: Yassein NA, Helali RGM, Mohomad SB (2017) Predicting Student Academic Performance in KSA using Data Mining Techniques. J Inform Tech Softw Eng 7: 213. doi: 10.4172/2165-7866.1000213

Association

Association is one of the best-known data mining techniques. In association, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as relation technique [6].

Often two or more techniques are combined together to form an appropriate process that meets the business needs. Researchers used various classification methods in their studies to predict students' academic performance, such as decision trees, classification and regression trees, logistic regression, bayesian classification, support vector machine, neural network. Among these, decision trees gain popularity in predicting students' performance [2]. A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm (e.g. ID3, C4.5 etc.). The final result is a tree in which each branch represents a possible scenario of decision and its outcome. Among the decision tree algorithms C4.5 gains popularity in terms of its higher performance in Classification accuracy [7].

Related Work

A number of studies addressed the analysis of educational data to get useful information that affect learning quality. Baker and Yacef [8] summarized the goals of educational data mining in to: Predicting student's future learning behaviour, discovering or improving domain models, studying the effects of educational support and advancing scientific knowledge about learning and learners.

The implementation of data mining methods and tools for analysing data available at educational institutions, defined as Educational Data Mining (EDM) [9] is a relatively new stream in the data mining research. Extensive literature reviews of the EDM research field are provided by Romero and Ventura [9], covering the research done in the area between 1995 and 2005, and by Nithya [2], for the period after 2005. It is remarkable that most often attracting the attention of researchers and becoming the reasons for applying data mining at higher education institutions are focused mainly on retention of students, improving institutional effectiveness and enrolment management.

Classification techniques were used for education data modelling. There is an increase in their application within the last six years [10]. Researchers prefer to apply a single technique in their studies on student evaluations. Mustafa A [11] research in educational mining focuses on modelling student's performance instead of instructors' performance. One of the common tools to evaluate instructors' performance is the course evaluation questionnaire to evaluate based on students' perception. In this study, four different classification techniques, decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis are used to build classifier models. Brijesh and Saurabh [6] used decision tree classification technique to evaluate student's performance and extract knowledge that describes students' performance in end semester examination. Their goal is to identify the dropouts and students who need special attention early before final exams.

In the same direction, authors in [5] presented a model to predict student performance. They evaluate student success by passing grade at the exam. Parameters addressed for prediction including students' socio-demographic variables, achieved results from high school, the entrance exam, and attitudes towards studying which can have an effect on success Ramesh et al. [12] presented a valuable study to figure out factors influenced student success. They focused on parents' occupation and school type. Their obtained results from hypothesis testing reveals that type of school is not influence student performance and parents' occupation plays a major role in predicting grades [12]. The ability to predict students' mark could be useful in a great number of different ways associated with university-level learning. Farhana et al. [3] proposed a predictive model aimed to investigate if first semester performance is the most informative for their first year performance.

Agathe and Kalinain [13] discussed how to use data mining algorithms to help discovering relevant knowledge contained in databases obtained from Web-based educational systems. These findings can be used both to help teachers with managing their class, understand their students' learning and reflect on their teaching and to support learner reflection and provide proactive feedback to learners. An attempt was made by Minaei et al. To presents an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system [14]. Their method provides considerable usefulness in identifying students at risk early, especially in very large classes. Zekić-Sušac et al. created a model for predicting students' performance using neural networks and classification trees decision-making, and with the analysis of factors which influence Students' success [15]. From the above studies, we concluded that the student performance could depend on different factors such as demographic, academic, socio-economic and other environmental factors. It was also learnt that the variation in the predictive accuracy could be correlated with the nature of student dataset and data size. The following section describes suggested predictive model and experiments.

Course Predictive Model

The work is divided into two parts: first part related to identifying factors that affect course success rate, second part related to finding out predictors of student performance. Predictive model consists of two main phases:

Data collection and preparation phase

The data set used in this study was obtained from community college, computer science and business administration department at Najran University. Available data contains total of 108 Records represent courses taught in a period of two semesters and 150 Records taken from student results for academic year 2014-2015 and 2015- 2016. For each course 12 attributes were selected including (Course id, Credit hours, Practical work (if any), Assignments, number of assignments, number of Mid exams if any, Number of final exam questions, Education type, Level, study field and success rate) for student also 7Attributes were selected and recorded. Attributes selection done based on their ability to provide acceptable predictability. The domain values for some of the selected variables were defined for the present investigation to facilitate analysis process. Table 1 below provide list of attributes and their domains. Table 2 show part of courses and students datasets used for analysis.

