Predicting Student Enrolments and Attrition Patterns in Higher Educational Institutions using Machine Learning

Samar Shilbayeh and Abdullah Abonamah
Business Analytics Department, Abu Dhabi School of Management, UAE

Abstract: *In higher educational institutions, student enrollment management and increasing student retention are fundamental performance metrics to academic and financial sustainability. In many educational institutions, high student attrition rates are due to a variety of circumstances, including demographic and personal factors such as age, gender, academic background, financial abilities, and academic degree of choice. In this study, we will make use of machine learning approaches to develop prediction models that can predict student enrollment behavior and the students who have a high risk of dropping out. This can help higher education institutions develop proper intervention plans to reduce attrition rates and increase the probability of student academic success. In this study, real data is taken from Abu Dhabi School of Management (ADSM) in the UAE. This data is used in developing the student enrollment model and identifying the student's characteristics who are willing to enroll in a specific program, in addition to that, this research managed to find out the characteristics of the students who are under the risk of dropout.*

Keywords: *Machine learning, predictive model, apriori algorithm, student retention, enrolment behaviour, association rule mining, boosting, ensemble method.*

Received March 10, 2020; accepted July 19, 2020
https://doi.org/10.34028/18/4/8

1. Introduction

Machine learning can play a critical role in building up a strategic enrolment plan for academic institutions. This includes detecting the enrolment trends over time, which gives the institution a better ability to manage student enrolments and align institutional resources more efficiently and effectively.

Higher education institutions are increasingly interested in identifying their enrolment pattern to understand the factors that maximize their enrolment numbers. These factors allow the student recruitment department to identify, attract, engage and enrol the right students. The direct result of this is to help the institutions to improve their financial and academic performance. In academic institutions, student attrition continues to be a significant and costly challenge. It is considered to be the main concern that negatively impacts the institution's reputation, ranking, and success. The attrition rate in the academic institutions is inspired by churn analysis that is usually performed in marketing analysis. It is found out that attracting a new customer in an organization is much more expensive than retaining existing ones. In the same way, for any academic institution, maintaining the current students is one of their higher concern. This can be accomplished by understanding the student attrition pattern by detecting the students who are under the risk of attrition, then building up a general model that can predict this incident in advance. In this way, they can develop proper strategies to retain those students and give them more support if needed.

This paper closely investigates the enrolment pattern that helps in attracting the type of students who are most likely to enrol in one of the Abu Dhabi School of Management (ADSM) academic programs and also detects the student attrition patterns.

2. Literature Review

There are many enrolment and attrition rate prediction studies that have utilized machine learning approaches to identify the student enrolment and attrition pattern. Petkovski *et al*. [7] conducted a study focused on the student's performance as an indicator for the student dropout rate in their freshman year. This study uses naïve Bayes, decision tree, and rule-based induction machine learning algorithm to build the best model that predicts the attrition rate with 80% accuracy. Yukselturk *et al*. [12] examined on their study the dropouts rate in an online program for 189 students who registered in the online information technology certificate program. This study uses four machine learning algorithms K Nearest Neighbour KNN, Decision Tree (DT), Naive Bays (NB) and Neural Network (NN), resulting in 87% accuracy for NN, and 80% accuracy for DT and finding out that the student demographics information plays a very critical role in their dropout rate. Rai and Jain [8] performed a study for the aim of understanding the student drop rate factors using machine learning algorithms utilizes ID3 and J48 decision trees using weka on 220 undergraduate students in Information Technology courses. In this research; it is found out that personal factors is the most important factor that effects the student attrition and weights for 28% of dropout rate. On the same study, the institutional factors such as the university environment, and the course cost weight for 17% of dropout rate, also it is noticed that few students are most likely to drop out due to homesickness and adjustment problems [8]. In the same prospect; kemper *et al*. [4] managed to predict the student dropout rate and success factor using logistic regression and decision trees on a resample examination data at the Karlsruhe Institute of Technology (KIT) with 95% accuracy after three semesters of the student enrolments.

