ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 12 (2022) No. 6, pp. 147-162
DOI:10.33168/JSMS.2022.0610

Detecting At-Risk and Withdrawal Students in STEM and Social Science Courses using Predictive and Association Rules Mining

Muhd Syazwan Aqrimi Bin Suhaimi, Amy Hui Lan Lim+, Hui-Ngo Goh

Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia

nawsay7@gmail.com, amy.lim@mmu.edu.my (corresponding author), hngoh@mmu.edu.m[3]

Abstract. This research aims to identify potential at-risk and withdrawal students to help these students in their studies. Interactions consisting of surfing behaviour in the Virtual Learning Environment (VLE) among two different groups of students namely disabled and non-disabled students for Social Science and STEM courses are analysed. Predictive analytics and association rule mining (ARM) analysis are performed. Predictive analytics is performed to predict students' likelihood of withdrawing from their registered courses. Among the students who choose to pursue their registered courses, predictive analytics is also used to predict at-risk students. Six predictive algorithms namely Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), K Nearest Neighbour (KNN), Random Forest (RF), and Support Vector Machine (SVM) are compared. FP-Growth algorithm is applied in ARM analysis. Predictive results show that DT is superior with the accuracy scores reaching 0.91. Most association rules are positively correlated, and they represent the set of commonly surfed pages by the potential at-risk and withdrawal students. The predictive results can help VLE developer to determine the possible algorithms to be used in the intelligent VLE to make accurate predictions based on students' interactions in the VLE. The results from ARM analysis prove that FP-Growth can also be included in the intelligent VLE. The intelligent VLE can assist the relevant staff in an education institution to provide timely and personalized support to students who are struggling in their studies. This research contributes to precision education through learning analytics.

Keywords: data science applications in education, distance education and online learning, evaluation methodologies

1. Introduction

United Nations Sustainable Development Goals (UN SDGs) consist of seventeen goals that aim to set the world to be a better place for mankind. Žalėnienė & Pereira (2021) have indicated that higher education institutions (HEIs) play an important role towards the achievement of UN SDGs namely Goal 1, Goal 3, Goal 5, Goal 8, Goal 12, Goal 13, and Goal 16. This puts HEIs in the spotlight of providing quality programmes to equip individuals with the necessary skills and knowledge as well as to shape their personalities and characters so that they are ready for the workforce. To cater for a variety of workforce demand, HEIs are not only accepting non-disabled students but increasingly HEIs are accepting disabled students to pursue programmes in HEIs.

In HEIs, courses are offered according to the nature of the programmes. Typically, these courses can be classified as Science, Technology, Engineering, and Mathematics (STEM) or Social Science. Regardless of whether these courses are STEM or Social Science, HEIs usually have their own Virtual Leaning Environment (VLE) to share learning materials and encourage interactivity between students and instructors. A VLE is a virtual space for teaching and learning, hosted by a website or through an application. Students will have to refer to the VLE to obtain learning materials, submit assignments and attempt assessments related to their registered courses within the semester.

There are three possible outcomes from the students' registered courses namely, withdrawal, pass, or fail. Our focus is to analyze the students' surfing behaviour and their tendencies towards withdrawal or failing from their registered courses. Withdrawal from courses is a concern as it prolongs their studies due to retake, thus delaying their graduation on time. The affected students will also suffer financial repercussions due to the need to pay for the tuition fees for courses that are withdrawn beyond the Add and Drop's deadline. When failing a course, the students will have to face consequences both before and after graduation. Failing a course can cause a dramatic drop in their grade point average (GPA) and cumulative grade point average (CGPA). If their CGPA is too low, they may even be put under probation or being terminated from the programme. If they graduate with a low CGPA, it may even limit their job opportunities as compared to graduates with higher CGPA.

As mentioned in earlier paragraph, HEIs are also accepting disabled students but under the universal academic system, disabled students might have more learning challenges in contrast with their non-disabled peers. Precision education (Cook et al., 2018) is necessary and being able to timely detect at-risk students and students who are likely to withdraw from their courses, allows initiation of early intervention, customized to assist students in different groups. It can also help instructors in predicting the students' future performance in certain courses and provide them with a personalized learning experience based on their needs. Not only does it benefit students, but instructors and researchers can also use this data to decide on ways to revise teaching and learning strategies to reduce the number of failures and withdrawals.

