Predicting Student Success for Programming Courses in a Fully Online Learning Environment

Neil Arvin BRETANA[a*], Mehdi ROBATI[a], Aastha RAWAT[b], Aashi PANDEY[b], Shreya KHATRI[b], Kritika KAUSHAL[b], Sidarth NAIR[b], Gerald CHEANG[b], Rhoda ABADIA[a]

[a]*UniSA Online, University of South Australia, Australia*

[b]*UniSA STEM, University of South Australia, Australia*

*neil.bretana@unisa.edu.au

Abstract: The emergence of online learning environments is important for teaching programming courses. In this study, demographic and performance-related data from two programming courses of a fully online learning platform, UniSA Online, were explored. Statistically significant features were identified using Varian Inflation Factor and Chi-Square test. Four prediction models were trained and tested using four sets of features: demographic, performance, statistically significant features, and all available features. The model trained using demographic features yielded an accuracy of 45.45%. The models trained usind performance-related features, statistically significant features, and all features yielded an accuracy of 86.86%, 86.53%, and 86.53%, respectively. This highlights the importance of performance-related data in predicting student success outcomes in learning programming via a fully online learning environment.

Keywords: Online learning, model, classifier, student success

1. Introduction

Online learning has become a viable alternative for learners who are unable to participate in a traditional face-to-face university (Zhou, 2016). Online learning environments are presumed to be more inclusive relative to traditional learning environments as it allows participants of all ages, education levels, diverse regional backgrounds and even those whose performance may be limited due to accessibility needs. It is important to improve student success and learning experience and increase student retention to help achieve the goals of both the students and the learning providers. In order to achieve these, learning providers must understand their students, be able to intervene challenges early on, engage effectively with the students, and support the students through content and delivery (Stone, 2017). Aslanian and Clinefelter (2013) have shown that the educational outcomes in online learning environments are linked to the reputation of both facilitators and institutions.

Given the increase in the popularity of online learning platforms, it has become crucial to understand what characteristics affect student success in online learning environments to help guide facilitators. The growing amount of educational data in online learning provides challenges for online facilitators to organize and understand these large complex data set. With limited resources available, this large amount of data is making it increasingly difficult to monitor, identify, and solve learning issues ahead of time (Howarth, 2019; Koutropoulos & Zaharias, 2015).

Using predictive analytics in online learning can assist in predicting student performance that can assist in academic retention and student support efforts, especially for at-risk students. The use of predictive analytics approach can help understand what characteristics affect student success in online learning environments. Previous studies have shown a strong association between online learners' performance and their demographic characteristics, such as regional belonging, socio-economic standing, education level, gender and age (Rizvi, Rienties, & Khoja, 2019). Many studies have also observed learners' behavioural changes in online learning environments over time (Kloft, Stiehler, Zheng, & Pinkwart, 2014; Nguyen,

Huptych, & Rienties, 2018). However, these mainly relied on clickstream information from the respective week using data obtained from Massive Open Online Courses (MOOC). Although clickstream information can provide information about how a student navigates through and interacts with online education resources, studies that use clickstream data focus only on understanding student's self-regulatory behaviours, i.e., how students are using the online resources to improve instructional design (Bodily & Verbert, 2017; Diana et al., 2017; Paechter & Maier, 2010; C. Shi, Fu, Chen, & Qu, 2015) and identifying "stop-out classifiers "to prevent students from quitting (Whitehill, Williams, Lopez, Coleman, & Reich, 2015). There were also studies on the early detection systems for poor course performance (Baker, Lindrum, Lindrum, & Perkowski, 2015) but these predictive models are based on indicators that are meaningful only to academic staff who are teaching history courses and may not be useful for other courses. For instance, these wouldn't be applicable to analyzing online programming courses, which are major courses in many online learning environments. Where other courses look at high activity in discussion forums as formative assessments, programming courses give more value to practical coding formative assessment than postings in discussion forums (Azmi, Ahmad, Iahad, & Yusof, 2017; Restrepo-Calle, Ramírez Echeverry, & González, 2019).