On the other hand, multiple authors and studies use machine learning predictive methods and data science techniques to predict student behaviour and performance in educational settings. The researchers Mulugeta and Berhanu [5] performed three different machine learning algorithms: mainly J48, naïve bayes, and neural network to build three prediction models that predict the student's enrolment at the department level for higher education institutions in both private and public universities. The developed model in [5] shows that the highest model accuracy is reached by applying neural network machine learning algorithm over J48 and naïve Bayes. In the same prospect; Nakhkob and Khademi [6] conducted a study that utilized neural network, bagging, boosting, and naïve bayes to predict the rate of student enrolment in higher education in different universities in Iran. In the same study, a comparison between those four models is performed using model accuracy, and ROC curve. The result shows that bagging outperforms neural network, boosting and naïve Bayes.

Another study conducted by Waters and Miikkulainen [11] utilized machine learning algorithms to predict how likely 588 Ph.D. applicants will be admitted to their colleges based on their information provided in their applications, the result is used to reduce the admission process time.

Two machine learning approaches: logistic regressions, and decision trees are performed to predict student dropout at Karlsruhe Institute of Technology (KIT), both used models yield high prediction up to 95% after three semesters, this study is done by Kemper *et al*. [4]. For the same aim; a study is done by Al-Shabandar *et al*. [2]. This study utilized two machine learning models to find out the students who under the risk of failure and withdraw at the early stages of online courses. Random forest on decision tree and gradient boosting machine are used for this aim and result in 89 % and 95% accuracy respectively.

3. Research Methodology

3.1. Data Description and Understanding

Data understanding and description is an important stage in data analysis, which includes gathering, understanding, cleaning, describing and verifying the data. More accurate, comprehensive and higher quality data will result in more efficient output. The data used for this research is taken from the Student Information System (SIS) database of Abu Dhabi School of Management (ADSM). This particular dataset is used to capture information about the students who enrolled in ADSM along with their demographics and academic background information for 1600 students enrolled from (2013-2018). Student dataset variables are described in Table 1.

Data Cleansing process includes filling up the missing values, deleting the outliers or incorrectly input data to make sure that the models will be built on clean, proper data that can be converted into interesting results.

Table 1. Student dataset variables.

Variable name	Description
Program	This includes four master programs available at ADSM
Workplace	The student workplace
UGSChool	Student undergraduate school
UGMajor	Student undergraduate major
Status	Student enrolment status (active, inactive, under process, admitted)
UGCountry	Student undergraduate country
Experience Year	Number work experience years
Age	Student age
WorkPlaceSector	Student workplace sector divided into (private, public)
Position	Working student current position
UGGPA	Undergraduate GPA
Emirates	The emirate region (state)
Nationality	Student Nationality
ELR Score	English Test Score
Sponsor	The name of the organization sponsoring the student (if any)
Start Year	The student starting year in the academic program

3.2. Predictive Analysis

For the aim of developing the proposed models, this research is divided into three stages, the following is the description for each stage with its results:

- *Stage* 1: For the aim of predicting the number of enrolment for the coming four years (2019-2022), a boosted regression tree algorithm is applied to six years of data on 1600 students. The boosted tree is compared with the single regression tree. The boosting method initially is defined by Schapire and Freund [10] as a classification supervised ensemble method. Boosting is one type of ensemble approaches and defined as the method of applying the same machine learning algorithm (regression decision tree in our case) in the whole dataset several times. On each learning stage; the weight of correctly classified instances is decreased, and the weight of incorrectly classified instances is increased to give those that are incorrectly classified in the previous learning stage more attention in the next learning process [3]. Figure 1 shows the stage 1 framework.