There has been a lot of active research in analyzing students' behaviour using publicly available datasets or proprietary datasets related to students. However, to the best of our knowledge, there is no work on analytics that does both predictive analytics and associative analysis to understand students' surfing behaviour in VLE based on specific groups of students (disabled students and non-disabled students) and types of registered subjects (STEM and Social Science).

This research aims to study two different groups of students namely disabled and non-disabled students and how these groups of students behave and perform when they have registered for Social Science and STEM courses. Firstly, for each group of students and type of courses, this research aims to predict students who are likely to withdraw from their registered courses. Secondly, for each group of students and type of courses, this research aims to predict whether students who did not withdraw a course would subsequently pass or fail their registered courses. Thirdly, for each group of students and the type of registered courses with the information of the outcome of their registered courses (withdrawal or fail), this research aims to derive associative surfing patterns consisting of common set of web pages that are surfed in the VLE.

The dataset that will be used is the Open University Learning Analytics Dataset (OULAD). It is a comprehensive learning analytics dataset consisting of 22 courses, 32593 students, their assessment results, and logged interactions in VLE (Kuzilek et al., 2017). We will only focus on attributes related to interactions in VLE. Further descriptions about the data will be explained in Section 3.

2. Related Works

Data mining and machine learning algorithms have been proposed to assist decision-making in domains such as cybersecurity, finance, and many other domains (Gurumurthy et al., 2022; Nunez & Gatica, 2022; Sivakami et al., 2022). This section reviews some of the studies in learning analytics where machine learning or data mining techniques are used to understand students' behaviour and characteristics in the learning domain.

Heuer & Breiter (2018) categorize the attributes and apply four algorithms namely Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM). The study shows that attributes related to students' daily activities are important indicators to student performance and can be used to predict a pass or a fail for a course. However, the study does not explore more complex models and perform any hyperparameter optimization. The authors have also applied K-Means clustering algorithm to cluster students according to similar characteristics.

Jha et al. (2019) have categorized the attributes and predicts ‘dropout’ and ‘results’ using ensemble, deep learning, and regression techniques. Four algorithms are used namely Deep Learning, Gradient Boosting Machine, Distributed Random Forest, and Generalized Linear Model and applied to each category with all algorithms having high AUC scores when they are applied to category of attributes related to VLE.

He et al. (2020) propose Recurrent Neural Network (RNN)-Gated Recurrent Unit (GRU) joint neural network to predict students’ performance in a specific course. The proposed method is derived by choosing the better baseline methods among simple RNN, GRU, and Long Short-Term Memory (LSTM). The model has recorded an accuracy of 80% in predicting the students’ performance in a specific course.

Aljohani et al. (2019) apply a deep LSTM model to predict students’ performance in a course. The authors transform the interactions in the VLE into a weekly basis. The proposed method is compared with LSTM, Logistic Regression (LR), and artificial neural network.

Wang et al. (2022) apply convolution residual RNN on a clickstream data to predict at-risk students for a specified duration and obtain particularly high precision in four STEM courses in the dataset. Similar predictive analysis on OULAD dataset is performed and compared using various machine learning algorithms with RF producing the highest F-measure in a balanced dataset (Hlioui, 2021).

Demographic-based indicators are used to predict the learning outcome as pass, fail or distinction (Rizvi et al., 2019) but the study does not indicate the techniques that have been used to manage the imbalanced data.

Pilevari et al. (2021) propose Adaptive Neuro Fuzzy Inference System (ANFIS) to predict the e-learning resilience during the Covid-19 pandemic. A total of 22 features from 5 factors are considered as input into ANFIS. The 5 factors are individual, technology, content, agility, and assessment/ support. The authors have reported that changes to the factors have significant impact to the agility factor. The proposed model has been implemented in an Iranian university and can be adopted by any education institutions.

