Predicting Student Performance Using Data Mining Approach: A Case Study in Oman

Sultan Juma Sultan Al Alawi[1], Jastini Mohd Jamil[2], Izwan Nizal Mohd Shaharanee[3]

School of Quantitative Sciences, Universiti Utara Malaysia

sultanjuma44@gmail.com

jastini@uum.edu.my

nizal@uum.edu.my

Article Info
Page Number: 1389-1398
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract

Oman Education Portal (OEP) storage large amounts of raw data which invariably contains usable information not yet discovered. Data mining give us techniques which can be used to analyze data to discover unseen information and pattern. The aim of this paper is to predict student performance by analyzing Oman Education Portal data using prediction analysis by employing J48 Decision Tree algorithm. Case study results on Oman Education Portal data show J48 to be efficient in predicting student performance.

Keywords: Data mining, Decision tree, J48 algorithm, Student performance.

Introduction

The academic performance of students in Oman has recently come under the spotlight. According to a report from Trends in International Mathematics and Science Study (TIMSS), The results of three TIMSS test (TIMSS 2007, TIMSS 2011, and TIMSS 2015) indicated that Oman students performed badly in all aspects of the tests (Mullis, 2015). This report has run alert at Oman government to come and discovered that there was a decline in the quality of education.

Improving student performance is a long term goal for Oman government. The costs of student failure rate are significant for Oman government. In 2009, recurrent expenditure on education accounted for 37% of all civil ministries’ recurrent expenditure (Education, 2013). As Oman government has spent a lot every year for education sectors the high failure rate could increase the cost. Oman government put both educator and decision makers in education institutions under pressure to come up with policies that could decrease failure rates on their institutions.

The main objective of education institutes is to provide quality education to its students and to improve the quality of managerial decisions (Council, 2014). One way to achieve highest level of quality in education system is by discovering knowledge from educational data to study the main attributes that may affect the students’ performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in education institutes such as, to enhance their decision-making process, to improve students’ academic performance and trim down failure rate, to improve teaching, to better understand students’ behavior and many other benefits.

A number of studies have been investigated the reason behind the issue of student failure (Khoshgoftaar et al., 2010, Mlambo, 2012). Most of these studies have been carried out to identify causal factors of low academic performance in a number of institutions worldwide. Their studies focus on the three elements that intervene, that is, students (personal causal factors), parents (family causal factors), and teachers (academic causal factors) (Al-Barrak & Al-Razgan, 2016). The combination of these factors influencing academic performance, however, varies from one academic environment to another, from one set of students to the next, and indeed from one cultural setting to another.

Oman Education Portal (OEP) platform is the only data storage that store large data about students' information and student learning process. This paper investigates the student performance using OEP data by developing decision tree model.

Related Works

Predicting student performances, in order to take actions and prevent failure, is an important issue for researcher in the educational data mining area. Final exam marks and assignment grades as well as student demographic, such as age, live location, gender, family background, students' personal behaviors, motivations and learning approaches, are characteristics that have been frequently used by researchers in predicting student performance. Online learning platforms and e-learning provide goldmine data about student learning style that can analysis to predict student performance. For example, the study by Jishan et al. (2015) predicted student performance using data from North South Bangladesh University that contains the students' demographics and academic records. They employed various classification techniques, including Naive Bayes, Decision tree and Neural Network in conducting this research. Yadav & Pal (2012) applied C4.5, ID3 and CART decision tree algorithms on engineering student's data to predict their performance in the final exam. The result show that decision tree can help to identify weaker students and to improve education strategy. Ramaswami & Bhaskaran (2010) applied CHAID decision tree algorithm on a dataset extracted from the higher secondary school education in India. They had identified student at risk of fail and the influence of the academic achievement. They concluded that CHAID decision tree is an efficient algorithm to generate rules of prediction model.

In summary, there is mixed evidence on whether the contribution of demographic variables to the early prediction of student success is significant or not. Therefore, student demographic data in OEP have been applied in this research to test whether the data can give significant result to predict student performance.