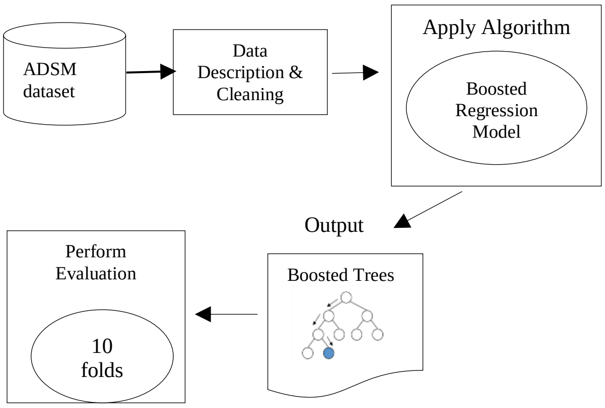


Figure 1. Stage 1 framework.

The final result of the boosted regression model prediction is the average of those different created models. In our work, 500 regression trees are created. The final result of applying boosted regression model is shown in Figure 3. The boosted regression model is tested using 10 folds cross validation and gives 89% accuracy which outperforms the single regression decision tree that gives only 76% accuracy tested using 10 folds cross-validation as well.

- *Stage* 2: The association rule Apriori algorithm is applied to the given dataset for the aim of revealing interesting patterns about the characteristics of the students who are most likely to enrol in one of the school academic programs. Association rules mining is one of the machine learning techniques that is used to reveal interesting and meaningful hidden patterns between different dataset variables. The developed rules are interestingly unexpected for business stakeholders. Apriori algorithm is defined by Agrawal *et al*. [1]. Apriori algorithm is used to uncover interesting relations between different items in the market basket analysis. In Apriori algorithm, a predefined interestingness measurements called support and confidence are identified to limit the number of rules generated [9]. The generated rule could be written as: If {X} Then {Y}.

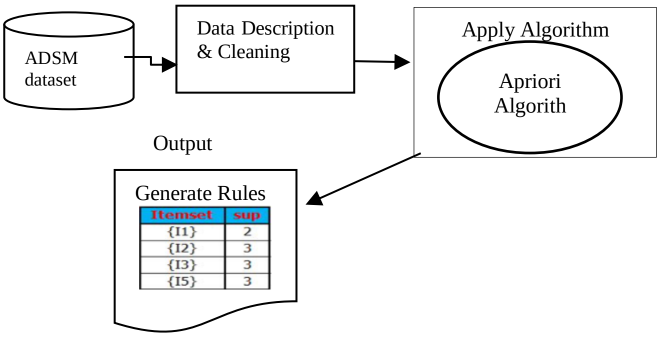
The If part of the rule (the {X} above) is known as the antecedent and the THEN part of the rule is known as the consequent (the {Y} above). The antecedent is the condition and the consequent is the result. The support is defined as the number of transactions that include items in the $\{X\}$ and $\{Y\}$ parts of the rule as a percentage of the total number of transactions N. It is a measure of how frequently the collection of items occur together as a percentage of all transactions. Support formula is shown in Equation (1)

$$Support = \frac{Frequency(X,Y)}{N} \tag{1}$$

Confidence is defined as is the ratio of the number of transactions that include all items in $\{Y\}$ as well as the number of transactions that include all items in $\{X\}$ to the number of transactions that include all items in $\{X\}$. Confidence formula is shown in Equation (2)

$$Confidence = \frac{Frequency(X,Y)}{Frequency\ (X)} \tag{2}$$

The two measurements support and confidence reveal the degree of interestingness for the generated rules. The user should define a minimum threshold that meets his/her requirements for every set of generated rules. The generated rules should assure minimum support and minimum confidence that is greater than the user predefined thresholds. Figure 2 shows stage 2 framework.



Itemset	sup
{I1}	2
{I2}	3
{I3}	3
{I5}	3

Figure 2. Stage 2 framework.

