Education and Information Technologies (2021) 26:69–88
https://doi.org/10.1007/s10639-020-10260-x

A two-phase machine learning approach for predicting student outcomes

Omiros Iatrellis[1] · Ilias K. Savvas[1] · Panos Fitsilis[1] · Vassilis C. Gerogiannis[1]

Received: 11 May 2020 / Accepted: 22 June 2020 / Published online: 25 June 2020

Abstract

Learning analytics have proved promising capabilities and opportunities to many aspects of academic research and higher education studies. Data-driven insights can significantly contribute to provide solutions for curbing costs and improving education quality. This paper adopts a two-phase machine learning approach, which utilizes both unsupervised and supervised learning techniques for predicting outcomes of students following Higher Education programs of studies. The approach has been applied in a case-study which has been performed in the context of an undergraduate Computer Science curriculum offered by the University of Thessaly in Greece. Students involved in the case study were initially grouped based on the similarity of specific education-related factors and metrics. Using the K-Means algorithm, our clustering experiments revealed the presence of three coherent clusters of students. Subsequently, the discovered clusters were utilized to train prediction models for addressing each particular cluster of students individually. In this regard, two machine learning models were trained for every cluster of students in order to predict the time to degree completion and student enrollment in the offered educational programs. The developed models are claimed to produce predictions with relatively high accuracy. Finally, the paper discusses the potential usefulness of the clustering-aided approach for learning analytics in Higher Education.

Keywords Learning analytics · Unsupervised learning · Supervised learning · Higher education

✉ Omiros Iatrellis
iatrellis@hotmail.com

[1] University of Thessaly, Larissa, Greece

1 Introduction

Graduation and time to degree completion are always high concerns in Higher Education Institutes (HEIs) and have attracted the attention of policy makers, educators and researchers in recent years (Yue and Fu 2017). Any delay in degree completion represents a waste of resources both for learners and the HEIs, thus affecting the returns of investment in Higher Education (Iatrellis et al. 2020). Particularly in Greece, as reported by the Hellenic Quality Assurance and Accreditation Agency (HQA), many students at Greek HEIs, which are highly publicly subsidized, do not graduate on time (HQA 2017). As a result, there is an imperative need for insightful planning procedures to identify factors, which possibly contribute to a delayed time to completion. Learning analytics presents a potential solution to be used as a practical methodology for building appropriate prediction models that can support management of HEIs to effectively and efficiently plan their educational services.

The research aim of the current study focuses on predicting student outcomes with reference to computer science undergraduate curricula offered by HEIs in Greece. In particular, this paper presents a machine learning approach that can be used to make predictions for the student's time to degree and student's potential enrollment in another program of studies (for example enrollment in postgraduate studies).

In HEIs, we can identify four categories of student outcomes which include: a) Academic outcomes, b) Learning outcomes, c) Financial outcomes, and d) Perceptual outcomes. Being more specific, the academic outcomes mainly represent the student progress in terms of completion of the requirements for competing the degree. The learning outcomes are associated with the knowledge and skills acquired by the students as outcomes when they follow learning processes, which may include various extracurricular activities in diverse settings, varying from conventional face-to-face to fully online and distance-learning (synchronous or asynchronous) educational processes. Financial outcomes cover the financial value creation and transactions which are performed in the context of an academic program; financial outcomes are used to assess the efficient usage of all HEI resources. The perceptual outcomes might be the most intangible set of outcomes of an educational process, which can represent a student's satisfaction with education received and its faculty members. The scope of student outcomes addressed by the paper can be considered to fall within the academic and financial categories.

In this paper, we started with the aim of grouping students from a data-driven perspective. In this regard, in phase one, unsupervised machine learning approach using the K-Means algorithm was applied. The resulted clusters were then utilized in phase two to train the prediction models for students who share a common set of characteristics or achieved similar outcomes (such as student's personality, high school type among others). Furthermore, the approach adopted by the study was compared to other non-clustering guided prediction models. In this respect, this paper implements a comparative analysis to estimate the magnitude of improvement in prediction accuracy using our approach against other simpler approaches.

The rest of the paper is organized as follows. After presenting the related work in the area of our interest and an overview of the Higher Education system in Greece in Section 2, the EDUC8 framework is presented in Section 3, which will utilize the proposed approach. In Section 4, the significance of the study is analyzed. Section 5 overviews the proposed approach, while Section 6 describes in detail the methodology we followed. Experimental results and comparative analysis between the clustering-aided and the non-clustering approach are presented in Sections 7 and 8 respectively. The final section concludes the work presented in this paper combined with our thoughts for future work.

2 Background

2.1 Related work

The term "Learning analytics" was defined during the 1st International Conference on Learning Analytics and Knowledge[1] as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. Learning analytics constitutes a multi-disciplinary application field where approaches from various domains can be utilized (such as artificial intelligence, machine learning, information retrieval, statistics, and visualization) (Chatti et al. 2012). It is considered that is closely related to educational data mining. However, the former aims at improving the student learning process, while the latter aims at detecting patterns of interest in educational data (Aldowah et al. 2019).

Several research studies in learning analytics explore the potential of machine learning approaches dealing with diverse objectives, such as prediction of student's learning performance, recommendation of resources, evaluation of learning material and services, feedback to all HEI's stakeholders and student behavior modeling (Papamitsiou and Economides 2014).

An important parameter in most relevant studies is the adopted techniques for analyzing the collected data. Among these techniques, the most popular are logistic regression, clustering, support vector machines, decision trees, random forest, naïve Bayes, and artificial neural networks. Table 1 presents a classification of recent studies according to the research objectives and the technique they use.

Some papers propose hybrid approaches by presenting an amalgamation of various data mining techniques. For example, a prediction algorithm for evaluating student's performance in academia has been developed based on both classification and clustering techniques (Francis and Babu 2019). In phase 1, the study carried out the experiments using Support Vector Machines, Naïve Bayes, Decision tree and Neural Network classifiers in order to identify which of the features give best results. In phase 2, the obtained features are passed into K-Means clustering algorithm to acquire clusters features indicating high, medium and low performing students.

