Universidad Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Analyzing the Behavior of Students Regarding Learning Activities, Badges, and Academic Dishonesty in MOOC Environments

Autor:

José Antonio Ruipérez Valiente

Director y Tutor:

Dr. Pedro José Muñoz Merino

DEPARTAMENTO DE INGENIERÍA TELEMÁTICA

Leganés, Abril de 2017

Universidad Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Analyzing the Behavior of Students Regarding Learning Activities, Badges, and Academic Dishonesty in MOOC Environments

Autor: *José Antonio Ruipérez Valiente*

Director: **Dr. Pedro José Muñoz Merino**

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Secretario:

Calificación:

Leganés, de de

Analyzing the Behavior of Students Regarding Learning Activities, Badges, and Academic Dishonesty in MOOC Environments

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Prepared by
José Antonio Ruipérez Valiente

Under the advice of
Dr. Pedro José Muñoz Merino

Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid

---

Date: abril, 2017

Web/contact: joseantonio.ruiperez@imdea.org
http://people.networks.imdea.org/~joseantonio_ruiperez/

This work has been supported by IMDEA Networks Institute.

Acknowledgements

Primero me gustaría expresar mi gratitud a mi tutor Pedro J. Muñoz y a Carlos Delgado por haberme brindado la posibilidad de hacer esta tesis doctoral y por todas las oportunidades que me han proporcionado durante mi investigación. Me gustaría agradecer a todos los maravillosos compañeros de departamento y laboratorio con los que he compartido estos momentos y parte de mi vida, algunos como Guille, Dmitry, Jaime, Valentín, María, Alicia, Jose A. Gascón, Jose A. Díaz, Ruth, Estrella, David, Carlos Alario, Derick, Isra, Ramona, Salva y muchos más. En especial a mis dos pilares en esta etapa, Patri y Diego, gracias por haber estado siempre ahí.

My most sincere gratitude to the people I spent time with during my two research stays. My deepest thanks to David E. Pritchard to provide me with the opportunity to spend 4 months in Boston at MIT, it has been an honor to work with you. To Giora Alexandron for his wonderful mentoring and companionship during those months and after. My gratitude to Dragan Gašević for enabling my visit to the University of Edinburgh and his always wise feedback, I understand now why you are such a renowned researcher in learning analytics. Special thanks to Srecko and Vita for their friendship and making me feel at home during my stay in Edinburgh, and also to the rest of wonderful PhD students I met there Wannisa, Mihaela and Nora'ayu.

Al centro de yoga 'Aquí y Ahora' que me ha dado tanto balance en los últimos años durante la tesis y a las maravillosas personas que allí he conocido como Vero, Sonia, Alberto y especialmente Mónica. Me gustaría recordar en este agradecimiento a mis compañeros del Grado en la UCAM y del Máster en la UC3M, porque con vosotros empezó este camino. Al igual que a todos los amigos que conservo del colegio de Murcia y muchos de ellos desde los 4 anõs, desde luego ellos sí que han visto todo el camino. A mi madre, padre y hermana, porque de lo poco inamovible y claro que tengo en mi vida, es que son lo más importante que tengo y siempre están ahí para apoyarme. A mi familia perruna Yuma y Lolita, y especialmente Sera, por todo lo que un perro ha enseñado a un humano, porque día tras día me recibes con tu mejor actitud y siempre me acompañas tanto en los malos como en los buenos momentos. A mi abuela Queta, porque la última vez que te vi fue en mi graduación del Grado y estabas muy feliz y sonriente, así te recordaré para el día que defienda esta tesis y me gradúe como doctor. A mi abuelo Juan Antonio, si a alguien le tengo que dedicar esta tesis es a él, porque sé que le hace inmensamente feliz y orgulloso. A mi abuela Conchita y mi tía Ana, porque vosotras me criasteis y me hicisteis quien soy, sois un ejemplo de motivación. También quiero extender este agradecimiento al resto de mi familia.

Abstract

The 'big data' scene has brought new improvement opportunities to most products and services, including education. Web-based learning has become very widespread over the last decade, which in conjunction with the Massive Open Online Course (MOOC) phenomenon, it has enabled the collection of large and rich data samples regarding the interaction of students with these educational online environments.

We have detected different areas in the literature that still need improvement and more research studies. Particularly, in the context of MOOCs and Small Private Online Courses (SPOCs), where we focus our data analysis on the platforms Khan Academy, Open edX and Coursera. More specifically, we are going to work towards learning analytics visualization dashboards, carrying out an evaluation of these visual analytics tools. Additionally, we will delve into the activity and behavior of students with regular and optional activities, badges and their online academically dishonest conduct. The analysis of activity and behavior of students is divided first in exploratory analysis providing descriptive and inferential statistics, like correlations and group comparisons, as well as numerous visualizations that facilitate conveying understandable information. Second, we apply clustering analysis to find different profiles of students for different purposes e.g., to analyze potential adaptation of learning experiences and pedagogical implications. Third, we also provide three machine learning models, two of them to predict learning outcomes (learning gains and certificate accomplishment) and one to classify submissions as illicit or not. We also use these models to discuss about the importance of variables.

Finally, we discuss our results in terms of the motivation of students, student profiling, instructional design, potential actuators and the evaluation of visual analytics dashboards providing different recommendations to improve future educational experiments.

Keywords: Learning analytics; educational data mining; information visualization; MOOCs; SPOCs; behavioral modeling; machine learning.

Resumen

Las novedades en torno al ‘big data’ han traído nuevas oportunidades de mejorar la mayoría de productos y servicios, incluyendo la educación. El aprendizaje mediante tecnologías web se ha extendido mucho durante la última década, que conjuntamente con el fenómeno de los cursos abiertos masivos en línea (MOOCs), ha permitido que se recojan grandes y ricas muestras de datos sobre la interacción de los estudiantes con estos entornos virtuales de aprendizaje.

Nosotros hemos detectado diferentes áreas en la literatura que aún necesitan de mejoras y del desarrollo de más estudios, específicamente en el contexto de MOOCs y cursos privados pequeños en línea (SPOCs). En la tesis nos hemos enfocado en el análisis de datos en las plataformas Khan Academy, Open edX y Coursera. Más específicamente, vamos a trabajar en interfaces de visualizaciones de analítica de aprendizaje, llevando a cabo la evaluación de estas herramientas de analítica visual. Además, profundizaremos en la actividad y el comportamiento de los estudiantes con actividades comunes y opcionales, medallas y sus conductas en torno a la deshonestidad académica. Este análisis de actividad y comportamiento comienza primero con análisis exploratorio proporcionando variables descriptivas y de inferencia estadística, como correlaciones y comparaciones entre grupos, así como numerosas visualizaciones que facilitan la transmisión de información inteligible. En segundo lugar aplicaremos técnicas de agrupamiento para encontrar distintos perfiles de estudiantes con diferentes propósitos, como por ejemplo para analizar posibles adaptaciones de experiencias educativas y sus implicaciones pedagógicas. También proporcionamos tres modelos de aprendizaje máquina, dos de ellos que predicen resultados finales de aprendizaje (ganancias de aprendizaje y la consecución de certificados de terminación) y uno para clasificar que ejercicios han sido entregados de forma deshonesta. También usaremos estos tres modelos para analizar la importancia de las variables.

Finalmente, discutimos todos los resultados en términos de la motivación de los estudiantes, diferentes perfiles de estudiante, diseño instruccional, posibles sistemas actuadores, así como la evaluación de interfaces de analítica visual, proporcionando recomendaciones que pueden ayudar a mejorar futuras experiencias educacionales.

Palabras clave: Analítica de aprendizaje; minería de datos educacionales; visualización de información; MOOCs; SPOCs; modelado de comportamiento; aprendizaje máquina.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Contents

This initial Chapter 1 establishes the research framework of this dissertation. First in Section 1.1 we present the initial motivations to carry out our research. Section 1.2 states the objectives that we aimed to accomplish during this dissertation. Finally, in Section 1.3 we describe the outline of the rest of the document.

1.1. Motivation

Over the last decade the production of data has expanded at a stunning fast pace. This has been due to, among other reasons, switching from analog to digital technologies as well as the increased data generation of corporations and individuals (CSC, 2012). With this expansion new terms appeared such as big data which is originally attributed to John R. Mashey (Mashey, 1997), and addresses the three *V*s problems of analyzing large magnitudes of data (*volume*), data is much more diverse taking different forms and coming from different sources (*variety*) and data is generated in real time and might need to be processed immediately (*velocity*) (Intel, 2013). With the big data phenomenon many research lines have arisen to try to get advantage of the data explosion. Some application examples include the acceleration of value and innovation in healthcare now that all medical records and pharmaceutical information are being digitalized (Groves, Kayyali, Knott & Kuiken, 2016), regarding how to use Internet of Things and big data to build the next generation of smart cities (Strohbach, Ziekow, Gazis & Akiva, 2015), to improve web advertising (Chandramouli, Goldstein & Duan, 2012), and also in terms of security regarding how to effectively protect all these data (Mahajan, Gaba & Chauhan, 2016). It is not only research but from banks to retail we can see how all industry areas are starting to take advantage of their data to improve their revenue, quality of their business and satisfaction of their users and customers, such as via personalized product recommendations (Amazon), audiovisual content recommendations (Spotify or Netflix), self-awareness of your own activity (Fitbit), targeted advertising (AdWords or ExoClick), contact suggestions (Facebook, Twitter or LinkedIn), weather forecasting (The Weather Company), to enhance loyalty programs (Kroger) or detection of image types (Google Images), among many others.

The potential areas of improvement that big data can bring to education are wide and can affect different end-users. For example, administrators can benefit from a better resource allocation and academic programming, students with adapted learning pathways and personalized feedback or instructors with information to improve teaching and enabling them to detect students at risk (Daniel, 2015). Additionally, we should note out the importance also of e-learning initiatives for corporate training (Urdan & Weggen, 2000), and the potential impact that big data can have to help evolve this corporate training [1]. However, the integration of data-driven approaches in education has evolved more slowly than in other areas. The initial data mining approaches in education were based on analyzing demographics and performance in different subjects, for example, to create meaningful learning outcome typologies or to predict academic outcomes that can enable interventions (Jing, 2004). As web-based education, such as the use of Learning Management Systems (LMSs) and other Virtual Learning Environments (VLEs), became more frequent, the amount of data available grew exponentially in size, but also became much more rich in terms of the available details (i.e., all interactions of students with the VLE are stored as logs). Some of these initial Educational Data Mining (EDM) studies using data from web-based systems were related to the recommendation of activities and other educational resources or to visualize student activity (Romero & Ventura, 2007). Over the last years a new disrupting phenomenon in online education and distance learning started as what have been commonly denominated as MOOCs.

MOOCs are defined as courses with a structured start and finish date, which might have a high number of learners (Massiveness), where the registration, access and participation of the activities is free (Open) and the whole course and interaction with the components and courseware is through the Internet (Online) (Siemens, 2013). MOOCs had quite a disruptive effect on online education when they emerged. The New York Times even dubbed 2012 as "The Year of the MOOC" [2] and they received a lot of publicity worldwide. Many MOOC platforms started offering courses from leading universities such as Coursera [3] or edX [4] but also in collaboration with industry partners such as the case of Udacity [5]. MOOCs have many positive features and potential to be one of the main possibilities for learning from high quality universities, such as for example in developing countries (Liyanagunawardena, Williams & Adams, 2013). The massive amount

[1] https://elearningindustry.com/impact-of-big-data-changing-corporate-training
[2] https://nyti.ms/2mdCNxY
[3] https://www.coursera.org/
[4] https://www.edx.org/
[5] https://http://udacity.com/

of data generated in MOOCs have facilitated the proliferation of new research studies with large data samples, sometimes even above 100.000 students in the same MOOC (Breslow, Pritchard, DeBoer, Stump, Ho & Seaton, 2013) or working with data of many different MOOCs at the same time e.g. (Brinton, Chiang, Jain, Lam, Liu & Wong, 2014). The MOOC explosion has widened the possibilities to collect large samples of data that can be more independent (students from all over the globe), in different data format (unstructured, structured, text log files), from different sources and functionalities (forums, wikis, exercises, videos, gamification and other external tools) and in many different topics (computer science, philosophy, art, life sciences, etc). Additionally, the web platforms where these courses are offered can adapt its functionality for each course, enabling a good framework for experimental design e.g., to implement A/B testing with different functionalities or course designs and analyze which of them can lead to better learning outcomes. These large data samples also provide the opportunity to establish deep and meaningful statistical significances and be moderately confident about the lessons learned. To analyze all these data we need a combination of theory, design and data mining techniques, and in order to fulfill these requisites the field of Learning Analytics (LA) as an intersection between data science and learning sciences (Gašević, Kovanović & Joksimović, 2017) has been gaining a lot of notoriety over the last years.

The potential possibilities regarding the data analysis that can be performed with these huge data samples are immense. Nonetheless, let us not forget that LA should be focused on the learning process and therefore it also should be in line the existing framework of educational research (Gašević, Dawson & Siemens, 2015). Some possibilities include the implementation of algorithms for behavioral modeling that can lead to a better understanding of what students are doing during their learning process and why. This can help to delve into what behaviors might be positive and which of them might be negative for their learning, but also to know the preferences of students in terms of their favorite types of activities, learning goals, learning habits and motivations in order to use all that information to adapt their learning paths and maximize the chances of providing an engaging and joyful learning experience. It is possible to use all this information to analyze the relationship of those variables with learning outcomes, e.g., if a student is going to achieve a passing grade in a course or if the student will dropout from the MOOC, among many other target objectives. This type of detectors can enable early interventions or automatic warning systems if the students' behaviors are not positive for their learning process (e.g., performing illicit behaviors). Another potential area of improvement can be instructional design based on data-driven approaches, to automatically detect problems in resources or to use A/B testing to analyze which specific course design elements can lead to better learning outcomes, which was a complicated task to perform before in more traditional educational face-to-face formats, but now with web-based applications is easier to perform experimental design. Additionally, it is also possible to analyze in smaller scale each one of the design decisions, e.g. to analyze student engagement based on video production (Guo, Kim & Rubin, 2014), to improve the overall learning experience based on smaller findings. All these potential applications can be embedded as LA systems within the VLEs. Ultimately, this can facilitate the interaction of the different stakeholders with the information for decision making or other purposes.

Our motivations for the research in this dissertation are diverse. Data analysis in MOOC environments is still in its early stage. Therefore we aim to carry out different studies to delve into different aspects of the behavior of students in these platforms, obtain relationships with other indicators and learning outcomes, so that we can obtain conclusions that can be used to improve the learning process. Furthermore, we want to work towards the evaluation of visual analytics dashboards in those environments to have a more clear evidence about which visualizations are usable, useful and effective for instructors and students. We want to delve into the activity and behavior of students with regular activities (e.g., exercises, videos, etc), optional activities (e.g., goals, setting up an avatar image, etc) and gamification features (e.g., badges, points, etc). We want to analyze the relationship between different indicators in order to delve into understanding student behavior and we will also look into clustering students based on these indicators, which can be helpful to profile them and use the different preferences to adapt their learning experiences. We would like to delve into the motivations of students based on these indicators, e.g., to analyze if students are earning badges on purpose or not, or why they are using optional activities. Additionally, we want to develop prediction models of learning outcomes that can help to understand which of these variables and behaviors are positively or negatively correlated to learning achievement. These findings might be useful for future early recommendation or warning system that can be part of the learning analytics platforms.

Furthermore, we want to delve into online academic dishonesty and illicit collaborations in MOOCs. Since these courses are completely online students are able to commit behaviors that are prohibit such as creating several accounts or sharing their solutions with peers. MOOCs deliver certificates of accomplishment to show that a student has achieved the level of proficiency and competencies given by a certain course. Nevertheless, many of these students might be deceiving the system and achieving these certificates illicitly which can cause a serious and grave problem to the certificate system. Additionally, the learning indicators of these students might represent outlying behaviors that are actually strongly correlated with success, hence leading to systematic bias in educational research studies. Moreover, literature reports that cheating behaviors lead to poor learning. Therefore, we think that is necessary to develop of algorithms and detectors that can help to detect students committing illicit behaviors to enable interventions, but also to remove their data when performing educational research. Finally, it would be valuable to provide instructional guidelines that can be used to reduce the level of cheating. In next Section 1.2 we describe more specifically the objectives that we state for this dissertation.

1.2. Objectives

As we mentioned in the motivations, this dissertation is focused on contributing to the evaluation of LA dashboards in online learning, analyzing the activity and behavior of students with regular activities, optional activities and badges, and analyzing online academic dishonesty behaviors and illicit collaborations, this research will have a strong emphasis on MOOC platforms. The potential outcomes of this research will be in the direction of learning analytics dashboards and behavioral algorithms, as well as many lessons learned in terms of understanding student behavior to infer their activity and motivations, instructional design recommendations, student profiling for adaptation purposes or guidelines to develop actuators such as recommendation or other warning systems, among other findings. More specifically, the objectives that we establish are as follow:

- In the context of learning analytics dashboards that compute different learning indicators and show them in the form of visualizations, we will work towards the evaluation of the usability, usefulness and effectiveness of these dashboards.

- Analyze the activity and the behavior of students when interacting with the platform, with emphasis on the use of regular and optional activities:

  - Analyze the effectiveness and behavior of students with educational resources such as problems or videos.
  - Analyze the use of optional activities finding also associations between the use of the different optional activities.
  - Analyze the relationship of regular and optional activities with learning outcomes such as proficient exercises and learning gains, as well as with other metrics related to the learning process.
  - Compare the use of regular and optional activities in self-regulated learning environments.
  - Apply clustering to identify common profiles of students in terms of their use of regular and optional activities.
  - Build prediction models of learning gains and certificate accomplishment based on the interaction and behavior of students with the platform.

- Analyze the activity and behavior of students with badges:

  - Overview of the use of badges in self-regulated educational experiences.
  - Analysis of the influence of factors associated with the amount of badges triggered.
  - Analysis of the badge metrics and behavior of students with badges to find the relationship with other variables and learning outcomes.
  - Apply clustering to identify profiles of students regarding their interaction with badges analyzing the potential outcomes towards their learning process.

- Analyze academic dishonesty and illicit collaboration behaviors in MOOC environments:
    - Design and implement algorithms to detect academic dishonesty and illicit collaboration behaviors that are applicable to online environments.
    - Compare students who are performing some unethical behavior with the rest of students in the course in terms of their indicators related to the learning process.
    - Analyze different profiles of students committing unethical behaviors in online environments, as well as the potential motivations and impact on their learning process.
    - Provide guidelines supported by our results in order to decrease the prevalence of academic dishonesty in MOOCs based on instructional design
    - Develop a machine learning classification model that can serve as a first step for a run-time detector as well as to analyze the importance of the different student and problem features.
- The last step is to use all lessons learned in the previous objectives to provide recommendations that can be used to improve learning processes. These recommendations should be related to the amount of activity and motivations of the behavior of students, about student profiling and how to use it for adaptation purposes, regarding potential improvements based on instructional design and some ideas toward building actuator systems.

1.3. Dissertation Outline

The remainder of the dissertation is organized as follows:

Chapter 2 analyzes different studies related to the areas in relationship with this dissertation. We review basic concepts of MOOCs, LA and EDM. We explore different studies that implemented or evaluated visualization dashboards, and also studies regarding the use and behavior of users with regular and optional activities, gamification and academic dishonesty.

Chapter 3 sets up the framework of the dissertation. We describe the different tools that we use, the case studies that we analyze and the selected indicators that we use during this dissertation.

Chapter 4 is the first chapter of the results and presents some exploratory analysis of our findings such as visualizations, descriptive statistics and correlations of the different areas that we investigate during this dissertation.

Chapter 5 presents the clustering results to analyze the different profiles of students in terms of their behavior presenting some visualizations and examples of archetypal students.

Chapter 6 presents three different machine learning models, two of them are related to predicting learning outcomes (learning gains and certificate accomplishment) and the other to classify submissions as cheated or not.

Chapter 7 discusses some of the results and provides potential recommendations regarding how to use the findings of this dissertation to improve learning processes.

Chapter 8 finishes the dissertation with some final remarks and limitations of our work and some future work ideas.

Part I

Background and Method

Chapter 2

Related Work

Contents

This chapter introduces the related work that is connected with this dissertation. The chapter is divided into different sections that cover the different areas that have been researched and are in connection with the objectives. Sections 2.1 and 2.2 give an initial and general overview related to MOOCs, SPOCs, LA and EDM that will help to frame the general context of this dissertation. Section 2.3 describes several visualization dashboards in educational contexts and evaluation results of some of them. Section 2.4 presents different studies that analyze learning outcomes in educational environments whereas Section 2.5 describes studies in the area of measuring use and effectiveness of students with regular and optional activities. Section 2.6 presents different studies in the area of gamification in education, focusing on the use and behavior of students with badges. Finally Section 2.7 finishes presenting an overview of academic dishonesty, gaming the system and collaboration, focusing on contemporary research in online environments.

2.1. MOOCs and SPOCs

Masters (2011) sees MOOCs as the fourth stage in the progress of online education, which follows the previous stage that had LMSs as their central element. Some of the differences in the MOOC era is that teachers are not going to be monitoring all the actions of their students due to the massiveness, that learners must have a more active and independent learning, and the effort and intensity of their interaction with the MOOC platform. Literature shows that MOOCs frequently combine learning technologies with learning activities. For example the use of short video lectures and automatic graded assignments is common (Voss, 2013; Nicoară, 2013). The massiveness of MOOCs makes manual exercise grading difficult for instructors, therefore another viable option that is very extended in MOOCs is the use of peer review systems (O'Toole, 2013). Additionally, we can find that MOOCs frequently enable additional communications tools, e.g., internal messaging but also external tools such Twitter or Facebook groups (Nicoară, 2013). It is also extended the use of LA functionality (Voss, 2013) and gamification approaches to motivate and engage students (Vaibhav & Gupta, 2014). There are different MOOC providers available, such as Udacity, edX or Coursera. EdX is one special case since it is the only provider that has open sourced the software that they used as a collaborative project called Open edX. From now on, the term edX is used to address the institution and Open edX the software. In our case during this dissertation we have explored data from Khan Academy, Open edX and Coursera. We believe that exploring data from different platforms can enable us to generalize better and to analyze functionalities only available in certain platforms.

SPOCs (Fox, 2013) use similar technologies and courseware items than MOOCs. However, the number of users is usually smaller and the access to these courses is private and controlled by the platform administrators. SPOCs can be used as supplement to classroom teaching, for on-campus courses or even professional training. SPOCs are usually applied in blended methodologies where part the classroom lectures are available online but students still attend face-to-face classes for problem solving sessions and to solve doubts. This marriage of face-to-face classes and MOOC materials has the potential to maximize the strengths of traditional face-to-face classes while minimizing the potential weaknesses of pure MOOCs (Burge, Fox, Grossman, Roth & Warren, 2015). The incorporation of such models have been successful in different studies. Students usually recognize the benefit of these technologies, but the figure of the instructor still needs to establish a balance between the two approaches (Bansal & Singh, 2015). During this dissertation we analyze both MOOCs and SPOCs targeting different areas of interest. Analyzing these two educational contexts allow us to obtain different conclusions.

The massiveness and the fact that each student generates a large amount of events, provides the opportunity to analyze huge datasets about the interaction of students with these online educational platforms with the objective of improving the learning process. During this process, the raw data that comes from learning environments can be processed and converted into potential information that can have an impact on educational research and practice (Romero & Ventura, 2010). We discuss about this analytical process that transforms raw data into intelligent insights in next Section 2.2.

2.2. Learning Analytics and Educational Data Mining

Within this intersection between learning sciences and data science, LA and EDM emerge as the areas that study this phenomena. Both fields reflect the importance of analyzing data in education, though there are many overlaps, a key difference might be that EDM is more focused in automatic discovery, adaptation and specific models, whereas LA is more designed to inform and empower instructors and students, relying on them for final decisions (Siemens & Baker, 2012). More specifically, LA was defined during the 1st International Conference on Learning Analytics and Knowledge as the "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environment in which occurs" [1] whereas EDM was defined by Romero & Ventura (2013) as "developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist". In other words, the two research areas can be complementary to understand the whole picture, as they have the same ultimate goal which is to improve learning (Papamitsiou & Economides, 2014). LA has emerged as a different field than academic analytics, where the latter is more focused on the political and economic challenges of education, e.g. improving educational opportunities and results at a national or international level, based on data analysis (Ferguson, 2012). We can find several reviews in the state of the art collecting the different works in LA and EDM (Berland, Baker & Blikstein, 2014; Ferguson, 2012; Papamitsiou & Economides, 2014; Romero & Ventura, 2010). Based on that we divide the different studies according to the learning setting, the analysis method applied or research objectives (Papamitsiou & Economides, 2014) and then we compare them with the research performed in this dissertation.

The learning environment and specific settings can be quite different in each study. There are studies in the area of classical VLEs and LMSs (Lee, Recker, Choi, Hong, Kim, Lee, Lefler, Louviere & Walker, 2016; Młynarska, Greene & Cunningham, 2016), MOOCs (Sharma, Alavi, Jermann & Dillenbourg, 2016; Khalil, Kastl & Ebner, 2016), SPOCs (Fox, 2013; Delgado Kloos, Muñoz-Merino, Muñoz-Organero, Alario-Hoyos, Pérez-Sanagustin, Parada G., Ruipérez-Valiente & Sanz, 2014), more specific educational environments like Intelligent Tutoring System (ITSs) (Feng, Heffernan & Koedinger, 2006; Kelly, Arroyo & Heffernan, 2013), using mobile devices (Aljohani & Davis, 2012; Tabuenca, Kalz, Drachsler & Specht, 2015) or learner data from different modalities (Worsley & Blikstein, 2015; Ochoa, Worsley, Weibel & Oviatt, 2016). During this dissertation we focused on exploring data from MOOCs and SPOCs using Khan Academy, Open edX and Coursera platforms. Additionally, each study can also have specific and different research objectives.

These objectives can be very diverse, some of the most common goals are student behavioral modeling (Qiu, Tang, Liu, Gong, Zhang, Zhang & Xue, 2016; Wen, Yang & Rose, 2014), prediction of performance (Elbadrawy, Studham & Karypis, 2015; Anozie & Junker, 2006), prediction

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

of dropout and retention (Chaplot, Rhim & Kim, 2015; Whitehill, Williams, Lopez, Coleman & Reich, 2015), decision-making support and self-reflection (Littlejohn, Hood, Milligan & Mustain, 2016; Vatrapu, Teplovs, Fujita & Bull, 2011), recommendation systems (Hu, Lo & Shih, 2014; Dwivedi & Bharadwaj, 2015) or improvement of assessment and feedback (Suen, 2014; Maggs, 2014). Our work has dealt with different research questions within these areas with special emphasis on student behavioral modeling, visual analytics and prediction of learning outcomes.

Finally, the techniques also change in each study. Some of the most common techniques are classification (Whitehill et al., 2015; Hu et al., 2014), regression (Elbadrawy et al., 2015; Koedinger, Kim, Jia, McLaughlin & Bier, 2015), text analytics (Wen et al., 2014; Tucker, Pursel & Divinsky, 2014), discovery based on rule models (Lee, Yu, Lee, Tseng, Chang & Chen, 2014; Aleven, Mclaren, Roll & Koedinger, 2006), social network analysis (Joksimović, Dowell, Skrypnyk, Kovanović, Gašević, Dawson & Graesser, 2015; Cela, Sicilia & Sánchez, 2015) and visual analytics (Qu & Chen, 2015; Coffrin, Corrin, de Barba & Kennedy, 2014). In our work we designed and implemented several algorithms for behavioral modeling based on rule models. We worked on the evaluation of visual analytics. We applied classification and regression for the prediction of learning outcomes and clustering for student profiling. In the following sections we discuss with more emphasis each one of our areas of interests during this dissertation.

2.3. Educational Visualization Dashboards and its Evaluation

Different studies have approached the process of transforming raw data into indicators and parameters related to learning. We can find a review of different LA indicators presented by Dyckhoff, Lukarov, Muslim, Chatti & Schroeder (2013). Depending on the platform and the specific research work, different parameters might be available. For example, specific indicators such as hint abusing might be available only in some environments like Khan Academy (Muñoz-Merino, Ruipérez Valiente & Kloos, 2013), while others more general like resource coverage or access time patterns are generally available in all environments (Zhang, Almeroth, Knight, Bulger & Mayer, 2007). The use of visual analytics is one of the most common techniques to transfer information regarding students' actions to stakeholders. Visualizations can be used for a big variety of reasons. Generally speaking, in education visualizations are very useful for self-awareness and self-reflection in the case of students, and in the case of instructors or other interested stakeholders, as a data-driven support for decision making. Despite their usefulness, we find that most VLEs do not include any kind of LA dashboards with basic visualizations. One of the pioneer educational platforms in providing strong LA support was Khan Academy. This platform has individual and class visualizations about factors such as time spent in exercises and videos, progress over time and the specific skills achieved by each student.