Hu (2022) applies convolutional neural networks (CNN) on a set of activities that are logged in Learning Management Systems to predict at-risk students. The data consists of recorded interactions such as total downloads per materials, etc. The model achieves accuracy of at least 81% when it is implemented on a specific course in the year 2019.

The study by OuahiMariame et al. (2021) uses a combination of Sequential Forward Selection as feature selection technique and Naïve Bayes as classification technique on the OULAD dataset. The results show that the best accuracy is at 58%.

Cooper et al. (2022) attempt to predict at-risk students from the past academic records consisting of intermediate and final grades of a programming course. NeuroSolutions Professional by NeuroDimensional is used and the authors have reported that Probabilistic Neural Network provides the accuracy of 91.30%. However, the data that is used in this study is limited to grading marks.

The following describes some of the studies that use association rule mining (ARM) to solve problems in the learning domain. Jawthari & Stoffa (2022) investigate the association between students' demographic features and their engagement level in a VLE in the OULAD dataset. A two-level clustering model is used as it has the best performance among other models in terms of separation with silhouette coefficient. Apriori algorithm is used to generate a set of association rules that relates students' demographic features to students' engagement level. The results associate gender, highest education, studied credits, and the number of previous attempts with high engagement levels. However, according to the authors, the absence of features such as navigation sequences and a penalty for late submission limits the ability to accurately measure the level of engagement in VLE.

ARM is applied to discover relations among attributes that represent the underrepresented in two STEM courses namely Course A and Course B (Valdiviejas & Bosch, 2020). The authors share the set of association rules with lift values of more than 1.00 and less than 0.89 which represent patterns that describe the characteristics of students in both courses.

On the other hand, Apriori algorithm is applied to study the characteristics of potential undergraduate students who apply for admission to a tertiary education institution in Indonesia. The authors have identified several related attributes that can be used as indicators and their relations with the registration status (Nugraha & Hadi, 2022). This study can assist the marketing team in designing targeted marketing campaigns to attract more new students.

An intelligent platform having association rules algorithm is designed to suggest summary sheets that should be bought by students given that a particular summary sheet is bought. According to the authors, summary sheets are helpful for students during preparation before the final exams especially when the students are taking several courses (Binchai et al., 2022). Although the current e-commerce platforms do include recommenders that recommend similarly purchased products, the current e-commerce platforms do not solely focus on academic-related products.

Riofrío Calderón et al. (2021) apply Apriori algorithm to identify factors that contribute to a high dropout in a Massive Open Online Courses (MOOC) course. A survey is conducted among participants that register for "Conventional Clean Energy and Its Technology" course. The results reveal a list of association rules with confidence values up to 86%.

There are other studies that describe the use of clustering or combination of several algorithms to solve problems in the learning domain. Ahmed et al. (2022)

use Expectation-Maximization (EM) algorithm to identify the ideal number of clusters to group students for academic advising. A total of 146 records with students' assessment marks are used in this study. The results reveal that it is reasonable to group students into 7 groups for academic advising.

Kokoç et al. (2021) use a combination of clustering, Markov Chains and ARM to analyze the submission details that are logged in Moodle. The results reveal that it is possible to predict students' academic performance at the end of the term based on the assignments' submission behaviour in Moodle.

Most of the research studies are not segregating the data belonging to the disabled and non-disabled students when the researchers are doing analysis on the interactions in the VLE. There are studies that focus on analysis on specific courses. To the best of our knowledge, there is little work done on analysis based on specific group of students and specific type of courses.

3. Research Methodology

3.1. Data Description

The OULAD dataset is stored within seven tables which can be linked. Table 1 lists the attributes in each table in the OULAD dataset.