Another interesting hybrid approach is presented in (Bhogan et al. 2017). This approach applied an enhanced K-strange points clustering algorithm as well as a Naïve Bayes classification algorithm. The authors compared their solution with another hybrid approach which combines K-means clustering algorithm with decision trees. The implementation results were used to predict the students' performance in

[1] https://tekri.athabascau.ca/analytics/

Table 1 Classification of recent studies according to the research objectives and analysis method

Research objectives (goals)	Methods	Authors & Year (Paper Ref.)
Prediction of student's learning performance	Decision Trees, Support Vector Machines, Artificial Neural Networks	(Xu et al. 2019)
	Artificial Neural Networks	(F. Yang and Li 2018)
	Logistic Regression	(Lee 2018)
	Clustering	(Oyelade et al. 2010), (Anand et al. 2018)
	Logistic Regression, Artificial Neural Networks	(Umair and Majid Sharif 2018)
	Support Vector Machines, Clustering	(Al-Shehri et al. 2017)
	Artificial neural networks, support vector machines, logistic regression, Naïve Bayes, decision trees	(Hussain et al. 2019)
	Random Forest	(Abubakar and Ahmad 2017)
	Support Vector Machines, Naïve Bayes, Decision tree, Artificial Neural Network, Clustering	(Francis and Babu 2019)
	Clustering, Naïve Bayes, decision trees	(Bhogan et al. 2017)
Recommendation of resources	Naïve Bayes	(Muñoz-Merino et al. 2018)
	Decision Trees, Random Forest	(Pliakos et al. 2019)
	Decision Trees	(Asif et al. 2017)
Evaluation of learning material and services	Logistic Regression, Naïve Bayes, Decision Trees, Random Forest	(Abidi et al. 2018)
	Artificial Neural Networks	(Zhang et al. 2019)
Feedback to HEI stakeholders	Random Forest	(Chung and Lee 2019)
	Random Forest, Artificial Neural Networks, Decision Trees, Random Forest	(Gray and Perkins 2019)
	Clustering	(Kizilcec et al. 2013), (Iatrellis et al. 2020), (Bharara et al. 2018)
	Artificial Neural Networks, Logistic Regression	(Mason et al. 2018)
	Logistic Regression	(Burgos et al. 2018)
	Support Vector Machines	(Pang et al. 2017), (Cardona and Cudney 2019)
Student/Student behavior modeling	Naïve Bayes	(Ruiperez-Valiente et al. 2019)
	Clustering	(Pasina et al. 2019), (Fan and Sun 2017), (Nájera et al. 2017)
	Artificial Neural Networks	(T. Y. Yang et al. 2017)

examination in advance, so that necessary measures can be taken to improve on their performance so as for students to score better marks.

In this paper, we embraced an approach of clustering guided predictions performed in two phases aligned with the computer science (CS) programs of studies in

Greece. In the first phase, we divide the data set into smaller groups using the K-Means algorithm to classify features related to the basic cycle of the CS program. At the end of this phase, a set of coherent student clusters is obtained. In the second phase, the discovered clusters are utilized to train prediction models for students in the specialization cycle who naturally share a common set of characteristics and parameters. By exploiting the strengths of both unsupervised and supervised methods in two distinct phases, the developed models claim to produce better results in predicting student outcomes.

2.2 Overview of the higher education system in Greece

This section aims to deliver a brief overview of the application domain endorsed by the present research work. As alluded to earlier, the current research work is focused on undergraduate CS programs.

In Greece, the students who want to have access in the tertiary education, must participate in national examinations. These exams are held after the students have received their secondary education certificate from an academic high school (GEN) or from a vocational high school (EPAL). Higher tertiary education is provided by universities and polytechnics and the offered undergraduate courses typically last 4 or 5 years.

The exemplary curriculum of the CS university department which has been considered in our research work comprises of a basic cycle (semesters 1-4) and three parallel specializations (semesters 5-8). The basic academic cycle is common to all students and provides the necessary background knowledge in mathematics, physics, computer science, informatics, electronics and telecommunications. After the basic cycle, students enroll in one of three distinct areas of specialization, namely Software Engineering, Network Engineering, or Computer Engineering, where they attend more applied and technology-oriented courses in CS, most of which are related to the area of specialization they have chosen. Every year, the CS department used in our study admits a student intake of approximately 250 students. Among them, 20% of the students are originating from EPAL schools.

3 Significance of the study

According to Yue and Fu there are two definitions for Time to Degree (TTD) (Yue and Fu 2017). The first way to look TTD is the elapsed time between admission and graduation without taking into account any stop time, while the second is calculated by considering only the enrolled semesters for each student. In any case, the TTD metric allows HEIs to measure whether students are taking longer than average to complete their degrees. The longer students take to earn their degree, the more it costs themselves, their families and the state. Any delays also contribute to the misallocation of HEI's tangible and intangible resources, which potentially means that fewer students can be served because of limited capacity of the HEI organization. At the same time, HEIs are increasingly interested in identifying applicants who have higher tendency to enroll at their programs and accordingly to better schedule and allocate their resources, scholarship and financial aid.

To address the above mentioned challenges the current paper builds upon our previous research work which implemented EDUC8 (EDUCATE), an integrated information technology which recommends the dynamic and personalized composition of students' learning pathways during execution phase (Iatrellis et al. 2019a) (Iatrellis et al. 2020). Adding to our previous work, the current study proposes the utilization of unsupervised and supervised machine learning as an assistive artifact within the context of developing learning pathways with the aim of optimizing educational services in conjunction with the minimization of costs for tertiary institutions. In this regard, insights learned from the two-phase machine learning models are fed as input into the EDUC8 multi-layered integrated software environment with the aim to achieve predictions with better results on student outcomes.

The proposed hybrid machine learning models are capable of predicting TTD and student enrollment in programs with higher accuracy compared to other simpler approaches (Oyelade et al. 2010), (Anand et al. 2018). Moreover, the discovery of coherent clusters of students can extend opportunities to answer questions, or raise further intriguing questions for future studies such as:

- How do student clusters vary with respect to specific characteristics such as type of originating high school or academic performance?
- How do student clusters vary with respect to learning outcomes in terms of TTD and academic destination after graduation?
- Is it possible to correlate student outcomes and other education-related factors, such as "Time to Complete Basic Cycle" for example?