Visual analytics has been used in educational research for many different purposes such as to visualize patterns of engagement and performance in MOOCs (Coffrin et al., 2014), to explore the activity of students with software engineering subjects (Conde, García-Peñalvo, Gómez-Aguilar

& Therón, 2015), to visualize video click-stream data and detect problems in these educational resources (Shi, Fu, Chen & Qu, 2015), to create tag clouds based on the forum posts of students and analyze its evolution over time (Peñalvo, Conde, Bravo, Gómez & Therón, 2011), to provide topic-wise visualizations regarding the content and classification of discussion threads in MOOCs (Atapattu, Falkner & Tarmazdi, 2016) or to visualize different student profiles (Xu, Goldwasser, Bederson & Lin, 2014). The work by Duval (2011) shows several learning dashboards and recommender examples. Furthermore, they perform a comparison between educational and non-educational user tracking environments (Duval, 2011). We can find in the literature many visualization tools for different VLEs and using distinct types of plots and learning indicators. For example, TrAVis (May, George & Prévôt, 2011) is a tool that helps students to analyze and evaluate their own activities while learning online with computer mediated communication tools. TrAVis displays indicators in radar charts where students can compare their activity with others. Another visualization dashboard is GISMO for Moodle (Mazza & Milani, 2005), which provides graphics related to students activity in quizzes, forums and other learning resources. Another example of LA dashboard for Moodle is LAPA which had three segments: learning, prediction and action (Park & Jo, 2015). As we can see, due to the well-known popularity of Moodle as LMS there are many LA dashboard approaches for the tool. CourseVis (Mazza & Dimitrova, 2004) is another visualization tool, in this case for the WebCT; instructors can visualize different indicators, some of which are also similar to ours, such as number of accesses to each page of a course or progress with the course schedule. In the specific case scenario of Personal Learning Environments (PLEs), visualizations are particularly useful to enable self-reflection for students regarding their interaction with the PLE. CAMERA (Schmitz, Scheffel, Friedrich, Jahn, Niemann & Wolpers, 2009) is used for monitoring and reporting on learners' behavior enabling then the possibility of reflection with e.g. social network analysis. Another tool for a PLE is GLASS (Leony, Pardo, de la Fuente Valentín, de Castro & Delgado Kloos, 2012), which allows to capture events from different computer applications that students use during their practice hours providing afterwards visualizations regarding that interaction.

Although initially there were not too many initiatives for visualization dashboards on MOOC platforms, during the last few years there has been more research effort in this direction. MOOC providers are also understanding the importance of these LA dashboards and launching their own initiatives (e.g. edX Insights[2]). Due to the new technologies used in MOOC and also the massiveness, MOOCs present new challenges regarding technical design and visualizations. As an example of research studies in this direction, we can find Open-DLAs tool (Cobos, Gil, Lareo & Vargas, 2016) which was created by Universidad Autónoma of Madrid (UAM) and is able to digest edX interaction logs and provide visualizations and insights useful for the instructors. A similar approach was developed for FutureLearn data using a Shiny application for the visualization purposes (Chitsaz, Vigentini & Clayphan, 2016). A very interesting tool is PeakVizor (Chen, Chen, Liu, Shi, Wu & Qu, 2016) which enables an in-depth video peek visual analysis

[2] https://insights.edx.org/

supporting data from both edX and Coursera at the same time. In this direction, the author of this dissertation has been involved in the development of two LA dashboards. First, Add-on of the Learning Analytics support of the Khan Academy (ALAS-KA) platform (Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2013; Ruipérez-Valiente, Muñoz-Merino, Leony & Delgado Kloos, 2015; Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2015b), which extends the LA functionality of the Khan Academy platform with more than 20 new visualizations. Second, Add-on of the learNing AnaLYtics Support for open Edx (ANALYSE) which is a visualization dashboard for Open edX environment (Ruipérez-Valiente, Muñoz-Merino, Gascón-Pinedo & Delgado Kloos, 2016; Ruipérez-Valiente, Muñoz-Merino, Pijeira Díaz, Santofimia Ruiz & Delgado Kloos, 2017; Pijeira Díaz, Santofimia Ruiz, Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2016; Santofimia Ruiz, Pijeira Díaz, Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2014; Pijeira Díaz, Santofimia, Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2015) which includes 12 new visualizations and is designed to scale to the massiveness of MOOCs. During this dissertation we have used both ALAS-KA and ANALYSE in different case studies as support (e.g., to retrieve different indicators about the learning process) and to evaluate the learning process of students.

To validate the usability and effectiveness of these visualization tools, a study is often required for their evaluation to verify that the computer application can be used by non-technical users and can help to improve the quality of learning. Two of the main ways of testing the usability of an application are by preparing a usability survey (or using one of the ones available in the literature, like the System Usability Scale (SUS) questionnaire (Brooke, 1996)) and by preparing a set of tasks that respondents must perform in order to be able to answer questions. An evaluation of TrAVis (May et al., 2011) was performed where six students and one teacher answered the questionnaire. The authors noted that most of the comments about the usability and utility of TrAVis were positive. The LARAe platform (Charleer, Santos, Klerkx & Duval, 2014), which is also a teacher-oriented LA dashboard, was evaluated with six people with teaching responsibilities. The respondents tried to make sense of the data transmitted by visualizations in a survey obtaining a 4 in 5-scale Likert questions. They also included a SUS questionnaire, which had a score of 76. Similarly, the SAM tool (Govaerts, Verbert, Duval & Pardo, 2012) was also evaluated with 11 teachers. First, a series of tasks were proposed to the respondents, and then a set of open questions and a SUS questionnaire (with a score of 71.36) were performed. A prototype of the exploratory learning analytics toolkit (Dyckhoff, Zielke, Bültmann, Chatti & Schroeder, 2012) was evaluated by four teachers who were asked to perform tasks without giving a detailed explanation; the authors indicated that good usability results were achieved. The authors of LAPA (Park & Jo, 2015) performed an evaluation with 7 questions regarding conformity, 7 questions regarding perceived usefulness and 7 questions about the degree of understanding obtaining an average value for each set of questions of 3.70, 3.22 and 4.10 respectively. Very similar to the evaluation of this last study, during this dissertation we performed an extensive evaluation survey of ANALYSE (Ruipérez-Valiente et al., 2017). The evaluation contained 39 questions regarding the effectiveness, usefulness and usability and was taken by 40 respondents using ANALYSE with the role of an instructor. We describe in depth this evaluation in Section 7.5.

2.4. Analysis of Learning Outcomes

In the field of education there is extensive work towards the analysis of learning outcomes. The specific target of these studies can differ, e.g. course dropouts (Kloft, Stiehler, Zheng & Pinkwart, 2014), predicting if a student is going to be successful or not in a degree to support decision making in college admissions (Nghe, Janecek & Haddawy, 2007), to predict if students are going to surpass a course or not (Calvo-Flores, Galindo, Jiménez & Piñeiro, 2006), to predict the major that a student is going to pick, before the student actually enrolls in college courses (Pedro, Ocumpaugh, Baker & Heffernan, 2014), to provide information about the performance of groups in collaborative learning environments (Perera, Kay, Koprinska, Yacef & Zaïane, 2009), to provide learning recommendations in educational systems (Salehi & Kamalabadi, 2013) or to predict the score of a test before actually doing it (Pardos, Gowda, Baker & Heffernan, 2010; Feng et al., 2006). Most of these studies are performed using data from VLEs, but we can also find studies that use data from traditional classroom settings such as high school education (Aguiar, Lakkaraju, Bhanpuri, Miller, Yuhas & Addison, 2015). There has always been interest in analytical studies using educational data, however since the MOOC phenomenon started, the amount of data available has dramatically increased, allowing for Machine Learning (ML) models to thrive in order to find hidden patterns that can reveal insight regarding what variables have an effect on learning outcomes. As an example, the work carried out by Brinton et al. (2014) analyses data from more than 100.000 distinct students from Coursera a single MOOC, which is a data sample size hardly available in any other educational context.

All these results can have a direct impact to create tools such as actuator systems that can improve the learning process e.g. if hint abusing behavior is found to be bad for learning, the system could send a warning advising the student not to abuse hints and make a better use of them. As a whole system example, Student Success System (S3) (Essa & Ayad, 2012) identifies students at risk by applying prediction modeling, then they design the interventions to mitigate that risk and finally they close the feedback loop by checking the effectiveness of the applied intervention. In this direction many studies have approached how to develop early warning systems that can enable intervention for students that are at risk of not successfully finishing a course e.g. (Macfadyen & Dawson, 2010). A full working example in this direction was developed by Hu et al. (2014). They presented an early warning system within a LMS that can provide timely and automatic predictions regarding which students are at risk of not passing the course and was successfully tested with a positive outcome. We also analyzed during this dissertation the prediction of learning outcomes and we discuss our models and findings in Chapter 6. In this dissertation, we explore the prediction of learning gains (which is very similar to predicting a test score) and also the achievement of certificate of accomplishment. Our study can be related to two of the main prediction objectives that were mentioned as some of the most common educational objectives in education in Section 2.2: one is the prediction of learning outcomes or test scores, the other one is related to dropout prediction.

In the first case, there are many studies that work towards the prediction of post-test transfer or learning gains. Both cases are quite similar as learning gains depend on the post-test score. In this direction, there are several research works on the ASSISTment system (Feng et al., 2006; Feng, Beck, Heffernan & Koedinger, 2008; Anozie & Junker, 2006; Kelly et al., 2013) which predict a student performance indicator at the end of a course by using the data generated by the system. The results of the works developed on ASSISTment utilize variables related to seek-helping behavior and others more general about time or percentage of correct items. Others similar works are based on an ITS for College Genetics where they also try to predict learning outcomes (Baker, Corbett, Gowda, Wagner, MacLaren, Kauffman, Mitchell & Giguere, 2010; Baker, Gowda & Corbett, 2011; Corbett, Kauffman, Maclaren, Wagner & Jones, 2010). Some studies using data from the College Genetics ITS reported that the developed detector needed only a limited amount of data (around the first 20% of a students' data) in order to predict with reasonably accuracy (Baker et al., 2010). This is very interesting as it would allow to intervene in the early stages of a course. Another work on the College Genetics ITS compares several bayesian knowledge tracing variants in order to see which one of them predicts better post-test performance (Baker et al., 2011). All these studies use data from ITS environments, but we can also find in the literature similar studies on traditional LMS environment. For example, to predict performance in Moodle course activities using a collaborative multi-regression model (Elbadrawy et al., 2015) or the performance in midterm and final exams using partitioning trees (Pardo, Mirriahi, Martinez-Maldonado, Jovanovic, Dawson & Gašević, 2016). The environment of ITS, LMS and MOOC platforms can be different. Consequently, it might not be possible to apply the same variables in all environments and the effect of the predictor variables might change from one case study to another. MOOCs are recent and there are not as many works on prediction as on the ITS or LMS field. One key feature of MOOCs is the social activity and the prediction of how social activity evolves and which variables are important (Brinton et al., 2014). An interesting approach is to predict if students are going to solve correctly a question using video-watching stream data (Brinton & Chiang, 2015). Other example is the prediction of students' knowledge status in MOOCs using Open edX (Guo & Wu, 2015).

One of the most problematic issues of MOOCs is the high dropout ratio. Recent reviews estimate the average completion ratios in MOOCs around the 7% (Khalil & Ebner, 2014). These studies are also related to the prediction of certificate of accomplishment, since the students that drop out a course do not manage to acquire one. We can find in the literature different studies targeting learning outcomes in different MOOC platforms such as Open edX (Balakrishnan & Coetzee, 2013) or Coursera (Kloft et al., 2014; Rosé & Siemens, 2014; Sinha, Jermann, Li & Dillenbourg, 2014; Chaplot et al., 2015). For example, the study of Balakrishnan & Coetzee (2013) applies hidden markov model to predict retention, the work by Rosé & Siemens (2014) considers only the information of a specific shared task between each couple of students to predict dropout, the research by Sinha et al. (2014) uses the clickstream data of the interaction of students with videos to predict attrition and the work by Chaplot et al. (2015) analyses the importance of sentiment analysis obtained from applying text analytics on the posts of students. It is also important to analyze this problem over the weeks to see how accuracy evolves when more data is available (Kloft et al., 2014).

Another important question that has been addressed in the field of prediction modeling on education is about the development of these techniques and algorithms. We can even find several papers that focus on comparing different techniques or variants of the same algorithm with the purpose of finding which one is the most effective to predict learning outcomes (Nghe et al., 2007; Baker et al., 2011; Koutina & Kermanidis, 2011). The research by Kotsiantis (2012) performs a review of the different machine learning techniques for educational purposes. Many authors apply linear regression (Feng et al., 2006, 2008; Grafsgaard, Wiggins & Boyer, 2014; Kelly et al., 2013). We also use it in Section 6.1 because we expect a linear relationship between the selected variables and students' learning gains. Other authors use different methods such as bayesian knowledge tracing model (Baker et al., 2011; Guo & Wu, 2015), 1-NN (Koutina & Kermanidis, 2011), neuronal networks using radial basis functions (Calvo-Flores et al., 2006), hidden markov models (Balakrishnan & Coetzee, 2013), support vector machines (Kloft et al., 2014), partitioning trees (Pardo et al., 2016) or C4.5 (Hu et al., 2014) among many others. Another interesting approach is to ensemble different prediction methods to achieve more robust results (Pardos et al., 2010; Essa & Ayad, 2012). In our work we have also explored how different algorithms perform in different contexts and also what happens when more data is available for the prediction of certificate accomplishment as we describe in Section 6.2. Finally, most of these studies use variables computed from the interaction of the student with the VLE, but it is also possible to use multimodal sources to improve those predictions e.g., gestures and postures (Grafsgaard et al., 2014) or eye-tracking (Sharma, Jermann & Dillenbourg, 2014). In our case, we have used variables related to the interaction of students with the platform, but also trying to delve into complex student behaviors as we explain in Section 6.1.

2.5. Use of Regular and Optional Activities

Since the take over of VLEs as the main environment for distance learning (Keppell, Souter & Riddle, 2011), learners have achieved a higher degree of freedom to control their learning process and adapt their learning experience to their own needs. Often, we can see how VLEs are used for personalization and self-regulated approaches (McLoughlin & Lee, 2010). VLEs permit students to access learning resources, with a potential amount of additional features and tools that might not be mandatory, and that allow them to personalize their learning paths. Furthermore, VLEs allow students to communicate and collaborate remotely on learning activities (Dabbagh & Kitsantas, 2012). Over the last years, pedagogy has increased the weight on giving more responsibility and control to learners (Kay, 2001) which is beneficial for their actual learning outcomes (Carneiro, Lefrere, Steffens & Underwood, 2012). The main idea with self-regulated learning is that students should master a process that involves goal setting and planning, monitoring and control processes, as well as reflection and evaluation processes (Schon, 1984; Bolton, 2010). Within this self-regulated settings students can decide which items or activities they want to use. We can roughly divide it in regular or mandatory activities, that are those required to be completed by students in order to achieve a passing grade (e.g. graded exercises or videos) and those who are completely optional and might not even be related to the learning process (e.g. setting an avatar picture). This degree of optionality depends on each specific case study, e.g., in some occasions forum activity might be mandatory, in other cases completely optional.

If we assume that one of the main objectives of MOOCs and SPOCs is that students complete the proposed courseware in a correct way (e.g., their interaction with videos or educational activities). Therefore, it is necessary to define metrics that can accurately measure the effectiveness of students with the courseware. These metrics can help to determine how students progress in the course according to the proposed are activities. Nonetheless, in the literature we find that most metrics that are used to evaluate the effectiveness of students are very simple (e.g, number of videos completed or number of exercises solved correctly) and usually these metrics are not adapted to the specificities of the educational context. For example, the study carried out by Dyckhoff et al. (2013) shows a compilation of indicators used in different studies in the literature, showing that most of them are simple indicators such as number of threads started by a student, number of assignments submitted or number of pages viewed. These indicators do not take into account how educational resources and activities were structured or how they are related to each other.

The traditional educational literature defines the concept of effectiveness from a perspective of amount of learning, if we quote the work of Hiltz & Arbaugh (2003) the definition is as follows, “how much did the students learn, how well did they master skills and how well can they apply knowledge”. The concept of effectiveness applies for different educational settings such as face to face lessons or blended learning, but it might have become even more important for pure online learning courses where instructors cannot establish physical bonds and analyze the behavior of students in class so easily. Consequently, there is a need to design alternative methods to measure students’ effectiveness (Ni, 2013; Swan, 2003). One of the most common methods to measure learning effectiveness is the application of achievement tests or surveys (Moody & Sindre, 2003). Nevertheless, this might not be always available. In addition, each environment might need specific definitions to measure the effectiveness e.g., Serrano-Laguna, Torrente, Moreno-Ger & Fernández-Manjón (2012) uses the source of data from an educational game to feed a LA system to infer knowledge about the effectiveness of the students. One possibility is to analyze the effectiveness separately as suggested by Swan. Another possibility by Rourke, Anderson, Garrison & Archer (2007) is to measure effectiveness in terms of interactivity with peers (social presence), with instructors (teaching presence) and with contents (cognitive presence). This categorization of effectiveness can be a good match for MOOC and SPOC environments, where the social and teaching presence can be inferred from the activity in forums and the cognitive presence from the interaction with the courseware.

A deeper analysis regarding how students use the different educational activities and courseware can help to delve into different behavioral profiles such as for example 'copy and paster', 'hint abuser', 'hint avoider', 'student misuse', 'video avoider', 'unreflective user' or 'procrastinator' (Blikstein, 2011; Aleven, McLaren, Roll & Koedinger, 2004; Aleven et al., 2006; Muñoz-Merino et al., 2013; Baker, Corbett, Koedinger & Wagner, 2004; Baker, Corbett & Koedinger, 2004; Tervakari, Marttila, Kailanto, Huhtamäki, Koro & Silius, 2013). We can analyze the influence of these different indicators on different outcomes, e.g., factors that might affect teaching effectiveness (Kyriakides, Christoforou & Charalambous, 2013), factors of student persistence (Hart, 2012), relationship of different behaviors with learning gains (Aleven et al., 2006) or analysis of what items can increase student engagement (Wankel & Blessinger, 2012). However, most e-learning platforms are still providing just rough insight (usually just the number and the grade of the activities completed) regarding the interaction of students with the educational resources. There is a need for more precise strategies to measure the effectiveness of students that can take into account the structure of the activities and other specificities such as the relationship between the different items in a course. As part of this dissertation, we have analyzed the relationship between the effectiveness of students with other variables and also with the purpose of student profiling (Muñoz-Merino, Ruipérez-Valiente, Alario-Hoyos, Pérez-Sanagustín & Kloos, 2014; Muñoz-Merino, Ruipérez-Valiente, Alario-Hoyos, Pérez-Sanagustín & Delgado Kloos, 2015).

Additionally to the aforementioned learning activities, there are other activities that might not be mandatory or required to effectively complete the learning process. These activities can be defined as optional for students. For example, Muñoz-Merino, Delgado Kloos, Seepold & García (2006) analyzed which tools and functionalities that are provided by the VLEs Moodle [3] and .LRN [4] are the most important regarding students' perception. Some of the most highly rated were optional activities such as the use of forums or visualizations regarding their status. This shows that students also care about extra functionalities. Koedinger et al. (2015) compared the effect of passive and active learning. They found that only watching videos can be predictive of dropout and those who completed activities were more successful than just watching videos or pages. In addition, they also found that the combination of both passive and active learning lead to the highest success rates. Santos, Klerkx, Duval, Gago & Rodríguez (2014) analyzed the activities conducted by learners in two language MOOCs and they found that a higher activity in the forum correlated with students' success. This is in line with the findings of the study conducted by Cheng, Paré, Collimore & Joordens (2011) with over 2.000 students that found that students who participated voluntarily in forums also performed better in the course. Other works that explored activities that can be regarded as optional, are for example the one carried

[3] https://moodle.org
[4] http://www.dotlrn.org/

out by Gašević, Mirriahi & Dawson (2014) in terms of video annotation. They compared two courses, in the first one annotations were graded and in the second annotations were non-graded. Their findings suggest that students in the group of graded annotations, were able to use and develop more complex language indicators as a result of a potential more complex cognitive process. The study by Coetzee, Fox, Hearst & Hartmann (2014) with a reputation system for forum activities, suggested that students who were actively using the forum performed better and at the same time the use of the reputation system produced faster and more numerous post responses. On the other side, a study by Davies & Graff (2005) suggested the contrary, that forum activity alone is not enough to lead to higher grades, at least in their context. A study by Muñoz-Organero, Muñoz-Merino & Kloos (2010) found that participating in e-learning activities (such as forums) or uploading a profile photograph was positively correlated with the motivation and final grade of students. During this dissertation we analyze the relationship between the use of optional activities and learning outcomes (Ruipérez-Valiente, Muñoz-Merino, Delgado Kloos, Niemann & Scheffel, 2014; Ruipérez-Valiente, Muñoz-Merino, Delgado Kloos, Niemann, Scheffel & Wolpers, 2016). More specifically, we look into the use of feedback, votes, badge display, avatar image and setting up learning goals in Khan Academy (this optional activities are described in Subsection 3.1.1.1). We present the relationship between using certain regular and optional activities with other learning indicators. We also delve into how the behavior and activity of students might relate to learning outcomes.

2.6. Gamification and Use of Badges

The use of serious games is widely spread among different contexts. This technique encourages the use of game elements for educational purposes in order to provide a more immersive learning flow and improve engagement (Arnab, Berta, Earp, De Freitas, Popescu, Romero, Stanescu & Usart, 2012). Many studies apply also game components in non-game contexts applying what is known as gamification (Deterding, Dixon, Khaled & Nacke, 2011). These techniques have been used in many different contexts; for example the introduction of game achievements in a photo sharing service (Montola, Nummenmaa, Lucero, Boberg & Korhonen, 2009) or the inclusion of gamification elements in eco-driving (Magaña & Organero, 2014) showing a positive correlation with the use of the proposed eco-driving tips. Gamification has been tested in different e-learning experiments, reporting positive results. Some of the reasons to use gamification elements in education is to improve the motivation and engagement of the student towards their learning goals. Potential applications of gamification can be to improve the engagement of students in engineering education that it is often regarded as more difficult than other degrees (Douglas, Iversen & Kalyandurg, 2004), or in MOOCs in order to improve the intrinsic motivation of students to lower the high attrition rates (Borras-Gene, Martinez-Nunez & Fidalgo-Blanco, 2016). Some examples include a gamified AutoCAD tutorial (Li, Grossman & Fitzmaurice, 2012), a gamification system for Blackboard (Domínguez, Saenz-de Navarrete, de Marcos, Fernández-Sanz, Pagés &

Martínez-Herráiz, 2013), or the interesting example of Septris and SICKO which combines both LA and gamification in order to improve medical education (Tsui, Lau & Shieh, 2014).

Hamari, Koivisto & Sarsa (2014) presented a literature review of the empirical studies on gamification analyzing 24 research works. The results indicated that gamification yielded positive effects as a general rule, but these effects were strongly dependent on the contexts and the users of the experiment. Some examples of successful specific gamification applications are with mathematics computer games which were able to improve the motivation of those students who used them (Kebritchi, Hirumi & Bai, 2010), in engineering classes where students were able to improve their learning achievement while reducing the stress of complex lessons (Kim, 2013), the successful case of Pex4Fun (Xie, Tillmann, De Halleux & Bishop, 2015), a gamified engineering software where students earn badges and duel each other while learning the contents of the course, or for teaching computer programming skills to new students avoiding part of the stress (Mladenović, Krpan & Mladenović, 2016). Nonetheless, there are handicaps in the use of gamification. Since it is naturally a extrinsic motivator, some students might lose track of the actual task in hand and undermine motivation (Deci, Koestner & Ryan, 2001). Additionally, there are reports indicating that some student felt discouraged and perceived gamification as unnecessary (Berkling & Thomas, 2013). That is why it is also important to prepare a carefully tailored gamification design that enhances the intrinsic motivation of students (Barata, Gama, Jorge & Gonçalves, 2013b), such as for example by improving the control of students, enhancing cooperation and the possibility to gain recognition (Zirk, 2014). Some works proposed frameworks in order to effectively design beneficial gamification experiences (Hamari & Eranti, 2011).

One of the most common elements used in gamification is the use of badges, which are virtual tokens that are delivered after completing certain actions and represent visual achievements or skills (Goligoski, 2012). There are some open frameworks such as Mozilla Open Badges [5], which provide the possibility of using a shared infrastructure for implementing them. Badges are commonly used to try to encourage desired behaviors of users (Gibson, Ostashewski, Flintoff, Grant & Knight, 2015). Several studies show proof that badges can increment user activity, and encourage more social interaction or other desired behaviors (Grant & Betts, 2013; Anderson, Huttenlocher, Kleinberg & Leskovec, 2013). In the case of education, the use of badges is strongly related to the reinforcement of achievement goal theory, trying to create a positive relationship between mastering a skill, receiving a badge and actual academic performance (Abramovich & Schunn, 2011). One of the main objectives of using badges in educational settings, is trying to increment the engagement of students with the platform and the learning flow, which can be measured with various metrics such as time or frequency of visits to the learning environment (Muntean, 2011). There are some successful studies using badges in education, for example achieving an increment in the social activity of 511% in term of replies and 845% in term of number of threads (Barata, Gama, Jorge & Gonçalves, 2013a). In addition, the TRAKLA2 online environment reported good results using badges to encourage desired behaviors for 281 students

[5] http://openbadges.org/

(Hakulinen, Auvinen & Korhonen, 2013). Another successful example is GRASS project [6] which has been focused on the use of open badges to represent the soft skills of learners. Again, although badges can indeed improve learners motivations (Abramovich, Schunn & Higashi, 2013), bad designs can lead to counter-effective systems that might interfere with the important goal which is the learning of students.

Nevertheless, it is not common to find gamification indicators included in research studies, e.g., the aforementioned review by Dyckhoff et al., does not include any indicator related to gamification. Therefore, it is still needed the analysis of student behaviors towards badges. Additionally, these indicators can help towards adapting the learning experience of students. We believe there is a knowledge gap here, and that is why in this dissertation we analyze the use of badges of students in educational environments (Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2016a). We could use this information to understand how to use badges and instructional design to improve engagement and motivation in learning experiences. Previous research about badges on Stack Overflow [7], found that badges can have an effect on the behavior of users e.g., to edit more posts (Grant & Betts, 2013). That is why we want to see the effect of badges on the behavior of students and we analyze different behaviors such as the concentration or intentionality of students towards badges and its relationship with other metrics (Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2016b, 2017).

2.7. Academic Dishonesty and Illicit Collaboration

Academic dishonesty can be defined as "any type of fraudulent action in an academic work" (Lambert, Hogan & Barton, 2003). Academic dishonesty is often denominated as 'cheating'. It has been one key issue in education since the early beginnings. Singhal (1982) described the area of academic dishonesty as one of the main problems in contemporary education; now with the proliferation of online education, new problems related to academic dishonesty are emerging as well. Some authors divide academic dishonesty in traditional methods (typical education in a classroom context) and new methods (those who include electronics, new technologies or Internet) (Palazzo, 2006). This issue has been an area of study for at least half a century. According to a report by McCabe, Treviño & Butterfield (2001), they rate that very high percentages of students that have attended college, have also engaged in some kind of academic dishonesty. Other studies have reported different metrics, such as that most students only cheat occasionally and that there are not many that cheat systematically (Witherspoon, Maldonado & Lacey, 2012), or that around 67% of students have cheated in at least one exam or more (Balbuena & Lamela, 2015). Despite the exact percentages can vary greatly from one study to another, we should keep in mind that this can be strongly affected by the specific criteria to consider actions as cheating or not.

Academic dishonesty and cheating can be influenced by many factors. It has been found in

[6] https://sites.google.com/site/llpgrassproject/
[7] http://stackoverflow.com/

different studies that students' demographics can play an important role, such as the educational level (Harding, Mayhew, Finelli & Carpenter, 2007), age (Anderman, Cupp & Lane, 2009) or gender (Harding et al., 2007; Bogle, 2000; Witmer & Johansson, 2015; Anderman et al., 2009). It was also found that academic dishonesty was influenced for large and public institutions against small and private ones, which can indicate that a more personalize learning environment battles cheating (Palazzo, 2006). There are some studies that also suggest a existing relationship between personality traits and having a dishonest behavior in academia (Anderman et al., 2009; Harding et al., 2007; Giluk & Postlethwaite, 2015; Sanecka & Baran, 2015; De Bruin & Rudnick, 2007; Jordan, 2001). One other important factor of influence is the peers in the environment, which sometimes can even help or be supportive during the cheating process (Payan, Reardon & McCorkle, 2010). Despite some students might have initially an ethical behavior and do not cheat, after seeing other peers succeed using cheating methods, they might feel in disadvantage and start committing unethical acts as well (McCabe & Trevino, 1993). Additionally, whenever they start cheating, their guilt might decrease and continue in the future (Shu & Gino, 2012).

Many other factors can also have an effect on academic dishonesty, such as the role of teachers (Broeckelman-Post, 2008; Anderman et al., 2009) or the features available in the educational software that students are using (Kauffman & Young, 2015). This last item connects well with the idea of 'gaming the system' in which students will try to succeed in a learning environment by exploiting some of its properties instead of actually learning the courseware (Baker, Walonoski, Heffernan, Roll, Corbett & Koedinger, 2008; Desmarais & Baker, 2012). One example of gaming the system could be that a student consumes all the available hints before even trying to read and solve the problem first. We can find in the literature different gaming methods such as help abuse, systematic guessing and checking or copying hints (Wood & Wood, 1999; Muldner, Burleson, Van de Sande & VanLehn, 2011). Gaming is different than cheating, since gaming behaviors might not be strictly against the established academic rules. Some studies have shown also that gaming the system might reduce learning. Hence, it is important to develop detectors that are able to identify this behavior in online environments (Walonoski & Heffernan, 2006; Fancsali, 2013).

In educational settings, students might tend to group together to carry out certain learning activities (Webb, 1989; Curtis & Lawson, 2001). However, in some occasions this can also be associated with illicit behaviors, such as sharing answers between peers (Chapman, Davis, Toy & Wright, 2004). The use of a Computer-Supported Collaborative Learning (CSCL) environments is wide-spread nowadays. It is a dynamic, interdisciplinary field of research focused in how the use of technology can provide a better environment in which peers can learn interacting together and create knowledge through their learning processes (Resta & Laferrière, 2007). We can find different ways to implement CSCL features such as the use of wikis for collaborative knowledge (Judd, Kennedy & Cropper, 2010) or the use of peer review to learn from revising the work of other peers (Eaton & Wade, 2014). One of the most common approaches is the use of discussion forums as online communities, that can provide means for communication between learners (Rabbany, Elatia, Takaffoli & Zaïane, 2014). The latter, is one of the most frequently used in

MOOC environments and has been studied with frequency (Joksimović et al., 2015; Cela et al., 2015). Similar to what has been found in face-to-face courses, MOOC research has also found that learners tend to group (Li, Verma, Skevi, Zufferey, Blom & Dillenbourg, 2014). Additionally, jointly registering to a course with a friend can have a positive influence towards completion (Brooks, Stalburg, Dillahunt & Robert, 2015). However, not all collaboration might be ethical, since there is no ID confirmation about who took an exam or what was done during the test, in these online environments the issue of academic dishonesty might even be more problematic (Harmon & Lambrinos, 2008). One important way to battle academic dishonesty is by creating an honest culture and moral beliefs that help students understand what is expected from them (Galbraith & Jones, 2010). Several methods have been reported as useful to decrease academic dishonesty and gaming the system, e.g., the use of honor codes (LoSchiavo & Shatz, 2011) or delaying help feedback (Baker et al., 2004, 2004).