Table 1: The list of tables in the OULAD dataset

Tables	Columns(attributes)
studentInfo	code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, final_results
courses	code_module, code_presentation, length
studentRegistration	code_module, code_presentation, id_student, date_registration
assessment	code_module, code_presentation, id_assessment, assessment_type, date, weight
studentAssessment	id_assessment, id_student, date_submitted, is_banked, score
studentVle	code_module, code_presentation, id_student, id_site, date, sum_click
vle	id_site, code_module, code_presentation, activity_type, week_from, week_to

The *studentInfo* table contains students' information and general academic records. The *courses* table contains information about the available courses. The *studentRegistration* table contains the students' course registration details. The *assessment* table contains the assessments that are prepared for each course. The *studentAssessment* table contains the students' assessment marks. The *studentVLE* table contains the total number of clicks that have been made by each student when he/she is surfing the VLE. Lastly, the *vle* table captures the activities and the duration of each activity in the VLE.

3.2. Datasets Preparation

Since the focus is on two groups of students registering for two different types of courses, four datasets are formed from the pre-processing of the OULAD dataset. Table 2 lists the four datasets that are formed from the OULAD dataset. Table 3 describes the attributes (column names) for each of datasets in Table 2.

Table 2: The four datasets that are prepared from the OULAD dataset

Dataset	Description of content of the dataset
D_STEM	This dataset contains the interactions in VLE by disabled students that registered for STEM
D_SS	This dataset contains the interactions in VLE by disabled students that registered for Social Science
ND_STEM	This dataset contains the interactions in VLE by non-disabled students that registered for STEM
ND_SS	This dataset contains the interactions in VLE by non-disabled students that registered for Social Science

Table 3: The attributes (column names) in each dataset

Column Name	Description of the attribute
id_student	ID of students
withdrawn	status of students' withdrawal
final_result	result of pass or fail
date_vle	date the VLE was accessed
sum_click	sum of clicks in a day
activity_type	activity type in VLE
week_vle	week the VLE was accessed
month_vle	month the VLE was accessed

3.3. Overview of the Experiments and Setups

Fig. 1 shows the total number of experiments that are conducted to achieve the objectives of this research. There are 4 datasets that are formed from the OULAD dataset. Predictive analytics and association rule mining (ARM) analysis are performed on each dataset. Since we would like to predict at-risk students and withdrawal students, two separate experiments on predictive analytics are conducted on each dataset.

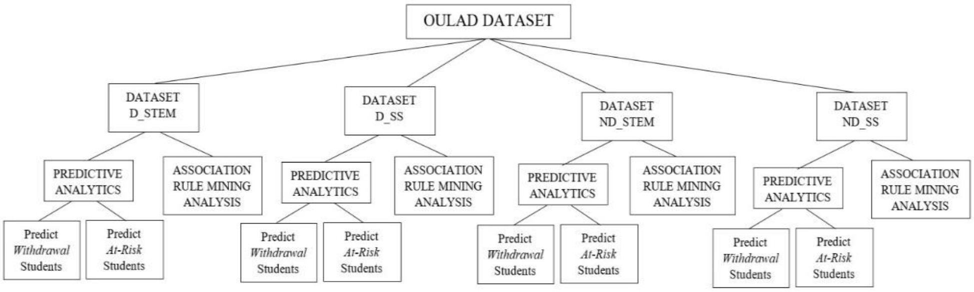
OULAD DATASET
DATASET D_STEM
DATASET D_SS
DATASET ND_STEM
DATASET ND_SS
PREDICTIVE ANALYTICS
ASSOCIATION RULE MINING ANALYSIS
Predict *Withdrawal* Students
Predict *At-Risk* Students

Fig. 1: Overview of the experiments

For each experiment on predictive analytics, the predictive algorithms that are chosen for comparative analysis are Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), K Nearest Neighbour (KNN), Random Forest (RF), and Support Vector Machine (SVM). These algorithms are chosen because they are popularly used to predict the students' learning outcomes.

Before performing predictive modelling, the ratio of data for train: test is fixed at 80:20. The imbalanced datasets are oversampled with SMOTE (Chawla et al., 2002) The predictive algorithms for classification are adopted from Python's scikit learn library. The experiments using predictive algorithms have been tweaked and tested to achieve the best results possible with the extracted datasets. The following table indicates the settings for the predictive algorithms.