Several studies have looked at predicting student success on a programming course using online data, but these studies were focused on a hybrid course delivery where students were required to be in the university and use an online learning environment to learn the course (Azcona, Hsiao, & Smeaton, 2019; Azcona & Smeaton, 2017; Carter, Hundhausen, & Adesope, 2017). Azcona et al. (2019) detected student at-risk by looking at students' demographics (age, travel distance from home to the university and basis of admission) and digital behaviour log. Azcona and Smeaton (2017) looked at student engagement and effort as predictors. And Carter et al. (2017) explored the relationship between the students' programming behaviours and course outcomes, and students' participation with the online social learning environment and course outcomes. While Yukselturk and Bulut (2007) studied students in fully online course, their analysis focused on identifying predictors of student's success and no predictive model was created to identify which students will most likely succeed or are at-risk of failure in the course.

In this paper, data from two fully online programming courses were considered. Data were collected from UniSA Online – a fully online learning platform operated by the University of South Australia. A model for predicting student success was trained and tested with the main aim of informing which features indicate likely success in a fully online programming course. The results of this work aim to further enable online academic staff in programming courses to identify and support students most at-risk of failing.

2. Methodology

2.1 Cohort

University of South Australia (UniSA) is one of Australia's largest online education provider. As part of its digital learning strategy, it launched the UniSA Online to provide 100% online degrees. One of the main online degrees offered through UniSA Online is the Bachelor of IT and Data Analytics. In this degree program, students study a series of programming courses in a fully online learning environment. Since the courses are fully online, learning is asynchronous. All course contents are made available to the students from the first day of the study period. Throughout the course, UniSA Online students get focused and personalized support from dedicated online academic staff via various forms of communication and quality assessment feedback presented in various forms. Students interact with their peers and the academic staff through course forums, live chats, regularly scheduled video conferencing sessions and course e-mails.

Two introductory programming courses in the online degree were chosen for this study: *Python programming* (Course ID: COMP 1043) and *Java and Object-Oriented Programming* (Course ID: COMP 1044). For COMP 1043, the final grade was calculated from the following: Programming Assignment 1 (10%), Programming Assignment 2 (15%), Programming Assignment 3A (15%), Programming

Assignment 3B (10%), and a final exam (50%). For COMP 1044, the final grade was calculated from the following: Continuous Assessment (10%), Assignment A (25%), Assignment B (25%), and a final exam (40%). For both courses, success is defined by having 45% and above score in the online exam and having a total final grade of greater than or equal to 50.

Although the programming language taught in these courses are different, the format of the online courses is highly similar. Content videos and code-along videos are presented to the students in addition to the e-readings. Students have regular online practical activities and regularly scheduled video conferencing sessions which are recorded and available for students to watch.

2.2 Data pre-processing

Data were collected from UniSA online students who completed COMP 1043 and COMP 1044 in consecutive study periods in 2018 and 2019. From a set of 341 student records, 33 student records showing withdrawal from the course have been removed, resulting in 308 students for the final data set. Combining all remaining records for the two courses, the final data set consisted of 297 records. As this study aims to look at the difference between passing and failing the course, the final grade for each student were recorded as a binary value. To be able to feed the information into the model, the features selected for this study were also converted into binary values.

A set of 24 features and 1 identifier were engineered for each student (Table 1). The set of features were categorized into demographics and performance-related attributes pertaining to quizzes, assignments, exam and programming exercises. It should be noted that the age bins followed the standard used by UniSA online based on student professional experience. The feature *Failed*, which represents whether or not a student failed to succeed in the course, was set as the target feature for the model. For this study, students who had to do a supplementary assessment at the end of the study period to increase their final grade to 50 were still assigned a *Failed* feature value of 1 to represent a grade below 50. This is because supplementary assessments are given to students on a case-by-case basis to ultimately pass the course if they have originally obtained a failing mark.

2.3 Identifying statistically significant features

A series of statistical tests were employed to identify statistically significant features associated with student success in this setting.

2.3.1 Variance Inflation Factor

Variance Inflation Factor (VIF) is a measure applied to check multi-collinearity between available features. Multi-collinearity, represented by a high variance, makes it difficult to differentiate between the effects of supposedly independent features on the target variable (Jayaprakash, Moody, Lauría, Regan, & Baron, 2014). It also gives rise to feature redundancy, which means that a feature shares the same linear dependency as other features and does not contribute to the improvement of model performance (Brooks, Thompson, & Teasley, 2015). Therefore, it's important to remove highly correlated features in order to enhance model performance and avoid feature redundancy. VIF achieves this by assigning a score to each feature as defined by the following equation:

$$VIF = \frac{1}{1 - R^2}$$

Table 1. *Feature list*

Feature	Type of Feature	Value	Description
Student_ID	Identifier	9 integers	A unique identifier for each student.
Gender_Male	Demographic	0 or 1	A binary value representing student gender: 0 (female), and 1 (male)
Age_Binned_1	Demographic	0 or 1	A binary value representing whether student age is 21 or under: 0 (no), and 1 (yes)
Age_Binned_2	Demographic	0 or 1	A binary value representing whether student age is 22 to 24: 0 (no), and 1 (yes)
Age_Binned_3	Demographic	0 or 1	A binary value representing whether student age is 25 to 29: 0 (no), and 1 (yes)
Age_Binned_4	Demographic	0 or 1	A binary value representing whether student age is 30 to 39: 0 (no), and 1 (yes)
Age_Binned_5	Demographic	0 or 1	A binary value representing whether student age is 40 to 49: 0 (no), and 1 (yes)
Age_Binned_6	Demographic	0 or 1	A binary value representing whether student age is 50 or above: 0 (no), and 1 (yes)
Location_NSW	Demographic	0 or 1	A binary value representing whether student location is in the state of New South Wales: 0 (no), and 1 (yes)
Location_SA	Demographic	0 or 1	A binary value representing whether student location is in the state of South Australia: 0 (no), and 1 (yes)
Location_NT	Demographic	0 or 1	A binary value representing whether student location is in the state of Northern Territory: 0 (no), and 1 (yes)
Location_QLD	Demographic	0 or 1	A binary value representing whether student location is in the state of Queensland: 0 (no), and 1 (yes)
Location_TAS	Demographic	0 or 1	A binary value representing whether student location is in the state of Tasmania: 0 (no), and 1 (yes)
Location_VIC	Demographic	0 or 1	A binary value representing whether student location is in the state of Victoria: 0 (no), and 1 (yes)
Location_WA	Demographic	0 or 1	A binary value representing whether student location is in the state of Western Australia: 0 (no), and 1 (yes)
Location_Overseas	Demographic	0 or 1	A binary value representing whether student location is outside of Australia: 0 (no), and 1 (yes)
Part_Time	Demographic	0 or 1	A binary value representing a mode of study: 0 (no), and 1 (yes)

Failed	Performance	0 or 1	A binary value representing student failure as an outcome of the course: 0 (no), and 1 (yes)
All_Assessment_Submitted	Performance	0 or 1	A binary value representing whether a student submitted all marked assessments for a course: 0 (no), and 1 (yes)
No_Assessment_Submitted	Performance	0 or 1	A binary value representing whether a student has submitted none of the marked assessment for a course: 0 (no), and 1 (yes)
All_Mandatory_Quizzes_Submitted	Performance	0 or 1	A binary value representing whether a student has submitted all mandatory quizzes for a course: 0 (no), and 1 (yes)
All_Mandatory_Assignments_Submitted	Performance	0	A binary value representing whether a student has submitted all mandatory assignments for a course: 0 (no), and 1 (yes)
Completed_Exam	Performance	0	A binary value representing whether a student has completed the exam for a course: 0 (no), and 1 (yes)
Acitivity_Count_0	Performance	0	A binary value representing whether a student failed to complete any non-mandatory activity for the course: 0 (no), and 1 (yes)
Activity_Count_1	Performance	0/1	A binary value representing whether a student was able to complete at least 1 non-mandatory activity for the course: 0 (no), and 1 (yes)

Table 2. *Summary of final grades per course and study period*

Course ID	Study period	Pass	Fail	Total
COMP 1043	1	42	23	65
COMP 1043	2	68	40	108
COMP 1044	1	36	5	41
COMP 1044	2	11	19	30
COMP 1044	3	27	26	53

For each feature, linear regression is performed against the other remaining features to get the value of R^2 (Kutner, Nachtsheim, Neter, & Li, 2005). Using this value, the VIF score is calculated using the equation above. The higher the value, the more correlated the feature is to the other features. A threshold of VIF greater than or equal to 10 has been selected to reject features (Alauddin & Nghiem, 2010; Midi & Bagheri, 2010). Features with a VIF score meeting this threshold were excluded. A recursive method was performed with the removal of features over the threshold at every step until all factors are given a score below the threshold.

2.3.2 Chi-square

Using the remaining features selected via the VIF test, a chi-square test was conducted to check for association with the target variable *Failed*. The null hypothesis of this test is that no relationship exists between the variables. However, if the p-value is higher than the defined alpha value, then there is enough evidence to reject the null hypothesis and state that there is a relationship between two values (Novaković, 2016).