Different detectors have been developed to detect academic dishonesty (Sheridan, Alany & Brake, 2005) or gaming the system (Muldner et al., 2011; Walonoski & Heffernan, 2006). Nonetheless, more work is required in this direction as it can seriously affect learning quality and educational research. During this dissertation we explore academic dishonest collaboration, by designing a method to detect student ties based on temporal proximity of their assignment submissions (Ruipérez-Valiente, Joksimović, Kovanović, Gašević, Muñoz-Merino & Delgado Kloos, 2017a). We will discuss this detector in Subsection 3.3.7.1. We also present one specific form of academic dishonesty that have been found is MOOCs that is known by the name of Copying Answers using Multiple Existences Online (CAMEO) (Ruipérez-Valiente, Alexandron, Chen & Pritchard, 2016; Alexandron, Ruipérez-Valiente & Pritchard, 2015; Alexandron, Ruipérez-Valiente, Chen, Muñoz-Merino & Pritchard, 2017; Northcutt, Ho & Chuang, 2016). In CAMEO, students use multiple accounts to harvest correct solutions and then insert the correct answers into their main account, which is used to earn a certificate. CAMEO is strongly related academic dishonesty, since students are breaking the agreed honor code specified by the online environments (e.g., terms of service of Coursera [8] or edX [9]). CAMEO is also related to gaming the system since students are exploiting certain features such as being able to create several accounts and receiving feedback from submissions. We discuss CAMEO detection algorithm in Subsection 3.3.7.2.

[8] https://www.coursera.org/about/terms
[9] https://www.edx.org/edx-terms-service

Chapter 3

Method

Contents

This chapter describes the method that we follow during the research process, which also resembles the outline of the dissertation. The overview of the different stages is presented in Figure 3.1. Now we describe the different phases within our approach:

1. **Educational setting and research questions**: The first phase is establishing an educational context, with the tools that we are going to use, the case study, the selected indicators that we want to explore and the research questions that need to be answered:

a) **Tools**: From the educational side we have used mostly MOOC platforms that are usually composed by different tools. Students interact with the learning environment and contents generating the educational data that later on is analyzed. In some cases we also had the support of LA dashboards that provide visualizations regarding the learning process of students in the platform. These LA systems compute most of the selected indicators that are used during this research. Finally, in order to perform this research it was necessary the use of different technologies, as well as statistical analysis software and methods. The different tools that are used during this dissertation are presented more deeply later on in next Section 3.1.

b) **Case studies**: We carry out some case studies based on the aforementioned tools, where students interact with the MOOC platform and the course contents generating data regarding all the actions that they perform. These studies can have different contexts. In our research, we focus on MOOCs and SPOCs, albeit other frequent contexts in the literature include online courses for credit or professional development. We present the different case studies in Section 3.2.

c) **Selected indicators**: The indicators vary from one case study to another as the research questions change as well. In our research, we focus on the use of regular and optional activities, the behavior with badges and online academic dishonesty and unethical collaboration. Nevertheless, many other indicators are used to answer the research questions established for this dissertation. The selected indicators and its implementation details are described in Section 3.3.

2. **Analysis and modeling of educational data**: During the second phase we perform the analysis to be able to understand the research questions. During this stage, it is also important to keep in mind the specific settings of each case study, e.g., topic, weight of each assignment, deadlines, and so on. Usually, this step is divided in an initial exploratory analysis and then more complex algorithms to model the educational data and delve into understanding student behavior:

a) **Exploratory analysis**: This is the first stage in order to understand the data and give answer to the research questions. During this stage we use common methods to provide descriptive statistics regarding the selected indicators that we are analyzing, different visualizations that are useful for interpretation, relationship mining or comparison between different groups. The specific analysis is detailed in Chapter 4

b) **Models and algorithms**: The second stage to really understand the data is to apply more complex algorithms and models so that greater insight can be obtained to answer the research questions. We divide this stage in two main chapters of the dissertation. Chapter 5 builds clustering models that can be used to personalize/adapt given the different learning profiles and for student profiling. Chapter 6 builds prediction

Figure 3.1: Overview of the method followed during the dissertation.

and classification models of learning outcomes, which are useful to understand the influence of the different selected indicators and perform recommendations.

3. **Outcomes, recommendations and feedback**: Finally, the last phase is to use all the lessons learned as part of the analysis and modeling, in order to provide feedback that can improve educational settings. These improvements and outcomes can have diverse types of applications, e.g., a diverse set of guidelines about which behaviors might be good or bad for the learning process of students, evaluation of the usefulness, usability and effectiveness of visualizations, actuator systems, use student profiling to personalize an adapt learning processes or the detection of problems in resources or course design. These findings are discussed in depth in Chapter 7.

3.1. Tools

This section presents an overview and brief description of the different tools and software used during this dissertation. More specifically, Subsection 3.1.1 presents the different MOOC platforms that are used to collect the educational data, Subsection 3.1.2 presents an overview of the two learning analytics dashboards that are used for support and research during this dissertation and Subsection 3.1.3 presents the different technologies and statistical analysis software and methods.

3.1.1. MOOC Platforms

We analyze data from different case studies using the following MOOC platforms: Khan Academy (Subsection 3.1.1.1), Open edX (Subsection 3.1.1.2) and Coursera (Subsection 3.1.1.3). We selected these 3 platforms since they are some of the most representatives within the MOOC panorama. Each one of these platforms might have different functionalities and data available which limits also the type of research that can be pursued.

3.1.1.1. Khan Academy

Khan Academy[1] is a not-for-profit educational institution founded originally by Salman Khan with the mission of providing free education accessible for every person in the world. They have already reached millions of students and are being translated into 36 languages. Khan Academy was one of the pioneer educational platforms to provide open content and it was originally founded by donations from organizations such as Google. Khan Academy incorporated since its beginning modern engaging ideas for gamification and learning analytics. During our research we use Khan Academy case studies to investigate the use of optional activities and badges, which are features that are not present in the other two MOOC platforms that we explore.

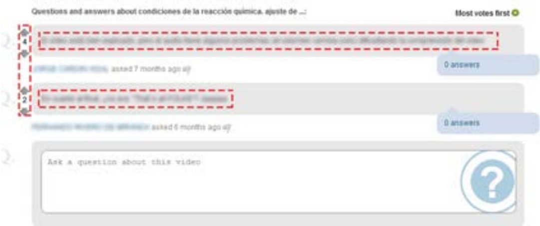
Optional activities As part of the dissertation, we analyze the use of optional activities. We understand that optional activities are those that can be used voluntarily and students are not oblige to use in order to pass the requirements of a course. Depending on each specific case study, activities might or might not be optional, e.g., in some courses social activity in the forum might be mandatory and graded, whereas in other courses might be completely voluntary. In our case, we analyze the use of the five activities that were optional in our experiments using Khan Academy. First, two of these activities are related to games and social networks:

1. **Profile avatar:** Students can change the default avatar of their profile. They have access to a selection of six different avatar images at the beginning of using the platform and can earn access to more images by acquiring points in their interaction with the platform functionalities as can be seen on the upper left side of Figure 3.2a.

2. **Badge display**: Students can personalize a selection of badges to be displayed on their personal profile. The badges that can be displayed are the ones that each student has earned

[1] https://www.khanacademy.org

(a)



(b)

(c)

Figure 3.2: This figure shows the different optional activities in Khan Academy. Subplot (a) shows avatar personalization and badge display, Subplot (b) shows feedback and votes functionality and Subplot (c) shows goals system.

previously. The upper right side of Figure 3.2a shows a portion of the personal profile where the badge display can be observed.

Then, there are three optional activities that are more related to learning:

3. **Feedback**: Comments that students post to videos of the course are considered as feedback (Figure 3.2b).

4. **Votes**: Students can vote down (-1), be indifferent to (0) or vote up (+1) the feedback that other students have posted to videos. Figure 3.2b shows an example of a comment that has some votes.

5. **Goal**: Students can set goals, i.e., they choose a selection of videos or exercises that must be completed by them. When they finish the goal, they obtain an additional amount of points. Figure 3.2c shows an example about how to set a custom goal.

Badges One big area that we analyze is the use and behavior of students with badges. Khan Academy incorporates a wide badge system that has been useful to analyze the behavior of students with badges. Table 3.1 shows all badge types split in four categories. Each category contains the different types of badges that can be acquired and next a description of the requisites for each type of badge is presented. Finally, the last column has the number of different badges of the same type that can be earned. As an example the ‘Streak’ type of badge have five different levels which are denominated as ‘Nice’, ‘Great’, ‘Awesome’, ‘Ridiculous’ and ‘Ludicrous Streak’ badge that are triggered when the student correctly solves 20, 40, 60, 80 and 100 exercises in a

Table 3.1: Classification of the different badge categories in our Khan Academy instance.

Badge category	Type of badge	Requisites	Number of different badges
Exercise	Streak	To solve several exercise correctly in a row of the same skill.	5
	Timed Problem	To solve a certain number of correct problems within a specific amount of time	5
	Exercise Completion	To complete a specific number of exercises	4
	Recovery Problem	To get exercise problems correct after having some problems solving exercises	2
	Unfinished Exercise	These badges are awarded when the user does not acquire a,proficient level but he is answering many exercises correctly	3
	Topic	These badges are awarded when achieving a proficiency level in a,subset of exercises	Different for each
	Challenge	These badges are awarded when achieving a proficiency level in a set of exercises which are related to the same topic of knowledge	course
Video	Topic Time	To watch a specific amount of videos in the same topic	4
	Video Time	To spend a certain amount of time watching videos	1
Social	Feedback	To receive up votes in your questions or answers	6
	Discussion	To flag or vote the questions and answers of other peers	4
	Profile	To customize your personal profile	1
General	Points	To earn a certain amount of points	2
	Power Time	To watch an specific amount of video and completing a certain amount of exercise problems within a set amount of time	3
	Consecutive Activity	To consecutively perform an activity on the site for a set of days in a row	3

row respectively. Additionally, for the specific research on the behavior with badges, we focus on the following two types of badges:

- Topic Badge (*topic badge*): These badges are awarded to students when they accomplish to earn proficiency in a set of exercises (skills). In our experiment, the required exercises to earn one badge are always different from the others. This means that each problem belongs only to the requirements of one *topic badge*. Each one of these badges can be earned only once to each student. As part of the experiment, the badge system was customized and new badges were added in the case of *topic badges*, to match the exercises that were developed for each one of the courses. A total number of 7, 12 and 16 *topic badges* were designed for the mathematics, physics and chemistry courses respectively. The amount of *topic badges* is in relationship with the amount of exercises in each course and the relationship between exercises, as related exercises were united to provide a *topic badge* about a specific area of knowledge.

- Repetitive Badge (*repetitive badge*): We classify within this category those badges that can be earned repetitively by the same student as many times as students want (as long as they keep fulfilling the required conditions). Specifically in our experiment, we have two types of badges that fall within this category, which are called as 'Timed Problem' and 'Streak' badges. The first ones are delivered when solving problems rapidly, and the second ones when solving several exercises correctly in a row. Each one of these two types of badges have 5 different levels. The different levels of the 'Timed Problem' type are quite similar as the former one. So there are a total number of 10 badges, but note that these badges can be earned repetitively.

3.1.1.2. EdX and Open edX

EdX is a not-for-profit venture with the general objective of improving online learning. It was initially founded by Harvard and Massachusetts Institute of Technology (MIT) but currently has a consortium of more than 90 global partners. More than 500 MOOCs have been delivered in edX with more than 5 million students. In June 2013, they open sourced the software that they use to run the MOOCs creating a collaborative project called Open edX [2] which at the moment it is being used in 18 languages, 35 countries, more than 200 external institutions reaching over 20 million students [3].

3.1.1.3. Coursera

Coursera is a for-profit educational company that was founded by Andrew Ng and Daphne Koller, two Stanford professors offering MOOCs from top universities of the world. The software is proprietary and it is not available for interested stakeholders. Coursera has already offered [4] more than 1600 MOOCs, has more than 145 university partners reaching over 22 million students and delivering more than 600.000 completion certificates.

3.1.2. Learning Analytics Dashboards

MOOC platforms such as the ones presented in previous Subsection 3.1.1 often generate large datasets that usually remain unused by the instructors and students that are taking those courses. The learning analytics support provided in terms of visualizations is often not enough to fulfill the requites of instructors and students. The author of this dissertation and other researchers at Universidad Carlos III of Madrid (UC3M) have been involved in the development of two learning analytics dashboards. One is ALAS-KA (Subsection 3.1.2.1) and the second one is ANALYSE (Subsection 3.1.2.2). These visualization dashboards provide additional indicators and information regarding the learning process for the instructors and students taking a course. This information can be used by instructors as a data-driven help for decision making and to keep track of their students whereas students can improve their self-awareness and self-regulated skills. We use these two applications for different purposes in our research.

3.1.2.1. ALAS-KA for Khan Academy

ALAS-KA (Ruipérez-Valiente et al., 2013, 2015, 2015b) was developed as part of the master thesis of the author of this dissertation and is used for research purposes during the studies of this dissertation that involved Khan Academy data. ALAS-KA is designed as a plug-in for the Khan Academy platform. The Khan Academy system as well as ALAS-KA run over the Google App Engine (GAE) architecture and use the GAE Datastore for data persistence. In addition, the underlying programming language is Python. ALAS-KA needs the data generated by students while interacting with Khan Academy to process it to obtain higher level information. Furthermore, we

[2] https://open.edx.org/
[3] https://con.openedx.org/
[4] https://about.coursera.org/

use the Google Charts Application Programming Interface (API) for visualizations. A more in-depth description of technology aspects has been addressed in previous work (Ruipérez-Valiente et al., 2013; Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2014).

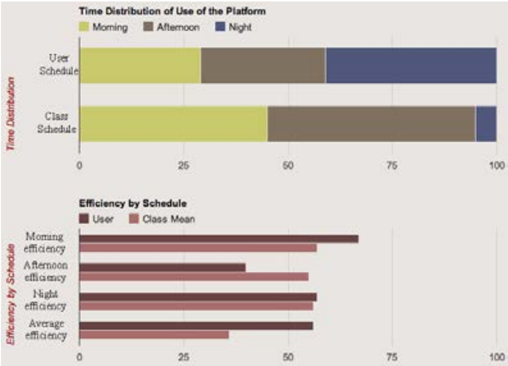
ALAS-KA has been open sourced in GitHub [5] and is therefore available for use for the community, it is also possible to consult a video online [6]. A total set of 20 different parameters have been introduced in ALAS-KA. Specific formulas for some of these metrics can be consulted as part of the publications (Muñoz-Merino et al., 2013). The metrics of ALAS-KA are divided into six functional modules (Ruipérez-Valiente et al., 2014) which are enumerated next:

- 'Total Use of the Platform' provides insight about the use that students have done in the platform.
- 'Correct Progress on the Platform' contains parameters which try to assess how good the interactions have been.
- 'Time Distribution of the Use of the Platform' focuses on analyzing the distribution of the time in which users have interacted with the platform.
- 'Gamification Habits' offers a couple of metrics to see if students are motivated by gamification elements.
- 'Exercise Solving Habits' analyzes users' behaviors when solving exercises such as hint avoidance, try abuse or unreflective.
- 'Affective States' (Derick Leony, Pedro J. Muñoz-Merino, José A. Ruipérez-Valiente, Abelardo Pardo, David Arellano Martín-Caro & Carlos Delgado Kloos, 2015) reports the levels of emotions of students.

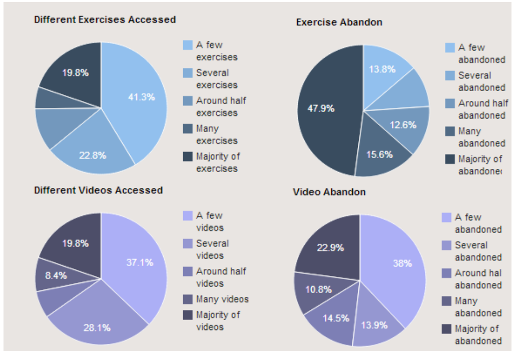
Each one of the parameters is represented by the two types of visualizations that follow:

- **Class visualizations**: These visualizations present an overview of the status of the entire class or a set of students. The main type of graphic used for class visualizations are pie charts because we want to give an overview of how the class is distributed for each metric. An example is shown in Figure 3.3a.
- **Individual visualizations**: In-depth visualizations enable teachers to analyze each student separately and self-reflection for students. Most of the visualizations use bar charts, and also establish a comparison with the average of the class; Figure 3.3b shows an example.

[5] https://github.com/jruiperezv/ALAS-KA
[6] https://www.youtube.com/watch?v=vDs1tt7siBA



(a)



(b)

Figure 3.3: Visualization examples of ALAS-KA, on the left regarding the time distribution of a single student and on the right the use of the platform of all the class.

3.1.2.2. ANALYSE for Open edX

ANALYSE is a project that has been developed within the Telematics Department at UC3M. Several students and researchers have collaborated towards its development, including the author of this dissertation. ANALYSE is a Django application embedded within the LMS of Open edX as one more of the installed applications. ANALYSE processes the data generated by the students and provides a visualization dashboard for instructors and students that adds information and graphics that are not enabled by default in Open edX. More information regarding technical details of the design are available in previous publications (Santofimia Ruiz et al., 2014; Ruipérez-Valiente et al., 2016). ANALYSE works within the scope of each course. This means that the metrics are generated per course and it is accessible by instructors and students by clicking on a new tab in the course contents. Therefore, instructors belonging to a course are able to access all the individual visualizations and aggregates of the entire class about the students that have enrolled for that course, whereas students can only access their own information. In the case that instructors or students are enrolled in several courses, they are able to access the information of each course separately, but they do not have access to information available from courses they are not enrolled in.

There are 12 different visualizations that have been grouped in those related to exercises, videos and course activity (Ruipérez-Valiente et al., 2017). The visualizations can be used for different purposes e.g., to detect problems in video resources (Pijeira Díaz et al., 2015). ANALYSE has been also used in different educational experiments (Redondo, Muñoz-Merino, Ruipérez-Valiente, Delgado Kloos, Pijeira Díaz & Santofimia Ruiz, 2015). The interface and setting of each visualization are similar. On the left we can see a description and selection boxes for the visualization options. The graphic is in the center of the visualization and on the right we can find the descriptive legend. Figure 3.4 shows an example of 'Course Summary' visualization within the dashboard for instructors.

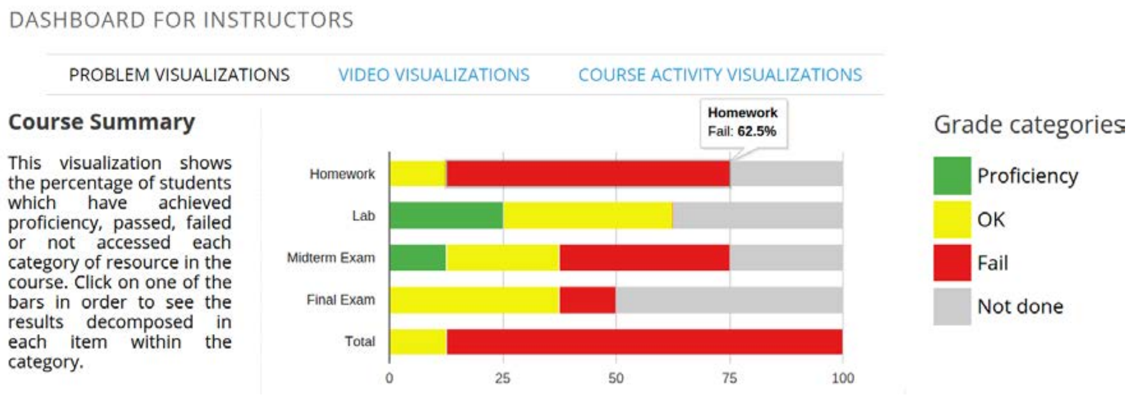


Figure 3.4: Interface of ANALYSE and visualization example of ‘Course Summary’ plot.

3.1.3. Technologies, Statistical Methods and Software

The main software, libraries and statistical methods that have been used during this dissertation are the next ones:

- **Web development**: Python with Django [7].
- **Data persistence**: MySQL, MongoDB, GAE Datastore, and text logs in different formats such as JavaScript Object Notation (JSON), eXtensible Markup Language (XML) or Comma-Separated Values (CSV) formats.
- **Statistical software**: R (*caret* and *dplyr* packages), Python (*pandas* and *scikit-learn* libraries) and SPSS.
- **Information visualization**: R (*ggplot2* package), JavaScript (Google Charts [8] and Highcharts [9] APIs), Gephi [10].
- **Group comparison**: Independent Student’s *t*-test [11], One-way Analyses of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA).
- **Supervised ML**: We have used algorithms such as linear and logistic regression (*glm* package), Random Forests (RF) (*randomForest* package), Support Vector Machine (SVM) (*svmRadial* function within *kernlab* package), k-Nearest Neighbours (kNN) (*knn* package) and Gradient Boosting Machine (GBM) (*gbm* package).
- **Clustering**: Two-Step Cluster Analysis [12].

[7] https://www.djangoproject.com/
[8] https://developers.google.com/chart/
[9] http://www.highcharts.com/
[10] https://gephi.org/
[11] https://en.wikipedia.org/wiki/Student%27s_t-test
[12] https://www.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.help/idh_twostep_main.htm

3.2. Case Studies

This section describes the different case studies and educational data that are analyzed in the dissertation. Subsection 3.2.1 describes the SPOC case studies that have been used as remedial courses at UC3M and using a local instance of Khan Academy platform. Then, we describe several MOOC case studies. Subsection 3.2.2 describes 'The Spain of Don Quixote' MOOC delivered on edX by UAM and Subsection 3.2.3 describes 'Mechanics Review' MOOC delivered also on edX by MIT. Finally, Subsection 3.2.4 describes the MOOCs 'Music Theory' and 'Introduction to Philosophy' delivered on Coursera by the University of Edinburgh.

3.2.1. SPOCs: Physics, Chemistry and Math with Khan Academy

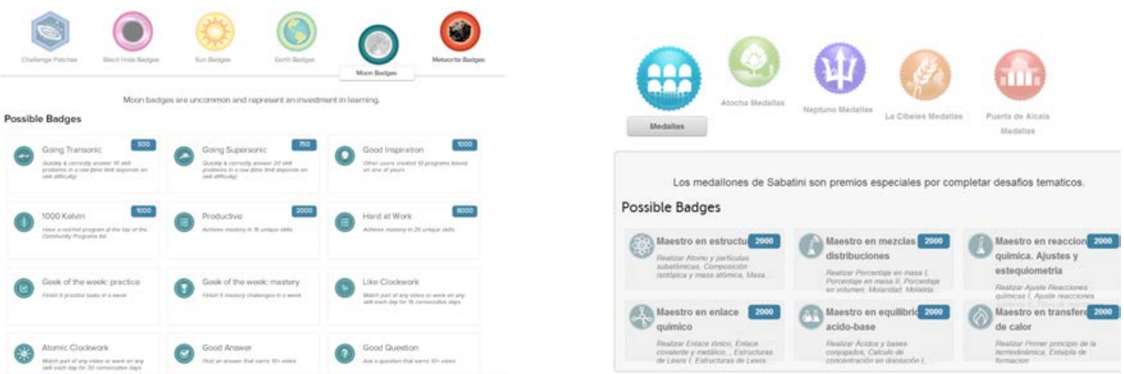
We analyze several SPOCs that lie within the context of the so called 0-courses at UC3M [13]. These courses are for first-year students who are entering a science degree and would like to review the concepts required for physics, chemistry and mathematics during their first year at the university, i.e., the courses are not mandatory for students entering at university but they can subscribe to them to reinforce their initial knowledge. Most of the participants are first year students around 18 years old. An 'inverted classroom' methodology (Lage, Platt & Treglia, 2000) is being used for those courses, that is, students first learn and review concepts at home by using an online platform during the month of August, and next take the face to face lessons in the university during the month of September. This initiative started with a pilot experience in the Summer of 2012 with physics, and expanded to chemistry and mathematics also next years. During the years 2012, 2013 and 2014 the university used Khan Academy as the support platform, during 2015 and 2016 the support platform was Open edX. Our findings suggest that students enjoyed the application of this initiative and also that it might improve the learning of students compared to more traditional approaches (Muñoz-Merino, Ruipérez-Valiente, Delgado Kloos, Auger, Briz, Castro & Santalla, 2016; Muñoz-Merino, Méndez Rodríguez, Delgado Kloos & Ruiperez-Valiente, 2017). During this dissertation we use the data that belongs to 2013 and 2014 0-courses with Khan Academy, data from 2012 was discarded as it was a pioneer initiative which had few students.

The learning resources and activities that are prepared by teachers for the online phase are composed of a set of videos and exercises. Although it is not mandatory for the students to access these online courses, it is strongly recommended; this is an important fact when measuring the use of the platform. Some of the students might have enrolled to several of these courses. It is important to note that when reporting descriptive statistics in the courses, we use the total number of students in the courses, i.e., students can be counted more than once, as each student might have behaved differently in each course and that is also valuable. However, when we perform inferential statistics such as correlation, we use the number of unique students i.e., every student is only counted once, in order to maintain the assumption of independence between cases for such

[13] https://www.uc3m.es/ss/Satellite/Grado/en/TextoMixta/1371213440582/Zero_courses

statistical techniques. This is due to the fact that we join data from all courses together.

The badge system of the Khan Academy was presented in Subsection 3.1.1.1. As part of the customization of the platform, the badge system was also adapted for both case studies of Subsection 3.2.1.1 and Subsection 3.2.1.2. The quantity of *topic badges* is different for each course. There is a base amount of common badges of 43 in all courses which is the sum of Table 3.1 badges and a specific amount of 7, 12 and 16 *topic badges* that were exclusively designed for the mathematics, physics and chemistry courses respectively.



(a) (b)

Figure 3.5: Subplot (a) shows an example of the original badge system of Khan Academy whereas Subplot (b) shows the adapted badge system for our case studies with the new *topic badges*.

3.2.1.1. Summer 2013

These SPOCs ran during the August month of 2013 and had 167 students in the physics course, 73 students in chemistry and 243 students in mathematics, that at least logged into the platform once. Thus, there is a total of 564 cases from the different courses. Additionally, as some of the students take more than one course, the number of unique students who participated in this experience was 372. The available courseware for these courses was 30 videos and 30 exercises in physics, 25 videos and 30 exercises in mathematics, and 22 videos and 49 exercises in chemistry.

3.2.1.2. Summer 2014

These second set of SPOCs run during the August month of 2014. In this second set we ran an experiment with a pre-test at the beginning and post-test after interacting with the platform to compute learning gains. These tests were available in physics and chemistry courses, thus we discarded math data from this dataset. The total amount of students who logged into the Khan Academy platform at least once was 156 for physics and 69 for chemistry. Additionally, the courseware contained 51 exercises and 24 videos for chemistry, and 33 for both exercises and videos for physics.

The purpose of enabling a pre-test and post-test is to be able to measure the learning achievement of students while interacting with the Khan Academy platform and to relate it with different indicators. We implement a pre-test and a post-test design for the second experiment. The pre-test and post-test were offered in the Moodle LMS. The pre-test aimed at measuring the students' prior knowledge in each one of the courses and was made available at the beginning of August.

Students had to complete it in order to be granted access to Khan Academy contents. At the end of August, the post-test was enabled, so that the students' knowledge after interacting with the Khan Academy platform could be measured. It is important to note that the post-test was not mandatory thus it was left undone by many students. The pre-test and post-test were a set of questions designed to have a similar level of difficulty. Then we can infer the learning gain of students ($LG = post_{test} - pre_{test}$).

The pre-test and post-test in physics had 10 questions each whilst the ones of chemistry had 21 as the contents which needed to be reviewed were broader. In order to guarantee that the difficulty of the pre-test and post-test was similar, the questions were pulled from a pool of similar difficulty. A total of 163 students in physics and 77 in chemistry completed the pre-test, but only 48 students in physics and 30 in chemistry also did the post-test. This was due to the fact that students had to do the pre-test in order to be able to access the Khan Academy contents, while the post-test was a voluntary activity (although emails were sent in order to encourage students to do both tests). In addition, not all the students who did both tests were included into the analysis. We added a condition that students needed to spend at least 30 seconds multiplied by the number of questions of the test in each test as this is the minimum estimated time for a student to read a question and answer it. We set this condition in order to remove those students that answered the test just randomly e.g., some students took only 1 minute or less to answer the complete test. With these restrictions, the total number of students that were considered for the analysis is 25 for chemistry and 44 for physics, which makes a total amount of 69 students. Based on this case study we define three variables that are the following:

- Pre Test Score (*pre test score*): Score of the student in the pre-test of this case study.
- Pre Test Time (*pre test time*): Time required to complete the pre-test of this case study.
- Learning Gain (*learning gain*): Variable defined as the difference between the *post test score* and the *pre test score*, providing an approximation about the learning achievement of the student while interacting with the platform.