Table 4: The settings for the predictive algorithms

Algorithms	Settings
DT	criterion = "entropy"
LR	solver = 'lbfgs', max_iter = 500

NB
KNN
	GaussianNB()
	n_neighbors=7

RF	n_estimators=100
SVM	LinearSVC()

In this study, we also consider ARM analysis. Traditionally, ARM (Agrawal et al., 1993) is used to understand shoppers' buying patterns. These patterns are used to design the store layout to increase shoppers' convenience and sales. The output from ARM analysis is a set of association rules having antecedent and consequent as shown in equation 1 where X and Y are referring to itemset.

$$X \rightarrow Y \tag{1}$$

The interestingness of an association rule is measured using some common metrics such as support, confidence, and lift. Support for an association rule refers to the total number of times that the itemset appear in the data. Support value ranges from 0 to 1 (both inclusive). An association rule with a high support means that the itemset appears very frequently in the dataset. Confidence for an association rule refers to the likelihood that the consequent of the rule will occur when the antecedent of the association rule is present. Similar to support, its value ranges from 0 to 1 (inclusive). An association rule with high confidence refers to the high probability that the consequent will occur given the presence of the antecedent. Lift refers to the degree of correlation among itemset. The value for lift ranges from negative infinity to positive infinity. However, a lift with the value of 1 indicates that the presence of the antecedent in the association rule has no impact to the presence of the consequent in the association rule. A high lift value indicates an existence of a high degree of correlation or dependence among items in the association rule.

In this research, the ARM analysis is programmed using Python’s mlxtend library. Frequent Pattern-Growth (FP-Growth) is chosen due to its storage and performance efficiency. The frequent itemset is set as 0.90 as the minimum support. However, for the dataset related to non-disabled students who have registered for STEM-related courses, 0.97 is set as the minimum support value. Association rules are then extracted using 0.90 as the minimum confidence value. The setting of high support and confidence values are necessary to extract strong association rules and avoid crashing in the experiments’ environment. The extracted association rules are observed and evaluated.

4. Results

This section reports the results from the predictive and ARM analysis. In Section 4.1, Table 5 reports the accuracy scores from the application of 6 predictive algorithms to predict the likelihood of withdrawal from courses. Table 6 reports the accuracy scores from the application of 6 predictive algorithms to predict whether the students will pass or fail courses. In Section 4.2, the results from ARM analysis are reported.

4.1. Results from Experiments on Predictive Analytics

Table 5 shows the accuracy scores that are recorded by the 6 predictive algorithms to predict the likelihood of withdrawing from courses.

Table 5: Accuracy scores of predictive algorithms to predict the likelihood of withdrawing from courses

	Disabled Students		Non-Disabled Students	
	STEM Courses	Social Science Courses	STEM Courses	Social Science Courses
Decision Tree (DT)	0.92	0.91	0.91	0.92
Logistic Regression (LR)	0.50	0.50	0.51	0.49
Naïve Bayes (NB)	0.40	0.44	0.37	0.87
K Nearest Neighbour (KNN)	0.89	0.87	0.90	0.91
Random Forest (RF)	0.62	0.72	0.53	0.57
Support Vector Machine (SVM)	0.15	0.89	0.89	0.44

Table 6 shows the accuracy scores that are recorded by 6 predictive algorithms to determine the likelihood of passing or failing from courses.

Table 6: Accuracy scores of predictive algorithms to predict the likelihood of passing or failing from courses

	Disabled Students		Non-Disabled Students	
	STEM Courses	Social Science Courses	STEM Courses	Social Science Courses
Decision Tree (DT)	0.83	0.81	0.86	0.89
Logistic Regression (LR)	0.56	0.59	0.56	0.58
Naïve Bayes (NB)	0.53	0.54	0.51	0.52
K Nearest Neighbour (KNN)	0.80	0.76	0.85	0.87
Random Forest (RF)	0.61	0.64	0.57	0.59
Support Vector Machine (SVM)	0.61	0.67	0.14	0.73

As seen from the results in Table 5 and 6, it is observable that Decision Tree algorithm produces the highest accuracy scores compared to the rest of the algorithms. The lowest accuracy score recorded by Decision Tree algorithm is 0.81. K Nearest Neighbour algorithm trails behind Decision Tree algorithm closely with the accuracy scores ranging from 0.76 to 0.91. Random Forest algorithm seems to be a promising predictive algorithm too when it is applied on the datasets consisting of interactions in the VLE by disabled students. Logistic Regression algorithm and Naïve Bayes algorithm perform badly, but Naïve Bayes algorithm achieves an accuracy score of 0.87 when it is used to predict the likelihood of non-disabled students withdrawing from their courses.