In a broader context, we believe that the current study can be considered within other comparable learning analytics problems in Higher Education. Specifically, the two-phase unsupervised-supervised approach adopted by the study can arguably draw further attention to the significance of student clustering while dealing with prediction-related problems. In this respect, the paper evidently demonstrates that the weight of features used for training our models remarkably varies from a cluster of students to another.

4 The EDUC8 framework

This section provides an overview of the EDUC8 framework, which implements and utilizes the proposed hybrid machine learning approach. As depicted in Fig. 1, the conceptual architecture of the EDUC8 environment comprises of distinct architectural layers. The lower layer of the EDUC8 architecture is called "semantic infrastructure layer" as it encloses the required semantic information structure. This layer represents the cornerstone of the integrated software environment since it realizes the respective business logic concerning the dynamic and personalized composition of students' learning pathways. It encloses 1) the Learner part, 2) the Learning pathway part, 3) the Business and Finance part (in HEIs) and 4) the Quality Assurance part in education provision environments. These four knowledge streams are intertwined and together with a set of semantic rules are utilized for the academic advising guidelines formalization, the modeling of learning pathways, and the recycling of knowledge through the dynamic generation of facts by the rule engine.

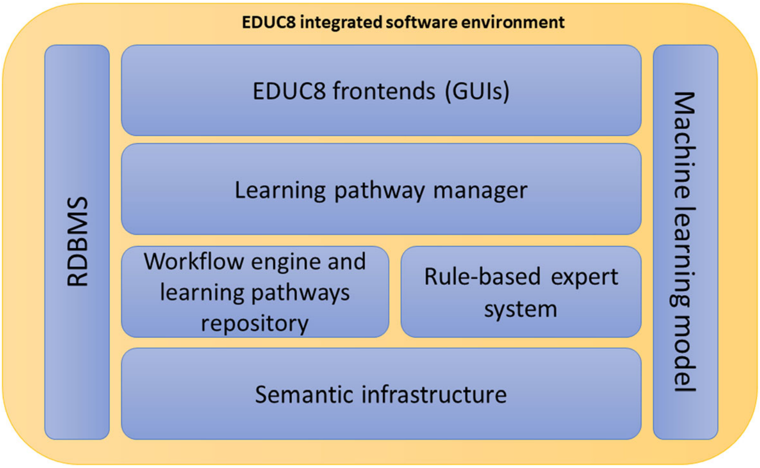


Fig. 1 EDUC8 conceptual architecture

The next layer of the implemented software infrastructure includes two mechanisms responsible for the execution of the learning pathways and the respective set of semantic rules. The first submodule contains the workflow-part of each learning pathway. The execution of a learning pathway is performed inside a workflow engine since it is considered as a higher education business process. The mechanism also encompasses a repository, which contains a set of available learning pathways parts so as to select the most appropriate pathway for each student during the evolvement of the educational program. In this way, the EDUC8 environment reasons over the stored knowledge at the decision points of each sub-process so as to select the best learning step taking under consideration the academic advising guidelines alongside with the academic background and personal characteristics of the student. The next module in the same layer encloses a rule-based expert system, which undertakes the task of executing the semantic rules and performing decision-making, concerning the suggestion for the appropriate next step of the learning pathway.

The next layer of the EDUC8 conceptual architecture is the "Learning pathway manager" layer. The "Learning pathway manager" module is responsible for updating the structure of the currently executed learning pathway by taking as input the decision and proposal from the expert system in the form of facts. It is also responsible for processing the business logic of EDUC8 components and for accessing the data layer to retrieve, modify and delete data to and from the relational database management system (RDBMS). The "Learning pathway manager" contains all the automatically generated forms that are needed for the process of data entry that correspond to learning and administrative tasks.

The upper layer allows the integration and presentation of several client-side components for various tools and applications of the EDUC8 software environment. It also serves for the triggering of applications, tools and services of the integrated EDUC8 software environment.

Finally, the conceptual architecture includes two modules positioned vertically, which can be accessed by various software components. The "RDBMS" module offers a management layer to save and retrieve data from a database during the execution of a learning pathway. Finally, the machine learning module is specifically applied to learn potential insights pertaining to student characteristics, education factors and outcomes, which can be used effectively to conceptualize the system's structure or behavior. The work presented in this paper attempts to extend the machine learning module by including a clustering guided approach that can contribute to improving the accuracy of predicting student outcomes.

5 Approach overview

In the case study that we have followed, the adoption of the unsupervised and supervised learning was conducted over two phases. Initially, our analysis aimed to discover the potential presence of some coherent clusters of students. In this regard, the K-Means algorithm was used to explore the student groups from a data-driven perspective. This phase realized the discovery of three well-separated student clusters as it is explained in the next paper section.

The second phase included the development of machine learning models that can make predictions on student outcomes. The prediction models were developed with respect to every cluster of students separately. Specifically, a regression model was trained for predicting the TTD, and another binary classifier was used for predicting Student Enrollment In Postgraduate Studies (SEIPS). We used one computational node (AMD@ FX (tm)-4100 Quad-Core Processorx4, with UBUNTU 16.04 LTS operating system, gcc 4.84, and Python 3.7) to train and test the prediction models. Figure 2 illustrates the steps of the presented approach.