3.2.2. MOOC 'The Spain of Don Quixote' on edX

UAM offered the first delivery of their MOOCs at edX platform in February 24th 2014 and this dataset belongs to one on these MOOCs, which is entitled 'The Spain of Don Quixote' - Quijote501x [14]. A total of 3.530 students enrolled in the course. However, only 1.718 students

[14] https://www.edx.org/course/la-espana-de-el-quijote-uamx-quijote501x-0

were actively involved with any of the course content of which 164 students obtained a grade of over 60% and thus received a certificate. Therefore, around 4.65% of the enrolled students earned a certificate, which is a completion rate similar to the ones reported in the literature. It is a 7-week course where every week there are multimedia resources, discussion forums, practical activities without evaluation, and also a final evaluation activity per week. The first and last week (seventh), students were evaluated with a peer review activity. For weeks 2 to 6, they were evaluated with a multiple choice test of 21-23 questions. Each weekly evaluation contributed a 14% to the final grade of the course. For the first three weeks of this course, the evaluation activities deadlines are four weeks after the release date. Then, from the fourth week to the end of the course the evaluation activities deadlines are three weeks after.

3.2.3. MOOC 'Mechanics Review' on edX

The case study is a MOOC in the topic of introductory physics called 'Mechanics Review' about Newtonian Mechanics [15] and run on edX by MIT faculty during the Summer of 2014. A total amount of 13.500 participants enrolled in the MOOC, and 502 of them managed to earn a certificate. The course lasted for 14 weeks and there were 12 mandatory units and two additional optional units on advanced materials. The course contained about 1.000 problems and 69 videos. These problems are organized as checkpoints embedded within e-text and videos, and homework and quiz problems which should be done at the end of each week. The weight of each type of assignment towards the final grade is different, being quizzes the most valuable and checkpoint the less valuable (Quiz > Homework > Checkpoints).

3.2.4. MOOCs 'Music Theory' and 'Introduction to Philosophy' on Coursera

This case study involves two different MOOCs run on Coursera platform and taught by faculty from the University of Edinburgh. The first one is 'Music Theory' [16] that was five weeks long, running from 14/07/14 to 18/08/14 and a total amount of 89.893 students signed up, 1 graded quiz per week with 10–14 questions each quiz. The second one is 'Introduction to Philosophy' [17] that run for 7 weeks from 15/09/14 to 10/11/14, had a total sign up of 33.446 students, 1 graded quiz per week with 6–12 questions each quiz. We used Coursera trace data in JSON format which contain records of all course events. From all the students that signed up, 2.359 and 5.159 students submitted all questions for philosophy and music respectively.

3.3. Selected indicators

In this section, we provide an overview of the different indicators we selected and use over the different case studies during this dissertation. We analyze data from different VLEs such as

[15] https://www.edx.org/course/mechanics-review-mitx-8-mrevx
[16] https://www.coursera.org/learn/edinburgh-music-theory
[17] https://www.coursera.org/learn/philosophy

Khan Academy, Coursera or Open edX, since each one of these platforms might have specific functionalities, some of the proposed indicators might not make sense in some platforms (e.g., those related to gamification only make sense in Khan Academy since there are not gamification features in Coursera or Open edX). Additionally, the data format for each platform is different and can also limit which of the indicators can be computed, since some data logs are richer than others. Furthermore, depending on the case study we have more interest in some indicators than others. Therefore, not all case studies use all the indicators described in this section. The organization of the indicators is as follows. We have grouped the indicators in those related to the use of the platform (Subsection 3.3.1), those related to the correct progress in the platform (Subsection 3.3.2), time in the platform and its distribution (Subsection 3.3.3), indicators that describe problems and submissions (Subsection 3.3.4), behavioral indicators about exercise solving habits (Subsection 3.3.5), indicators about badge behavior (Subsection 3.3.6) and finally we describe the two algorithms to compute the indicators regarding online collaboration and academic dishonesty (Subsection 3.3.7).

3.3.1. Use of the Platform

These features focus on how many learning items each student have interacted with. This is related to the number of videos and exercises a user has accessed, or the time a user has spent in the platform and on those different types of activities. These parameters do not take into account if a user has done very well or bad their exercises, but only the total use on the platform.

- Exercises Accessed (*exercises accessed*): The amount/percentage of unique exercises accessed by a given student.
- Videos Accessed (*videos accessed*): The amount/percentage of unique videos accessed or downloaded by a given student.
- Optional Activities (*optional activities*): This variable measures the number of optional activities (such as setting up an avatar or learning goals) that have been used by the student. More information regarding the optional activities can be found in Subsection 3.1.1.1 and about this metric in previous publications (Ruipérez-Valiente et al., 2014, 2016).
- Number of Active Days (*number active days*): Number of different days that the student logged into the platform and performed some action.
- Number of Active Sessions (*number active sessions*): Number of different sessions of the student with the course.
- Number of Events (*number events*): Number of events generated by the student during their interaction with the course.

- Number of Submissions (*number submissions*): The total number of problems that a particular student submitted.

- Number of Threads Viewed (*number threads viewed*): The total number of unique discussion topics accessed by a given student.

- Average Number of Hints (*average number hints*): Average number of hints asked by the student.

- Average Number of Attempts (*average number attempts*): Average number of attempts that a student makes trying to solve an exercise.

3.3.2. Correct Progress in the Platform

The features of this section describe indicators that represent how well users have interacted with the platform. This block does not take into account the total use but the performance of the student with the proposed materials.

- Proficient Exercises (*proficient exercises*): Percentage of exercises in which the student has acquired a proficiency level (from 0 to 100).

- Exercise Effectiveness (*exercise effectiveness*): This is a specific variable that computes a measure regarding the progress of students with exercises and might be specific for each course and platform. More information is available in previous work (Muñoz-Merino et al., 2015).

- Exercise Effectiveness with No Help (*exercise effectiveness no help*): This is the same previous measure, but adapted to consider only exercises solved without using hints.

- Video Effectiveness (*video effectiveness*): This is a specific variable that computes a measure regarding the progress of students with videos and might be specific for each case study and platform. More information is available in previous work (Muñoz-Merino et al., 2015).

- Completed Videos (*completed videos*): Percentage of videos completed by the student.

- Performance First Attempt (*performance first attempt*): Percentage of exercises that were solved correctly in their first attempt to the exercise.

- Average Time for Correct Answer (*average time correct answer*): Average amount of time required to provide a correct answer to an exercise.

- Number of Attempts per Correct Answer (*number attempts correct answer*): Average number of attempts required to provide a correct answer to an exercise.

- Certificate (*certificate*): Binary variable that represents the acquisition of a certificate and can take just two values: 0 (if it is not achieved) or 1 (if it is achieved).

3.3.3. Time in the Platform and its Distribution

These features describe the amount of time in the platform and how it has been distributed among the different activities.

- Exercise Time (*exercise time*): Summation of time invested between the access to a problem and the submission attempt to the problem. When there are several attempts, the time of each one is computed individually. This is computed separately for each exercise assignment.

- Video Time (*video time*): Summation of all the time invested in videos measured as the *play* time.

- Page Time (*page time*): Summation of time spent in course pages by the student.

- Total Time (*total time*): Summation of all the time invested interacting with the platform.

- Average Time per Day (*average time per day*): Average time spent in each day of the course.

- Dispersion of Time per Day (*dispersion time per day*): Dispersion measure of the time invested in each day of the course.

- Dispersion of Time per Exercise (*dispersion time per exercise*): Dispersion measure of the time invested in each exercise separately.

- Dispersion of Time per Video (*dispersion time per video*): Dispersion measure of the time invested in each video separately.

3.3.4. Problem and Submission Features

These features describe specific characteristics of a problem and about the submission attempts of students to problems. First the characteristics of problems that we use are:

- Location (*location*): Location of the problem or video within the course structure indicating with an integer the chapter where the problem is located.

- Type of Assignment (*type assignment*): Factor variable that indicates the type of assignment e.g., free text peer review, checkpoints, exam, laboratory, etc.

- Type of Response (*type response*): Factor variable that defines the type of response of each problem e.g., multiple choice, fill the gap, formula, etc.

- Show Answer (*show answer*): Factor variable that defines the configuration of the 'show answer' button. It can be available always, only after maxing all your attempts or only after the due date.

- Random (*random*): Binary variable indicating if the problem contains random variables or not.

- Max Attempts (*max attempts*): This variable specifies the maximum number of attempts allowed in the problem.

Then, the features related to the submission attempts of students are:

- Time to Deadline (*time to deadline*): Difference of minutes between the submission deadline for the problem and the actual timestamp when the student submitted the problem.

- Attempt Duration (*attempt duration*): Number of minutes elapsed between the event when the student accessed the problem and the submission of the problem.

- Attempts Required (*attempts required*): Number of attempts required to answer correctly the problem.

3.3.5. Behavior Solving Exercises

These parameters represent behaviors that student might do when interacting with exercises.

- Exercise Abandonment (*exercise abandonment*): Percentage of exercises that were started but the student never achieved mastery in them.

- Video Abandonment (*video abandonment*): Percentage of videos that were started by the student but never completed.

- Follow Recommendations (*follow recommendations*): This variable provides the percentage of exercises that were accessed by the student via a resource recommendation sent by the platform.

- Forgetful User (*forgetful user*): This variable provides information about the percentage of exercises that students failed to solve after solving an exercise of the same type correctly.

- Video Avoidance (*video avoidance*): Variable about users who failed to solve correctly an exercise and still they do not watch the video which is associated to that exercise.

- Hint Avoidance (*hint avoidance*): Variable about users who failed to solve correctly exercises and still they do not ask for hints.

- Unreflective User (*unreflective user*): Variable about students who attempt to solve an exercise too many times without reflecting.

- Hint Abuse (*hint abuse*): Variable about students who ask for too many hints without reflecting on the exercise statement or previous hints.

3.3.6. Behavior with Badges

This section describes the implemented indicators which model students' behavior towards badges. The objective is to propose some measures which can provide a deep insight about the interaction of users with badges. These four models take into account the specific *topic badges* and *repetitive badges* that were explained in Subsection 3.1.1.1. We provide here only a brief overview, more details can be found at (Ruipérez-Valiente et al., 2016b, 2017). We defined the following four indicators:

- Intentionality on Topic Badges (*intentionality topic badges*): The subset of exercises that are required to earn *topic badges* are always different. Therefore, we can try to infer if a student is trying to maximize the number of *topic badges* that he/she is acquiring, or if they are earning them as just part of the learning process. The implemented algorithm gets the number of problems that the student has achieved as well as the number of *topic badges* earned by the student. Then, it calculates the maximum number of *topic badges* that the student could have earned with that number of proficient problems in case he/she intended to do that.

- Intentionality on Repetitive Badges (*intentionality repetitive badges*): Upon achievement of proficiency, students receive a notification from the system, and they should stop doing exercises of this same skill and move on to the next one. In the case that students keep doing exercises of a skill in which they are already proficient, and they keep earning *repetitive badges* this way, we hypothesize that are earning those badges on purpose. Finally, we can compute a percentage on the amount of *repetitive badges* that were earned intentionally, and that provides information about if the student is trying to earn badges on purpose instead of as part of the learning process.

- Concentration on Achieving Badges (*concentration badges*): Students can devote all their consecutive actions into fulfilling the requirements of one badge. Another possibility is that students carry out different actions in the middle, which are not related to the requirements of that badge before actually receiving the badge. Since *topic badges* have as requisites a fixed set of exercises, we can track if students have done the required exercises in a consecutive way or others in the middle before earning a *topic badge*. Following this criterion we can infer the proportion of the previous exercises that a student attempted that actually belong to the requisites. This metric is applicable only to *topic badges* since *repetitive badges* by doing many times a single action.

- Time Efficiency in Badges (*time efficiency badges*): The time invested by students can be used to obtain measures which give insight about the number of badges they earn per unit of time. We define this measure as the total number of badges divided by total time in the platform.

3.3.7. Online Academic Dishonesty and Collaboration

As part of the dissertation we design and implement two novel algorithms to detect different types of academically dishonest behaviors. The first one is related to accounts that always submit their assignments very close in time (Subsection 3.3.7.1). The second one is a particular cheating method denominated as CAMEO where students create puppet accounts that are used to obtain the correct solutions that are later on used in their main account to gain a certificate (Subsection 3.3.7.2).

3.3.7.1. Close Submitters

We aim to detect user accounts of students in online courses that always submit their assignments very close in time. We want to address this issue by providing a systematic method and algorithm that can be easily applied to any online environment where students have to perform certain learning activities. This algorithm will be able to detect different associations between accounts such as unethical behaviors but also genuine beneficial collaborations between students. In order to build this algorithm let us start by defining a vector $\vec{s} = [s_1 \cdots s_N]$ representing all the students in a course and a vector $\vec{p} = [p_1 \cdots p_M]$ representing all the problems in a course. Then we can define a matrix SP with dimensions NxM where each row represents a student ($\vec{s}$) and each column represents a problem ($\vec{p}$) in the course, thus SP has the following shape:

$$\underline{SP} = \begin{bmatrix} sp_{11} & sp_{12} & sp_{13} & \dots & sp_{1M} \\ sp_{21} & sp_{22} & sp_{23} & \dots & sp_{2M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ sp_{N1} & sp_{N2} & sp_{N3} & \dots & sp_{NM} \end{bmatrix}$$

where the value of each cell e.g. sp_{jk} represents the timestamp of the last submission done by s_j (student j) to P_k (problem k). Additionally, $sp_{jk} = NA$, in the case that s_j did not submit p_k.

Then let us define a symmetric distance matrix D with dimensions NxN where both rows and columns represent the vector of students ($\vec{s}$) and it will store the distance between two students taking into account their submissions to all problems. Then D has the following shape:

$$\underline{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1N} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & d_{N3} & \dots & d_{NN} \end{bmatrix}$$

where the value of each cell d_{ij} represents the distance between all the submissions of s_i and s_j, more formally and applying for example a euclidean distance, d_{ij} is defined as follows:

$$d_{ij}^{EUC} = dist(i,j) = \sqrt{\sum_{k=1}^{M}(sp_{ik} - sp_{jk})^2} \tag{3.1}$$

Where the function $dist(i,j)$ can take different distance metrics and this can lead to different matrices depending on the distance metric. Since the matrix is symmetric, note that $d_{ij} = d_{ji}$. Additionally, note that the entries of the main diagonal are 0 ($d_{ii} = 0$ for all $1 \leq i \leq N$, thus $\underline{D}$ is also a hollow matrix). For those distances between students that cannot be computed, then $d_{ij} = d_{ji} = NA$.

Notice the complexity of computing this matrix is $O(N^2 * d)$, where d is the cost of computing a distance between two arrays of submission timestamps and depends on the distance metric and the number of submissions. Therefore the complexity grows exponentially with the number of students that we want to include in the distance matrix. Now we are interested in the distance values of the lower triangular part of the matrix without including its diagonal (that means every possible different combination of students) and we define this vector of distances as $\vec{d}$ which is has the following shape $[student_i, student_j, dist(i,j)]$ to allow identification of each distance.

From the previous general mathematical definition, there are certain decisions we take about the criteria for a practical implementation in the case study. These decisions are as follow:

- We keep only the assignments that are graded quizzes.
- We keep the last submission timestamp for a given problem and student, that is the timestamp that is stored in $\underline{SP}$.
- To include an account in the analysis we require that all the graded quizzes within the course have been submitted. This decision makes the implementation less computationally demanding because there are fewer students and increases certainty that distances are not small by mere chance.
- In terms of distance metrics we decide to use average distance metrics since this way is easier to compare different courses that can have different amount of assignments. Also, we consider that it is adequate to use squared metrics because we want to heavily penalize far distances. Therefore, we decide to use two distance metrics, Mean Absolute Deviation (MAD) and Mean Squared Deviation (MSD) that are defined as follow:

$$d_{ij}^{MAD} = \frac{1}{M}\sum_{k=1}^{M}|sp_{ik} - sp_{jk}| \tag{3.2}$$

$$d_{ij}^{MSD} = \frac{1}{M}\sum_{k=1}^{M}(sp_{ik} - sp_{jk})^2 \tag{3.3}$$

More information about this algorithm and method can be found in our publication (Ruipérez-Valiente et al., 2017a). Finally, we only need to empirically compute a threshold for $\vec{d}$ and all the triplets below that value will be labeled as close submitters. Following this method we define the following variables:

- Close Submitter (*close submitter*): Boolean variable indicating if a specific account was detected as *close submitter* by the algorithm or not.

- Order (*order*): Given two accounts in the course detected as *close submitters*, this variable ranges from -1 to 1 indicating the sign of the order of the submissions of one respect to the other i.e., when the variable takes the value of 1 it means that the first account always submitted first and when it is -1 it means that the second account always submitted first; values in between indicate middle situations.

3.3.7.2. Copying Answers using Multiple Existences Online

This algorithm aims to detect the use of multiple accounts by the same student for copying answers. The method works as follows. The student uses one or more harvesting accounts (the *harvester/s*) to obtain the correct answer, and then submits it in the *master* account, the account for which the student intends to earn a certificate. Finding the answer in the harvester account can be done either by asking to see the correct answer after using all the attempts ('show answer' on the edX platform), or by 'exhaustive search' (e.g., pure guessing for multiple choice questions) until the correct answer is found. This cheating method is known as CAMEO. The operationalization of the algorithm is based on the IP address of the submissions. Since the user accounts can have different IP's during a course (due to switching locations or other reasons). We define 'IP group' as all accounts linked through an IP. More specifically, it is a group of accounts that shared the same IP at least once in the course, or are connected through an account with whom both shared an IP (this criterion is applied recursively). This a graph theory problem which identifies all the connected components to build the IP groups.

The implemented algorithm searches for CAMEO events between all pairs of accounts in each IP group. It is composed of two main steps. The first step detects which events fulfill the general pattern of CAMEO events, which is that one account gets a solution to a problem, and then a second account that belongs to the same IP group submits a correct answer to the given problem shortly after. Students can obtain the correct solution to a problem either applying exhaustive search by doing different attempts in multiple choice questions until getting the correct one or by using the show answer button enabled in some exercises. This step generates a list of pairs of master and harvester accounts, and for each given pair, a list of questions which are suspected to be correct applying CAMEO. Then, the second step has the purpose of adding criteria in order to filter false positives and increase the reliability of the detected pairs. More specifically, the two steps are executed as follow:

1. ***First step***: For each account a_1, for each correct submission made by a_1 to a question q, we check whether any other account a_2 within the IP group of a_1 obtained the correct answer to q, in the previous 24 hours. If a match is found, we add $< a_1, a_2, q >$ to the list of potential CAMEO events. A graphical illustration of two kinds of CAMEO schemes are given in Figure 3.6.

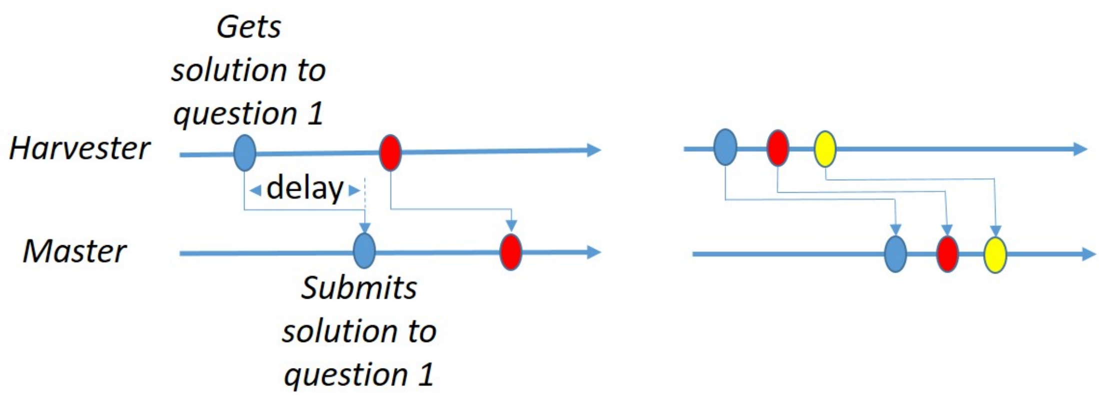


Figure 3.6: Diagram showing two different CAMEO patterns: *immediate* (left chart) and *batch* mode (right chart).

2. ***Second step***: To the events collected as part of the step 1 we apply the additional filters in the specified order:

   a) **The harvester account does not receive a certificate**. The rationale is that if the harvester account is used only to harvest solutions, it would not make sense that it gets a certificate, thus it might be a false positive.

   b) **Master-harvester pair harvests at least 10 questions**. This is a filter that seeks to remove noisy master-harvester couples that have performed CAMEO in a small number of questions. The rationale behind it is that real master-harvester pairs would exhibit this behavior on a significant amount of questions. The specific value has been found empirically by analyzing the accumulative distribution of CAMEO questions for each pair.

   c) **More than 5% of the master's correct submissions are potentially harvested**. The rationale is to have a 'significant level' threshold on the amount of questions that the master is suspected to harvest.

   d) **Evidence of 'inhumanly fast' submissions**. Previous work found that a very short delay between opening a problem and submitting a correct answers was related to cheating (Palazzo, Lee, Warnakulasooriya & Pritchard, 2010), and our findings suggest the same. Therefore, this filter is passed when a potential master have a number of inhumanly fast correct submissions. There are two parameters, the upper bound time to consider it 'inhumanly fast' which is established as 30 seconds based on previous research (Palazzo et al., 2010) and our own findings, and the number of questions which is established as a minimum of 6 events below 30 seconds.

e) **Harvester works for masters**. We require that most of the questions done by the harvester (more than 55%) were actually used by a master account. The purpose of the harvesting account is to support the master account, so it should do only helpful actions. The specific value (55%) was picked by observing an elbow in the graph based on the master-harvester pairs detected.

f) **The harvesting account must not exhibit 'master' behavior (and vice versa)**. The rationale is that an account that is a 'service' account is not likely to use harvester accounts, and that an account that is a master is not likely to 'service' other accounts. We note that such a behavior would be expected of two students who collaborate, but not in CAMEO.

Users whose master accounts pass these filters are termed CAMEO users. A more complete description of the algorithm can be consulted in our previous publications (Ruipérez-Valiente et al., 2016; Alexandron et al., 2017). Using this method we are able to detect CAMEO users and also define the following variable:

- Harvested (*harvested*): Boolean variable indicating if a specific correct submissions was cheated using CAMEO or not.

Part II

Analysis and Statistical Inference from Educational Data

Chapter 4

Exploratory Analysis

Contents

This chapter provides an exploratory analysis about the different learning indicators that we use in the posterior chapters. There is a special emphasis on those that we further analyze later in this dissertation. The chapter is divided as follows. First, we explore some indicators related to the use of the platform and behavior of students in Section 4.1. Second, we explore the interaction with regular courseware activities in Section 4.2 and with optional activities in Section 4.3. Next, we explore badge activity and the behavior of students with badges in Section 4.4. Finally, we explore online academic dishonesty and collaboration in Section 4.5 in the specific cases of CAMEO (Subsection 4.5.2) and *close submitters* (Subsection 4.5.1).

4.1. Use of the Platform and Student Behavior

The analysis that we present in this section is based on the dataset of Case Study 3.2.1.1. First, we explore those indicators related to the use of the platform in Figure 4.1. We plot the density distribution of the entire population for each indicator, for *exercises accessed* and *videos accessed* (top left), for *exercise abandonment* and *video abandonment* (top right), for *optional activities* (bottom left) and for *total time* (bottom right). The plot regarding *exercises accessed* and *videos accessed* show as a very similar distribution for both metrics with an average value of 38.73% for *exercises accessed* and 39.71% for *videos accessed*, which seems to indicate that the percentage of accessed resources is similar for both exercises and videos. In the case of the abandon metrics, *video abandonment* (average of 40%) is higher at the beginning and *exercise abandonment* (average of 48%) at the end of the distribution. These distributions seem to indicate that students tend to abandon exercises more often than videos, which makes sense since it might be harder to solve an exercise correctly rather than just finish a video. The *total time* indicates that the average student invested 247 minutes in the platform, however there are many students that invested more than 16 hours in the platform, which we consider a very high amount of time taking into account that the use of Khan Academy was not mandatory.

Figure 4.1: Density distribution of the indicators related to the use of the platform.

Additionally, while students solve those exercises they can perform different behaviors. Figure 4.2 shows the density distribution of the exercise solving behaviors. Since students that hardly interacted with the platform would have behavioral indicators around 0%, only those students who interacted at least for 60 minutes have been included, so that we can have a more realistic visualization of the distribution of the behaviors. The *follow recommendations* indicator is the most uniformly spread, where some students follow recommendations frequently while others never do that. The rest of the behavioral indicators *forgetful user*, *video avoidance*, *hint avoidance*, *unreflective user* and *hint abuse* have a similar density distribution, where most students are placed in the lower mid area of the function ($< 50\%$), but some of them have high values as well.

Figure 4.2: Density distribution of indicators representing the behavior of students when solving exercises.

4.2. Regular Courseware Activities

In this section we explore the interaction of students with regular learning activities (such as videos and exercises) and analyze the indicators that measure the effectiveness of students with these activities. First we provide an overview of these effectiveness indicators 4.2.1 and second we analyze its relationship with other metrics 4.2.2. This section also uses the dataset of the Case Study 3.2.1.1.

4.2.1. Overview of the Effectiveness with Exercises and Videos

Figure 4.3 shows the density distribution of the effectiveness metrics (*exercise effectiveness* and *video effectiveness*) on the left and the completed resources metrics on the right (*proficient exercises* and *completed videos*). Both plots have in common that the effectiveness and the percentage of completed resources is a bit higher for videos, again probably related to the fact that advancing in videos might be easier and less demanding that completing exercises correctly. However, we can see that the proposed definition that we did in Subsection 3.3.2 regarding the effectiveness, makes less abrupt the difference between *exercise effectiveness* and *video effectiveness* (on the left) compared to *proficient exercises* and *completed videos* (on the right). Although we can see many students with very low rates of effectiveness and completed resources, many others are situated at the end of the distribution meaning that they completed all the available resources.

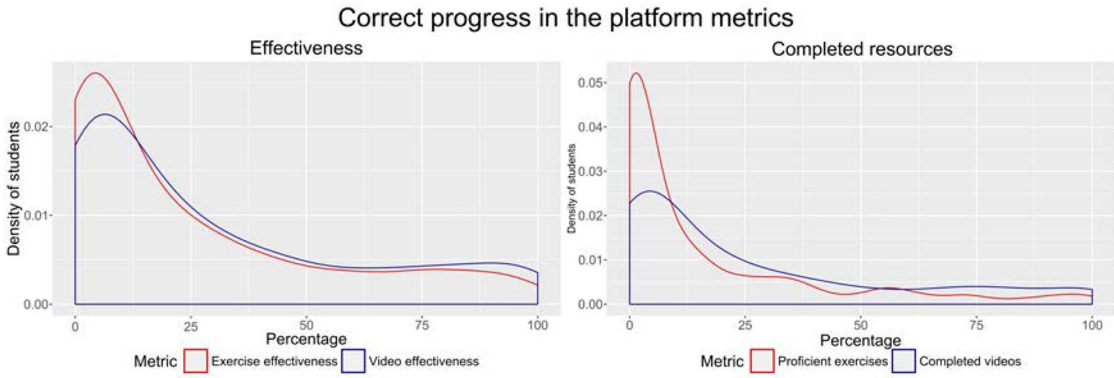


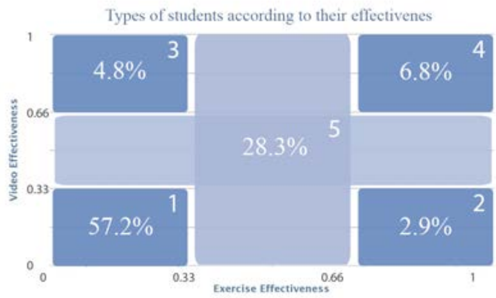
Figure 4.3: Density distribution of the effectiveness and the completed resources among the different students.

Figure 4.4a represents the effectiveness of each student with videos and exercises in the three SPOCs. Each point represents the effectiveness of a particular student with videos (y-axis) and exercises (x-axis). This way, teachers can know students' interactions with educational resources and activities, and recognize at a glance different types of students. Figure 4.4b classifies the students in Figure 4.4a in five profiles regarding their level of effectiveness with videos and exercises. This visualization can be particularly useful to detect resources and activities that are poorly balanced (high proportion of students in areas tagged '2' and '3'). We can have five different types of students:

- Do nothing (neither exercises nor videos) or very little (area tagged '1').
- Do everything or almost everything (area tagged '4').
- Do only or mainly videos (area tagged '3').
- Do only or mainly exercises (area tagged '2').
- Do some videos and some exercises (area tagged '5').

4.2.2. Relationship of Effectiveness with Other Metrics

Table 4.1 shows the Pearson correlation of *exercise effectiveness* and *video effectiveness* indicators with other metrics and themselves. We find that *exercise effectiveness* and *video effectiveness* are strongly correlated (0.63). This correlation indicates that active users, who interact a lot with one of these types of contents, will potentially interact also a lot with the remaining



(a) (b)

Figure 4.4: Effectiveness scatterplot and types of students according to their effectiveness.

types of contents. Besides the obvious relationships of effectiveness with other metrics of use of the platform, there are also several statistically significant relationships with students' behavior metrics. We find a low and negative correlation (-0.115) between *exercise effectiveness* and *follow recommendations* which seem to indicate that students who followed the recommendations provided by the Khan Academy platform were less effective when interacting with exercises. We also find significant but low correlations between *video effectiveness* with *forgetful user*, *video avoidance* and *hint avoidance* (0.124, -0.234, 0.128). The most interesting one is between *video effectiveness* and *video avoidance* since it indicates that students who avoid watching videos, are less effective when interacting with the videos they access (as one could expect).

4.3. Optional Activities

In this section we analyze the use of optional activities, that can be defined as those that students can perform voluntarily. The section is divided as follows: Subsection 4.3.1 presents an overview of the use of the optional activities available in Khan Academy. Subsection 4.3.2 analyzes the relationship of the use of optional activities with other metrics and learning outcomes. Finally, Subsection 4.3.3 presents a categorical variable analysis of optional activities. This section uses the two SPOCs datasets of years 2013 and 2014 described in Case Study 3.2.1.

4.3.1. Overview of the Use of Optional Activities

Students did not use optional activities a lot in their interaction with the platform. The distribution of *optional activities* indicator can be seen in previous Figure 4.1. Actually, only 23.2% of the students made use of at least one of the five optional activities available. We can see the use of each optional activity separately in Table 4.2. These results take into account all the students who logged in at least once to the Khan Academy platform. Consequently, some of these students did not interact much with the system, neither with optional activities nor with learning activities.