Support Vector Machine demonstrates inconsistent results. When looking at the datasets individually, there are some promising results. Support Vector Machine can achieve an accuracy score of 0.89 when the algorithm is applied to the dataset consisting of interactions in the VLE by disabled students who register for Social Science courses. The same accuracy score is reported when the method is applied to predict the likelihood of non-disabled students withdrawing from STEM courses.

There are two cases where Support Vector Machine produces poor accuracy scores. The accuracy score of 0.15 is obtained when the method is applied to predict the likelihood of disabled students withdrawing from STEM courses. The accuracy score of 0.14 is recorded when the method is applied to predict whether the non-disabled students will pass or fail their STEM courses. The low accuracies can be attributed to the combination of the following two factors.

Firstly, the datasets that are used for the above two cases are highly imbalanced datasets. The ratio of *withdrawal*: *non-withdrawal* in the dataset that is used to predict the likelihood of disabled students withdrawing from STEM courses is approximately 1:10 which can be interpreted as for every 10 instances of not withdrawing from courses, there is only one instance of student withdrawing from courses. Similar condition exists in dataset that is used to predict whether the non-disabled students will pass or fail the STEM courses. The ratio of *Fail*: *Pass* is approximately 1:6 where for every 6 instances related to passing a course, there is only one instance of student failing a course.

Secondly, each dataset is randomly partitioned into 80% for training and 20% for testing before the predictive algorithms are applied. The randomness may have impacted the content of the training and testing sets such that less instances related to withdrawal or failing a course are selected to be part of the training sets as compared to the test sets. Although SMOTE is applied on the training sets to solve the imbalanced issue, the Support Vector Machine algorithm may not learn the instances in the training sets well enough, due to lesser variations of instances related to withdrawal or failing a course, thus generating a model that cannot generalize well when the model is tested on the test sets.

Overall, Decision Tree algorithm can be programmed into VLE because it generates predictive models of highest accuracy scores compared to other predictive algorithms.

4.2. Results from Experiments on ARM Analysis

Based on ARM analysis, it can be observed that all the association rules are strong association rules, having high support and high confidence values. All the association rules have support values of at least 0.90 and confidence values of 0.92. Most of the association rules are positively correlated.

Each association rule represents the pages being surfed. Due to large number of association rules that are generated, equation (2) shows an example of an association rule that describes the surfing pattern of disabled students that withdraw from STEM courses. The equation (2) hopes to give the readers a better clarity on the content of the association rule that is generated.

$$\text{ouwiki} \rightarrow \text{oucollaborate}, \text{oucontent} \tag{2}$$

The association rule in equation (2) has the highest lift value of 1.08. This indicates that these three pages are frequently surfed in combination by the disabled students who are withdrawing from STEM courses. The support and confidence values are above 0.90. The high support value indicates that these three pages in combination are frequently occurring in the dataset. The high confidence value indicates that when a student surfs the *ouwiki* page, there is a high likelihood that he or she will surf the *oucollaborate* and the *oucontent* pages.

Since we are interested to identify commonly surfed pages by students who withdraw or fail from courses, we focus on association rules with highest lift values and extract frequent patterns among them. Table 7 shows a list of commonly surfed pages among different groups of students that withdraw from STEM and Social Science courses. Table 8 shows a list of commonly surfed pages among different groups of students that fail STEM and Social Science courses.