6 Methodology

6.1 Data description

The main source of the data used in the case study is from the CS Department of the University of Thessaly (UTH),in Greece. UTH is a HEI offering degrees in the

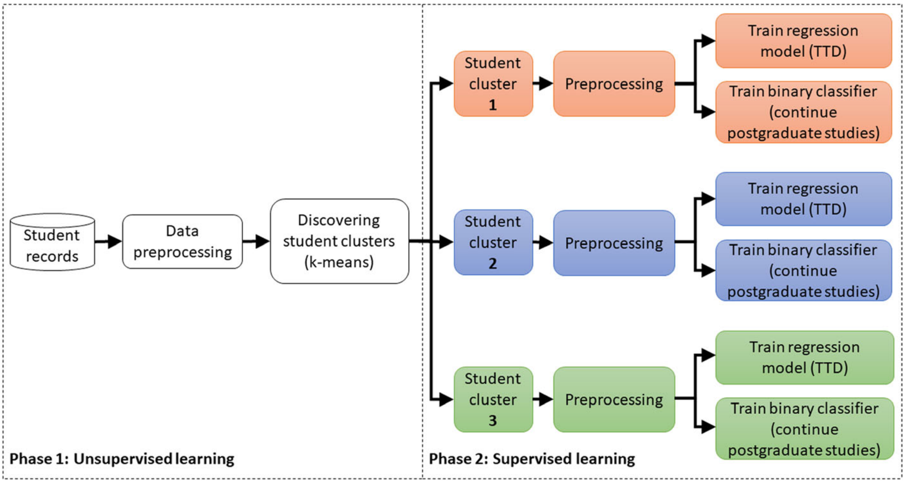

Student records
Data preprocessing
Discovering student clusters (k-means)
Student cluster 1
Preprocessing
Train regression model (TTD)
Train binary classifier (continue postgraduate studies)
Student cluster 2
Preprocessing
Train regression model (TTD)
Train binary classifier (continue postgraduate studies)
Student cluster 3
Preprocessing
Train regression model (TTD)
Train binary classifier (continue postgraduate studies)
Phase 1: Unsupervised learning
Phase 2: Supervised learning

Fig. 2 Approach overview

traditional, face-to face way. We acquired a dataset extracted from the student university database. This dataset included records about graduates excluding students who graduated on time. The data comprised 1100 records over six academic years, from 2007 to 2012. All records were completely anonymized for privacy purposes.

The dataset records contained ample information about a student's studies from admission to graduation. Specifically, a typical student record included multiple data fields such as type of high school, date of admission and Time to Degree. Initially, we explored the variables that can serve as features for training our machine learning models, most of which were derived from the learner model implemented in the EDUC8 framework (Iatrellis et al. 2019b). Based on our intuition, many irrelevant variables of the student data could be obviously excluded (e.g. admission date, region). Table 2 lists the variables that were initially considered as possible features for the clustering and prediction models as well.

Our emphasis was placed on students who delay the completion of their studies by considering the resulting consequences of these delays on the HEI's educational policy. Thus, we were concerned with students who have completed their studies in more than five (4+1) years. We have used learning analytics with the aim to offer the University's executives with some practical suggestions for specifying evidence-based directions for future strategies.

Table 2 Student variables explored as possible features

Variables explored	Description
Grade Point Average (GPA)	A 10-scale number representing the average value of the accumulated final grades earned in courses
Specialization field	Software engineering, Network Engineering or Computer Engineering
Capstone project grade	10-scale grade for the capstone
Time to Complete Basic Cycle	The time it takes for a student to complete the basic cycle
First Year Performance in Core Subjects	A 10-scale number representing the average value of the academic performance in the first year's core subjects
High School	Academic high school (GEN) or Vocational high school (EPAL)
Rank-in-class	A numerical representation of a student's academic achievements in comparison to those of their peers
Student mobility program	It represents whether a student has participated in a mobility program (Boolean)
Internship	It represents whether a student has completed internship (Boolean)
Awards	It represents whether a student has received any award (Boolean)
Grades in high school	20-scale grade in high national school
Grades in entrance exams	20-scale grade in entrance exams
RIASEC/Holland code	RIASEC codes derived from the Holland's theory (Nauta 2010) and utilized by EDUC8 platform in order to provide an avenue for recommending learning pathways based on student's personality

6.2 Data preprocessing

This section describes the data preprocessing procedures conducted prior to training our machine learning models including the clustering and prediction models. The preprocessing step included three steps: A) Removing outliers, B) Feature scaling, and C) Extraction of features that are considered as indicators of the offered education quality. The preprocessing was implemented using the Python language, gnuplot and MS Excel.

The existence of outliers was largely unavoidable. An outlier is an object that deviates significantly from the rest of the objects (Malhotra 2014). Outlier analysis is essential for identifying those data points that are over influential and must be analyzed for exclusion from the data set. In our case, data preprocessing included analysis and removal of outlier data as it is explained in the following sections.

6.2.1 Removing outliers

Sometimes students are interrupting their studies or even withdrawing from their studies without taking a license to have a "formal" break. Therefore, we considered only the students whose TTD were no longer than 10 years and did not miss two consecutive semester enrollments. The excluded outliers represented approximately 2% of the overall dataset. Figure 3 plots a histogram of the TTD which has been used to identify outliers.

6.2.2 Feature scaling (min-max normalization)

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing phase.

In this regard, the min-max normalization method was used which resulted in rescaling the range of features to a range in the [0, 1] interval. All data values were normalized and scaled through the following formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

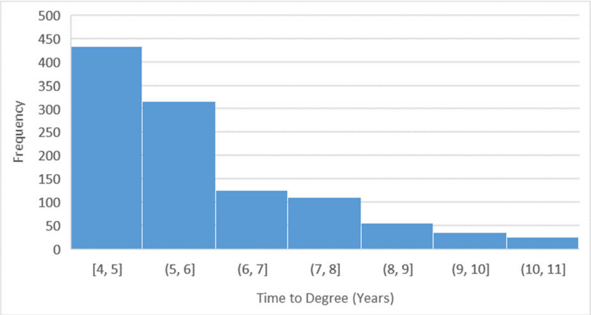


Fig. 3 Histogram of the TTD variable

Table 3 K-Means algorithm parameters

Parameter	Value
Number of clusters	2–5
Centroid initialization	Random
Similarity metric	Euclidian distance

where x is an original data value and x' is the normalized value.

6.2.3 Feature extraction

Feature extraction for the clustering phase was focused on features for which the data can be collected before the student's specialization declaration time. Initially the clustering model included TTD and Time To Complete Basic Cycle (TTCBC) as features. Subsequently, more features were extracted from the dataset, which are related to some quality measures as proposed in relevant research works (Kappe and Van Der Flier 2012), (McKenzie and Schweitzer 2001). In particular, two quality measures could be captured from the student database aligned with the following hypothesis:

- Higher grades in High school and entrance examinations are often associated with better university results and higher grades
- Performance in first year core subjects is often one of the most commonly found metrics for student success

The student database did not include fields which explicitly represented all required features. However, all required features could be derived based on specific SQL queries that were executed on the database. In this regard, two new features were selected named as Grade in Entrance Exams (GIEE) and First Year Performance In Core Subjects (FYPICS). Eventually, only the FYPICS was included, because the GIEE contained a significant amount of missing values. Instead, the type of student's high school (GEN or EPAL) was used as a factor that could have potential influence on education outcomes.