Correlation is done between course and student attributes separately to find out the feature/s which has the obvious effect on success rate. For course features the correlation results reveals that there is a relation between practical sessions, assignments and success rate. That mean if course contains practical part success rate will be better from courses depends on theoretical part only also, courses contains assignments has better success rate than other courses. On the other hand, number of assignments has a negative correlation with will According to student correlation result we find that most attribute correlated to success

Course Features		
	Attributes	**Domain**
1	Course id	Numeric value
2	Credit hours	Numeric value
3	Practical work	{yes, no}
4	Assignments	{yes, no}
5	If yes how many assignments	Numeric value
6	Mid exams	{yes, no}
7	If yes how many Mid exams	Numeric value
8	Number of final exam questions	Numeric value
9	Education type	{Uniformity, Electronic, Both }
10	Is there a course description	{yes, no}
11	Study field	{Computer, Business Management }
12	Success rate	Numeric value
Student features		
1	Did student done courses assignments	{yes, no, part of assignments}
2	Student Attendance	{Good, Boor, Excellent}
3	Attending lab work	{yes, no}
4	exam marks	{Mark=60 pass, mark <60 fail, mark >60 v.good.}
5	Mid exam grades	{Good, Boor, Excellent}
6	Education type	{Uniformity, Electronic, Both }
7	Success rate	Numeric value

Table 1: Course/ student attributes.

ID	Assignments	Lab	Attendants	ExamM	MidGrade	Ed type	Success R
Sd1	1.00	2.00	2.00	34.00	8.00	2.00	72.00
Sd2	3.00	2.00	1.00	0.00	0.00	2.00	0.00
Sd3	1.00	2.00	2.00	32.00	7.00	2.00	65.00
Sd4	2.00	2.00	1.00	20.00	5.00	2.00	45.00
Sd5	1.00	2.00	1.00	0.00	4.00	2.00	0.00
Sd6	3.00	2.00	2.00	28.00	6.00	2.00	63.00

Table 2: Collected courses/student data.

rate is exam marks then mid grades, attendance, lab work then finally student performance in assignments. Tables 3 and 4 bellow show part of correlation results.

Data analysis phase

Our analysis is divided into two main steps. The first step is to determine the suitable mining technique to build the predictive models both classification and clustering in addition to feature selection techniques is suggested. For classification we benefit from previous studies findings.C5.4 algorithm was selected for classification and two-step clustering technique selected for clustering phase. The reason of selecting two-steps is its ability to divide data based on similarity without need to determining number of required clusters. This algorithm in addition to feature selection algorithm used to grantee proper attributes selection. The second step is applying the selected techniques on data then results are recorded. Both student and course data are divided into training and testing sets. Training set is labeled according to success rate. Two classes were suggested high and low. Figure 1 shows steps of course and student prediction model.

As mentioned total number of records fed to classifier is 108 course records and 150 student record. Records are randomly split into two sets, a training set and a testing set. The training set used to create the mining model. The testing set used to check model accuracy. Training data represents 40% of total data/records. Results are listed in the following section.

Results and Discussion

For predicting course success rate courses records fed firstly to

Success rate		
	P-value	**R-value**
Is there is assignments	0.002	0.035
#Assignments	0.054	-0.07
Education type	0.521	0.076
Level	0.00	0.402
# exam questions	0.231	0.141
Practical work	0.002	0.358

Table 3: Part of correlation results of course features.

Success rate		
	P-value	**R-value**
Assignments	0.134	0.165
Attendance	0.00	0.721
Lab work	0.00	0.513
Mid grades	0.00	0.793
exam marks	0.00	0.959

Table 4: Part of correlation results of student features.

	High	Low
Predicted high	24	0
Predicted Low	0	474

Table 5: Course prediction details.

	High	Low
Predicted high	24	0
Predicted Low	0	474

Table 6: Student prediction details.