Table 7: Commonly surfed pages by students that withdraw from courses

Students	Courses	Pages that are commonly surfed together
Disabled Students	STEM Courses	oucollaborate, ouwiki
	Social Science Courses	url, resource, oucontent, subpage
Non-Disabled Students	STEM Courses	externalquiz, glossary
	Social Science Courses	oucollaborate, url

Table 8: Commonly surfed pages by students that fail courses

Students	Courses	Pages that are commonly surfed together
Disabled Students	STEM Courses	externalquiz, oucollaborate
	Social Science Courses	subpage, oucontent, url
Non-Disabled Students	STEM Courses	glossary, ouwiki
	Social Science Courses	quiz, url

Most of the commonly surfed pages are unique, however, noticeable patterns can be seen among disabled students who withdraw or fail the Social Science courses. These students have similar surfing patterns: they surfed *url*, *subpage* and *oucontent*. If we compare among their non-disabled peers, the common page being surfed is *url*.

Disabled students who are at-risk or likely to withdraw from STEM courses will surf common page, which is *oucollaborate*. Non-disabled students who are at-risk or likely to withdraw from STEM courses will surf common page, which is *glossary*.

Overall, it is possible to identify commonly surfed pages by two groups of students who are potentially at-risk or likely to withdraw from their courses.

5. Discussion

This study focuses on identifying potential at-risk students and potential students that are likely to withdraw from their courses. Using historical log data consisting of interactions that are performed by disabled and non-disabled students for their registered STEM and Social Science courses, predictive analytics and ARM analysis are performed. Based on the results as shown in Table 5 and Table 6, models that are derived using Decision Tree algorithm have the highest accuracy scores. Complimenting these models are the lists of commonly surfed pages by disabled and non-disabled students who withdraw or fail different types of courses. Table 7 and Table 8 show the lists of commonly surfed pages that are obtained from Association Rule Mining analysis. The models that are derived using Decision Tree algorithm and the lists of commonly surfed pages can be embedded into the VLE's knowledge base and thus, upgrades the existing VLE to an intelligent VLE. The intelligent VLE can track students' interactions in the VLE. If a student's surfing behaviour in the VLE matches the models or the list of commonly surfed pages, he/she is identified to be a potential at-risk student or potential student who is likely to withdraw from courses.

The intelligent VLE will notify relevant staff in an educational institution when there is a potential student who are at-risk or likely to withdraw from their respective courses. It can help an instructor to initiate early intervention programme for that specific course. Depending on the number of students that meet the above characteristics, suitable, customized remedial classes or periodical short meetings can be arranged to assist these groups of students.

In addition, if the intelligent VLE has identified a significant number of potential at-risk students or a significant number of potential students who are likely to withdraw from a specific course, an instructor will have to review the effectiveness of the current teaching and learning strategies being implemented in that course.

For the faculty or the school, the total number of potential at-risk students or total number of potential students who are likely to withdraw courses can serve as an input to course and programme revisions for quality improvement. Furthermore, it can also be used as an input to a more effective counselling or academic advising session to help these students.

6. Conclusion

This research has proved the viability of applying both predictive analytics and ARM analysis to understand students' behaviour in VLE based on specific target groups of students (disabled and non-disabled) and types of subjects (STEM and Social Science) that they have registered. For predictive analytics, the aim is to predict the learning outcomes, i.e.: predict students who are likely to withdraw and for those who choose to pursue the registered courses, predictive analytics are performed to predict whether they are potential at-risk students. The OULAD dataset is used for both predictive analytics and ARM analysis. Decision Tree algorithm displays superior performance compared to other algorithms with its accuracy score reaching 0.91. This helps to narrow down the possible algorithms to be used in the intelligent VLE by VLE developer to make accurate predictions based on students' interactions in the VLE. Using FP-Growth algorithm in ARM analysis, most of the association rules generated are positively correlated association rules that represent the list of commonly surfed pages. The results have proved that Decision Tree algorithm and FP-Growth algorithm can be used by the VLE developers, in designing an intelligent VLE that benefits staff and students in an educational institution.