Table 4.1: Bivariate Pearson correlation of the effectiveness and use of optional activities with the rest of indicators.

Bivariate Pearson Correlation[1]	Exercise Effectiveness	Video Effectiveness	Optional Activities
Exercises Accessed	0.901**	0.678**	0.429 **
Videos Accessed	0.654**	0.937**	0.419 **
Exercise Abandon	-0.288**	-0.044	-0.259 **
Video Abandon	-0.168**	-0.418**	-0.155 **
Total Time	0.795**	0.833**	0.491**
Proficient Exercises	0.910**	0.517**	0.553**
Completed Videos	0.612**	0.983**	0.435**
Follow Recommendations	-0.115*	-0.101	-0.002
Forgetful User	0.029	0.124*	0.007
Video Avoidance	-0.078	-0.234**	-0.051
Hint Avoidance	0.061	0.128*	0.053
Unreflective User	0.035	0.044	0.039
Hint Abuse	-0.111	-0.103	-0.089
Optional Activities	0.485**	0.438**	1
Exercise Effectiveness	1	0.630**	0.485**
Video Effectiveness	0.630**	1	0.438**

[1] N = 291 students from Case Study 3.2.1.1.
[*] Correlation is significant at the 0.05 level (2-tailed).
[**] Correlation is significant at the 0.01 level (2-tailed).

From Table 4.2 we can extract the following main conclusions. The optional activities used the most are the customization of a profile avatar and the badge display. Although the exact percentage numbers differ from one course to another, on average the results show that 10.8% customized their avatar and 12% their badge display i.e., they are by far the most used optional activities in all courses. A possible reason for this could be that these students, who are aged around 17-19 years, are comfortable using activities that come from a social network or gaming contexts. On the other side, optional activities that are related to learning (feedback, vote and goal) have been used much less (4.1%, 6.6% and 6.2%, respectively). The activity which has been used the least is feedback. A reasonable argument is that writing a feedback answer about a video generally requires a greater effort than just simply changing an avatar.

It is also interesting to look at more specific details about students' behavior in some optional activities. For example, we can focus on the ratio of finished goals and the type of votes. The number of students who set goals was 30, setting up a total number of 55 goals when taking into consideration all courses. The minimum number of goals set by a student was 1 while the maximum was 3. Taking into account all goals, 28 of them (50.9%) were reached. This finishing ratio seems to be rather high. However, the goal setting is optional but the selected goal e.g., finishing an exercise, might be crucial for understanding the topics that the course is covering. Furthermore, we assume students that use the optional functionality of setting goals to be highly self-motivated and confident about reaching a goal when selecting it which might bias the finishing ratio. Moreover, the number of students who submitted a vote is 32 with a total number of 40 votes in all courses: 26 of them were positive (65%), 13 of them were indifferent (32.5%) and only one of them was negative (2.5%). These results indicate that most of the users vote for positive reasons given these conditions and it is very unlikely that they vote negatively on other students.

Table 4.2: Percentage of students who used each optional activity.

Optional activity	**Feedback**	**Vote**	**Goal**	**Profile avatar**	**Badge display**
Percentage	4.1%	6.6%	6.2%	10.8%	12%

Table 4.3: Comparison between the use of regular learning activities versus the use of optional activities

Type of activity	**Percentage of activities accessed**				
	0%	1-33%	34-66%	67-99%	100%
Regular learning activities	2.48%	51.55%	23.19%	18.84%	3.93%
Optional activities	76.81%	18.43%	4.14%	0.41%	0.21%

We can establish a comparison between the access to regular learning activities, such as exercises and videos, and to optional activities. This comparison is presented in Table 4.3. It allows us to get a sense of how many students have used regular activities in comparison to optional activities. We divide the use of regular and optional activities in five intervals and we show the percentage of students from all courses in each interval. The first detail to notice is that only 12 students (2.48%) who logged in on the platform did not use any of the regular learning activities while 76.81% did not use any of the optional activities. This is a huge difference that already gives insight about the low use of optional activities compared to the use of regular learning activities. On the other end, we can notice that 19 students (3.93%) used all the regular learning activities while only one of the students (0.21%) used all the optional activities. We should also keep in mind that the amount of learning activities is above 40 in all courses whereas the number of optional activities taken into account in the study is only five. Finally, we can see that the use of activities in the 1-99% interval declines gradually and is always superior for the regular activities.

4.3.2. Relationship of Optional Activities with Other Metrics

In this section, we analyze how the use of optional activities is related to other metrics, and more importantly to learning outcomes. Table 4.1 showed the correlation between *optional activities* and other metrics. The data shows that the most significant correlations are with the *total time* (0.491) and *proficient exercises* (0.553). We further explore this relationship with *proficient exercises* later. Some other correlations are also high such as with *exercises accessed*, *videos accessed* and *completed videos* (0.429, 0.419 and 0.435). The results show that the use of *optional* *activities* is also correlated with the effectiveness in educational resources (0.485 for *exercise effectiveness* and 0.438 for *video effectiveness*) These results make sense, as the more time a student spends on the platform, the more videos, exercises and optional activities might do. In addition, another significant but negative and low correlation exists for *exercise abandonment* and *video abandonment* (-0.259 and -0.155). This negative correlation means that users who abandon exercises and videos use a bit less optional activities than others. Finally, the results indicate that other behavioral indicators (*follow recommendations*, *forgetful user*, *video avoidance*, *hint avoidance*, *unreflective user* and *hint abuse* indicators) are not significantly correlated with *optional activities*. We found that there is no relation between *follow recommendations* indicator and *optional activities*, although we initially thought that there could exist a relation due to the fact that the use of optional items can be regarded as an exploring behavior.

As next step we delve into understanding the relationship between the use of optional activities and learning outcomes. Table 4.4 provides the Pearson correlation and partial correlation of both *proficient exercises* and *learning gain*, with *optional activities* and each optional activity separately. The first row of Table 4.4 shows the Pearson correlation with *proficient exercises*. The most significant correlation (0.553) is with the global measure *optional activities*. This strong correlation points out that the use of optional activities might be used as an indicator to know how well students have mastered the exercises. Avatar and display badge (0.415 and 0.418) are the optional activities that have been most highly correlated with the percentage of *proficient exercises*, whereas feedback and vote (0.205 and 0.243) have been the least. This might be surprising at first sight because feedback and vote are supposed to be related to the learning process and one might thus think that they should have a higher correlation with solving exercises correctly than avatar and display badges which are not related to the learning process. However, the use of the avatar and display badge are moderately related to the total time spent on the platform (correlations of 0.28 and 0.24 respectively) and students that spend time on the platform are related to perform better (correlation of 0.70) when solving exercises, so the cause of an improvement in *proficient exercises* might not be the use of optional activities by itself. In order to gain more insight about the results, the second row of Table 4.4 shows the partial correlation between the same indicators taking out the effect of the rest of the variables considered in the study. After controlling the effect of the rest of the variables, the significant correlation disappears in the case of *proficient exercises* with feedback and votes, and decreases in the case of *proficient exercises* with *optional activities* (0.282), goals (0.250), avatar (0.235) and display badges (0.229). Indeed, these are low levels of relations. Therefore, when removing the effect of other variables, the relation between *proficient exercises* and *optional activities* is not so strong. However, there is some relation between *optional activities* and *proficient exercises*, taking out third variables like the effect of total time spent.

The last two rows of Table 4.4 show the correlation of *optional activities* with *learning gain*, which we note that has been performed using the data from the Case Study 3.2.1.2 since it is the only one where we have learning gains available. The third row shows the Pearson correlation where there are significant relations of *learning gain* with *optional activities* (0.293), use of vote (0.333) and use of display badges (0.296). However, the level of relation might be due to third variables, such as for example *total time* or *proficient exercises* that had a moderate/high correlation as we saw in previous correlations. For example, the more time a student spent on the platform, the more probable it is that he/she votes or changes the badges, does more activities of all the types and thus learns more. The last row of Table 4.4 presents the partial correlation of *learning gain* and *optional activities* taking out the effect of the rest of the variables considered. The objective is to remove the possible influence of the other variables to better understand the relation of *learning gain* and *optional activities*. When removing the effect of the other variables in the partial correlation, only one significant correlation remains which is with the use of display badges (0.261).

Table 4.4: Pearson and partial correlations of proficient exercises with optional activities.

Metric	Optional Activities	Goal	Feedback	Vote	Avatar	Display badges
Pearson correlation[1] Proficient exercises	0.553**	0.384**	0.205**	0.243**	0.415**	0.418**
Partial correlation[2,*] Proficient exercises**	0.282**	0.25**	-0.04	-0.031	0.235**	0.229**
Pearson correlation[1] Learning gain	0.293**	0.102	0.219	0.333*	0.221	0.296**
Partial correlation[2,*] Learning gain**	0.142	-0.07	0.124	0.214	0.17	0.261*

[1] N = 291 students from Case Study 3.2.1.1.
[2] N = 69 students from Case Study 3.2.1.2.
[*] Correlation is significant at the 0.05 level (2-tailed).
[**] Correlation is significant at the 0.01 level (2-tailed).
[***] Controlling for the all the variables considered in the study.

4.3.3. Categorical Variable Analysis

We analyze the relationship among all the categorical variables that represent the use of each optional activity. To this end, we apply a log linear analysis which allows the comparison of three or more categorical variables in order to determine if there is an association between two or more of them. The factors of the test are the use of each optional activity separately (yes or no) for each student. Table 4.5 shows the cell count of a log linear analysis of only those associations where the observed count is above or equal to 1 percent of the cases.

Table 4.5 shows which ones are the most typical associations in percentage. The higher counts are the use of display badge (4.1%), the use of avatar (2.9%), the use of both display badge and avatar (2.9%) and the use of votes (3.1%). The data indicate that there are probably underlying associations between the use of these activities, consequently we check other tests to see if it is really significant. The z-score values show that the most significant relations are between the

Table 4.5: Log linear analysis of optional activities as categorical variables

Used goal?	Used feedback?	Used vote?	Used avatar?	Used display badges?	Observed Count	Observed Percentage
No	No	No	No	No	371	76.8 %
No	No	No	No	Yes	20	4.1 %
No	No	No	Yes	No	14	2.9 %
No	No	No	Yes	Yes	14	2.9 %
No	No	Yes	No	No	15	3.1 %
No	Yes	No	No	No	8	1.7 %
Yes	No	No	No	No	5	1.0 %
Yes	No	No	Yes	Yes	7	1.4 %

use of avatar and display badges ($z = 2.68, p = 0.007$), between the use of feedback and votes ($z = 2.26, p = 0.008$) and also between the use of goal and avatar ($z = 2.1, p = 0.036$). These results make sense because an association between the use of avatar and display badge is related to activities that come from customizing your personal profile, and the association between the use of feedback and votes are activities related to participation in a forum. In addition, there is a three-way significant relation between the use of goals, avatar and display badge ($z = 1.96, p = 0.05$), which is also interesting because these three activities are related to gaming or social networks environments.

4.4. Use of Badges and Gamification Behavior

This section explores the use of badges and the behavior of students with them. First, we present an overview of the use of badges in Subsection 4.4.1. Second, we analyze the factors that can influence the number of badges delivered in Subsection 4.4.2. Third, we explore the distribution of the badge metrics in Subsection 4.4.3. Finally, we present the correlation of badge metrics with other metrics in Subsection 4.4.4. The analysis of this section uses the dataset of Case Study 3.2.1.1.

4.4.1. Overview of Badge Activity

This subsection presents a general analysis of the achievement of badges by the students of the three courses. As some badges are quite straightforward to acquire with the interaction with the platform, then most students have obtained some of them (even if they did not have the intention to get them). Nevertheless, there are others that are very hard to earn. The total number of badges delivered is 1153, 1609 and 4773 for the chemistry, physics and mathematics courses respectively. Taking into account the number of students in each course, the number of badges per student is 15.8, 9.6 and 19.64 respectively.

Figure 4.5 shows on the left a histogram chart of the total amount of badges earned by each user taking into account all the courses. This graph describes the distribution of badges earned by each user and can be used to inspect where most of the population is concentrated. For example, we can see an important peak at the beginning of the distribution which are those students who interacted very little with the platform. In addition, it is interesting to see how there are many students in the interval from 100 to 1000 which has achieved a big amount of badges. Analogously Figure 4.5 shows on the right a histogram of the quantity of different badge types earned by each user considering all courses. The vast majority of the population is concentrated in the interval from 1 to 10 different badges. The rational is that most students earn the same types of badges repetitively. There are important differences with Figure 4.5 on the left, since the previous analysis took into account all those badges that can be earned repeatedly and this plot considers each badge type once. This distribution has a more abrupt descendant curve that the other one, because most users earned few different badge types.

Figure 4.5: Histogram representation of the amount of badges (left) and amount of different badge types (right) earned by each student.

Additionally Figure 4.6 represents a boxplot visualization of the percentage of badges acquired by students split by course and by the different badge categories as described in Table 3.1. The black dashed line represents the sample mean. Students beyond the end of the whiskers are considered outliers and plotted as black dots. This data confirms that not many social badges were delivered, as social activity within the platform was not very widespread. Video badges have the highest median since there were only 5 different video badges, and some of them were easy to acquire. We can find some interesting outliers such as some students that achieved more than 75% of all the exercise badges.

4.4.2. Influence of Factors on Badge Count

Some badges are triggered when solving exercises or when watching videos, thus it is interesting to analyze which exercises and videos trigger the biggest amount of badges. The causes

Figure 4.6: Boxplot visualization representing the percentage of badges earned by each student divided by badge category (x-axis).

can be very diverse, for example an easy exercise can be used to obtain many correct exercises in a row or a difficult topic might trigger more video badges because students need to watch the video more than once. There are some important differences between those exercises and videos that triggered the biggest amount of badges and those which triggered the least. For example for the math course, the exercise which triggered more badges is the 'Biquadratic equation' exercise with a 301 badge count whilst 'Basic operations with complex numbers triggered' only 31 badges. These differences are also applicable for the other courses and in videos as well. We analyze the possible reasons for these results taking into account the *location* of the exercise and the following two additional variables:

- Percentage Correct of Exercise (*percentage correct exercise*): This numeric variable represents the correctness ratio of each type of problem and we use it to operationalize the difficulty of an exercise, hypothesizing that easier exercises are used to obtain more badges.
- Video Duration (*video duration*): This variable represents the time length of a video.

We found a positive and moderate correlation between the badge count and *percentage correct exercise* ($0.45, p < 0.00$) which indicates that easier exercises trigger more badges. This finding makes sense as it is more accessible for students to solve those exercises correctly. We also found a correlation between badge count and *video duration* ($0.5, p < 0.00$) which might mean that longer videos trigger more badges, and the rationale behind this result can be that students need to spend more time on the video, thus it is more probable that they earn some of the 'Video Time' badges. These correlations are presented for all the exercises and videos from the three courses, but are also maintained when performed within the data from each course separately.

We make a more in-depth and graphical analysis for the case of badges triggered by exercises in Figure 4.7. We explore how the normalized badge count is affected by *percentage correct exercise* and *location* variables, separating also by course. We express the badge count as normalized z-scores ($z = \frac{x-\mu}{\sigma}$), otherwise the difference between the number of students in each course would complicate comparing the amount of badges triggered. The plot in the top of the figure shows a visualization where each point is characterized in the y-axis by the normalized badge count and in the x-axis by the *percentage correct exercise*. Additionally we draw the regression line with the standard error (gray shadow) which shows in all courses a positive tendency suggesting that, the higher is *percentage correct exercise* variable, the more badges are triggered by the exercise. The bottom visualization of Figure 4.7 shows a line plot representing the *location* of the exercise on the x-axis, being the left side the first exercise and the right side the last one in the course structure. The visualization shows that those exercises located at the beginning of the course triggered more badges than those at the end, except for a peak in the middle-end of the chemistry course. The exercises that caused the peak in the chemistry course are 'Le Chatelier Principle' and 'Lewis Structure'. These exercises have been solved correctly many more times (around twice) than others located nearby within the course structure. Although we cannot establish the causes with certainty, we can hypothesize that maybe the difficulty was easier and students used these exercises to earn more badges or that the topic was appealing for students increasing the amount of activity. These results are aligned with the negative and significant correlation between the *location* of the exercise and the badge count ($-0.46, p < 0.00$), which seems to indicate that as the location of the item within the course structure advances, it will trigger less badges.

4.4.3. Distribution of the Badge Indicators

In this subsection we analyze the distribution of the badge metrics, where we have included students who interacted at least 60 minutes for more realistic results. Figure 4.8 shows on the top left the density distribution of both intentionality indicators for all students. The 1st quartile of *intentionality topic badges* and *intentionality repetitive badges* indicators is 0. This means that there are a big percentage of people who did not show much interest on earning badges, especially *topic badges*. On the other hand, the median for *intentionality repetitive badges* is 48.44%, which seems to indicate that students show more interest in *repetitive badges*. The mean value of *intentionality repetitive badges* indicates that the average user earns 39.52% of *repetitive badges* intentionally, which we think is a high percentage. We must state too, that probably many of the students near the 0% of both indicators, probably did not interact a lot with the platform, as a result, they might end up classified as having no interest for badges. We can see that for *intentionality topic badges*, a higher amount of the population is accumulated in the low values of *intentionality topic badges* distribution, and there are not many users between 50-100% of *intentionality topic badges* interval. Nevertheless, we can see a small peak at 100%, who are the cohort of students showing a lot of interest. In the case of *intentionality repetitive badges*, there is a valley between 10-30% who are probably those students who interacted with the platform, but did not show interest for *repetitive badges*. Also we can see a moderate peak between 50-75%, whom are students showing a moderate interest in *repetitive badges* and 75-100% whom are those showing a high interest. We should note out that, to be able to acquire in *intentionality repetitive*

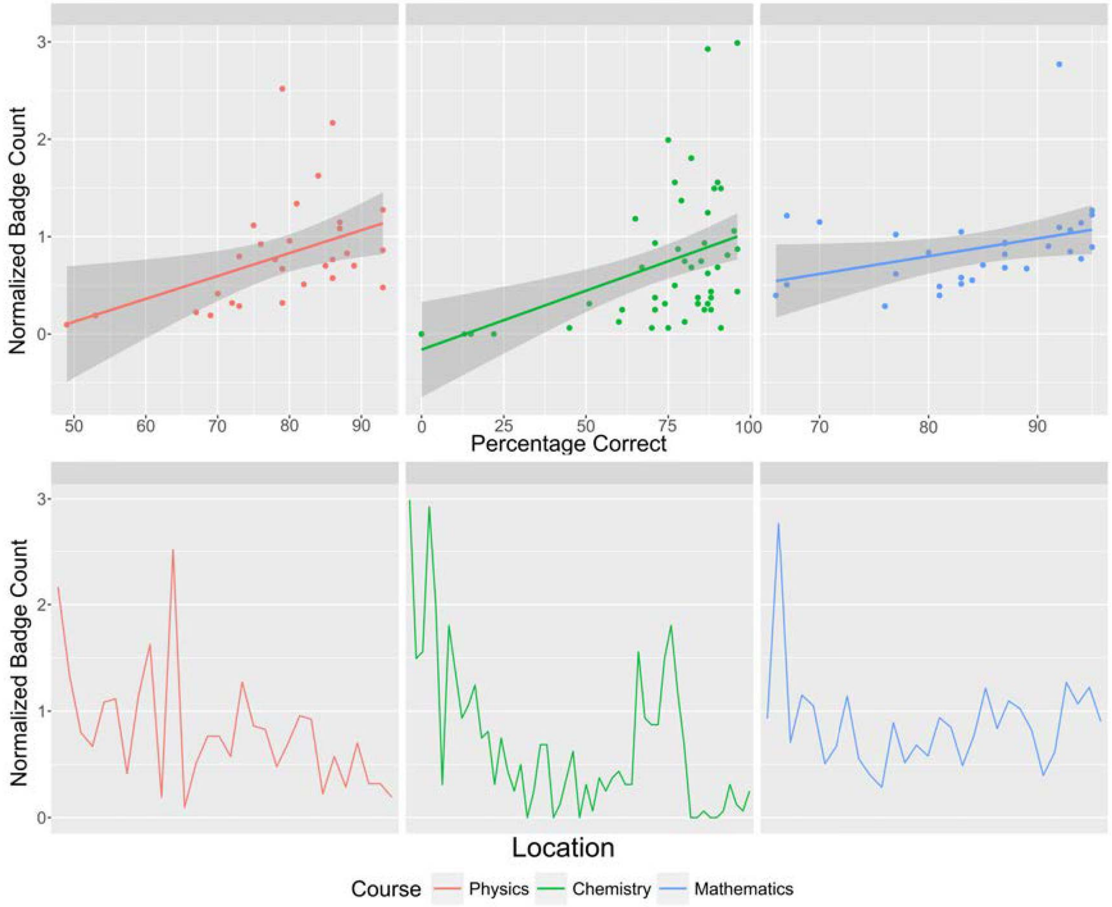


Figure 4.7: Influence of factors (percentage correct and location within course structure) in amount of badges triggered by exercises.

badges such high values, you must really put a lot of interest in these badges. Overall, it looks like students felt more motivated towards *repetitive badges* rather than *topic badges*, but we should also mention that as these badges are easier to earn, and that might be why students might feel more motivated towards them.

The top right of Figure 4.8 shows the density distribution for *concentration badges* indicator. Most of the students have low values ($\leq$ 15%) of *concentration badges* indicator, however we can find also some students with high values ($\geq$ 50%) that show that their actions were really targeting the acquisition of badges. Finally the bottom of Figure 4.8 shows an histogram of *time efficiency badges* indicator. Although most of students are placed in less than 10 badges per hour, some of them have higher values that seem to indicate that they earn many badges in short intervals of time.

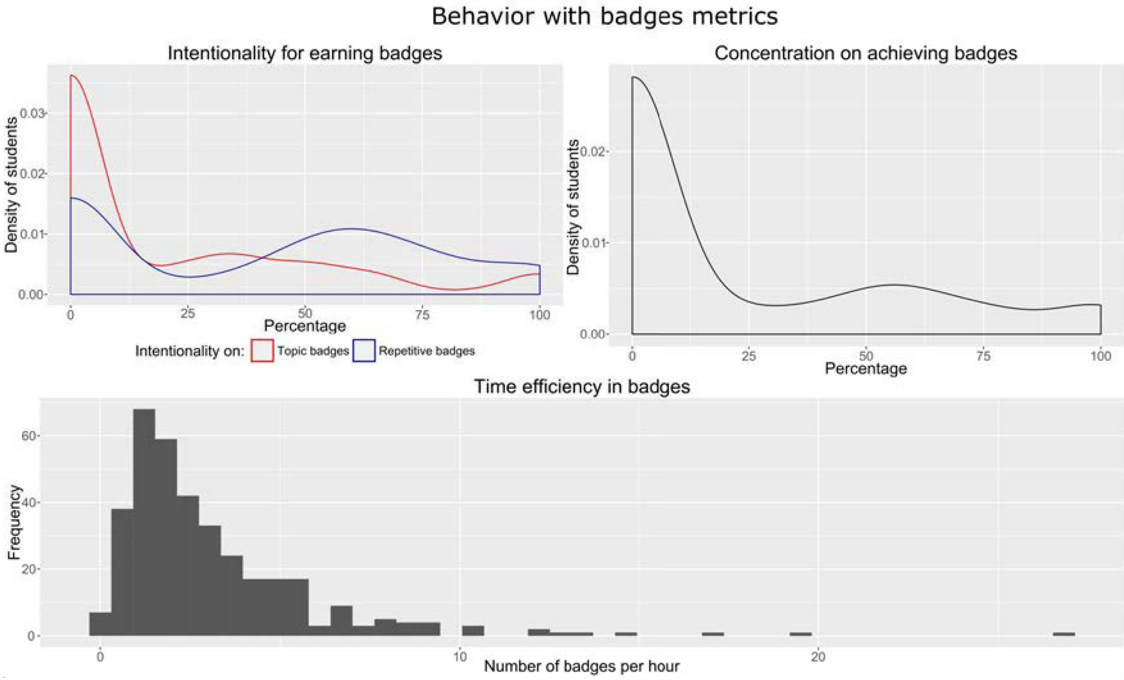


Figure 4.8: Distribution of the different badge metrics.

4.4.4. Relationship of Badge Indicators with Other Selected Indicators

Finally, we analyze some correlations between badge metrics and others. Table 4.6 shows the correlation results where the first set of indicators are composed by *exercises accessed*, *videos accessed*, *exercise abandonment*, *video abandonment*, *total time* and *optional activities* which are related to the use of the platform. As we can see, most of these indicators have been found statistically significant with badge metrics, which makes sense and we can hypothesize that the more use of the platform students do, the more badges they will earn. This is due to the fact that many of the badge requirements are related to the amount of activity of students with the contents (see Table 3.1). The *exercise abandonment* and *video abandonment* indicators have been found negatively and significantly correlated with several badge metrics. Although this correlation is low, we believe that it might indicate that engagement plays an important role in the amount of badges received. The indicators which were less correlated to badge metrics are *videos accessed* and *video abandonment*, which are coherent results taking into account that most of badge metrics does not take into account video activity. Finally, *exercises accessed*, *exercise abandonment*, *total time* and *optional activities* are strongly correlated with most of the badge metrics. We also believe that it makes sense that the use of optional activities is related to interest in badges, as both of them represent non-mandatory activities for the student.

The second set of correlations contains *proficient exercises* and *completed videos* indicators. The correlation with *completed videos* is significant but low, as there are less video badges than those who are earned by solving exercises. The correlation with *proficient exercises* is the highest of all indicators which is probably related to the big amount of exercise badges that can be earned repeatedly and also due to *topic badges*. Two of these correlations are especially significant. First with *intentionality topic badges* ($0.737, p < 0.000$), which makes sense since students who have more *proficient exercises* are more likely to earn more topic badges as well. Second with *time efficiency badges* ($0.625, p < 0.000$), which also makes sense as the more *proficient exercises* as student has mastered, the more badges the student will earn, hence *time efficiency badges* will also be higher. The third set contains all the exercise solving behavior indicators and only *follow recommendations* was found to be slightly correlated with *intentionality topic badges* (0.169) and *concentration badges* (0.202).

The last section of the table, which is separated by a double line, presents the correlations among the badge metrics with themselves. All the correlations have resulted to be statistically significant probably due to the fact that when one student shows interest towards earning badges it will be reflected in several of these indicators. We should point out the correlation between *intentionality topic badges* and *concentration badges* ($0.859, p < 0.000$), which is the highest of all the correlation analysis. One hypothesis is that this correlation is very high as the students who are concentrated earning *topic badges* are also probably maximizing and earning as many *topic badges* as possible.

Table 4.6: Bivariate Pearson correlation of the badge metrics with the rest of indicators.

Bivariate Pearson correlation[1]	**ITB**	**IRB**	**CAB**	**TEB**
Exercises Accessed	0.456**	0.464**	0.361**	0.438**
Videos Accessed	0.305**	0.322**	0.228**	0.225**
Exercise Abandon	-0.456**	-0.327**	-0.399**	-0.352**
Video Abandon	-0.177**	-0.125**	-0.168**	-0.049
Total Time	0.51**	0.409**	0.372**	0.338**
Optional Activities	0.489**	0.358**	0.345**	0.393**
Proficient Exercises	0.737**	0.511**	0.563**	0.629**
Completed Videos	0.333**	0.293**	0.259**	0.219**
Follow Recommendations	0.169**	0.041	0.202**	0.024
Forgetful User	0.01	-0.053	0.01	-0.047
Video Avoidance	0.004	-0.047	-0.012	0.019
Hint Avoidance	-0.027	-0.025	-0.063	0.021
Unreflective User	0.032	0.024	0.027	0.027
Hint Abuse	-0.06	-0.031	-0.015	-0.065
Intentionality on Topic Badges (ITB)	1	0.445**	0.859**	0.567**
Intentionality on Repetitive Badges (IRB)	0.445**	1	0.417**	0.488**
Concentration of Achieving Badges (CAB)	0.859**	0.417**	1	0.456**
Time Efficiency in Badges (TEB)	0.567**	0.488**	0.456**	1

[1] N = 291 students from Case Study 3.2.1.1.
** Correlation is significant at the 0.01 level (2-tailed).

4.5. Online Academic Dishonesty and Collaboration

This subsection explores the results of the two proposed algorithms to detect online academic dishonesty and collaboration. Subsection 4.5.1 presents the results after applying the *close submitter* algorithm and Subsection 4.5.2 describes the findings related to CAMEO algorithm.

4.5.1. Close Submitters

In this subsection, we apply the algorithm for the detection of *close submitters* as described in Subsection 3.3.7.1 to the dataset of both MOOCs presented in Case Study 3.2.4 and explore the basic results.

4.5.1.1. Overview and Distance Distribution

We computed the two $\underline{SP}$ submission matrices, $\underline{SP}_{mus}$ for music MOOC with a shape of 89896 rows x 5 columns and $\underline{SP}_{phi}$ for philosophy MOOC with a shape of 53531 rows x 7 columns. We use these two matrices as the entry to the algorithm and compute two distance matrices for each course using the distance metrics specified in Equation 3.2 and 3.3. Hence, we obtain $\underline{D}_{mus}^{MAD}$ and $\underline{D}_{mus}^{MSD}$ with a shape of 5159 rows x 5159 columns for music and, $\underline{D}_{phi}^{MAD}$ and $\underline{D}_{phi}^{MSD}$ with a shape of 2359 rows x 2359 columns for philosophy. We note again that we apply the criteria as specified in Subsection 3.3.7.1, therefore 5159 and 2359 represent for music and philosophy respectively the number of accounts that submitted all assignments in the course. Next, we compute the vectors of distances that contain every possible different combination of distances obtaining $\vec{d}_{mus}^{MAD}$ and $\vec{d}_{mus}^{MSD}$ with a length of 13.305.061 triplets for music and, $\vec{d}_{phi}^{MAD}$ and $\vec{d}_{phi}^{MSD}$ with a length of 2.781.261 triplets for philosophy. Finally, Figure 4.9 shows the histogram distribution of the distances of each one of these vectors.