2.4 Model training and testing

2.4.1 Support Vector Machine

Support Vector Machine (SVM) was employed to build the models for this study. SVM maps nonlinear data to a higher-dimensional linear space where data can be linearly classified by hyperplane (Z. Shi, 2011). For this study, the SVM model implementation in RapidMiner (Mierswa & Klinkenberg, 2018) was utilized with a linear Kernel type and a cost of misclassification value of 0.01.

2.4.2 Training

Four training sets were prepared based on the type of features. The first set included only demographic features, the second set included only performance features, the third set included only statistically significant features resulting from the mutual info gain test, and the last set included all features. Four models were trained based on these sets.

2.4.3 Testing

To test the predictive performance of the constructed models, a 10-fold cross-validation was carried out. Each training set was divided into 10 groups by splitting each data set into 10 approximately equal-sized subgroups using stratified sampling. During cross-validation, each one of the 10 subgroups was regarded as the validation set in turn, and the remainder was regarded as the training set. The average of each run was calculated.

The following measures of predictive performance of the trained models were calculated: Precision (Pre) = TP/(TP+FP), Sensitivity (Sn) = TP/(TP+FN), Specificity (Sp) = TN/(TN+FP), and Accuracy (Acc) = (TP + TN)/(TP+FP+TN+FN), where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively.

3. Results

3.1 Data summary

From the final data set of 297 records, there were 184 with passing final grades, and 113 with failing grades. The data set contained records from 77 female students and 220 male students. Student ages were normally distributed, with a peak at the 30-39 age group. The majority (28%) of the students were based in South Australia, followed closely by students in New South Wales (21%). Table 2 shows a summary of final grades per course and its corresponding study period.

3.2 Statistically significant features

As shown in Table 3, a VIF test resulted to 15 demographic and 4 performance-related features after excluding other features sharing high multi-collinearity with the others. A chi-square test applied to the remaining features in Table 3 further revealed statistically significant features (*P*-value < 0.05) in relation to the target variable *Failed*. As listed in Table 4, completing the final exam (*P*-value: 2.15 e-35), submitting all assessments (*P*-value: 1.48 e-24), and failing to submit any of the marked assessments (*P*-value: 2.07 e-09) were found to be associated with successful outcome.

Table 3. *Summary of features selected via VIF*

Feature	Type of feature	VIF Score
Part_Time	Demographic	3.86
Gender (Male)	Demographic	4.03
Location (New South Wales)	Demographic	7.02
Location (Northern Territory)	Demographic	1.82
Location (Queensland)	Demographic	4.76
Location (South Australia)	Demographic	8.14
Location (Tasmania)	Demographic	2.03
Location (Victoria)	Demographic	5.12
Location (Western Australia)	Demographic	4.60
Location (overseas)	Demographic	1.26
Age (22-24)	Demographic	3.6
Age (25-29)	Demographic	5.68
Age (30-39)	Demographic	9.06
Age (40-49)	Demographic	4.42
Age (50 and above)	Demographic	2.27
Completed_Exam	Performance	8.66
All_Assessment_Submitted	Performance	4.36
No_Assessment_Submitted	Performance	1.47
Activity_Count_0	Performance	1.34

Table 4. *Summary of statistically significant features selected via chi-square test*

Feature	Type of feature	*P*-value
Completed_Exam	Performance	2.15 e-35
All_Assessment_Submitted	Performance	1.48 e-24
No_Assessment_Submitted	Performance	2.07 e-09

3.3 Model performance

As shown in Table 5, the model trained using only demographic features was able to identify 64 records with a final passing mark correctly and 71 records with a failing final mark correctly. The model trained using only performance-related features was able to identify 182 records with a final passing mark correctly and 76 records with a failing final mark correctly. The model trained using statistically significant features was able to identify 184 records with a final passing mark correctly and all records with a failing final mark correctly. The model trained using all features was able to identify 181 records with a final passing mark correctly and 76 records with a failing final mark correctly.