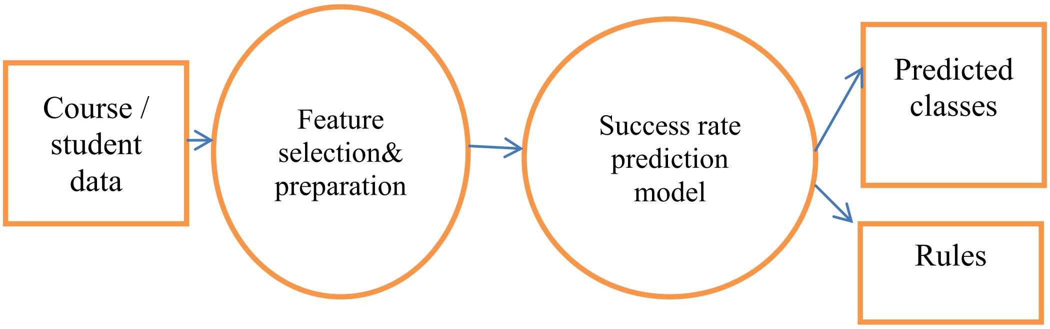
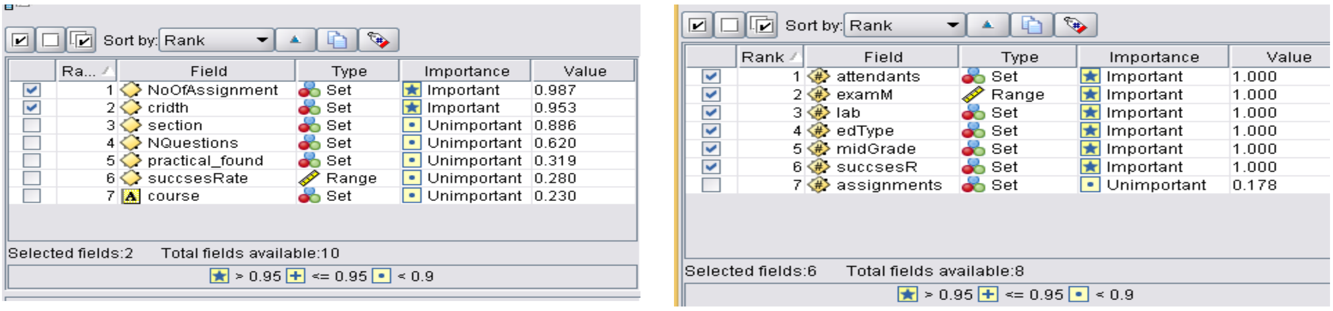


Figure 1: Prediction model.



Sort by: Rank

	Ra...	Field	Type	Importance	Value
☑	1	NoOfAssignment	Set	Important	0.987
☑	2	cridth	Set	Important	0.953
□	3	section	Set	Unimportant	0.886
□	4	NQuestions	Set	Unimportant	0.620
□	5	practical_found	Set	Unimportant	0.319
□	6	succsesRate	Range	Unimportant	0.280
□	7	course	Set	Unimportant	0.230

Selected fields:2 Total fields available:10

★ > 0.95 + <= 0.95 • < 0.9

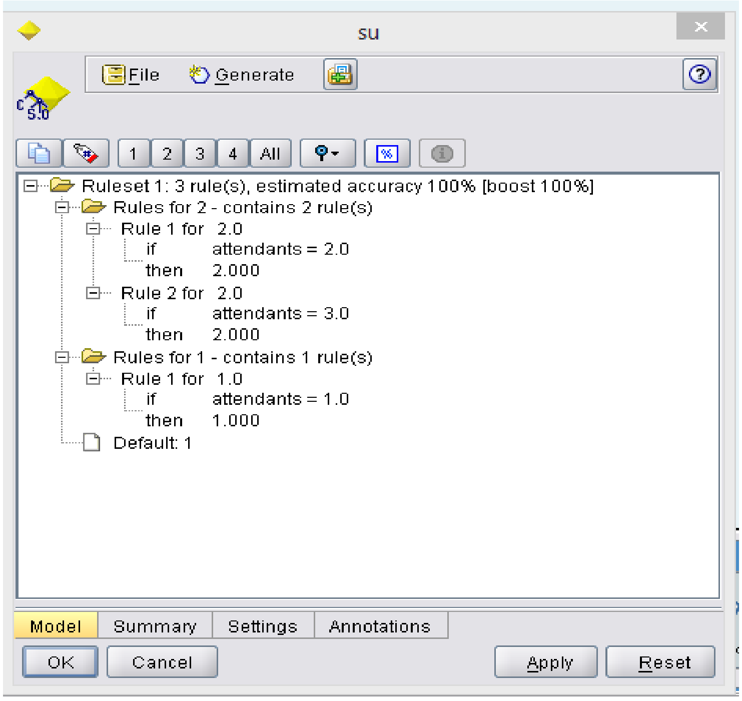
Sort by: Rank

	Rank	Field	Type	Importance	Value
☑	1	attendants	Set	Important	1.000
☑	2	examM	Range	Important	1.000
☑	3	lab	Set	Important	1.000
☑	4	edType	Set	Important	1.000
☑	5	midGrade	Set	Important	1.000
☑	6	succsesR	Set	Important	1.000
□	7	assignments	Set	Unimportant	0.178