The two top visualization of Figure 4.9 show the distribution of the MAD distances for both courses and the distributions look skew-normal. We believe that the left-skewness distribution is due to the effect of the deadlines i.e., students are more likely to submit close together since they are more active when the deadline is closer. The two visualizations in the bottom of Figure 4.9 represent the MSD distances and they look like decreasing exponential distributions due to the effect of the squared metric. We can see that the variance of the distances in the philosophy MOOC is higher, probably due to the fact that they had 7 graded quizzes instead of the 5 graded quizzes of music MOOC, thus increasing the variance of the distance distribution.

4.5.1.2. Detection of Close Submitters

For the detection of *close submitters*, we need to establish a criteria based on the vector of distances. We establish a threshold, and we consider distances below that threshold to be values abnormally similar. Then, we categorize those accounts as *close submitters* which are carrying

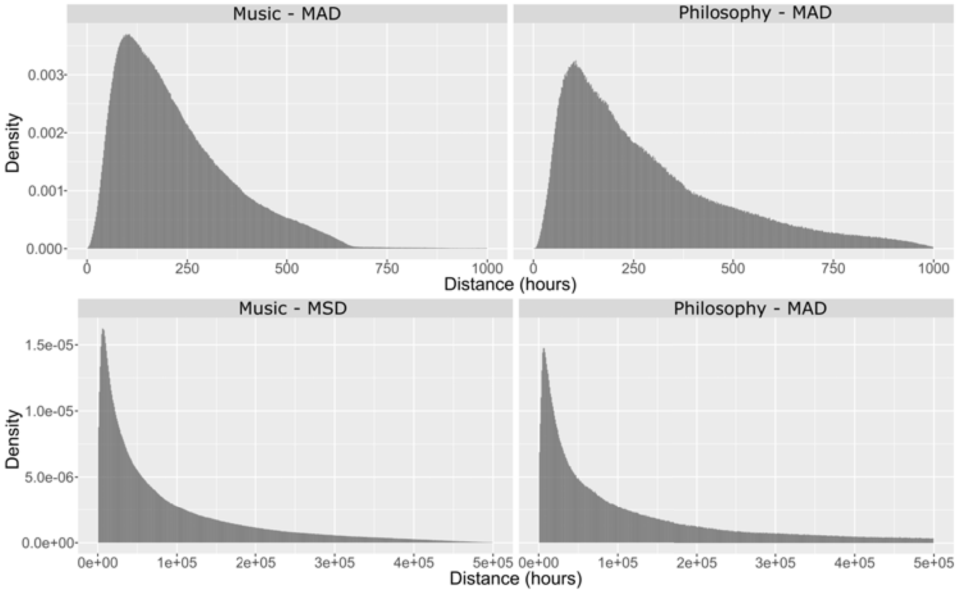


Figure 4.9: Density histogram distribution of the distance vectors $\vec{d}_{mus}^{MAD}$ (upper left plot), $\vec{d}_{mus}^{MSD}$ (bottom left plot), $\vec{d}_{phi}^{MAD}$ (upper right plot) and $\vec{d}_{phi}^{MSD}$ (bottom right plot).

out some collaboration or academic dishonest strategy. We follow the next steps in order establish that threshold:

1. MAD metric is easy to interpret since it is expressed in the same time units whereas MSD is more problematic to interpret because it is expressed in quadratic time units. Therefore, we apply a MAD threshold by common sense which is 0.5 hours. We consider this threshold to be quite strict as it implies that on average, the distance in time of the submissions of $student_i$ and $student_j$ is below 30 minutes to be considered as *close submitters*.

2. We compute the quantile of the distribution represented by this threshold value, which is 4.81e^{-06} quantile for music and 5.75e^{-06} quantile for philosophy.

3. We establish other thresholds based on the value of this previous one that we have obtained by common sense. We choose as quantiles, 6e^{-06}, 1e^{-05} and 5e^{-05} and we can apply the three different quantiles to $\vec{d}_{mus}^{MSD}$ and $\vec{d}_{phil}^{MSD}$ distributions. The rationale to use now MSD distance distributions is because we want to penalize more heavily large distances.

4. Finally, we apply the three quantiles and compare the number of triplets detected by each one of them. We also explore how the MAD and MSD threshold values change with each quantile. These results are represented in Table 4.7.

The rationale behind this approach is to make detection more independent from the course. For generalization and reproducibility purposes, we want to provide a criteria that can be applied to other courses even when there might be different amount of exercises in each quiz and the complexity of those exercises might change as well. Therefore, the use a quantile for the distance distribution instead of a fixed value of MAD can help. If we inspect the results of Table 4.7, with the most restrictive quantile ($6e^{-06}$) there are 78 and 17 triplets below that quantile for music and philosophy respectively. We can also see that both the MAD (0.61 h for music and 0.57 h for philosophy) and MSD (0.51 h^2 for both music and philosophy) remain very similar and that is a good sign since it might indicate that we are detecting a similar population. The results of the other two quantile values ($1e^{-05}$ and $5e^{-05}$) show values of MAD too high and very different one course from the other.

Table 4.7: Number of triplets, MAD and MSD values after applying different quantiles as threshold of the distance distribution for music and philosophy courses.

Quantile	6e-06			1e-05			5e-05		
	#Pairs	MAD tresh	MSD tresh	#Pairs	MAD tresh	MSD tresh	#Pairs	MAD tresh	MSD tresh
Music	78	0.61 h	0.51 h^2	132	0.9 h	1.15 h^2	664	2.9 h	10.94 h^2
Philosophy	17	0.57 h	0.51 h^2	28	1.25 h	1.98 h^2	140	4.98 h	38.13 h^2

Choosing a higher threshold might reveal more true positives but also increasing the amount of false positives, hence we prefer to keep a high precision with a safe threshold rather that increase recall and decrease precision. Therefore we decide to stick to $6e^{-06}$ since this allow us to be quite secure about our precision and then we can better characterize the group of *close submitters* in Subsection 4.5.1.3. Additionally we represent a histogram density distribution of the MAD distance of those triplets below the threshold (thus categorize as *close submitters*) in Figure 4.10. We can see that although the MAD threshold for music is 36.6 minutes and for philosophy is 34.2 minutes, there are many triplets with much lower distances. For example, 12 out of the 17 triplets (70%) in philosophy and 23 out of the 78 triplets (30%) in music are below a MAD distance of only 10 minutes. Such low MAD distance values seem to indicate that these results are far from just mere chance.

Figure 4.10: Density histogram distribution of the MAD distance of the triplets below the threshold (close submitters) separated by course.

Finally, we can apply basic graph theory to detect communities among students who are detected as *close submitters*. Each one of the different accounts is a node of the graph, and from each one of the *close submitters* triplets ($[student_i, student_j, dist(i,j)]$), $student_i$ and $student_j$ are the unordered pair that link together in the graph the two accounts (nodes). Applying this criteria, we detect the couples and communities of Figure 4.11. The number and size of the communities that we find is as follows:

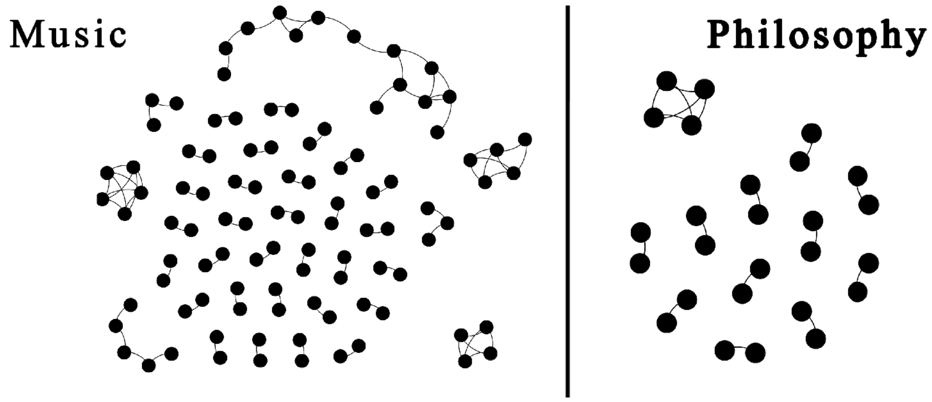


Figure 4.11: Couples and communities detected from the population of accounts detected as close submitters after applying graph theory.

- Music: 30 couples, two 3-communities, one 4-community, three 5-community, one 14-community, therefore 99 different students
- Philosophy: 11 couples, one 4-community, therefore 26 different students.

As we can see, most of the communities detected are only couples of accounts. However, we can see several bigger communities where all the nodes are connected e.g., the community of 4 accounts in philosophy. Due to the restrictive threshold that we have applied we might miss the detection of additional nodes as well as links betweens nodes e.g., when using the higher quantile $1e^{-05}$ the community of 14 accounts that we can see in Figure 4.11 would become a much bigger community of 34 accounts after some of the smaller communities merge it. Therefore, we cannot establish final conclusions about the shape of most communities.

4.5.1.3. Comparison of Close Submitters and Rest of Accounts

This section focuses on analyzing differences between the population of *close submitters* detected in Subsection 4.5.1.2 and the rest of accounts of the course. Table 4.8 shows a cross-tabulation of the course, if the account was detected as *close submitter* and if the account earned a certificate. Note out again that we only keep those accounts who submitted all graded quizzes in each course.

We want to compare the population of *close submitters* with the rest of accounts and we keep for this comparison only students who got a certificate, since we expect them to have made a

Table 4.8: Cross tabulation of the variables certificate, close submitter and course. Only accounts that submitted all quizzes are included.

		Close submitters		Course
		False	True	
Certificate	False	794	21	Music
	True	4262	78	
	False	105	6	Philosophy
	True	2228	20	

similar effort on average to acquire enough score to get a certificate. As specified in Table 4.8, there is a total amount of 4340 (78 *close submitters*) accounts for music and 2248 (20 *close submitters*) for philosophy. The next boxplot visualization in Figure 4.12 shows this comparison for the indicators *exercise effectiveness*, *number submissions*, *number active days*, *videos accessed* and *number threads viewed* between the two populations for each course separately.

Figure 4.12: Boxplot visualization comparing the student features of the close submitters and the rest of accounts separate by course.

First, Figure 4.12 allows to see that both populations have a similar grade distribution, which makes sense because they need to pass a threshold in order to get the certificate. For the rest of the indicators we can see a clear difference between the two populations where the median for *close submitters* is further below than the rest of the accounts. We can check that these differences between populations are statistically significant by applying an unpaired t-test for each one of the variables per course, as we can see the next Table 4.9. Additionally to the independent t-

Table 4.9: Independent unpaired *t*-tests for the different variables of the study comparing close submitter and the rest of accounts.

Variable	Unpaired t-test	
	Music	Philosophy
number submissions	t = 10.14, p = 3e-16	t = 6.33, p = 3e-06
number active days	t = 6.29, p = 1e-08	t = 4.88, p = 9e-05
videos accessed	t = 7.73, p = 3e-11	t = 3.84, p = 0.001
number threads viewed	t = 5.74, p = 1e-07	t = 15.45, p = 2e-16

tests, we perform a MANOVA which allows to test for difference in the means with two or more vectors of means at the same time. The results of the MANOVA show that there is a statistically significant difference in means in both music ($F = 55.74, p = 2e-16$) and philosophy ($F = 15.6, p = 1e-12$) testing all variables at the same time. Therefore, accounts in the *close submitter* population are able to earn a certificate with a lower number of submissions, being active less days, watching few videos and without having much forum activity. These findings confirm that we are indeed identifying a different population of accounts as the distribution of their variables clearly shows.

4.5.2. CAMEO

The exploratory analysis of CAMEO is divided in several subsections with different purposes. The initial Subsection 4.5.2.1 presents an overview of the amount of CAMEO found in the course, Subsection 4.5.2.2 compares CAMEO accounts with the rest of accounts. Subsection 4.5.2.3 analyzes the evolution of CAMEO over the course timeline. Subsection 4.5.2.2 deepens into the different profiles of CAMEO accounts that we have found. Finally, Subsection 4.5.2.5 studies which factors are related to CAMEO. This section is based on the the data from the Case Study 3.2.3.

4.5.2.1. Overview of the Amount of CAMEO

After applying the algorithm for the detection of CAMEO described in Subsection 3.3.7.2, we analyze first the total amount of CAMEO accounts including both certificate and non-certificate earners in the course, we can see these results in Table 4.10. It shows that our algorithm detected 65 master accounts, 12.9% of the certificate earners in the course. These accounts operated 78 harvesters. It is important to note out that some master accounts will use several harvesting accounts to increase their number of attempts available. These masters harvested 17350 correct answers, 4.3% of all the correct answers submitted by certificate earners (including non-CAMEO

Table 4.10: Amount of CAMEO by certificate and non-certificate earners

	#Master accounts	**#Harvester accounts**	**#Harvested answers**
Certificate earners	65 (12.9%)	78	17350 (4.3%)
Non-certificate earners	84 (7.7%)	74	12438 (5.1%)

users). The table also shows the results for the non-certificate earners. The table shows that 84 (7.7%) of the non-certificate earners were master accounts, and that these accounts operated 74 harvesting accounts and copied 12438 answers. This is possible when several masters are getting answers from the same harvester.

Next, we look on how the CAMEO events are distributed between the master accounts. This is shown in Figure 4.13. The figure shows the percentage of the certificate earners (x-axis) who harvested at least y% of their correct answers. The point (3.7, 50) means that 3.7% of the certificate earners used CAMEO to obtain more than 50% of their correct answers. As can be seen in the graph, the CAMEO events are distributed unevenly between the accounts.

Figure 4.13: Amount of CAMEO among students.

4.5.2.2. Feature Comparison: CAMEO Accounts and Rest of Accounts

In this subsection we compare CAMEO accounts with the rest of accounts in terms of two variables, *performance first attempt* and *average time correct answer*. Our findings suggest that masters have high success rate with *performance first attempt* values and are able to solve questions correctly very fast leading to high *average time correct answer* values. For this analysis we focus only on certificate earners since they answered a similar amount of questions. Figure 4.14a shows in the x-axis the *performance first attempt* indicator, and the y-axis shows the *average time* *correct answer* indicator. Master accounts are marked in red, with the size of the circle proportional to the amount of CAMEO. In general, the trend is that the bigger the red circle (more CAMEO), the higher the success rate and the faster the submission. Also, the top performers of the course in terms of low *average time correct answer* and high *performance first attempt* are master accounts. Additionally, we also found that it was statistically significant that master accounts have higher *performance first attempt* ($r = 1.86, p = 0.03$) and *average time correct answer* is smaller ($r = 3.98, p = 4.6e^{-06}$).

Additionally, we want to compare the distribution of *performance first attempt* among masters, harvesters, and the rest of the students, and this probability density comparison is presented in Figure 4.14b. In terms of average values, masters have a *performance first attempt* of 78.9%, harvesters have a *performance first attempt* of 39.7%, and the rest of accounts have a *performance first attempt* of 61.6%. The results of an ANOVA test ($F = 103.6, p < 2e^{-16}$) confirm that *performance first attempt* of masters is greater than the rest of students which is greater than harvesters (masters > rest of students > harvesters). We further discuss the implications of these findings later in the discussion of Section 7.1.

(a) (b)

Figure 4.14: Subplot (a) shows average time for correct attempt vs. performance at first attempt for certificate earners; size of red dot indicates amount of CAMEO. Subplot (b) shows the density distribution of the performance at first attempt for the master, harvester and rest of accounts separately.

4.5.2.3. Distribution of CAMEO Over Course Timeline

This subsection focuses on the analysis of how the CAMEO submissions performed by certificate earners evolves over the course timeline. We partitioned the 13 chapters of our course into 10 sections. Some chapters were joint together because they have common quiz and homework (the combined chapters are 1 and 2, 4 and 5, and 9 and 10). Then for each section, we compute the amount of CAMEO performed by certificate earners, by mapping the problems to each section. Most of the students usually perform the activities of the chapters in a linear way, therefore this binning into chapters help us see the evolution over time, and we can refer to it as a 'temporal' analysis. It will also reflect the progress of students in terms of the graded cumulated after each chapter, which will help shed light on students' behavior. The findings of this temporal analysis are presented in Figure 4.15. The figure shows the percentage of the questions in each section that were attempted, correct, and harvested (for all the lines, the 100% baseline is the total number of questions in the section). Additionally, the figure also shows the percentage of students that achieved enough grade to earn a certificate of accomplishment in each chapter (i.e., achieved more than 60% of total points in the course). The graphs hows that about 85% of the certificatees passed the certification point in section 7 (chapters 9+10). We can clearly see that from that chapter the percentage of questions attempted, correct and harvested drop importantly. This finding strongly supports the hypothesis that for most students using CAMEO, the main motivation is obtaining enough credit for a certificate.

Figure 4.15: Evolution of the amount of attempted, correct and harvested questions, as well as percentage of certificates earned, over the course timeline.

4.5.2.4. Comparison Among CAMEO Student Profiles

In this section we compare and find differences between different CAMEO profiles. First, we present a comparison between certificate earners and non-certificate earners. Analogously to the figure in previous subsection, Figure 4.16 represents the percentage of correct and harvested answers over the course timeline. Figure 4.16 is divided into two parts. The first one for certificate earners (bottom), and the second one for non-certificate earners (above).

The figure clearly shows that non-certificate earners harvest a higher fraction of their correct answers than certificate earners, and that this fraction increases to 100% before the harvesting non-certificate earners drop-out from the course (the decreasing curve of success rate reaches zero), which is why they did not earn a certificate. Altogether, non-certificate earners harvested 68% of their correct answers, whereas certificate earners harvested 44% of their correct submissions, and a *t*-test confirms that it is statistically significant that non-certificate earners CAMEO copy a higher percentage of their correct submissions ($p < 0.0001$). We believe that the most reasonable explanation for this pattern is that the non-certificate earners who harvested are students who started the course in the purpose of 'harvesting for the certificate', but dropped-out, maybe because they found that our course contains a lot of questions, many of them randomized (thus, quite problematic to CAMEO).

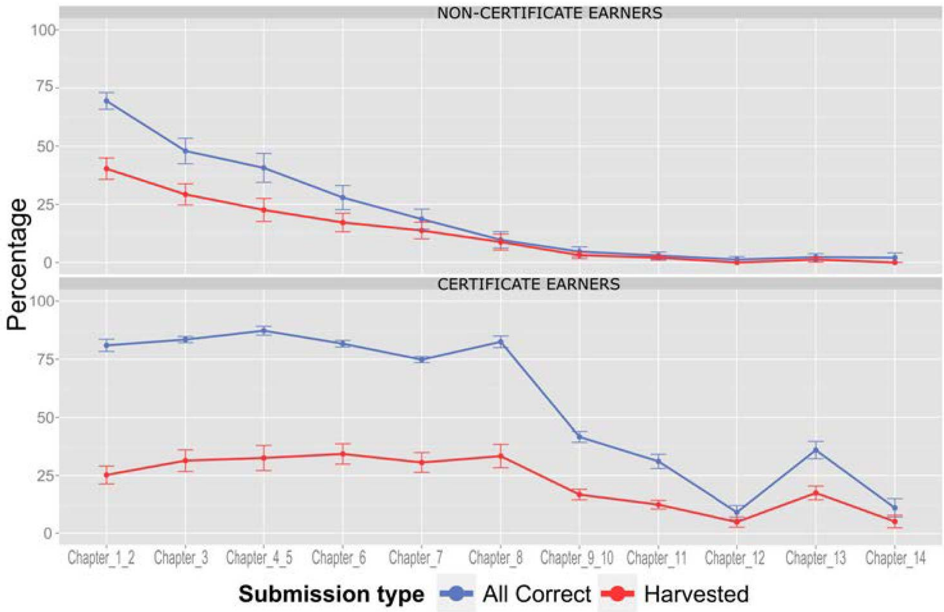


Figure 4.16: Comparison of master accounts certificate earners vs. non-certificate earners in terms of correct and harvested questions, over the course timeline.

Following the aggregated analysis, we look on individual profiles, and classify the students in the group of the certificate and the non-certificate earners into four different profiles that can be archetypical. This is illustrated in Figure 4.17. Per each student we see again the percentage of correct and harvested submissions in each chapter:

1. ***Student A (certificate earner). High and stable***: This profile describes a user who uses CAMEO extensively from the beginning of the course at least till the account qualifies for a certificate. This is the more severe use of CAMEO.

2. ***Student B (certificate earner). Mild and stable***: This profile describes a user who is using CAMEO on a small fraction of the questions that he/she submits, but does so in a relatively stable manner from the beginning of the course. This student is probably using CAMEO as a help seeking strategy.

3. ***Student C (certificate earner). Start low and increase before certificate***: Users following this pattern do not, or rarely use, CAMEO at the beginning of the course, but this use increases significantly towards the last chapters. These might be users who use CAMEO

when they get stressed as the course progresses, or that what we observe here is the 'birth' of new CAMEO users during our course.

4. ***Student D (non-certificate earner). Start high and dropout***: This represents a typical pattern for non-certificate earners. Students following this pattern apply CAMEO in most of their correct answers, but at some point of the course the dropout and do not accomplish enough points to get a certificate. We consider that this type of students got tired of doing CAMEO and decided to dropout the course.

Figure 4.17: Selected students that represent archetypal CAMEO behavioral profiles.

4.5.2.5. Factors Associated with CAMEO

Several factors have been analyzed to see their relationship with CAMEO:

1. ***High-stake questions***. The course contains three different categories of assignments which are quizzes, homework and checkpoints. Each one of this assignments have a different contribution towards the final grade, quizzes have the highest weight, then homework questions and then checkpoints. Additionally this information is available in the syllabus of the course, therefore students are aware of these details. We compute the fraction of CAMEO submissions with respect the total amount of correct submissions in each category, finding on average a percentage of 7.25%, 5.65% and 5.09% of CAMEO for quizzes, homework and checkpoints respectively. Therefore, we can find more CAMEO on high-stake questions.

2. ***Delayed feedback***. We have several places where questions that are similar in terms of weight are accompanied by different levels of feedback regarding the correct answer. The 'show answer' feature was not available before the deadline for quizzes except for one quiz (due to a mistake in settings). On this specific quiz, harvesting accounted for 6.4% of

correct answers, vs. 3.83% on those without show answer. Similar findings were obtained when looking at the midterm and final exams. On the midterm, with the ‘show answer’ disabled, the average amount of CAMEO was 3.43%, while on the final exam, on which ‘show answer’ was enabled, the average amount of cheating was 6.35%. These results clearly show that limiting the feedback reduces CAMEO. It does not go totally down since students can still use exhaustive search utilizing the correct/wrong feedback that is always provided by the platform.

3. ***Randomized variables***. There are some questions of the course where the question variables are randomized meaning that each account will receive slightly different variables, thus final solution for a sample problem. We found that there is less cheating on randomized questions, both globally when considering the total amount of cheating events, and also after normalizing by the total number of submissions to the problem. For randomized questions, the percentage of cheating is 4.06%, while for non-randomized questions, it is 6.07%. The normalized results are statistically significant ($p < 0.01$), for both the normalized and the absolute comparisons.

Chapter 5

Clustering and Student Profiling

Contents

This chapter focuses on clustering students according to different metrics and exemplifying their behavior with different student profiles. This can help deepen into how students behave with respect different items in VLEs and use this knowledge towards personalizing learning experiences to each student profile. This chapter is divided in three sections. Section 5.1 clusters students based on their use of regular and optional activities. Section 5.2 groups students according to their behavior with badges. Ultimately, Section 5.3 profiles couples and groups of students according to the type of collaboration that they are performing. In all these sections we apply a Two-Step cluster algorithm leaving the number of groups to be determined by the execution of the algorithm automatically, we decided to use this algorithm as we do not know beforehand how many clusters we shall find in our data sample.

5.1. Based on Use of Regular and Optional Activities

We cluster students according to their interaction with regular and optional learning activities using the dataset of Case Study 3.2.1.1. We use as input variables of the algorithm *exercise effectiveness*, *video effectiveness* (for the use of regular learning activities) and *optional activities* (for the use of optional activities). We use the effectiveness metrics as these have been specifically designed for this case study. We apply the Two-Step algorithm obtaining four clusters with a cohesion quality of 0.5 in a quality ranging from -1 to 1, which is a good cohesion value. The cluster with more students has 141 (48.5%) and the one with less only 29 students (10%), the other clusters contain 38 (13.1%) and 83 (28.5%) students. This provides a size ratio between the biggest and smallest of 4.86. The variable importance is 1.0 for *optional activities*, 0.98 for *exercise effectiveness* and 0.77 for *video effectiveness*. Therefore, all of them provide valuable information for the clustering model.

Figure 5.1: Boxplot visualization of the clustering results based on the use of regular and optional activities.

Figure 5.1 shows a boxplot visualization regarding the distribution of each cluster, we can describe them as follows:

- Cluster 1: This is the smallest cluster that includes 29 students (10%) and has as key feature that is the one with the highest use of the *optional activities* indicator showing an average value of 50%. Additionally these students have also high values of *exercise effectiveness* (average 67.5%), *video effectiveness* (average 68.6%) and *total time* (average 697 min). Therefore, they interacted a lot with the platform showing interest both for regular and optional activities.

- Cluster 2: This cluster contains 83 students (28.5%) that have shown a moderate interest in the platform in terms of regular activities as the indicators *exercise effectiveness* and *video effectiveness* have an average value of 27.3% and 35.9% respectively, also the average total time is 270 minutes. However, most of them did not show interest on optional activities since the average value of *optional activities* is 6.4%.

- Cluster 3: This cluster contains 38 students (13.1%) and the indicators *exercise effectiveness* (average 69.7%), *video effectiveness* (average 75%) and *total time* (average 631 min) have similar average values than the students in cluster 1. However, cluster 3 indicators have a lower variance than the ones of cluster 1. Additionally, we can see that the use of optional activities is very low despite they invested a lot of effort in the platform with an average value of *optional activities* of 4.6%. Consequently, the key difference between students from cluster 1 and cluster 3 is the use of optional activities.

- Cluster 4: The biggest cluster involves 141 students (48.5%) that have not made much use of the platform and neither showed interest in the regular nor in the optional activities as all the indicators are quite low.

Additionally, we use a high dimensional data visualization denominated as 'parallel coordinates' (Inselberg & Dimsdale, 1990) to represent the tendency of each one of the students in the different clusters. Parallel coordinates represents consecutive parallel axes in which an n-dimensional point will be represented as a polyline with vertices on the parallel axes. This type of visualization can be used to detect 2D patterns and it is often use for clustering purposes. Figure 5.2 shows the parallel coordinates visualization where each polyline represents a student characterized by its indicators, and the vertical facet and color represent the cluster. This visualization helps see more clearly differences among the cluster populations. For example, we are able to see the higher variance in terms of *exercise effectiveness*, *video effectiveness* and *total time* of cluster 1 when compared with cluster 3, where most of the students follow a more alike distribution.

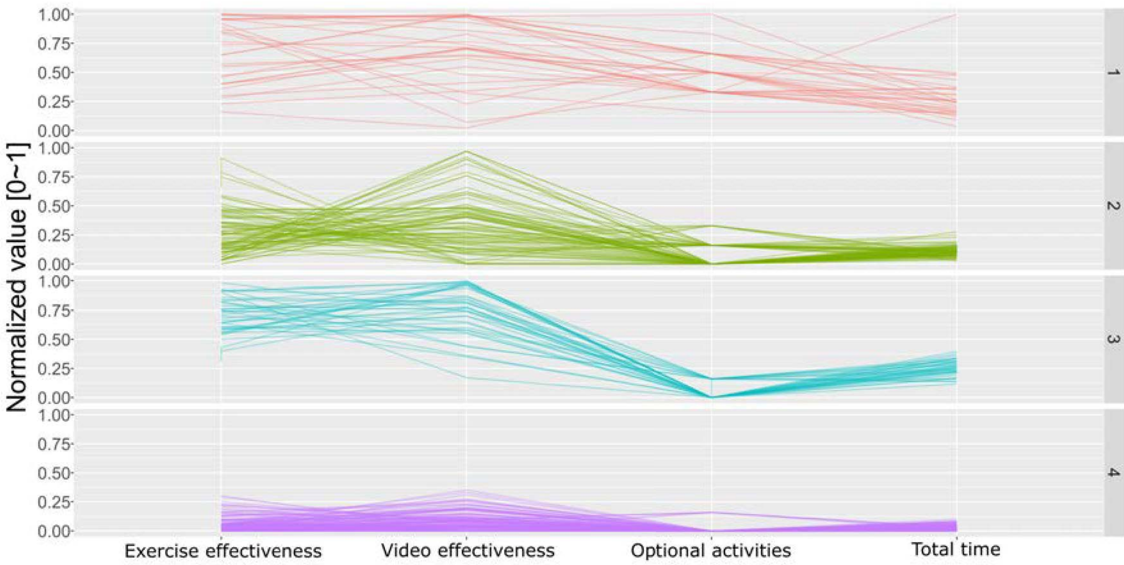


Figure 5.2: Parallel coordinates visualization of the use of regular and optional activities after clustering.

Finally, Figure 5.3 shows a line chart representing the x-axis indicator has the value of the y-axis for the student represented by a specific color; this way we are able to plot some examples of student profiles. Student A shows very high values of *exercise effectiveness*, *video effectiveness* and *optional activities*, hence Student A showed interest in regular and optional activities in the platform. On the contrary, we can see that Student B also showed interest in regular activities with high values of *exercise effectiveness* and *video effectiveness* but never used any optional activity. Additionally, Student C did not use much regular activities with low values of *exercise effectiveness* and *video effectiveness*, but *optional activities* is moderately high. Other different profiles are Student D which shows low interest in *exercise effectiveness* but high interest in *video effectiveness* and *optional activities*, so this is maybe a visual learner exploring the contents, and Student E that which did not use any optional activity of video, but still managed to solve most exercises correctly as the high value of *video effectiveness* shows, hence we can hypothesize that Student E knew most of the contents before accessing the platform. Therefore, we just showed different examples of students' behavior with respect the use of regular and optional activities that can be later on used to customize learning experience. We delve into this discussion in Section 7.2.