Table 6 shows a summary of model performance for each of the 4 models. The model trained using only demographic features yielded an accuracy of 45.45%. The model trained using only performance-related features yielded an accuracy of 86.86%. The model trained using statistically significant features yielded an accuracy of 86.53%. The model trained using all features yielded an accuracy of 86.53%

Table 5. *Average result form 10-fold cross-validation for each model*

Model	True Negative	False Negative	True Positive	False Positive
Demographic	64	42	71	120
Performance	182	37	76	2
Statistically significant	184	40	73	0
All features	181	37	76	3

Table 6. *Summary of model performance*

Model	Precision	Sensitivity	Specificity	Accuracy
Demographic	37.17%	62.83%	34.78%	45.45%
Performance	97.43%	67.25%	98.91%	86.86%
Statistically significant	100%	64.60%	100%	86.53%
All features	96.20%	67.25%	98.36%	86.53%

4. Discussion

This study demonstrates the ability of limited demographic and performance-related data to predict student outcome in a fully online programming course. Based on the results, demographic features, specifically the ones included here were found to be weak predictors of successful outcomes in learning programming in a fully online environment. This result is consistent with other studies that looked at predicting student learning outcomes using learning analytics for other courses in online learning environments (Hu, Cheong, Ding, & Woo, 2017). It should be noted that the locations included in this study were all in Australia. But with an increase in the number of student enrollments from locations outside Australia, it would be interesting to look at how this can impact the model.

The model performance results, however, reveal how engagement features could predict student success in this type of learning environment. Adding performance-related features to demographic features was able to significantly improve the predictive performance of the trained model from 45.45% to 86.53%, indicating the importance of these data in training such models. Specifically, completing the online final exam and summative assessments were found to be statistically associated with successful outcome in learning programming in a fully online environment. Moreover, training a predictive model based solely on these 3 features demonstrated its reliability in predicting success outcome in this case as supported by an 86.53% prediction accuracy.

It is important to note that student engagement in terms of completing non-mandatory assessments was found to have no statistically significant association with student outcome for fully online programming courses. This result is different from studies that looked at MOOCs general courses and not specific to programming courses (Baker et al., 2015) where early access to resources, constant access to the courses and performing well in the formative non-marked activities are the indicators for the students' successes. There are several factors that explain why the results for the predictors online programming courses are different. First, programming needs constant practice in coding and the number of times a student accesses the courses without "doing the coding" does not affect the student's final grade. Second, in the case of the online students for the programming courses in this study, most students do not usually engage in formative non-marked learning activities and usually only view this as additional workload. This is the reason why continuous summative assessments were introduced in these courses for UniSA Online. Since non-mandatory assessments were not accounted for in the computation of the final grade, it is understandable that this shows no statistically significant association with passing the course. Lastly, with regard to other engagement methods, the online programming courses used in the study only use discussion forums primarily as a tool for asking question. This is similar to how MOOCs EdX programming courses use forums (Waller, 2019). Because of this, it is unknown if engagement in the forums is a good predictor of a successful outcome in the course. For future studies, text analysis of the discussion forums can be investigated if it can help identify students at-risk.

For online learning facilitators, this study reiterates the need to monitor student engagements in the submission of summative assessments (Baker et al., 2015). This also informs how learning facilitators can adjust how they monitor groups of students, especially those at risk of failing. For

instance, additional reminders can be set to ensure students do not miss submitting assessments and completing exams required for the course. By improving monitoring check-ins especially before assessment due dates, early intervention for non-submission of initial assessments can be prevented.

It should be noted that this study is not without any limitations. First, the features included were limited to general demographic data (e.g., name, age, gender location, type of study). Analysis of other factors such as employment status, type of work, and basis of entry can be added to further see if demographics is a possible predictor of successful outcomes in an online environment, especially for programming courses. Gender data is still presented as only two binary choices, and it is possible that a third option may have been present but not shown in the reported data. Second, the performance features included in this analysis requires completion of the whole course before a reliable prediction could be made using the models. Therefore, the models cannot be utilized prior to the course being run or during the early phase of the course. Other data that can be used in future studies is looking at the student's personality (self-efficacy, self-regulation) and previous performances from previous online courses. It is also important to investigate the generalizability of these models for other courses and online learning environments. This study advocates for the collection of finer and more specific student demographic, personality, academic, and behavioural data in a fully online learning environment to enable prediction of success outcomes early on.

Acknowledgements

The authors would like to acknowledge the UniSA Online programming students and online course facilitators for the data used in this study.

References

Alauddin, M., & Nghiem, H. S. (2010). Do instructional attributes pose multicollinearity problems? An empirical exploration. *Economic Analysis and Policy, 40*(3), 351-361.

Aslanian, C., & Clinefelter, D. (2013). Online College Students 2013. Comprehensive data on demands and preferences. The Learning House. *Inc, Louisville, KY*.

Azcona, D., Hsiao, I.-H., & Smeaton, A. F. (2019). Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. *User Modeling and User-Adapted Interaction, 29*(4), 759-788.