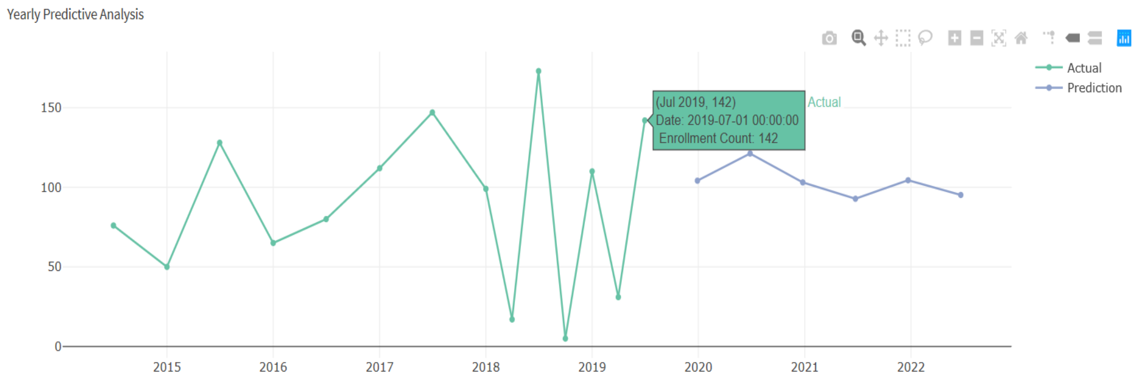


Figure 3. Enrollments number prediction using boosted regression trees (2019 -2022).

Table 2 shows the minimum thresholds defined for each ADSM programs, the measurements vary according to the number of the students enrolled in each program.

Table 2. Minimum association rules thresholds defined for each ADSM program in stage 2.

Program	Minimum Support	Minimum Confidence	Rule Generated
MBA (Master of Business Administration)	0.5	0.95	20
MSBA (Master of Science of Business Analytics)	0.05	0.90	10
MSLOD (Master of Science in Leadership and Organizational Development)	0.4	0.90	15
MSQBE (Master of Science in Quality and Business Excellence)	0.3	0.95	10

Table 3 shows the sample of the association rules that are revealed from our dataset after specifying the previous threshold for the student's characteristics in MBA Program.

Table 3. Sample of rules generated using apriori algorithm for the student's characteristics in the MBA Program.

LHS	RHS	Support	Confidence
{Status=Graduated, workplaceSector=Government}	{MBA}	0.6708861	0.9814815
{UGCountry=United Arab Emirates, workPlace=OilGasSector}	{MBA}	0.5063291	0.9756098
{UGSchool=UAE UNIVERSITY,workplace=Education}	{MBA}	0.5000000	0.9753086
{gender=male, Nationality=United Arab Emirates, position=Admin, age = [30-35] }	{MBA}	0.6582278	0.9904762

Figure 4 is the sample of the association rules that are revealed from our dataset after specifying the previous threshold for MBA.

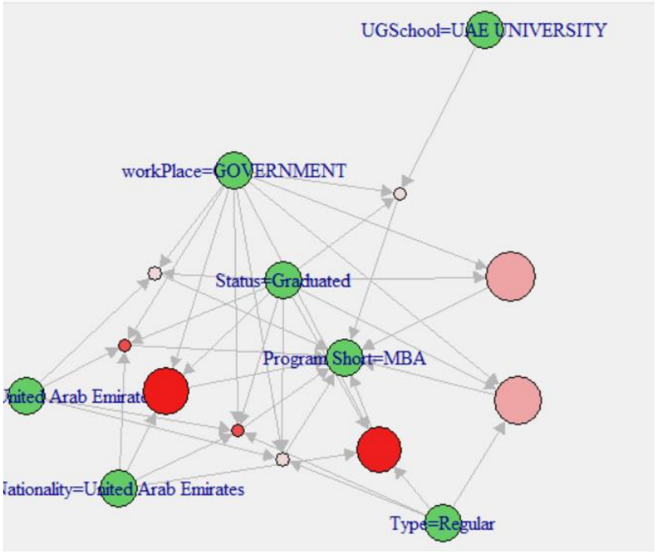


Figure 4. Sample of rules generated using apriori algorithm for the students in the master of Business Administration Program (MBA) program.

- *Stage* 3: For the aim of detecting the students who are under the risk of attrition, association rule Apriori algorithm is applied to the given dataset with the following minimum threshold defined in Table 4. The measurements vary according to the number of the students drop out from each program in the past five years. Stage 3 framework is shown in Figure 5

Table 4. Minimum association rules thresholds defined for each ADSM program in stage 3 for attrition rate detection.