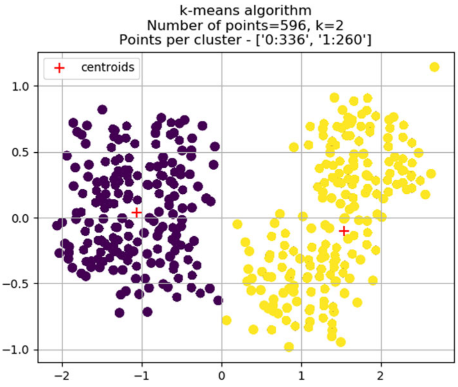


Fig. 4 Clustering experiments with two clusters (K = 2)

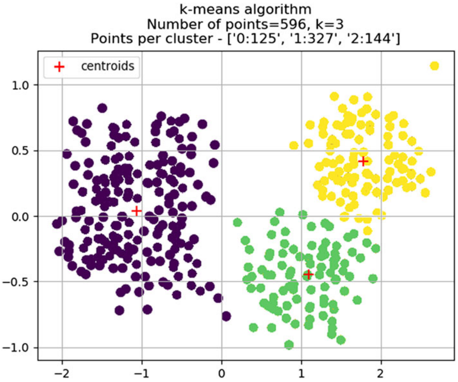


Fig. 5 Clustering experiments with three clusters (K = 3)

6.3 Clustering approach

One of the simplest and most popular clustering techniques is K-means algorithm (MacQueen et al. 1967). K-means requires as input the number of clusters (K), and starts by choosing K initial centroids. Then, K-means assigns the data points to their closest centroid by calculating their distance. In our study, we have used the Euclidean distance between any two points. After that, the K-means algorithm recalculates the centroids and the algorithm terminates when the centroids do not change position (or when some convergence criteria are fulfilled).

In this study, we implemented the K-means algorithm for K values varying from 2 to 5 and in addition we applied Principal Component Analysis (PCA) (Pearson 1901) in order to reduce the dimensionality of data and increase its visualization. The quality produced by PCA was 91.06% by reducing the dimension from 3D to 2D, thus the results obtained were extremely accurate considering their quality. Table 3 summarizes the parameters used during our clustering experiments.

As it is already mentioned, the clustering experiments were examined using a number of clusters ranging from 2 to 5. In this manner, the appropriate number of clusters was decided based on the experimental results.

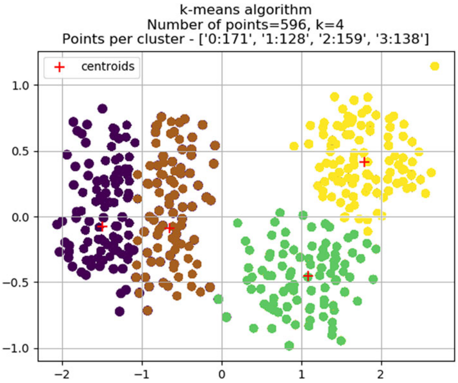


Fig. 6 Clustering experiments with four clusters (K = 4)

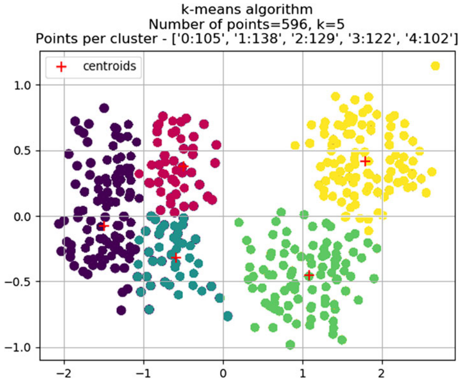


Fig. 7 Clustering experiments with five clusters (K = 5)

With the aid of PCA analysis, the computed clusters were projected into two dimensions as shown in Fig. 4, 5, 6, and 7. These figures represent the output of a corresponding clustering experiment using a different number of clusters (K). Initially for K=2, a promising tendency of clustering was indicated, where the data space was obviously separated into two big clusters. Similarly for K=3, the clusters were still well-separated. However, when K is greater than 3, the produced clusters only separate well defined groups so it was clear that this was the most significant division of the data set. In light of that, it turned out that there were three potential clusters of students that best separated the dataset.

The specific education related factors for each of the three resulted clusters are summarized in Table 4.

6.4 Random Forest algorithm

In phase 2, we applied Random Forest, a tree-based classification and regression method, developed by Leo Breiman (Breiman 2001). The performance of the Random Forests method is quite competitive with other supervised learning algorithms (Caruana and Niculescu-Mizil 2006). The Random Forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or the mean prediction of the individual trees. During the training phase, each tree selects appropriate test functions and labels leaf node probabilities. For the evaluation, a test sample x is propagated through each tree leading to a classification probability $p_t(k|x)$ of the t-th tree. A forest's joint probability is presented as follows:

Table 4 Summary of cluster characteristics

	~ % of dataset	High school		Avg. FYPICS	Avg TTCBC (Years)	Avg. TTD (Years)
		% EPAL	% GEN			
Cluster 1	54.87%	17.30%	82.70%	8.1	2.5	5.5
Cluster 2	24.16%	18.50%	81.50%	7.1	3.4	6.1
Cluster 3	20.97%	19.40%	80.60%	6.2	4.8	7.6

Table 5 Parameters of random forests models

Number of decision trees	8
Maximum depth of decision trees	64
Number of random splits per node	128

$$p(\mathrm{k}|\mathrm{x}) = \frac{1}{H} \sum_{k=1}^{H} p_{\mathrm{h}}(\mathrm{k}|\mathrm{x})$$

Therefore, given x, the predicted class y' is:

$$argmax\, p(\mathrm{k}|\mathrm{x})$$

$$k = 1, 2 \ldots H$$

Using Random Forests, two models were developed for each cluster of students. More specifically, the prediction models included a regression model for predicting the TTD, and a binary classifier for predicting SEIPS. The predicted SEIPS represented either that the student will continue for the Master's degree or not. Table 5 presents the parameters used for training the random forests.