Oman Educational Portal

In line with Oman's "National Strategy" for a digital society, the Ministry of Education in 2007 has created a comprehensive Electronic Management System called Oman Education Portal (OEP) (MOE, 2009). The portal serves as Oman's educational gateway by providing access to a group of program and services through the use of the internet. It is intended to serve as the entry point and one-stop site for everyone interested in education - parents, students, teachers, administrators, entrepreneurs or another ministry (MOE, 2009).

OEP is an electronic communication system that provides high-speed transmission for the exchange of information, ideas, experiences and views on the educational process. The portal, which represents a quantum leap in the use of modern information technology in education, involves the development of a comprehensive system that will complement the ministry's vision for deep-rooted educational reform in Oman. OEP used by ministry of education to facilitate all aspects of learning: creation and shearing of educational content, collaboration and communication, class monitoring and administration. The development of the portal seeks to improve communication, which in the past was hindered by complex administrative procedures.

OEP keeps lots of records for every student, and that information is kept in various formats on different systems. There is data related with the registration process with detailed information about each student, as well as large amounts of data animatedly created during the academic year related to student's activity. Data that is stored is a valuable resource, whose value can be further increased with the development of data mining techniques. It can offer new insights and give lots of opportunities to make better evaluation of students. It can also enable educators and decision makers to improve and enhance teaching methods, techniques and quality of offered study materials.

Data Mining Process

Many organizations in the mainstream of business, industry, and the public sector already rely heavily on the use of data mining as a way to search for relationships that would otherwise be "hidden" in their transaction data (Kovačić, 2010). In past few years there was a rush to develop data mining algorithms that were capable of solving all the problems of searching for knowledge in data. From the viewpoint of Data Mining & Knowledge Discovery process models, the year 2000 marked the most important milestone: SEMMA (Milley et al., 1998) was published in 1998 developed by the SAS Institute. SAS Institute further divides data mining into five stages that are represented by the acronym SEMMA. A general process of SEMMA usually followed as sample, explore, modify, model and assess (Corrales et al, 2015). Fig. 1 is an overview of each step in the SEMMA methodology:

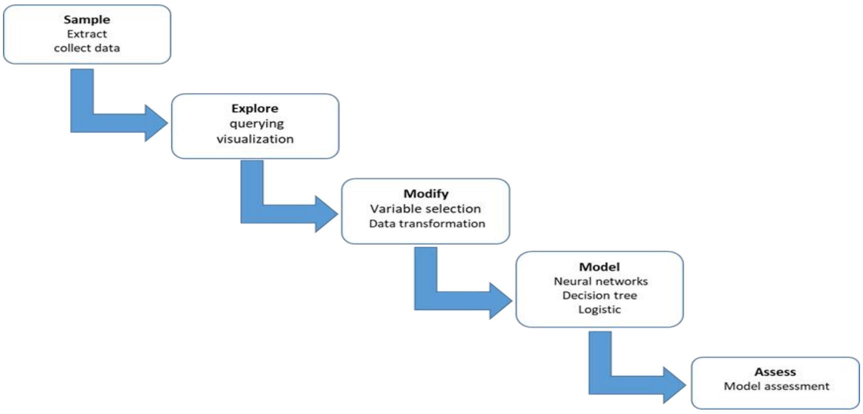

Fig. 1 Steps in the SEMMA Methodology

Sample

The first step of SEMMA data mining process is extracting portion of significant information from huge datasets where this sub dataset will be containing the important information that can be easy to manipulate. In the case of very large datasets, mining a representative sample instead of the whole volume may drastically reduce the processing time required to get significant information. If general patterns appear in the data as a whole, these will be traceable in a representative sample. It is important the sub data should be reflected the big picture of datasets.

Explore

The next step is to explore the data visually or numerically for inherent trends or groupings. Exploration helps refine and redirect the discovery process in order to gain understanding and ideas. If visual exploration does not reveal clear trends, one can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering. Exploring can discover richer patterns that may not be strong enough to be detected if the whole dataset is to be processed together.

Modify

In this stage the user selects and transforms the variables in order to build reliable and robustness model. The modify stage focus on the model construction process. By using the patterns that discovering in the exploration phase, one may need to manipulate data to include information such introduce new variables or grouping the users. It may also be necessary to look for outliers and reduce the number of variables, to narrow them down to the most significant ones.