This research has limitation because we choose to compare among 6 predictive algorithms for predictive analytics and use FP-Growth algorithm for ARM analysis. Future work can migrate to use different machine learning techniques or ensemble methods to improve the accuracy scores in predictive analytics. Furthermore, FP-Growth algorithm can be further enhanced so that the algorithm can deal with spurious number of association rules.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*

Ahmed, A. A., Mohammed, A., Osman, A. N. & Saeed, A. (2022). Measuring students' academic performances: A data clustering approach. *Journal of System and Management Sciences*, 12(2), 137-152

Aljohani, N. R., Fayoumi, A., & Hassan, S. U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability*, 11(24), 7238

Binchai, N., Songmuang, V., Worrawat, K., & Chondamrongkul, N. (2022). Applying association rules for summary sheet marketplace. *In 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357

Cook, C. R., Kilgus, S. P., & Burns, M. K. (2018). Advancing the science and practice of precision education to enhance student outcomes. *Journal of School Psychology*, 66, 4-10

Cooper, C. I. (2022). Using machine learning to identify at-risk students in an introductory programming course at a two-year public college

Gurumurthy, S., Hemalatha, K. L., Pamela, D., Roy, U., & Vishwanath, P. (2022). Hybrid pigeon inspired optimizer-gray wolf optimization for network intrusion detection. *Journal of System and Management Sciences*, 12(4), 383-397

He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G., & Jiang, B. (2020). Online at-risk student identification using RNN-GRU joint neural networks. *Information*, 11(10), 474

Heuer, H. & Breiter, A. (2018). Student success prediction and the trade-off between big data and data minimization. *DeLFI 2018-Die 16. E-Learning Fachtagung Informatik*

Hlioui, F., Aloui, N., & Gargouri, F. (2021). A withdrawal prediction model of at-risk learners based on behavioural indicators. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, 16(2), 32-53

Hu, Y. H. (2022). Using few-shot learning materials of multiple SPOCs to develop early warning systems to detect students at risk. *International Review of Research in Open and Distributed Learning*, 23(1), 1-20

Jawthari, M. & Stoffa, V. (2022). Relation between student engagement and demographic characteristics in distance learning using association rules. *Electronics*, 11(5), 724

Jha, N. I., Ghergulescu, I., & Moldovan, A. N. (2019). OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques. *In CSEDU*, (2)

Kokoç, M., Akçapınar, G. & Hasnine, M. N. (2021). Unfolding students' online assignment submission behavioral patterns using temporal learning analytics. *Educational Technology & Society*, 24(1), 223-235

Kuzilek, J., Hlosta, M. & Zdrahal, Z. (2017). Open university learning analytics dataset, *Scientific Data*, 4(1), 1-8

Nugraha, F. S. & Hadi, W. (2022). Apriori implementation to find the association rules of the new student admission data of STMIK AMIKOM Surakarta. *SISFOTENIKA*, 12(1), 114-124

Nunez, N. A. & Gatica, G. (2022). Applying profit-driven metrics in predictive models: A case study of the optimization of public funds in Peru. *Journal of System and Management Sciences*, 12(2), 52-65

OuahiMariame, S. K. (2021). Feature engineering, mining for predicting student success based on interaction with the virtual learning environment using artificial neural network. *Annals of the Romanian Society for Cell Biology*, 25(6), 12734–12746

Pilevari, N., Memarian, S., & Shokouhifar, M. (2021). Evaluation of distance learning resilience during COVID-19 pandemic using ANFIS. *Journal of Logistics, Informatics and Service Science*, 8(2), 103-118

Riofrío-Calderón, G., Ramírez-Montoya, M. S., & Rodríguez-Conde, M. J. (2021). Data analytics for predicting dropout. *In Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, 137, 32-47

Sivakami, R., Uday Kiran, G., Arun, M., David, L.G., Manohar, M. (2022). Dirichlet feature embedding with adaptive long short-term memory model for intrusion detection system. *Journal of System and Management Sciences*, 12(4), 398-412

Valdiviejas, H. & Bosch, N. (2020). Using association rule mining to uncover rarely occurring relationships in two university online STEM courses: A comparative analysis. Grantee Submission

Wang, X., Guo, B., & Shen, Y. (2022). Predicting the at-risk online students based on the click data distribution characteristics. *Scientific Programming*

Žalėnienė, I. & Pereira, P. (2021). Higher education for sustainability: A global perspective. *Geography and Sustainability*, 2(2), 99-106