6.5 Feature selection

As mentioned in subsection 7.1, the utilized dataset initially contained several features, however not all of them were relevant or usable. Intuitively irrelevant feature were simply excluded. To assess individual feature importance, we have applied the Permutation Feature Importance (PFI) technique (Fisher et al. 2019).

Table 6 presents the set of features used by both models.

It was observed that the importance of features remarkably varied from a student cluster to another. For instance, the student GPA scored the highest importance within the first cluster while training the "Time to Degree" regression model. On the other hand, the "Time to Complete Basic Cycle" feature had the highest importance in the third cluster. Table 7 lists the importance of features scored while training the TTD prediction model for every cluster. This can be interpreted as a relative disparity

Table 6 Selected features for the prediction models

Prediction model	Selected features
TTD regression model	GPA, Specialization field, Capstone project grade, Time to Complete Basic Cycle, First Year Performance in Core Subjects, High School, Rank-in-class
SEIPS classification model	GPA, Specialization field, Capstone project grade, Time to Complete Basic Cycle, First Year Performance in Core Subjects, High School, Rank-in-class, TTD

Table 7 TTD Prediction Model: Importance of selected features with respect to the three student clusters

Feature	Feature importance score (~)		
	Cluster 1	Cluster 2	Cluster 3
GPA	0.84	0.41	0.38
Specialization field	0.13	0.24	0.18
Capstone project grade	0.33	0.27	0.27
Time to Complete Basic Cycle	0.76	0.85	0.54
First Year Performance in Core Subjects	0.43	0.59	0.27
Rank-in-class	0.49	0.15	0.05
High School	0.13	0.23	0.55

between student clusters, which obviously reflected on the varying importance of the selected features.

7 Experimental results

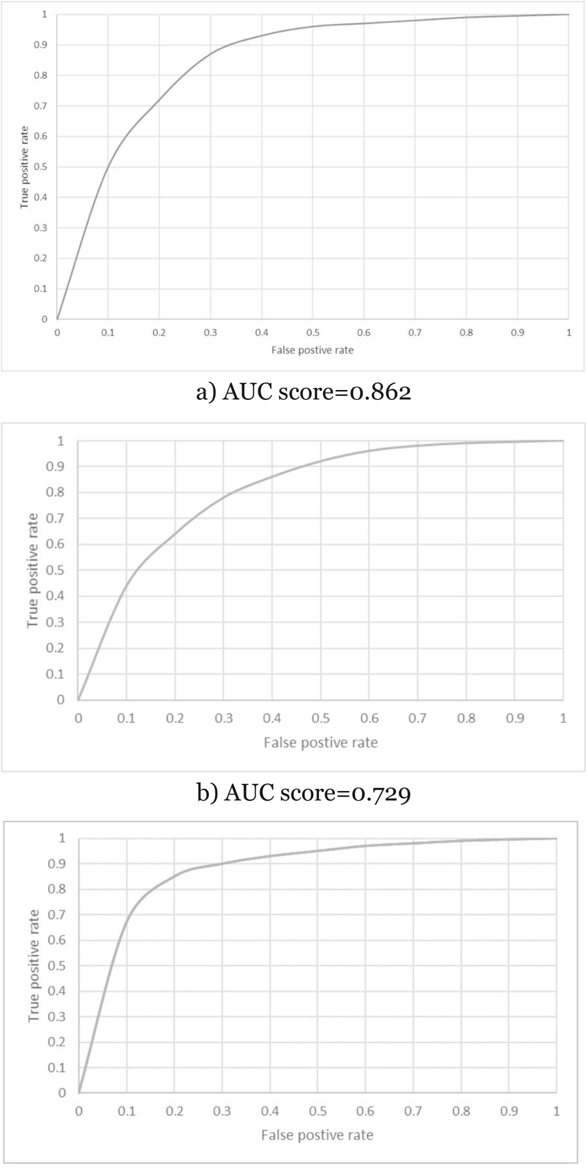
Each predictive model was tested using a subset from the dataset described in section 7.1. In this regard, the student data were separated randomly into two independent datasets: training and test sets. The training set comprised of 60% of the entire dataset. The prediction error of each model was internally calculated using 10-fold cross validation. Table 8 provides evaluation metrics for the TTD regression models while Fig. 8 shows the Area under the Receiver Operating Characteristics (ROC) curve (AUC) of the SEIPS classifiers. Furthermore, Table 9 shows the measures of precision, recall and accuracy of the SEIPS classifiers. According to Table 8, cluster 1 scored the best value for the coefficient of determination, which has the value of 0.87. Concerning the SEIPS classifier, the results for Cluster 1 and Cluster 3 were very similar and shared higher values for accuracy and AUC score.

8 Comparative analysis

It was important to evaluate the clustering guided approach compared to other traditional approaches. In this respect, we applied comparative analysis to assess the level of

Table 8 The average 10-fold cross-validation accuracy of the TTD predictor

TTD predictors	Relative absolute error	Relative squared error	Coefficient of determination
Cluster 1	0.26	0.13	0.87
Cluster 2	0.29	0.18	0.82
Cluster 3	0.27	0.17	0.83



a) AUC score=0.862

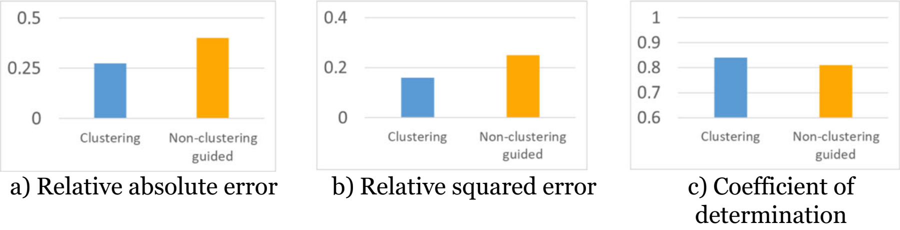
b) AUC score=0.729

c) AUC score=0.872

Fig. 8 Accuracy of the SEIPS binary classifier

Table 9 The average 10-fold cross-validation accuracy of the SEIPS classifier

SEIPS classifier	Precision	Recall	Accuracy
Cluster 1	0.852	0.891	0.821
Cluster 2	0.713	0.752	0.729
Cluster 3	0.812	0.878	0.832



a) Relative absolute error b) Relative squared error c) Coefficient of determination

Fig. 9 TTD Prediction Accuracy: Clustering–guided predictions against non-clustering

improvement in prediction accuracy using our approach against non-clustering-based predictions.