Model

After preparing the data, models that explain patterns in the data will be constructed. Modeling techniques in data mining include artificial neural networks, decision trees, rough set analysis, support vector machines, logistic models, and other statistical models – such as time series analysis, memory-based reasoning, and principal component analysis. Each type of model has particular strengths and is appropriate within specific data mining situations depending on the data.

Assess

The final stage of the SEMMA is assess the model. The model assessment done by the user to estimate how well model performs. A common strategy of assessing a model is by dividing the data into training and testing data, training data used to build the model and testing data used to assess the model. The aim of assessing stage is to evaluate model accuracy and consistency with the results obtained for the training data set. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model.

Methodology

The aim of this research was to predict student performance using J48 decision tree algorithm on data extracted from Oman Educational Portal (OEP) Database. OEP include student data, such as demographic information and academic information.

Because ministry of education need to understand factors that impact in student performance, OEP dataset may be offer worth data that can help ministry to achieve their goal. Understanding students' population and patterns in the OEP data becomes necessary before developing a predictive model. In this phase, the following questions will be solved: what is the profile of a student who successfully pass the final exam? Can we detect successful vs unsuccessful student by using of demographic variables (such as gender, age or nationality) or academic variable (such as student final exam marks)?

Sampling

Data analyzed in this study was extracted from tracking data captured from OEP database in academic year 2019/2018. OEP database have large dataset including administrator data, teacher data, parent data, schedule data and many more. The data can be downloaded in an Excel format. The student dataset contains 49588 student records with demographic and academic information.

Explore and Modify

In this stage, preprocessing techniques will be conducted in order to obtained complete dataset with better condition. First step is omitted irrelevant attributes. OEP database contained huge data and to reduce size of data it is important to select only related attributes to build the model. For example, this study used "studentID" to identify students and omitted other attributes like "first name", "Second Name", "Family Name".

OEP dataset as real data set was contained both numeric and nominal attributes, a feature which usually complicates analysis. In second step, numeric and nominal will be transformed to categorical variable in order to be easy for data understanding. For example, we changed age variables from date form in to three groups (under class age, in class age, above class age). For nationality variable we changed in to two groups (Omani and non-Omani). For size of school and class we used three groups (low, medium, high). For target variable which present student performance, we considered only two possible outcomes, labeled as: Pass and Fail. Students labeled Pass successfully pass the final exam and students labeled Fail unsuccessful fail in final exam. As the total final exam marks in Oman is 1100 the two possible values are Pass from 1100 to 450 and fail under 450. The attributes description presents in the Table 1.

Table 1. Variables Description

Variables	Values	Variable Description
Region	urban, rural	The location of living student
School Shift	morning, evening	School work time at morning or evening

Nationality	Omani, non-Omani	Student nationality
Gender	Male, female	Student gender
Religion	Muslim, non-Muslim	student religion
Age	under class age, in class age, above class age	Students' age in group
Attendance	normal, medium, high	student absences in year
Class size	low, medium, high	Number of students in the class
School Size	low, medium, high	Number of students in school
Teachers ratio	low, medium, high	student/teacher ratio
Student Performance	Pass, fail	Final exam marks

As part of the data understanding phase, we carried out the cross-tabulation for each variable. The exploratory analysis gives a clear understanding of how the attributes are distributed in the student dataset. The Table 2 reports the results. For instance, student's categorization based on region, 80.7% is from rural area and 19.3% is from urban area. For student nationality, 96.4% was Omani and 3.6% non-Omani. In gender variable we found that, 50.7% are male and 49.3% are female. The distribution of each variable used in this study can be found in Table 2.