Program	Minimum Support	Minimum Confidence	Rule Generated
MBA (Master of Business Administration)	0.01	0.8	28
MSBA (Master of Science of Business Analytics)	0.05	0.90	10
MSLOD (Master of Science in Leadership and Organizational Development)	0.4	0.90	15
MSQBE (Master of Science in Quality and Business Excellence)	0.3	0.95	10

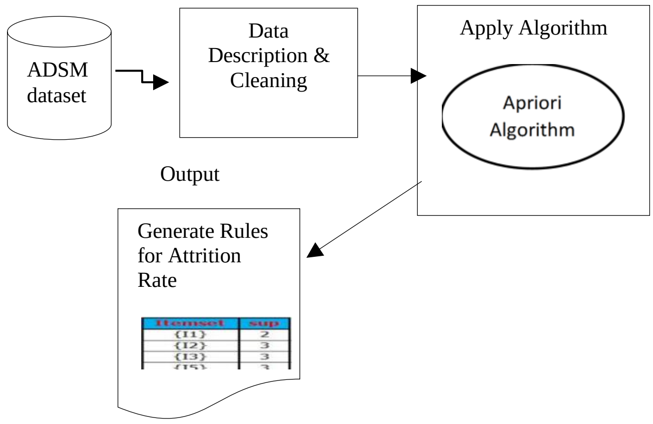


Figure 5. Stage 3 framework.

Table 5 shows the sample of the association rules that are revealed from our dataset after specifying the previous thresholds for the student's attrition in MBA program.

Table 5. Sample of rules generated using apriori algorithm for the students attrition rate in MBA program.

LHS	RHS	Support	Confidence
{Age=30-40,Experience.Years=< 2,UGCountry=United Arab Emirates,UGGPA=2.5-3,ELR.Score=4-5,UGcollege=College of Business and Economics}	{AttritionStatus= Attrition}	0.012	1.0
{UGGPA=2.5-3,ELR.Score=4-5,UGSchool=UAE UNIVERSITY,UGcollege=College of Business and Economics}	{AttritionStatus= Attrition}	0.012	1.0
{Programs=MBA,Age=30-40,gender=Male,UGGPA=2.5-3,ELR.Score=4-5,position=NOT EMPLOYEE}	AttritionStatus= Attrition}	0.014	0.818

3.3. Results Comparison

Given the fact that the admission cycle for year 2019 and 2020 is over. In this section the boosted regression model accuracy is computed by comparing the value of predicted and the actual enrolments that is done over 2019 and 2020 academic year. The boosted regression model that is developed for the aim of predicting the number of enrolments students for the academic years (2019 to 2022) and explained in section 3.1 stage 1.

It is found out that the model results in 86% accuracy in 2019, and 78% accuracy in 2020. The drop in the model accuracy in 2020 may occurred because of unexpected corona virus incident that effects the student's enrolment pattern.

4. Conclusions

The models we presented in this paper have broad significance for the higher educational institutions. They attempt to answer two important questions facing almost every institution of higher education, these are: What type of student the academic institution should attract and maintain to keep their academic standards?" and "What type of students are most likely to drop out during their academic journey?"

In the context of our work, we managed to answer these two questions using a set of machine learning approaches. The work described in this paper was held at ADSM and is aimed to identify and prioritize students who are at risk of attrition. Although the work in this paper is aimed to predicting students who are likely to enrol at specific academic program, we believe that this solution (problem formulation, the feature extraction process, data pre-processing, predictive analysis, and evaluation) applies and generalizes to other academic outcomes as well, such as predicting student behaviour.

We differentiate our work in this paper from the above other researches in enrolment and attrition rate prediction by predicting the number of enrolment of students in the coming three years and also we managed to generalize interesting patterns for the student who is interested in enrolling in a specific program. Besides, the size of our dataset is significantly larger (1600) than those used in previous studies, and the model performance is significantly improved by using the ensemble approach.