To achieve this, we implemented two additional models including a regression model for predicting the TTD, and a binary classifier for predicting SEIPS. Both models were trained using the Random Forest learning algorithm, and parameters as mentioned earlier in subsection 7.5. However, the new models were trained and tested with respect to the whole dataset without considering student clusters. The comparative analysis clearly showed that the clustering guided approach can make finer-grained predictions of TTD and SEIPS. Figures 9 and 10 demonstrate the comparative analysis results. It is important to mention that the clustering guided evaluation metrics are calculated as the median of the three models trained for student clusters.

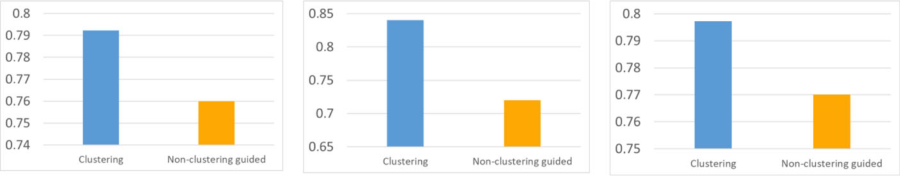
9 Limitations

We acknowledge that the students in this study were clustered based on a mere data-driven standpoint. It should be taken into account that adding faculty members perspective (e.g. academic advisors diagnosis) may group students differently (Iatrellis et al. 2017) and lead to different results.

In order to alleviate these limitations, EDUC8 framework integrates the two-phase data-driven approach presented in this paper with a semantic enhanced knowledge-based approach for making better predictions on student outcomes and better decisions regarding their academic plans.

10 Conclusions

In an increasingly competitive higher education sector, universities face significant challenges when it comes to improving student outcomes. By improving the time to



a) Precision b) Recall c) Accuracy

Fig. 10 SEIPS Destination Prediction Accuracy: Clustering–guided predictions against non-clustering

degree completion and by identifying student tendency to enroll at their programs, HEIs can improve retention rates, which affects almost all segments of HEI metrics: reputation, financials, ranking etc. In this regard, the proposed approach based on machine learning models presents an enhanced assistive artifact to exploit the data repositories used in EDUC8 so as to provide insights for educational improvement, cost cutting or identifying students who are academically at risk. In a broader context, we believe that the study can make an important contribution toward improving student outcomes with learning analytics in Higher Education. In order to achieve the aforementioned challenges, the proposed learning analytics approach utilizes a two-phased approach which applies both unsupervised and supervised techniques in an attempt to leverage the significance of student clustering while dealing with prediction-related problems.

Initially, the paper evidently demonstrated a relative disparity of the selected features from clustering of students to another. Accordingly, data clustering was applied to learn potential insights pertaining to student characteristics, education factors and outcomes. The experimental results proved that the proposed approach yielded clear evidence that can contribute to improving the accuracy of predicting student outcomes compared to non-clustering based predictions.

Our intentions for further work will be twofold. First, we will apply specialized algorithms to identify precisely the outliers like the Local Outlier Factor (LOF), which in addition will be trained with sophisticated clustering techniques like DBSCAN. Secondly, we will explore comparatively the outcomes of different machine learning techniques for the two phases of the current approach.

References

Abidi, S. M. R., Hussain, M., Xu, Y., & Zhang, W. (2018). Prediction of confusion attempting algebra homework in an intelligent tutoring system through machine learning techniques for educational sustainable development. *Sustainability (Switzerland), 11*(1). https://doi.org/10.3390/su11010105.

Abubakar, Y., & Ahmad, N. B. H. (2017). Prediction of students ' performance in E- learning environment using random Forest. *International Journal of Innovative Computing, 7*(2), 1–5.

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics, 37*(April 2018), 13–49. https://doi.org/10.1016/j.tele.2019.01.007.

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., … Olatunji, S. (2017). Student performance prediction using support vector machine and K-nearest neighbor. *Canadian Conference on Electrical and Computer Engineering*, 1–4. https://doi.org/10.1109/CCECE.2017.7946847.

Anand, V. K., Abdul Rahiman, S. K., Ben George, E., & Huda, A. S. (2018). Recursive clustering technique for students' performance evaluation in programming courses. *Proceedings of Majan international conference: Promoting entrepreneurship and technological skills: National Needs, global trends, MIC 2018*, 1–5. https://doi.org/10.1109/MINTC.2018.8363153.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers in Education, 113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007.

Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies, 23*(2), 957–984. https://doi.org/10.1007/s10639-017-9645-7.

Bhogan, S., Sawant, K., Naik, P., Shaikh, R., Diukar, O., & Dessai, S. (2017). Predicting student performance based on clustering and classification. *IOSR Journal of Computer Engineering, 19*(03), 49–52. https://doi.org/10.9790/0661-1903054952.

Breiman, L. (2001). Random forests. *Machine Learning*, 1–122. https://doi.org/10.1201/9780367816377-11.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering, 66*, 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005.

Cardona, T. A., & Cudney, E. a. (2019). Predicting student retention using support vector machines. *Procedia Manufacturing, 39*, 1827–1833. https://doi.org/10.1016/j.promfg.2020.01.256.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. https://doi.org/10.1145/1143844.1143865.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning, 4*(5–6), 318–331. https://doi.org/10.1504/IJTEL.2012.051815.

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review, 96*, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030.