Table 2. Distribution of OEP dataset by each variable

Variables	Values	Number of instances	Proportion (%)
Region	rural	40028	80.7
	urban	9560	19.3
School Shift	Evening	573	1.2
	Morning	49015	98.8
Nationality	Omani	47790	96.4
	non-Omani	1798	3.6
Gender	female	24423	49.3
	male	25165	50.7
Religion	Muslim	49576	99.98
	Non- Muslim	12	0.024
Attendance	High	621	1.3
	Medium	794	1.6
	Normal	48173	97.1
Age	under class age	19	0
	in class age	43495	87.7
	above class age	6074	12.2
Class Size	High	11846	23.9
	Low	665	1.3
	Medium	37077	74.8
School Size	High	19346	39

	Low	4988	10.1
	Medium	25254	50.9
Teachers Raito	High	17104	34.5
	Low	3054	6.2
	Medium	29430	59.3
Performance	Pass	48476	97.8
	Fail	1112	2.2

Model

The next step in SEMMA is to build the classification model using the decision tree method. The decision tree is a very good and practical method since it is relatively fast, and can be easily converted to simple classification rules (Li et al., 2015). The decision tree method depends mainly on using the information gain metric which determines the attribute that is most useful. The information gain depends on the entropy measure.

In this study, J48 Decision Tree Classifier have been used to predict student performance. J48 algorithm is a learning algorithm that builds decision trees using an entropy-based splitting criterion stemming from information theory (Seiffert, Khoshgoftaar, Hulse& Van., 2009). It improves upon ID3 by adding support for both tree pruning and for dealing with missing values and numeric attributes (Yadav & Pal, 2012). J48 is the Weka implementation of C4.5 models were built using the default parameters provided by Weka (Al-Barrak & Al-Razgan, 2016).

Results

Fig. 2 present tree diagram of J48 that was generated by WEKA. The highest information gain is used at the root of the tree. The procedure is repeated until the leaf node is created for the tree specifying the class attribute that is chosen.

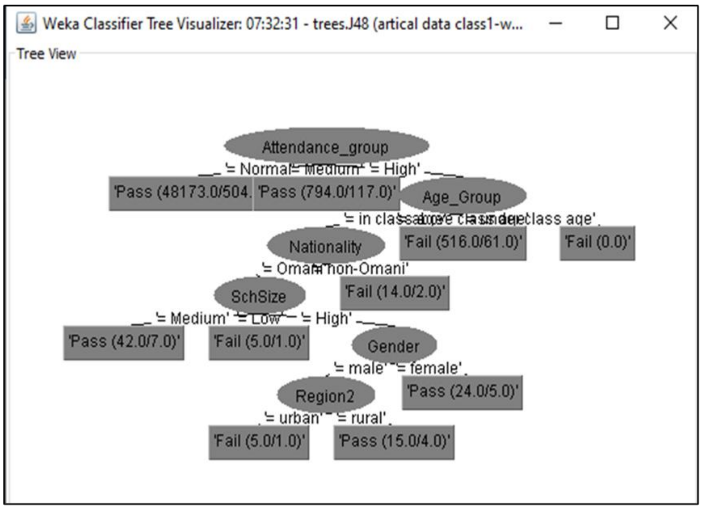

Fig. 2 J48 Tree Diagram

Figure 2 depicts the rules that resulted from applying the J48 classification algorithm on the final exam marks of the student as a target class. As it is seen from the figure, the attributes that influence the category of the student performance are the Attendance, Age, Nationality, School size, Gender and Region.

Table 3. Performance Results from J48 Classification Algorithm – Cross Validation

Stratified cross-validation	%
correctly classified instances	98.6
incorrectly classified instances	1.4
Kappa statistic	0.56
Mean absolute error	0.03

From Table 3 It was found that the overall model prediction accuracy of J48 prediction model was 98.6% and it indicated that the J48 model could correctly classify 48877 students among 49588 students with (98.6%) accuracy and only (1.4%) for incorrect instance. In addition, the Kappa statistic= 0.5614 which presented observed Accuracy with expected Accuracy and Mean Absolute Error= 0.0271 which present average of the absolute error between observed and forecasted value. The model presented a high accuracy at 98.6% for predicting the students' performance.