Selected fields:6 Total fields available:8

★ > 0.95 + <= 0.95 • < 0.9

Figure 2: Ranking results.



```
Ruleset 1: 3 rule(s), estimated accuracy 100% [boost 100%]
    Rules for 2 - contains 2 rule(s)
        Rule 1 for  2.0
            if      attendants = 2.0
            then    2.000
        Rule 2 for  2.0
            if      attendants = 3.0
            then    2.000
    Rules for 1 - contains 1 rule(s)
        Rule 1 for  1.0
            if      attendants = 1.0
            then    1.000
    Default: 1
```

Figure 3: Part of prediction rules.

both tow-step clustering algorithm and feature selection algorithm. Feature selection technique ranked the given attributes/ features based on stored data and out put the most important features with their ranks descending. For courses 7 features were ranked based on their importance and 5 features were taken as neutral based on their values. For student data 6 featured were ranked as important and 1 as unimportant and 1 as neutral. It is important to note feature selection algorithm assign important if the rank is near or equal to 1 is this is not the case it labeled as unimportant. Figure 2 shows ranking results.

Clustering was done by applying the selected algorithm on dataset. The algorithm classifying data successfully in to two clusters by using the above features that confirms the suitability of selected attributes for prediction purpose. For supervised classification data were labeled according to success rate to either high or low. If course success rate is bigger than 65% then it labeled as high and if less that that labeled as low. Data are further fed to C5.4 classifier to find out which of selected attributes can be used as predictor to success rate. The previous steps are repeated for student success rate prediction. Learned rules show the existence of strong relation between practical work and success rate of courses and between student attendance and student success rate. Tables 5 and 6 shows rules reached from classification process. Prediction accuracy is estimated by used mining algorithm equal to 100% for both student success rate prediction and course prediction. Figure 3 below shows some of the prediction rules learned from classification algorithm.

Accuracy = acc =100%

Probability of false alarm = pf = 0%

Probability of detection = pd = recall = 474/474=1

Precision = prec = 474/474 = 1

Accuracy = acc =100%

Probability of false alarm = pf = 0%

Probability of detection = pd = recall = 474/474 = 1

Precision = prec = 474/474 = 1

Conclusion

The main objective of the study is the identification of highly influencing predictive variables on both the course and student performance and to reveal the high potential of data mining applications for university management, referring to the optimal usage of data mining methods and techniques to deeply analyse the collected historical data. Data mining techniques are widely used for extracting previously unknown patterns and finding relationship between different features. In this paper, a simple data mining based prediction model were presented. Model employs both classification and clustering techniques to identify features affecting student performance in selected course/s in order to assist academic stakeholders to improve academic performance which is the main goal of study.

References

1. Baradwaj BK, Pal S (2012) Mining educational data to analyze students' performance. IJACSA 2: 63-69.
2. Nithya P, Umamaheswari B, Umadevi A (2016) A survey on educational data mining in field of education. J Comput Sci Softw Dev 1: 1-6.
3. Sarker, Farhana, Thanassis T, Hugh CD (2013) Student's performance prediction by using institutional internal and external open data sources. CSEDU: 5th International Conference on Computer Supported Education, Germany.
4. Klosgen W, Zytkow J (2002) Hand book of data mining and knowledge discovery. Oxford University Press, New York.
5. Osmanbegovic E, Mirza S (2012) Data mining approach for predicting student performance. J Econ Bus 10: 3-12.
6. http://www.zentut.com/data-mining/data-mining-techniques/
7. Verma K, Singh A, Verma P (2016) A review on predicting students performance using data mining techniques. IJCESR 3: 127-132.
8. Baker, Ryan SJD, Yacef K (2009) The state of educational data mining in 2009: A review and future visions. JEDM 1: 3-16.
9. Altaher A, BaRukab O (2017) Prediction of student's academic performance based on adaptive neuro-fuzzy inference. IJCSNS 17: 165-169.
10. Kabakchieva D (2012) Student performance prediction by using data mining classification algorithms. IJCSMR 1: 686-690.
11. Agaoglu M (2016) Predicting instructor performance using data mining techniques in higher education.
12. Ramesh V, Parkavi P, Ramar K (2013) Predicting student performance: A statistical and data mining approach. IJCA 63: 35-39.
13. Agathe M, Yacef K (2005) Educational data mining: A case study. AIED, pp: 467-474.
14. Minaei-Bidgoli B, Kashy DA, Kortemeyer G, Punch WF (2003) Predicting student performance: an application of data mining methods with an educational web-based system. Frontiers in education.
15. Zekic-Susac M, Frajman-Jaksic A, Drvenkar N (2009) Neuron networks and trees of decision-making for prediction of eficiency in studies. Ekonomski Vjesnik 22: 314-327.