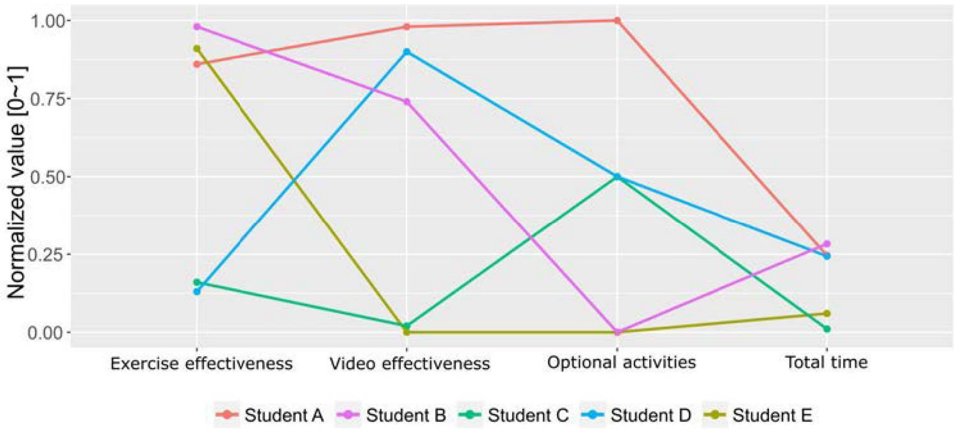


Figure 5.3: Line visualization for selected students that represent different interaction profiles.

5.2. Based on Behavior with Badges

In this section, we cluster students according to their behavior with badges and we use again the dataset of Case Study 3.2.1.1. Therefore, we use as input features for the Two-Step algorithm the badge indicators *intentionality topic badges*, *intentionality repetitive badges*, *concentration badges* and *time efficiency badges*. We also use *proficient exercises*, *completed videos* and *total time* indicators as evaluation fields to support the interpretation of the results from the algorithm output.

The Two-Step algorithm selects three clusters providing a good cluster quality (0.75) in terms of cohesion and separation. The smallest cluster has 70 students (24.1%) whilst the largest has 149 students (51.2%) providing a size ratio of 2.13; the middle-sized cluster has 72 students (24.7%). The predictors' importance for the four continuous variables has been 1.0, 0.94, 0.79 and 0.31 for *intentionality repetitive badges*, *concentration badges*, *intentionality topic badges* and *time efficiency badges* respectively.

Analogously to previous section, Figure 5.4 shows a boxplot with the indicator separated by cluster. The upper plot represents the four badge input metrics that were used to cluster students, and the bottom plot represents the evaluation fields.

Figure 5.4: Boxplot visualization of the indicators separated by cluster. The upper plot shows the badge metrics and the bottom plot the evaluation fields.

We use the information provided by Figure 5.4 to learn what type of students compose each group as we describe next:

- Cluster 1: The first cluster is composed by the 24.7% of students. We can rapidly perceive that students who belong to this cluster are those who have put the greatest effort in the platform in terms of amount of *proficient exercises*, *total time* and *completed videos*. The mean value of *proficient exercises* is 46.88%, for *completed videos* 46.76%, and for *total time* is 489.2 min per user on average, which are all high values. In addition, they have high values in all badge metrics when compared to the rest of the clusters. The average user of this cluster made an important investment on time, as well as progress in exercises and videos, showing also interest in the badge system.

- Cluster 2: The second cluster is composed by the 51.2% of the students and it is quite the contrary of the first one. These students did a small effort using the platform. We can see that on average they invested 125.8 min per user obtaining only 1.72% in *proficient exercises* and 18.93% in *completed videos*. In addition, they did not show interest in any of the badge indicators.

- Cluster 3: The third cluster is less clear than the others two being composed by the 24.1% of the population. We can see that students within this cluster invested a decent amount of time with 314.8 minutes per student. In spite of that, their progress is not so good with only 17.91% of *proficient exercises* which is much lower than cluster 1 and 39.07% in the case of *completed videos*, which is lower than cluster 1, but not that low. The badge metrics show that *intentionality topic badges* and *concentration badges* have very low average values (1.94% and 3.20%) but *intentionality repetitive badges* is even higher than in the in cluster 1. Finally, *time efficiency badges* indicator shows a moderate value. Therefore, this cluster concentrates students who have very low *intentionality topic badges* and *concentration badges*, thus they are not doing an organized effort towards achieving badges. Furthermore, the very high *intentionality repetitive badges* value demonstrates that they are very eager to earn those *repetitive badges*, consequently they are interested in the badge system. These students have invested a moderate amount of time but they have not achieved a great progress, additionally they have shown low *intentionality topic badges* and *concentration badges* but the highest average *intentionality repetitive badges* indicator of all clusters.

The interpretation of these results is further discussed in Section 7.2. Analogously to previous subsection, Figure 5.5 shows a parallel coordinates visualization, where students within clusters 2 and 3 are very alike as there are almost no outliers and they show a very clear distribution of indicators. Cluster 1 is composed of a more diverse source of students but all of them showed interest in the gamification indicators.

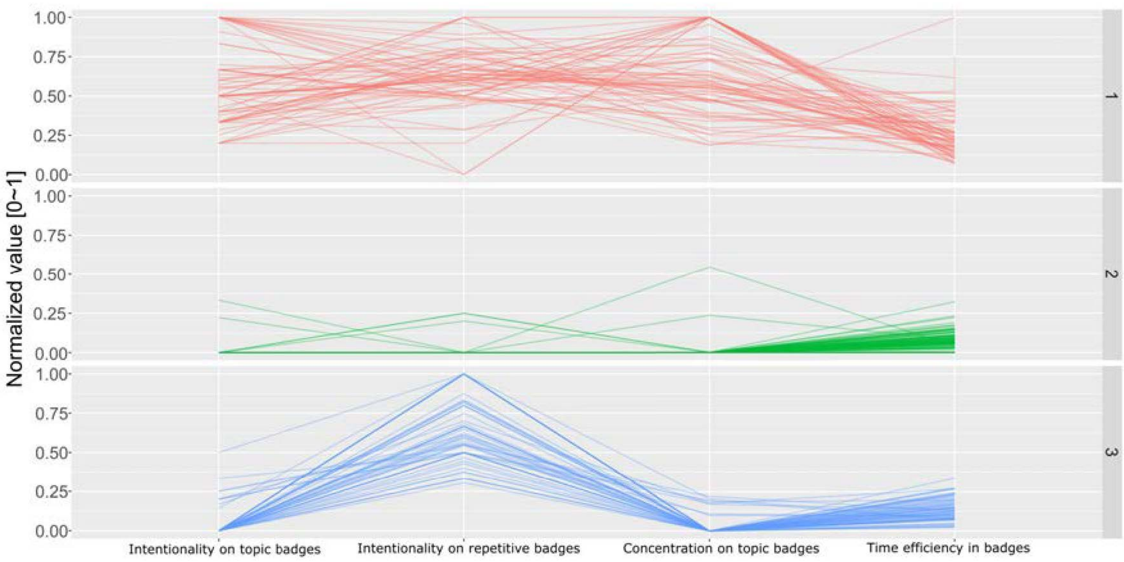


Figure 5.5: Parallel coordinates visualization of the badge metrics of all students separated by cluster.

Finally, with the purpose of student profiling and exemplifying the behavior of students with several specific cases, we provide a radar chart in Figure 5.6 to visualize five students and compare their indicators. Figure 3 shows the radar chart of Students A, B, C, D and E. The values of the indicators represented in the plot are also normalized in order to be in the interval [0-1]. Figure 5.6 represents starting from the bottom and going counterclockwise the four badge indicators: *time efficiency badges* (TEB), *concentration badges* (CAB), *intentionality repetitive badges* (IRB), and *intentionality topic badges* (ITB) in that order. Additionally, we present the indicators *completed videos* (CV), *proficient exercises* (PE) and *total time* (TT) to help us have an idea about the interaction with the platform done by those students.

Student A devoted a great amount of time in the platform, actually we can see that *total time* indicator goes straight to 1, which is the maximum normalized value, meaning that is the student who spent most time in the platform (2458 min). As additional note, that amount of time doubles the one of the second student in terms of *total time* in the experiment. Additionally, we can also see that Student A completed all videos and achieved proficiency in almost all exercises. The student was able to acquire a total amount of 934 badges, and 43 different types of badges. Student A showed an impressive interest on the badge system, as we can see in both *intentionality topic badges* and *intentionality repetitive badges* indicators. However, the student did not do the actions necessarily in a consecutive way to achieve the badges i.e., the student did several activities in parallel to achieve different types of badges. In addition, Student A was not very efficient in achieving badges, this can also explain the big amount of time that the student spent on the platform. Our hypothesis is that Student A probably invested that impressive amount of time because he was very interested in earning badges and he just kept going and going becoming the top earner among all students. However, he did not achieve the badges following a specific order.

Student B devoted less time (1168 min) but still managed to complete a high percentage of the course (100% of videos completed and 60% of exercises), and showed a high *intentionality repetitive badges* and *concentration badges*. However Student B did not have a high *intentionality topic badges*. This might be caused due to the fact that achieving these type of badges is more difficult and requires more effort. Student C spent a similar amount of time (762 min) and made a similar progress (80% of the videos completed and 40% of the exercises) as Student B, but his badge indicators are very low which indicates that Student C did not have a high interest in badges.

Student D devoted a good amount of time (927 min) completing almost all videos and exercises in the course. The results also show that Student D has average badge indicators which pointed out that he was not exceptionally motivated by badges, but made use of them. Finally, Student E invested a low amount of time (249 min), showing low progress where he only completed 16% of the exercises and 20% of the videos. Nonetheless, the results point out that he was very interested in badges as both intention and concentration indicators are almost in the maximum.

Figure 5.6: Radar chart representing five students with different profiles of badge behavior.

5.3. Based on their Online Collaboration

This section focuses in clustering the population of *close submitters* that we reported in Subsection 4.5.1. We also analyze the different associations between accounts from a pedagogical point of view. Subsection 5.3.1 applies the clustering algorithm to the *close submitters* population. Subsection 5.3.2 performs an analysis of the associations between couples of accounts from a pedagogical point of view. Finally, Subsection 5.3.3 presents the example of two bigger communities of accounts with the different pedagogical implications.

5.3.1. Clustering Analysis

We apply the Two-Step Cluster analysis with the input variables *exercise effectiveness*, *number active days*, *videos accessed* and *number submissions*, which are the ones that we found to be significant when comparing *close submitters* with the rest of accounts. We note again that we use the dataset from Case Study 3.2.4. Figure 5.7 shows a boxplot visualization with the clustering results where each input indicator is separated by cluster (on the x-axis) and by course (top subplot for philosophy and bottom subplot for music). The first noticeable detail is that the distribution of the different indicators in the three clusters is very similar for both courses, this a good sign indicating that we might be detecting genuine and real profiles of accounts that could potentially be found in other courses. The variable importance for the clustering is in descending order *videos accessed*, *exercise effectiveness*, *number active days* and *number submissions* for both courses.

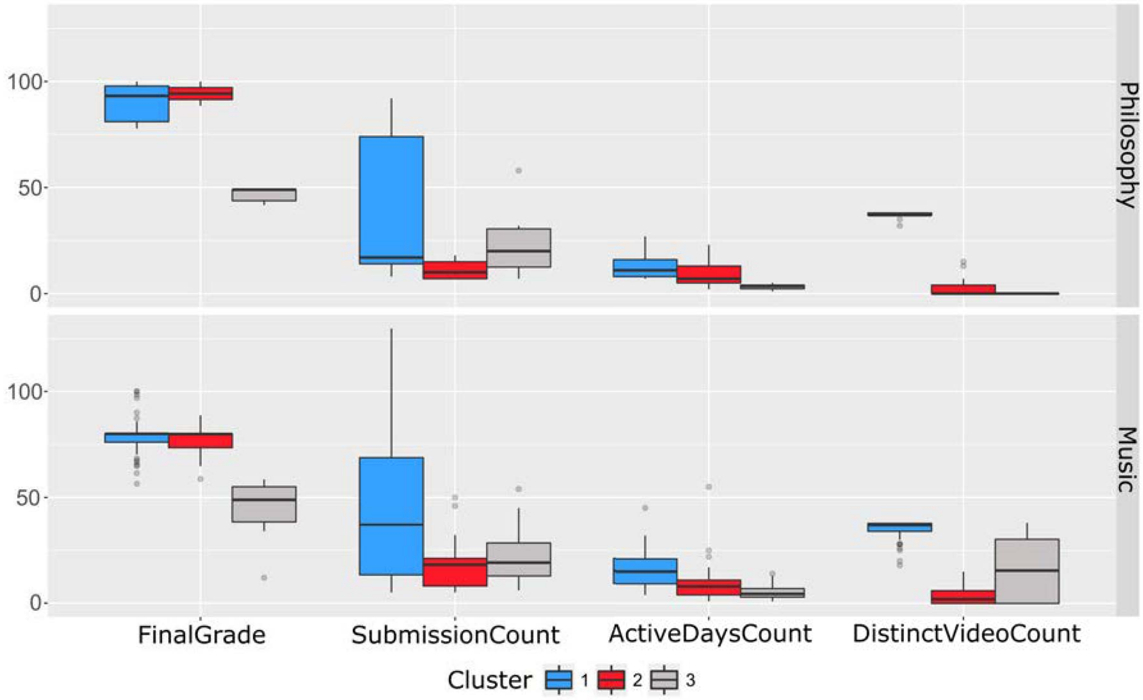


Figure 5.7: Clustering results showing a boxplot visualization of the input variables separated by cluster and course.

The highest influence for clustering lies in the two first variables *videos accessed* and *exercise effectiveness* and this might be associated with the different account profiles that we discuss later. The lowest importance is for *number submissions*, and as we can see in the plot the variance of this variable is the highest of all, thus it is not the real one defining the clusters. Therefore, from these cluster results we can infer the following summary for the majority of accounts belonging to each cluster:

- Cluster 1 (N = 34.6% for philosophy and N = 41.41% for music): The accounts that belong to this cluster have a high *exercise effectiveness* and the highest median in terms of *number active days* and *videos accessed*. Additionally, the variable *number submissions* has a very high variance, thus there are different types of accounts regarding the amount of submissions. Overall, since this cluster has the highest values for the two activity variables (*number active days* and *videos accessed*) and also a high *exercise effectiveness*, these accounts invested an important effort and time in the course by accomplishing a high grade and a certificate.

- Cluster 2 (N = 42.3% for philosophy and N = 42.42% for music): This cluster contains accounts that also have a high *exercise effectiveness*. Nevertheless, there are important differences with cluster 1 regarding the rest of the variables. Most importantly, we can see that in terms of *videos accessed*, accounts in cluster 1 have a very high use, whereas in cluster 2 this is quite the opposite case scenario, where most accounts watched very few videos. Additionally, the value of *number submissions* required and *number active days* are also lower than in cluster 1. As a summary, the accounts in this cluster achieved a high grade and certificate, but they were able to accomplish this achievement, by watching very few videos, being active less days and with fewer submissions than accounts in cluster 1. Therefore, either the students running these accounts had already knowledge regarding the topic of the course, or they might have been performing some illicit actions that facilitated their way into achieving a certificate without much effort.

- Cluster 3 (N = 23.1% for philosophy and N = 16.16% for music): The key difference of accounts in cluster 3 is that the *exercise effectiveness* is much lower with a median value of 50%. This means that most accounts in this cluster did not achieve a certificate of accomplishment. The value of *number active days* is also the lowest one of all clusters, with very few days active. It is also interesting to see that the median value of *number submissions* is the highest for both clusters in philosophy and higher than cluster 2 in the case of music. Therefore, despite these accounts did not achieve a certificate and were active only very few days, they did a lot of submissions, which are quite surprising results. Finally, for *videos accessed* variable, in the case of philosophy the median is 0 and none of those accounts watched any videos, in the case of music the variable has a high variance and the median is above cluster 2. Our hypothesis is that this cluster of accounts represents the harvesting accounts that we reported as part of the CAMEO cheating method in Subsection 4.5.2. These accounts are created for the mere purpose of harvesting correct solutions by using a exhaustive search (taking into account that questions have several attempts available). This hypothesis makes sense since the accounts in cluster 3 did not achieve a certificate, were not very active in the course but still did a lot of attempts to the quizzes showing a low performance.

Finally, Figure 5.8 shows the different couples and bigger communities that are detected by the algorithm and the color of each circle represents the cluster assignment. This way we are able to see the different cluster associations in the communities. Next Subsections 5.3.2 and 5.3.3 analyze respectively the couples and communities detected based on their cluster assignment and associations between accounts. Additionally, more discussion regarding this findings is included in Section 7.2.

5.3.2. Analysis of Couples

This subsection analyses the different associations between the couples of accounts based on their cluster assignment. We provide hypotheses regarding the pedagogical implications that each cluster association might have. Additionally, Table 5.1 presents some specific examples for each cluster association.

- Cluster 1 and cluster 1 (philosophy 3/11, music 5/30): This association represents two

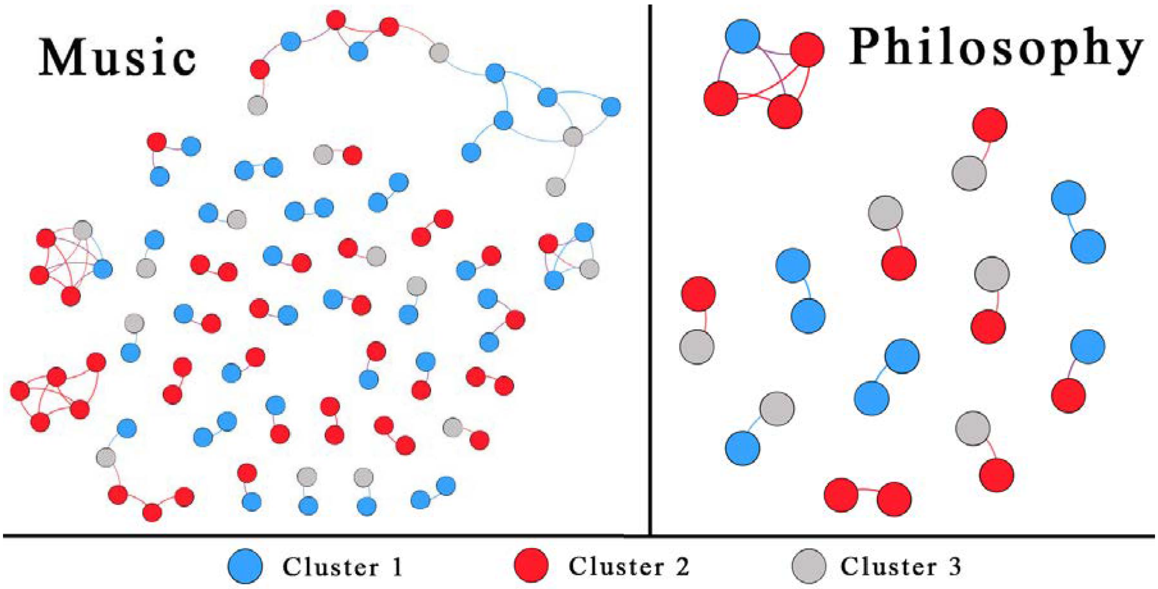


Figure 5.8: Visualization of the couples and bigger communities detected by the algorithm and colored based on their cluster assignment.

students from cluster 1 working together, as we reported in previous subsection, students from cluster 1 did an important effort in the platform achieving a certificate, with high values of *number active days* and *videos accessed*. Therefore, this association might represent two students that are taking the course seriously, and are also collaborating reciprocally one with each other in order to achieve a better grade. Couple 1 in Table 5.1 shows an example of this association. We can see that both account members of the couple have actively participated in the course with high values of *number active days* and *videos accessed*. As we can see, both of the accounts got a perfect score of 100%, which is well aligned with our hypothesis regarding this association representing students trying to improve their grades. The value of MAD is 2.65 minutes, which represents the average time difference between their submissions, thus being very low. Finally, *order* has a middle value in this case, which might indicate that the association is reciprocal.

- Cluster 1 and cluster 2 (philosophy 1/11, music 11/30): This association represents one student of cluster 1 and one of cluster 2. This might represent a genuine association between two real students that might not be reciprocal. The student of cluster 1 invests an effort in the platform, whereas student of cluster 2 does not make an effort but still gets a certificate with the help of student of cluster 1. They might be able to do that since student from cluster 1 might solve a quiz first, and then exchange the solutions with student from cluster 2. Table 5.1 exemplifies this association in Couple 2. As we can see the account from cluster 1 watched all videos (37) and did a high amount of attempts (74) achieving a *exercise effectiveness* of 81%, whereas account from cluster 2 watched only one video and with only 16 submissions was able to achieve a *exercise effectiveness* of 98.6%. Additionally, *order* is

+1, which means that the account of cluster 1 always submitted the assignments before the account of cluster 2. Therefore, we can hypothesize that this association was not reciprocal, and the account of cluster 1 is helping account of cluster 2, probably sharing the solutions of the quizzes after the feedback.

- Cluster 2 and cluster 2 (philosophy 1/11, music 5/30): In this association both accounts belong to cluster 2. Accordingly, this case scenario represents two accounts that did not perform a big effort in the course in terms of videos seen or active days but still were able to acquire a certificate of accomplishment. This association can potentially represent a couple of students collaborating in some academically dishonest way, such as sharing solutions based on the feedback that the platform provides, in order to receive a certificate while still investing little time and effort. Couple 3 in Table 5.1 represents an example for this association. This couple is quite a extreme case since they were active only during 5 days, once per week, and noting out that this couple belongs to philosophy, thus having 5 graded quizzes, they were active only to submit each one of the quizzes, and they did not watch any videos. Even so, they were able to achieve a high score and receive a certificate. MAD value for the couple is also low of only 2.64 minutes.

- Cluster 1 and cluster 3 (philosophy 1/11, music 6/30): This association represents one account from cluster 1 and one from cluster 3. Therefore, we have one account that achieved a certificate investing a great effort and a second one that can potentially be a harvesting account since it did not achieve a certificate, saw few videos and did many submission attempts. This association can potentially represent a CAMEO situation in which the student from cluster 1 is taking the course seriously and learning the contents but it is also using a the harvesting account to ease or ensure that he is able to get a certificate. We note that these two accounts can be presumably run by the same physical student. The couple 4 from Table 5.1 represents the example of this association. The account from cluster 1 was active for 11 days and watched all videos in the course whereas account from cluster 3 was active only 5 days (those days that the account from cluster 1 submitted the quizzes) and did not watch any videos. Additionally, we can see that the amount of *number submissions* of account from cluster 3 doubles the value of the account from cluster 1 despite the lower grade and being less active. Finally, the value of *order* is -1, which means that the account from cluster 3 always submitted first and the value of MAD is very low of only 1.21 minutes; these two results are inline with the expected behavior for this association.

- Cluster 2 and cluster 3 (philosophy 5/11, music 3/30): This association represents one account from cluster 2 that was able to achieve a certificate with a low effort and one from cluster 3 that can potentially be a harvesting account. As the previous case, this one can also presumably be a CAMEO association in which the account of cluster 2 is using the harvesting account to search the correct solutions and obtain a certificate

Table 5.1: Example couple for each one of the cluster associations found.

Couple	**Cluster**	**MAD**	***order***	***exercise effectiveness***	***number submissions***	***number active days***	***videos accessed***
1	1	2.65	+0.14	100	92	26	35
	1			100	12	16	32
2	1	17.07	+1	81	74	7	37
	2			98.6	16	19	1
3	2	2.64	+0.71	97.1	7	5	0
	2			91.4	18	5	1
4	1	1.21	-1	94	28	11	38
	3			49	58	5	0
5	2	1.27	-1	96.4	7	14	0
	3			48.5	32	4	0

without investing much time. Couple 5 in Table 5.1 exemplifies this association. None of the accounts of couple 5 watched any videos. The account from cluster 2 was active 14 days and did 7 submissions to obtain a grade of 96.4% whereas the account of cluster 3 was active only 4 days, did 32 submissions and the grade was only 48.5%. Similarly to what we say in couple 4, MAD value is very low of only 1.27 minutes and *order* is -1 meaning that the account from cluster 3 always submitted first. All the results agree with the typical CAMEO association where the two accounts are run by the same student with a very deliberate cheating behavior.

- Cluster 3 and cluster 3 (philosophy 0/11, music 0/30): We found no associations of two accounts from cluster 3. We believe this makes sense as we generally label accounts from cluster 3 as harvesting accounts, and therefore such association between two of them would not make sense from pedagogical point of view. In spite of we have found none associations for this case scenario, it would be possible to find them due to different reasons such as students dropping out or just due to wrong cluster assignments.

5.3.3. Analysis of Communities

The analysis of communities is more complicated that the previous section regarding the analysis of couples since the size and associations between the different members of the community vary from one case to another. Therefore, we cannot provide a systematic general approach to describe all communities. Instead, we delve into the specific patterns of two community examples that we believe might be representative of the population. The indicators of each account of the selected communities are represented in Table 5.2. We describe the two communities next:

- Community 1: The first community in Table 5.2 belongs to philosophy and is composed by three accounts from cluster 2 and one account from cluster 1. The account of cluster 1 watched all the videos in the course, whereas the rest of accounts fewer number of videos. All accounts have similar values of *exercise effectiveness*, *number active days* and *number submissions*. Additionally, we can support our hypothesis with Figure 5.9 which represents for each quiz, the time difference between the first submission (which has a value of 0) and the rest of the submissions of each account to that quiz. Therefore. we can see that for the Community 1 the submissions of all accounts for each quiz are always placed within a 5 minutes timeframe (except for the submission of Account 1 to Quiz 1). They always met one day each week (either on a Monday or a Tuesday) and solved together the weekly quiz. As a summary, it appears to be a community of students with similar indicators, meeting once per week to solve the quiz, and where it does not look like there is clear evidence about someone submitting always first to pass the solutions or something similar. These results might indicate that it is a genuine community of learning.

- Community 2: The second community represented in Table 5.2 was extracted from music course. There is one account from cluster 1, three from cluster 2 and one from cluster 3. Except for the account from cluster 1 that watched 20 videos, the rest of the accounts watched none or very few of them. Note out that the account from cluster 3, presumably a harvesting account, was active only one day, and still the algorithm detected that all quizzes were submitted close by all accounts, which implies that the five quizzes were solved the same day. To be more specific, the 25 submissions performed by the 5 accounts, were done in a interval of time of only 68 minutes. We can delve into the relationship between accounts by watching the visualization of Figure 5.9. There is a clear trend between Account 4 and Account 3, where the latter is always submitting few seconds after Account 4, which might indicate that both accounts are managed by the same student applying a CAMEO methodology. Additionally, Account 2 follows a similar trend but submitting always around 10 minutes later (except for Quiz 5), which might indicate that Account 2 is receiving the solutions from the student running Account 3 and 4, or that it is even the same persona running the three of them. Accounts 1 and 5 do not appear to have a clear relationship in terms of who answers first but seems to alternate. They might all be exchanging solutions and attempting quizzes at their own pace since all these happened within a 68 minutes timeframe. Therefore, we find here a community 'collaborating' towards achieving certification in a single day with an approach that looks illicit and combines different roles and associations between accounts.

Table 5.2: Description of the extracted indicators for each member of the two selected communities of accounts.

Community	**Cluster**	***exercise effectiveness***	***number submissions***	***number active days***	***videos accssed***
1	2	92.14	14	8	15
	1	91.79	14	8	38
	2	91.55	16	7	7
	2	88.57	14	12	13
2	1	56.47	71	8	20
	2	69.86	21	6	2
	2	79	5	10	0
	3	38.55	19	1	0
	2	80	27	25	4

Figure 5.9: Time difference between the submissions of each one of the members of the community for each quiz.

Chapter 6

Analysis of Learning Outcomes based on Machine Learning Models

Contents

This chapter focuses in the implementation of three machine learning models to predict learning outcomes. Section 6.1 targets the prediction of learning gains of students after interacting with the educational environment. Section 6.2 targets the prediction of certificate accomplishment in MOOC environments. Section 6.3 implements a classification model to assess which submissions have been harvested using CAMEO approach. The three sections of this chapter follow the following structure of contents. First, we describe the machine learning methodology applied to train the algorithm, and also the evaluation metrics and selected variables are presented. Then, we present the training and evaluation of the model while interpreting the results that have been obtained.

6.1. Prediction of Learning Gains

In this first section we approach the prediction of learning gains, which are computed as the difference between the post-test (done after finishing the interaction with the platform) and the pre-test (done before interacting with the platform). The section is divided in Subsection 6.1.1, which describes the machine learning methodology and variables used, and Subsection 6.1.2, which reports and interprets the results of the model. More information is available in our publications (Ruipérez-Valiente, Muñoz-Merino & Delgado Kloos, 2015a,).

6.1.1. ML Method and Training

The prediction of test scores (or similar) has been targeted in different educational studies before. Therefore, as first part of the analysis we perform a review of related work in prediction of learning outcomes in the specific case of learning gains and test scores. This analysis helps to select the initial set of variables that we consider for the study. Other specificities of the methodology followed for building this model are as follow:

- **Dataset**: For the training and evaluation of the model we use the dataset of Case Study 3.2.1.2 since it is the only one where we have available a pre-test and post-test.

- **Algorithm**: We are going to predict a continuous variable and we expect to find a linear relationship between the selected variables and *learning gain*. Therefore, we choose to use a multivariate linear regression.

- **Variable selection**: Since the size of this dataset is limited, we cannot include all the variables because that would lead to a potential overfitted model. To select the variables that we want to include, we perform the initial search of related work, an exploratory analysis by applying stepwise regression approach and correlation analysis. Based on this analysis, we determine which variables have the highest impact on the prediction model.