Table 4. Confusion Matrix – J48 Prediction Model

	Predicted	
Actual	Pass	Fail
Pass	48410	66
Fail	645	467

The Confusion Matrix has been presented in Table 4, which compared the actual and predicted classifications, the values represent following:

1. Number of correct predictions that the instance tested positive = 48410

2. Number of incorrect predictions that the instance tested negative = 645

3. Number of incorrect predictions that the instance tested positive = 66

4. Number of correct predictions that the instance tested negative = 467

Subsequently, the 10-fold cross method for the assess of the model was applied during J48 prediction model construction process. The Table 5 shows the accuracy as follows:

Table 5. Assessment of J48 decision tree

Assessment	TP	FP	Precision	Recall	F-Measure	ROC
J48	0.986	0.567	0.984	0.986	0.983	0.76

Table 5 shown that the overall model prediction accuracy of J48 prediction model was 98.6%. The values of Precision, Recall and F-Measure was 98.7, 98.6% and 98.3 respectively. Finally, the value of ROC was 76%. From the result it is clear that J48 decision tree have significant accuracy and suitable for predicting student performance.

Conclusion

The student achievement is an important real-world educational problem. Detection of student at risk of fail is the key for education enhancement. The aim of this study is predict student performance using J48 decision tree algorithm by analyzing OEP data set. J48 decision tree is suitable for data belonging in various distributions. This paper shows how J48 Decision Trees is used to build student prediction model. From the classifiers accuracy of J48 it is clear that the true positive rate of the model is 98.6% that means model is successfully identifying the students who are at risk of fail.

For sure more research needs to be done using this analytic tool in more populous and diverse learning settings, but this research provides strong evidence that the J48 algorithm based OEP data can effectively use the proposed predictors student achievement. This study will help to the educators and decision makers in ministry of education in Oman to improve the division of the student and take special attention to reduce fail ration and taking appropriate action. These students can be considered for proper counseling so as to improve their result.

Acknowledgment

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Mullis. (2015). TIMSS 2015 International Results in Mathematics.

[2] Education, M. of. (2013). Education in Oman :The Drive for Quality.

[3] Council, T. E. (2014). The Most Remarkable Projects Developed by The Education Council. The Education Council.

[4] Khoshgoftaar, T. M., Gao, K., & Seliya, N. (2010). Attribute selection and imbalanced data: Problems in software defect prediction. Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 1, 137–144. http://doi.org/10.1109/ICTAI.2010.

[5] Mlambo, V. (2012). An analysis of some factors affecting student academic performance in an introductory biochemistry course at the University of the West Indies. The Caribbean Teaching Scholar, 1(2), 79–92.

[6] Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees: A Case Study. *International Journal of Information and Education Technology*, *6*(7), 528–533. http://doi.org/10.7763/IJIET.2016.V6.745

[7] Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. Decision Analytics, 2(1), 1–26. http://doi.org/10.1186/s40165-014-0010-2

[8] Yadav, S. K., & Pal, S. (2012). Data Mining : A Prediction for Performance Improvement of Engineering Students using Classification. Wcsit, 2(2), 51–56.

[9] Ramaswami, M., & Bhaskaran, R. (2010). A CHAID Based Performance Prediction Model in Educational Data Mining. International Journal of Computer Science Issues, 7(1), 10–18. Retrieved from http://arxiv.org/abs/1002.1144

[10] MOE. (2009). ICT and Education in the Sultanate of Oman.

[11] Kovačić, Z. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. Proceedings of Informing Science & IT Education Conference, 647–665.

[12] Milley, A. H., Seabolt, J. D., & Williams, J. S. (1998). Data Mining and the Case for Sampling. A SAS Institute Best Practices. SAS Institute, 1–36. Retrieved from http://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf

[13] Corrales, D. C., Ledezma, A., & Corrales, J. C. (2015). A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal. Journal of Computers, 10(6), 396–405. http://doi.org/10.17706/jcp.10.6.396-405

[14] Li, Y., Sun, J., & Qiang, W. (2015). Application of Data Mining in Personalized Remote Distance Education Web System. The Open Cybernetics & Systemics Journal, 1769–1775.

[15] Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). Mining Student Data Using Decision Trees.

[16] Letouzé, E. (2012). Big Data for Development: Challenges & Opportunities. Global Pulse Is a United Nations Initiative, (May), 47. Retrieved from http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

[17] Seiffert, C., Khoshgoftaar, T. M., & Hulse, J. Van. (2009). Hybrid sampling for imbalanced data. *Integrated Computer-Aided Engineering*, *16*, 193–210. http://doi.org/10.3233/ICA-2009-0314