References

[1] Agrawal R., Mannila H., Srikant R., Toivonen H., and Verkamo A., "Fast Discovery of Association Rules," *Advances in Knowledge Discovery and Data Mining*, vol. 12, no. 1, pp. 307-328, 1996.

[2] Al-Shabandar R., Hussain A., Liatsis P., and Keight R., "Detecting At-Risk Students with Early Interventions Using Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 149464-149478, 2019.

[3] Freund Y., Schapire R., and Abe N., "A Short Introduction to Boosting," *Journal-Japanese Society for Artificial Intelligence*, vol. 14, pp. 1612, 1999.

[4] Kemper L., Vorhoff G., and Wigger B., "Predicting Student Dropout: A Machine Learning Approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28-47, 2020.

[5] Mulugeta M. and Borena B., "Higher Education Students' Enrolment Forecasting System Using Data Mining Application in Ethiopia" *HiLCoE Journal of Computer Science and Technology*, vol. 2, no. 2, pp. 37-43, 2013.

[6] Nakhkob B. and Khademi M., "Predicted Increase Enrolment In Higher Education Using Neural Networks and Data Mining Techniques," *Journal of Advances in Computer Research*, vol. 7, no. 4, pp. 125-140, 2016.

[7] Petkovski A., Stojkoska B., Trivodaliev K., and Kalajdziski S., "Analysis of Churn Prediction: A Case Study on Telecommunication services in Macedonia," *in Proceedings of 24th Telecommunications Forum*, Belgrade, pp. 1-4, 2016.

[8] Rai S. and Jain A., "Students' Dropout Risk Assessment in Undergraduate Courses of ICT at Residential University-A Case Study," *International Journal of Computer Applications*, vol. 84, no. 14, pp. 31-36, 2013.

[9] Sadatrasoul S., Gholamian M., and Shahanaghi K., "Combination of Feature Selection and Optimized Fuzzy Apriori Rules: The Case of Credit Scoring," *The International Arab Journal of Information Technology*, vol. 12, no. 2, pp. 138-145, 2015.

[10] Schapire R. and Freund Y., "Boosting: Foundations and Algorithms," *Kybernetes*, 2013.

[11] Waters A. and Miikkulainen R., "Grade: Machine Learning Support for Graduate Admissions," *AI Magazine*, vol. 35, no. 1, pp. 64-75, 2014.

[12] Yukselturk E., Ozekes S., and Türel Y., "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program," *European Journal of Open, Distance and e-learning*, vol. 17, no. 1, pp.118-133, 2014.

Samar Shilbayeh is Assistant Professor in Business Analytics program in Abu Dhabi School of Management, she worked as a senior data scientist and head of research Centre in Cognitro Analytics company experienced in extracting data, analyzing, findings and applying predictive modeling techniques, assisted in the development of an analytics framework for smart construction of a new type of predictive indictors “Cognitive Indicators”. Developed active learning algorithms. Developed a cost effective, expert intelligent system, which guides the data miner to optimize models selection. Her research interests include machine learning and AI approaches, cost sensitive machine learning algorithms, applying machine learning solutions in health, finance, telecom and insurance. Dr Samar has a PhD in Machine learning and AI from the university of Salford, Computer science and Engineering department, Manchester, UK.

Abdullah Abonamah is the President and Provost of the Abu Dhabi School of Management. From August, 2000 to October, 2007, he was a Professor and Director of the Institute for Technological Innovation at Zayed University and the Assistant Dean of the College of Information Systems. Before coming to the UAE, he was a Professor of Computer Science and the Computer Science Division Head at the University of Akron. Dr. Abonamah has a PhD in Computer Science from the Illinois Institute of Technology, Chicago, Illinois and an Executive Management Graduate Degree from Yale University School of Management. His research and teaching interests include strategy, technology management, and entrepreneurship and innovation. With over fifty publications in international journals and conferences and a US patent in reliable systems, Dr. Abdullah remains a strong advocate of strategic management, innovation, entrepreneurship, and proper technology-business integration.