Fan, Z., & Sun, Y. (2017). Clustering of college students based on improved K-means algorithm. *Proceedings - 2016 International Computer Symposium, ICS 2016*, 676–679. https://doi.org/10.1109/ICS.2016.0139.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(vi).

Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of Medical Systems, 43*(6). https://doi.org/10.1007/s10916-019-1295-4.

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers in Education, 131*(July 2018), 22–32. https://doi.org/10.1016/j.compedu.2018.12.006.

HQA. (2017). *Higher education quality report - 2017. HQA* (Vol. 1).

Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., & Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review, 52*(1), 381–407. https://doi.org/10.1007/s10462-018-9620-8.

Iatrellis, O., Kameas, A., & Fitsilis, P. (2017). Academic advising systems: A systematic literature review of empirical evidence. *Education in Science, 7*(4), 90. https://doi.org/10.3390/educsci7040090.

Iatrellis, O., Kameas, A., & Fitsilis, P. (2019a). A novel integrated approach to the execution of personalized and self-evolving learning pathways. *Education and Information Technologies (2019) 24:781-803*, *24*(ISSN 1360-2357). https://doi.org/10.1007/s10639-018-9802-7.

Iatrellis, O., Kameas, A., & Fitsilis, P. (2019b). EDUC8 pathways: Executing self-evolving and personalized intra-organizational educational processes. *Evolving Systems, 11*, 227–240. https://doi.org/10.1007/s12530-019-09287-4.

Iatrellis, O., Savvas, I. K., Kameas, A., & Fitsilis, P. (2020). Integrated learning pathways in higher education: A framework enhanced with machine learning and semantics. *Education and Information Technologies, 21*. https://doi.org/10.1007/s10639-020-10105-7.

Kappe, R., & Van Der Flier, H. (2012). Predicting academic success in higher education: What's more important than being smart? *European Journal of Psychology of Education, 27*(4), 605–619. https://doi.org/10.1007/s10212-011-0099-9.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *ACM international conference proceeding series*, 170–179. https://doi.org/10.1145/2460296.2460330.

Lee, K. (2018). Machine learning approaches for learning analytics: Collaborative filtering or regression with experts ? *Korea*, 1–11.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*(14), 281–297.

Malhotra, R. (2014). Comparative analysis of statistical and machine learning methods for predicting faulty modules. *Applied Soft Computing Journal, 21*, 286–297. https://doi.org/10.1016/j.asoc.2014.03.032.

Mason, C., Twomey, J., Wright, D., & Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a Backpropagation neural network and logistic regression. *Research in Higher Education, 59*(3), 382–400. https://doi.org/10.1007/s11162-017-9473-z.

McKenzie, K., & Schweitzer, R. (2001). Who succeeds at university? Factors predicting academic performance in first year Australian university students. *Higher Education Research and Development, 20*(1), 21–33. https://doi.org/10.1080/07924360120043621.

Muñoz-Merino, P. J., González Novillo, R., & Delgado Kloos, C. (2018). Assessment of skills and adaptive learning for parametric exercises combining knowledge spaces and item response theory. *Applied Soft Computing Journal, 68*, 110–124. https://doi.org/10.1016/j.asoc.2018.03.045.

Nájera, A. B. U., de la Calleja, J., & Medina, M. A. (2017). Associating students and teachers for tutoring in higher education using clustering and data mining. *Computer Applications in Engineering Education, 25*(5), 823–832. https://doi.org/10.1002/cae.21839.

Nauta, M. M. (2010). The development, evolution, and status of Holland's theory of vocational personalities: Reflections and future directions for counseling psychology. *Journal of Counseling Psychology, 57*(1), 11–22. https://doi.org/10.1037/a0018213.

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance, *7*, 292–295. Retrieved from http://arxiv.org/abs/1002.2425

Pang, Y., Judd, N., O'Brien, J., & Ben-Avie, M. (2017). Predicting students' graduation outcomes through support vector machines. *Proceedings - Frontiers in Education Conference, FIE*, *2017-Octob*, 1–8. https://doi.org/10.1109/FIE.2017.8190666.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systemic literature review of empirical evidence. *Educational Technology & Society, 17*(4), 49–64.

Pasina, I., Bayram, G., Labib, W., Abdelhadi, A., & Nurunnabi, M. (2019). Clustering students into groups according to their learning style. *MethodsX, 6*, 2189–2197. https://doi.org/10.1016/j.mex.2019.09.026.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559–572. https://doi.org/10.1080/14786440109462720.

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers in Education, 137*, 91–103. https://doi.org/10.1016/j.compedu.2019.04.009.

Ruiperez-Valiente, J. A., Munoz-Merino, P. J., Alexandron, G., & Pritchard, D. E. (2019). Using machine learning to detect "multiple-account" cheating and analyze the influence of student and problem features. *IEEE Transactions on Learning Technologies, 12*(1), 112–122. https://doi.org/10.1109/TLT.2017.2784420.

Umair, S., & Majid Sharif, M. (2018). Predicting students grades using artificial neural networks and support vector machine. *Encyclopedia of Information Science and Technology, Fourth Edition*. https://doi.org/10.4018/978-1-5225-2255-3.ch449.

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior, 98*, 166–173. https://doi.org/10.1016/j.chb.2019.04.015.

Yang, F., & Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers in Education, 123*, 97–108. https://doi.org/10.1016/j.compedu.2018.04.006.

Yang, T. Y., Brinton, C. G., Joe-Wong, C., & Chiang, M. (2017). Behavior-based grade prediction for MOOCs via time series neural networks. *IEEE Journal on Selected Topics in Signal Processing, 11*(5), 716–728. https://doi.org/10.1109/JSTSP.2017.2700227.

Yue, H., & Fu, X. (2017). Rethinking graduation and time to degree: A fresh perspective. *Research in Higher Education, 58*(2), 184–213. https://doi.org/10.1007/s11162-016-9420-4.

Zhang, H., Huang, T., Lv, Z., Liu, S., & Yang, H. (2019). MOOCRC: A highly accurate resource recommendation model for use in MOOC environments. *Mobile Networks and Applications, 24*(1), 34–46. https://doi.org/10.1007/s11036-018-1131-y.