Azcona, D., & Smeaton, A. F. (2017). *Targeting at-risk students using engagement and effort predictors in an introductory computer programming course.* Paper presented at the European Conference on Technology Enhanced Learning.

Azmi, S., Ahmad, N., Iahad, N. A., & Yusof, A. F. (2017). *Promoting students' engagement in learning programming through gamification in peer-review discussion forum.* Paper presented at the 2017 International Conference on Research and Innovation in Information Systems (ICRIIS).

Baker, R. S., Lindrum, D., Lindrum, M. J., & Perkowski, D. (2015). Analyzing Early At-Risk Factors in Higher Education E-Learning Courses. *International Educational Data Mining Society*.

Bodily, R., & Verbert, K. (2017). *Trends and issues in student-facing learning analytics reporting systems research.* Paper presented at the Proceedings of the seventh international learning analytics & knowledge conference.

Brooks, C., Thompson, C., & Teasley, S. (2015). *A time series interaction analysis method for building predictive models of learners using log data.* Paper presented at the Proceedings of the fifth international conference on learning analytics and knowledge.

Carter, A. S., Hundhausen, C. D., & Adesope, O. (2017). Blending measures of programming and social behavior into predictive models of student achievement in early computing courses. *ACM Transactions on Computing Education (TOCE), 17*(3), 1-20.

Diana, N., Eagle, M., Stamper, J., Grover, S., Bienkowski, M., & Basu, S. (2017). *An instructor dashboard for real-time analytics in interactive programming assignments.* Paper

presented at the Proceedings of the Seventh International Learning Analytics & Knowledge Conference.

Howarth, J. (2019). *MOOCs as a Pathway into Higher Education.* Charles Sturt University,

Hu, X., Cheong, C. W., Ding, W., & Woo, M. (2017). *A systematic review of studies on predicting student learning outcomes using learning analytics.* Paper presented at the Proceedings of the Seventh International Learning Analytics & Knowledge Conference.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics, 1*(1), 6-47.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). *Predicting MOOC dropout over weeks using machine learning methods.* Paper presented at the Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs.

Koutropoulos, A., & Zaharias, P. (2015). Down the rabbit hole: An initial typology of issues around the development of MOOCs. *Current Issues in Emerging eLearning, 2*(1), 4.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (Vol. 5): McGraw-Hill Irwin New York.

Midi, H., & Bagheri, A. (2010). *Robust multicollinearity diagnostic measure in collinear data set.* Paper presented at the Proceedings of the 4th international conference on applied mathematics, simulation, modeling.

Mierswa, I., & Klinkenberg, R. (2018). RapidMiner Studio (9.1)[Data science, machine learning, predictive analytics]. In.

Nguyen, Q., Huptych, M., & Rienties, B. (2018). *Linking students' timing of engagement to learning design and academic performance.* Paper presented at the Proceedings of the 8th international conference on learning analytics and knowledge.

Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research, 21*(1).

Paechter, M., & Maier, B. (2010). Online or face-to-face? Students' experiences and preferences in e-learning. *The internet and higher education, 13*(4), 292-297.

Restrepo-Calle, F., Ramírez Echeverry, J. J., & González, F. A. (2019). Continuous assessment in a computer programming course supported by a software tool. *Computer Applications in Engineering Education, 27*(1), 80-89.

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.

Shi, C., Fu, S., Chen, Q., & Qu, H. (2015). *VisMOOC: Visualizing video clickstream data from massive open online courses.* Paper presented at the 2015 IEEE Pacific visualization symposium (PacificVis).

Shi, Z. (2011). *Advanced artificial intelligence* (Vol. 1): World Scientific.

Stone, C. (2017). Opportunity through online learning: Improving student access, participation and success in higher education.

Waller, D. R. (2019). A Case Study of Discussion Forums in Two Programming MOOCs on Differ-ent Platforms. *American Society for Engineering Education*.

Whitehill, J., Williams, J., Lopez, G., Coleman, C., & Reich, J. (2015). Beyond prediction: First steps toward automatic intervention in MOOC student stopout. *Available at SSRN 2611750*.

Yukselturk, E., & Bulut, S. (2007). Predictors for student success in an online course. *Journal of Educational Technology & Society, 10*(2), 71-83.

Zhou, M. (2016). Chinese university students' acceptance of MOOCs: A self-determination perspective. *Computers & Education, 92*, 194-203.