- **Evaluation**: To assess the quality of the model, we report the the coefficient of determination R^2 which measures the amount of variability of the dependent variable predicted by the independent variables. We test if all the assumptions of the regression model are fulfilled for generalization purposes.

- **Variable importance**: We use the standardized coefficients of the regression model to have an idea about the importance of each variable.

The selected variables that we take into account to build the model are as follows:

- **Learning gain**: The continuous variable that we want to predict is the *learning gain* of students after the interaction of the course

- **Pre-test**: The first two variables are related to the pre-test that students completed before interacting with the platform, the two variables are the *pre test score* and *pre test time*.

- **Use of platform**: The variables that we use related to the use of the platform are *average number hints*, *average number attempts*, *exercise time*, *video time*, *total time* and *optional activities*.

- **Progress in the platform**: The variables related to how much students have progressed wit the platform are *exercise effectiveness*, *exercise effectiveness no help*, *video effectiveness*, *completed videos*, *proficient exercises* and *average time correct answer*.

- **Distribution of their time**: The variables that we use related to how the time is distributed are *average time per day* and *dispersion time per day*.

- **Behavior of students**: We also include variables related to the behavior of students such as *follow recommendations*, *forgetful user*, *video avoidance*, *hint avoidance*, *unreflective user*, *hint abuse*, *exercise abandonment* and *video abandonment*.

We apply a hierarchical method with three entry steps and a total of six independent variables (introducing two of them in each step). We select this method as we ran an exploratory analysis and reviewed related work before designing the model, and we use that outcome to define the model. In the first step we add as predictors *pre test score* and *average number attempts*. This decision was also supported by the work of Feng et al. (2006) where they made use of these variables too in their research. In the second step, we add the variables *exercise effectiveness no help* and *average time per day*. In the last step, we introduce the last two variables which are related to the behavior of students while interacting with the platform. As our sample size is small, we cannot include all the variables that we have, otherwise this can lead to a potential overfitting. Therefore, we follow the dummy rule of 10 cases per predictor variable [1]. We add two new variables related to the behavior of students. The first one collects the variables which have a significant prediction power predicting an increment of the *learning gain* and the other one has a decremental influence. The new variables included in the third step are the following:

- Total Abandonment (*total abandonment*): This variable combines both *exercise abandonment* and *video abandonment*, as they have a significant influence and incremental influence on the prediction of *learning gain*. This is interesting as we could guess that students who abandon exercises and videos would probably learn less.

- Negative Behaviors (*negative behaviors*): This variable combines *follow recommendations*, *forgetful user* and *unreflective user*. An interesting detail is that other behavioral variables such as *video avoidance*, *hint avoidance* or *hint abuse* were not as significant as the others, thus they were left outside the model. These variables have a decremental influence on the prediction model. This makes sense in the case of *forgetful user* and *unreflective user* but no so much about *follow recommendations*.

[1] `https://en.wikipedia.org/wiki/One_in_ten_rule`

6.1.2. Results and Interpretability

Next, we report a summary of the three models in Table 6.1. The first model provides a R^2 of 0.481. The second model introduces two new variables rising up to 0.616. Finally, the third model includes the two variables related to students' behavior and provides a R^2 of 0.68, which means that our final model is able to predict a 68% of the learning gain's variability. The standard error of prediction is 13.3. This implies that when making a prediction the average deviation from the real value is around 13.3 points. A first impression about the importance of each one of the variables can be obtained from Table 6.2 by checking the standardized coefficients. Equation 6.1 shows the complete prediction formula with unstandardized coefficients.

$$\begin{aligned} LG = \{&13.615 - 0.668 * pre\_test\_score + 6.426 * avg\_attempts \\ &+ 0.392 * correct\_exercises\_no\_help + 0.824 * avg\_day\_time \\ &+ 0.143 * exercise\_video\_abandonment - 0.721 * negative\_behaviors\} \end{aligned} \tag{6.1}$$

Next, we analyse each one of the model predictors and its importance separately:

- *pre_test_score*: This variable represents the most powerful predictor. The meaning of the negative sign is probably related to the fact that if the initial knowledge of students is very high, it is harder to improve that knowledge. For example, it is hard that a student who scores 90 in the pre-test goes to a 100 score in the post-test. However, it would be very probable that a student who scored 0 at the pre-test, will score higher at the post-test after using the platform. The higher value of the pre-test, the harder is to increase the post-test score with respect to the pre-test. For every point in the pre-test, the predicted learning gain decreases 0.668 points.

- *avg_attempts*: The average number of attempts that students make trying to solve an exercise reports a positive effect towards predicting a learning gain. The higher the average number of attempts the better. A possible hypothesis would be that students who cannot solve exercises, do not even attempt to solve them and just leave, so they do not increase learning. In addition, students who make a lot of attempts might learn by error and repetition and thus they can obtain a higher learning gain in this process than just students that answer the question directly. For every unit that the average number of attempts increases, the predicted learning gain raises 6.426 points.

- *exercise_effectiveness_no_help*: The percentage of correct exercises without use of hints and answering correctly at the first attempt represents one of the most important predictors of the model. This makes sense as the more exercises students are able to solve without help, more likely is that their knowledge is higher. For every point that this variable increases, the predicted learning gain increments 0.392 points.

Table 6.1: Model summary of the linear regression model.

Model	R	R Squared	Std. Error of the Prediction
1	0.693	0.481	16.42
2	0.785	0.616	14.34
3	0.825	0.68	13.3

- *avg_day_time*: The average number of minutes spent by the student each day is the second most important predictor of the model. It makes sense that the bigger is the amount of time invested by students in the platform, the higher is going to be the increment of their knowledge. However, there might be cases in which this relationship does not apply. For every minute that the average time per day increases, the predicted learning gain will increase 0.824.

- *total_abandonment*: This variable has the lowest weight on the prediction model, anyhow it also helped to improve the prediction power. Surprisingly, if the amount of exercises and videos that students abandon increases, the predicted learning gain will also increase. A possible explanation is that students that have high abandonment ratios might be abandoning those resources because they already have that knowledge, thus they will score high on the post-test later and that will result in a learning gain increment. For every point that this variable increases, the predicted learning gain will increase 0.143.

- *negative_behaviors*: The higher is this variable the lower is going to be the predicted learning gain. This relationship makes sense for *forgetful user* and *unreflective user*, as we would think that these behaviors do not represent good actions for learning. Students who forget how to solve exercises mean that they are not really correctly acquiring the knowledge and unreflective students do not have the knowledge to solve the exercises and they are not reflecting on their errors. However, the relationship with *follow recommendations* is not quite straightforward. A possible hypothesis could be that students who follow recommendations, do not have a good background knowledge about the topics covered and they are going step by step. On the other side, students with a good background might jump from one topic to another, exploring those topics that are more appealing for them.

All the assumptions of the regression model are fulfilled. The linearity and homoscedasticity assumptions are fulfilled by plotting the standardized residuals versus standardized predicted values. The normality of the residuals is tested by checking the histogram and normal probability plot of the residuals. We should also point out that there are zero cases with a standardized residual above ± 2, which means that the model is well fitted and there are no outliers. Therefore, under these circumstances the model should generalize well to predict other samples of the same population. We cannot however test this model with a cross-validation as the number of cases in the data sample is too small.

Table 6.2: Unstandardized and standardized coefficients of the regression models.

Model	Independent Variable	Un-std. Coeff.		Std. Coeff.
		B	**Std. Error**	**Beta**
1	Constant	38.556	7.88	
	pre_test_score	- 0.601	0.84	- 0.655
	avg_attempts	4.093	3.149	0.119
2	Constant	14.485	8.991	
	pre_test_score	- 0.646	0.076	- 0.703
	avg_attempts	5.362	2.776	0.156
	exercise_effectiveness_no_help	0.271	0.106	0.224
	avg_day_time	0.557	0.200	0.231
3	Constant	13.615	9.734	
	pre_test_score	- 0.668	0.071	- 0.727
	avg_attempts	6.426	3.142	0.187
	exercise_effectiveness_no_help	0.392	0.104	0.324
	avg_day_time	0.824	0.230	0.342
	total_abandonment	0.143	0.097	0.155
	negative_behaviors	- 0.721	0.223	- 0.264

We can establish a comparison of results with some of the related works in the area. Despite our research had several similarities with the study of Feng et al. (2006), some variables are not the same because of the different nature of the learning environment, but others are the same such as *pre test score* or *average number attempts*. Additionally, we have considered behavioral variables which were not present at all in the previous study (Feng et al., 2006).

The study by Kelly et al. (2013), which also made use of a linear regression analysis to predict standardized test scores, obtained a R^2 of 0.57. They used different variables except for *average number attempts*. The study by Grafsgaard et al. (2014) makes use of posture and gesture data provided by sensors, they obtained a R^2 of 0.38 predicting learning gains. Their educational environment is a Java Tutor, therefore the variables of their model were different as it was a programming environment. Anozie & Junker (2006) reported a regression model which is able to account for the 63.7% of the variability. They also make use *pre test score* and other variables related to time and percentage of correct exercises. One of the main differences of this work with others is that our learning environment was based on MOOC technologies and e.g., the course had an intense video activity, whereas none of the other works compared here used videos as part of their learning experience. Therefore, the considered variables change, as the context is different.

There are two other issues that we consider important to approach. First, how good would get the model if we could use more predictors? We set up a limit to the number of predictors since our data sample is small. A backward stepwise regression analysis reported that we could achieve a R^2 of 0.75 with the use of all the considered variables. As a result, we find an upper limit of measuring 75% of the learning gains variability with the use of more variables. That would provide an improvement of 0.07 points with respect to our design using only 6 predictors (so adding many more variables would not represent a big improvement).

A second interesting question is the effect of the pre-test variable on the prediction model. In the case that we remove this variable from the considered ones and repeat the backward stepwise regression, we achieve only a R^2 of 0.481 with the use of all variables. Hence, we can notice that the effect of using prior knowledge by using the pre-test score is highly important, as the influence of this variable in the prediction model cannot be covered by any of the rest of considered variables. A similar analysis of the influence of *pre test score* in their regression model was performed by Feng et al. (2006) with similar conclusions.

6.2. Prediction of Certificate Accomplishment

In this section we approach the early prediction of certificate accomplishment in a MOOC. The general objective is to see how the quality metrics of the machine learning model change as more data is available and also the performance of different algorithms, to assess what might be the best choices. Also, we look into how the importance of variable evolves over the course timeline. This section is again divided in methodology and selected variables in Subsection 6.2.1 and a report and interpretation of the results in Subsection 6.2.2. More information is available in our previous publication (Ruipérez-Valiente, Cobos, Muñoz-Merino, Andújar & Delgado Kloos, 2017).

6.2.1. ML Method and Training

In this subsection we describe the machine learning method and training of the model. First the details of the method are as follow:

- **Dataset**: The dataset that we use is collected from ‘The Quixote’ MOOC Case Study 3.2.2.

- **Algorithm**: As we want to test the performance of different models for an early prediction, we apply five different algorithms. The implemented models are RF, SVM, GBM, kNN, and a logistic regression.

- **Evaluation**: We report common metrics used in classifiers such as the confusion matrix, as well as sensitivity and specificity. Additionally, we also use other metrics that are better when assessing unbalanced data such Area Under the ROC Curve (AUC), F1-score is the harmonic mean of specificity and sensitivity, and finally we also use Cohen’s kappa coefficient which measures the inter-rater agreement for classification taking into account random guessing and also class ratios.

- **Variable importance**: We use the *varImp* function from *caret* package for this purpose and obtain a scaled (from 0 to 100) importance of each one of the variables of the model for each week. The relative variable importance metric that we use is the same that is reported by Friedman (2001) as the final selected machine learning model is the GBM.

The selected variables are a combination, some related to activity, to progress and to the distribution of the time of students. The specific variables are the following ones:

- **Certificate earner**: The binary dependent variable that we want to predict is if a student acquired or not a certificate (*certificate*).

- **Use of platform and activity**: The variables about the use of the platform and amount of activity that we use are *total time*, *exercise time*, *exercise time*, *number active sessions* and *number events*.

- **Distribution of their time**: The variables that we use related about how the time is distributed are *average time per day*, *dispersion time per day*, *dispersion time per exercise* and *dispersion time per video*

- **Progress in the platform**: The variables related the progress of students are *exercise effectiveness* and *video effectiveness*.

The specific steps that we follow for the training and the evaluation of the models are the following:

1. We divide the dataset in training (0.75) and test (0.25) maintaining the same ratio of the predicted variable in both datasets. This partition is shown in Table 6.3.

2. As part of the training pre-processing we scale and center all the numeric variables. We establish as the quality metric to maximize the Receiver Operating Characteristic (ROC), and it is going to be estimated through a 10-fold cross validation repeated and averaged three times. Additionally, we allow *train* function to automatically search for the best configuration parameters of each algorithm. We report the results of the 10-fold cross validation of the best model in terms of F1-score and AUC.

3. We use each one of the selected models of the previous step to predict on the data from the test dataset. We evaluate the results over the weeks using using F1-score and AUC.

4. We select the best model from the previous for the purpose of this research and further explore the results for the model, connecting also the results with the specific MOOC that we are analyzing. We also analyze the importance of variables of the selected model over the weeks.

After training the five models for each one of the seven weeks (which implies that a total number of 35 models have been trained) we find that the SVM model has a very strong tendency to predict the majority class like a baseline predictor, despite our use of ROC as the quality metric that we want to maximize in our machine learning method. Therefore, we remove the SVM classifier due to this bad performance. Next, Subsection 6.2.2 reports the results.

Table 6.3: Distribution of certificate and non-certificate earners in the train and test datasets.

	# Students	# Non-certificate earners	# Certificate earners
Train	1289	1166	123
Test	429	388	41

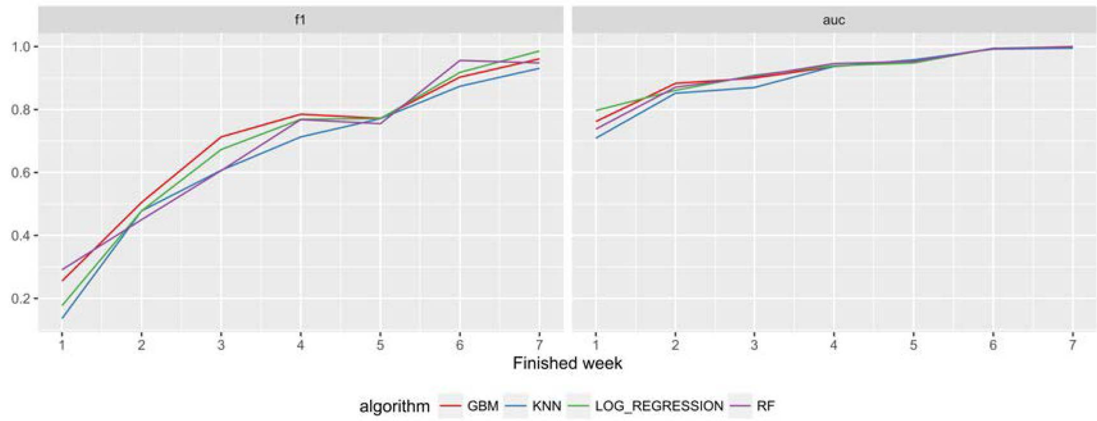


Figure 6.1: Evaluation results in the test dataset in terms of F1-score and AUC metrics for the models for each week.

6.2.2. Results and Interpretability

We evaluate each one of the models by predicting on the test dataset. Figure 6.1 represents the quality metrics i.e., F1-score (on the left) and AUC (on the right), over the weeks for each model. We can see that the performance after the first week in terms of F1-score is a bit higher for the RF model, and afterwards GBM model takes over as the best one. In terms of AUC in the first week log regression is the best, in the second week GBM has the best performance and afterwards the AUC values are very similar. We are looking for a stable model over the weeks, offering always a good performance and specially in the first four weeks, since those are the weeks in which we have chances of sending an early warning to avoid that a student misses the certificate. Considering these premises, in terms of the F1-score and AUC, we consider that the model that provides the best performance for this task is the GBM model, which always performs as the best or second best model over the four first weeks both in terms of F1-score and AUC, offering performance and stability.

Now we focus on the selected GBM model, which is the one that fits better the purposes of the study. Figure 6.2 shows the evolution over the weeks for the GBM model in terms of sensitivity, specificity, F1-score, Cohen's kappa coefficient, AUC and accuracy. Additionally, we add the baseline accuracy of the predictor that always classifies as non-certificate earners (0.904). We can see how the specificity remains high over the weeks, but the sensitivity is very low at the beginning, we are aware of these results but we think this is the correct approach. We want

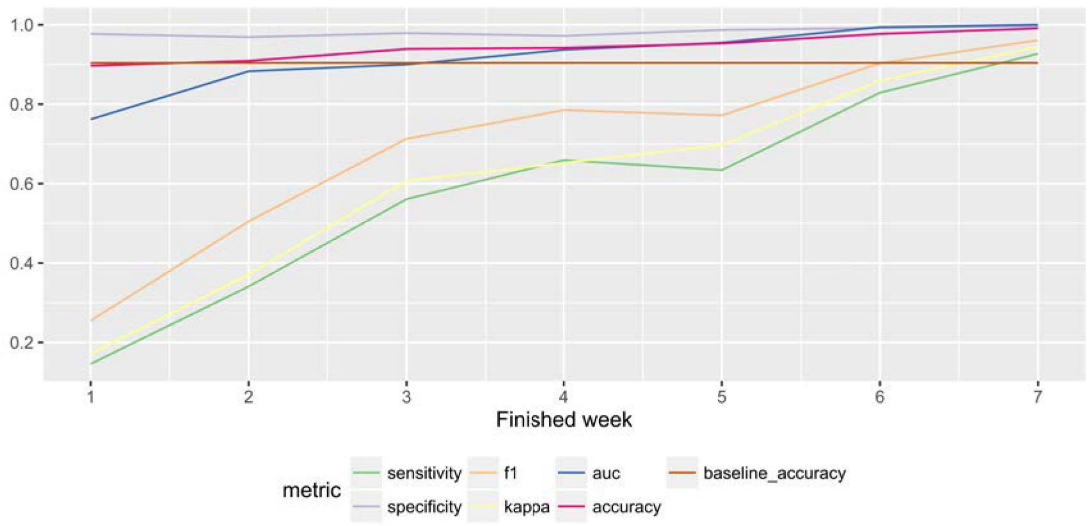


Figure 6.2: Evolution of the performance in the test dataset of the selected GBM model over the weeks.

to minimize false positives, since a false positive implies a student who is not going to achieve a certificate and still will not receive a warning by our system. We are less concerned about students who will get a certificate, but receive a warning regarding they are still in risk of not getting a certificate. F1-score increases in a similar trend than the sensitivity does, since F1-score is the geometric mean of sensitivity and specificity. Additionally, kappa coefficient also increases over time, the more that the predictor starts behaving differently than the baseline predictor (being able to detect both true negatives and true positives), the more the kappa coefficient increases.

One interesting detail which needs to be explored further, is the effect of the deadlines of week 5, where we can see that in terms of sensitivity, the predictor gets worse and accuracy improves little (this happened also with the rest of the algorithms). Then, after week 6 there is a big improvement in terms of accuracy and sensitivity. By the end of the last week the accuracy is really high (0.991), but the course is finished and the system can no longer send early warnings, that is why we should focus in the first three or four weeks.

We explore also the influence of the different variables of the GBM model over the weeks. We plot the variable importance results over the weeks in Figure 6.3. The results show an interesting trend where there is a lot of difference in the importance of variables during the first three weeks, where it is distributed among many variables, and at the end of the course, where *exercise effectiveness* is with much difference the most important. We can see that after week 3, the most important variable is *exercise effectiveness*, *exercise time* is the second most important one, and the rest of the variables have low relative importance. Nevertheless, during the first weeks the importance is more distributed among the different variables. The most interesting detail is that by the end of the first week, *number active sessions*, *exercise time*, *number events*, *dispersion time per video*, *total time* and *video effectiveness* have more importance towards the prediction than

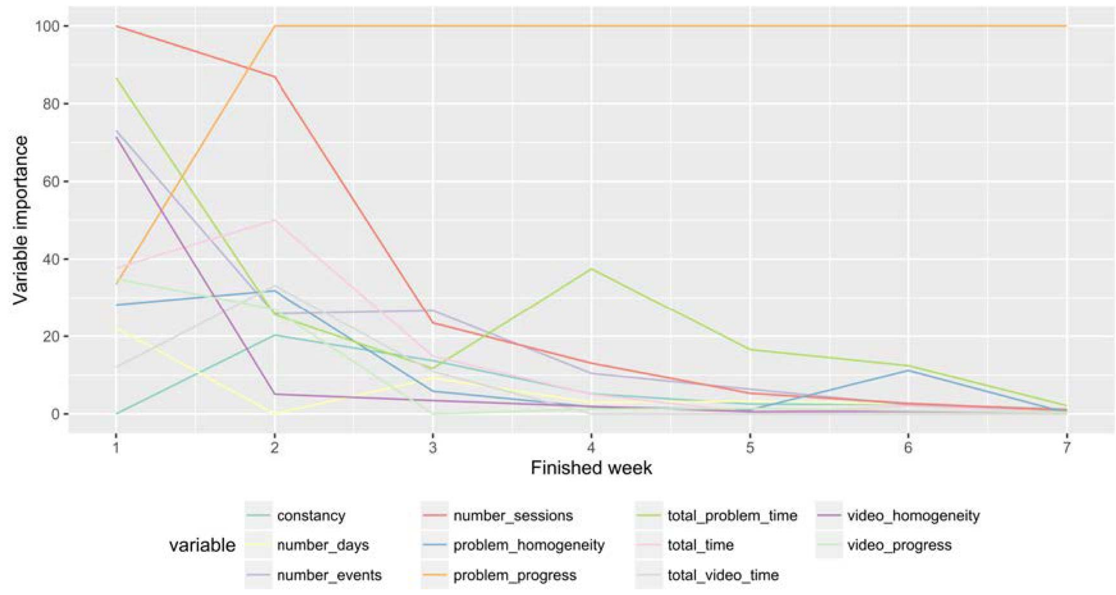


Figure 6.3: Evolution of variable importance of the GBM model in each week.

exercise effectiveness. Additionally, specially at the end of week two but also at the end of week three, some of these variables still have great importance (note *number active sessions* variable at the end of week two).

6.3. Classification of CAMEO Submissions

In this section we approach the prediction of which correct submissions have been copied applying CAMEO approach as described in Subsection 3.3.7.2. We delve into which features are more important for the machine learning model, as this can help understand if the specific characteristics of each problem or the student have a heavier influence on the amount of CAMEO. Additionally, such prediction models could be used as a run-time detector implement by MOOC platforms to detect and try to prevent cheating methods online. Analogously to the other two sections, Subsection 6.3.1 presents the methodology and training of the model, and Subsection 6.3.2 presents the evaluation of the model, assess the variable importance and discuss about these results as well. More information is available in our publication (Ruipérez-Valiente, Muñoz-Merino, Alexandron & Pritchard, 2017).

6.3.1. ML Method and Training

This section describes the machine learning methodology and training of the model. Next, we present some of the details of the method applied to build the model:

- **Dataset**: This study uses the introductory physics MOOC from MIT on edX (Case

Study 3.2.3). This is an adequate dataset for this study as it has many independent problems that make easier to detect CAMEO.

- **Algorithm**: We select RF algorithm (Breiman, 2001) as nowadays is considered one of the best classification algorithms, performing good with diverse types of data and also useful for raking the importance of predictors.

- **Evaluation**: Similar to the model for certificate accomplishment, we report confusion matrix, sensitivity and specificity on the test dataset. We also report AUC, F1-score, Kappa coefficient and compare with the baseline accuracy.

- **Variable importance**: To address the variable importance, we use VSURF [2] (Genuer, Poggi & Tuleau-Malot, 2015) algorithm which has been developed and optimized for rating variable importance using RF. Two metrics can be found in the literature to address the importance of variables in a RF model (Breiman, 2001). Mean decrease node impurity (Gini) and permuting out-of-bag (OOB) data. We use the second one that can be defined as follows. For each tree t belonging to the forest, we take the OOB_t sample (i.e., the cases not included in the boostrap data to construct t) and we denote the misclassification rate of tree t on OOB_t as $errOOB_t$. Next, we randomly permute the values of variable X^j in OOB_t to get a disturbed but realistic sample denoted as $\widetilde{OOB}_t^j$ with an associated $err\widetilde{OOB}_t^j$. Then, the variable importance of X^j is calculated as follows:

$$VI(X^j) = \frac{1}{ntree} \sum_{t}^{ntree} (err\widetilde{OOB}_t^j - errOOB_t) \tag{6.2}$$

- **Variable selection**: We also use VSURF algorithm to approach variable selection as support to see which variables are more important in the model.

Since in this study we delve into what variables have a higher influence on CAMEO, we divide the selected variables in these three groups, student, problem and submission features.

- **CAMEO submission**: The prediction target is the binary variable *harvested*, which indicates if the a student harvested or not a correct submission on a specific problem applying CAMEO.

- **Student features**: The student features that we include in the model are *performance first attempt*, *video time*, *page time*, *exercises accessed*, *number attempts correct answer* and *average time correct answer*.

- **Problem features**: These features describe characteristics of each problem, we use *type assignment*, *type response*, *show answer*, *location*, *random* and *max attempts*.

[2]https://cran.r-project.org/web/packages/VSURF/VSURF.pdf

Table 6.4: Distribution of non-CAMEO and CAMEO submissions in the train and test datasets.

	# Total submissions	# Non-CAMEO submissions	# CAMEO submissions
Train	376752	354966	21786
Test	94187	88741	5446

- **Submission features**: These features describe a specific correct submission to a problem, we have selected *time to deadline*, *attempt duration* and *attempts required* for this category.

The specific steps that we follow for the training and the evaluation of the models are the following:

1. We divide the dataset in training and test with a probability of 0.75 while maintaining the same ratio of the predicted variable. This partition is shown in Table 6.4.

2. We apply *train* function from *caret* package to implement the RF model. We perform a 10-fold cross validation and repeat 3 times to evaluate the results on the training set as well as select the tuning parameters for the RF model, the target quality metric that we seek to maximize is AUC. We also configure the *train* function to pre-process the features by scaling and centering the numeric variables. The selected model is implemented with 500 trees and 10 variables sampled at each split, the rest of configuration parameters are maintained as default. The resulting trained model has a AUC value close to 1 (0.99993) on the training set.

3. Next, we apply the model on the test dataset and report the different quality metrics of the model.

4. Finally, we apply VSURF algorithm for rating the importance of variables and discussing about these results to address which of the three categories of features is most important.

6.3.2. Results and Interpretability

We apply the model to the test dataset and Table 6.5 shows the percentage confusion matrix (N = 94187 submissions) and Table 6.6 shows some quality metrics regarding the model when applied to the train dataset. We report the AUC, sensitivity, specificity, Kappa coefficient and accuracy (although taking into account that data is really unbalanced, this is not a very reliable measure). We can see a clear improvement with respect to the baseline accuracy which would be the classification of all submissions as not CAMEO.

The results show that the model has very good quality metrics when applied to the test data with an AUC value close to 1, sensitivity (96.64%) and specificity (99.61%). These results are

Table 6.5: Confusion matrix applying the model to the test dataset.

Classification	**Reference**	
	Non-CAMEO	**CAMEO**
Non-CAMEO	93.852%	0.194%
CAMEO	0.366%	5.588%

Table 6.6: Quality metrics of the RF model applied to the test dataset.

Metric	**AUC**	**Sensitivity**	**Specificity**	**Kappa coefficient**	**Accuracy**	**Baseline accuracy**
Value	0.9993	0.9664	0.99611	0.9493	0.9944	0.9421

encouraging since they suggest that it would be possible to implement a detector that can predict CAMEO submissions with high reliability and without depending on the IP and also in real time (instead of as a retrospective analysis), which were the main handicaps of our previous detector. The next issue is to analyze how much each one the features is contributing to the model and if some features are redundant and could be removed without having a big negative impact on the model.

We apply VSURF algorithm to rate the importance of variables. The algorithm (Genuer et al., 2015) carries out as a first step a preliminary elimination and ranking of the variables, and second an analysis for variable selection. During its computation, the results of VSURF algorithm are averaged over many RF runs, which provides more certainty about the results, taking into account the intrinsic random factor of RF due to bagging (bootstrap aggregating). The algorithm provides three outputs, now we describe each one of these outputs providing the results and interpretation when applied to our model:

1. Sort the input features by variable importance ($VI(X^j)$) in descending order (averaged over 50 RF runs). It estimates a threshold of minimum VI (based on the VI standard deviation) and removes variables below the threshold, let m be the number of variables left. The m variables selected in descendent order of VI are shown in Figure 6.4. It is noteworthy to say that no variables were removed in this step, as none of them were below the threshold. A possible explanation is that no variables are redundant and all of them are able to convey some unique information for the prediction of CAMEO submissions, thus all variables are kept.

2. Constructs a nested collection of RF models involving the k first variables, for $k = 1$ to m. This means that in this collection, the first RF model constructed includes only the most important variable, and the last one includes all the variables. It selects the variables which provide the model with the smallest $errOOB$ (averaged over 25 RF runs). This leads to m' variables. The second step reveals that the best model is provided by removing the last three variables *max attempts*, *attempts required* and *random* as can be seen in Figure 6.5,

Figure 6.4: Descendent ranking of variables in terms of VI.

where the red line establishes the cutoff point.

3. The final step takes the m' variables, and constructs a new ascending sequence of RF models by introducing the variables following a stepwise procedure. More specifically, a variable is introduced into the model only if it decreases $errOOB$ more than the average variation provided by noisy variables. Finally, the variables of the last model are selected, Figure 6.6 shows each model built following the stepwise procedure. After this step, two more variables are removed since these did not improve the model enough. The removed variables in this step are *page time* and *exercises accessed*. These variables denote an indication of amount of activity and are correlated with other variables measuring student activity, thus not improving the model enough to be included.

The last checkup consisted in building a RF model with the final 10 variables selected by the VSURF algorithm, and compare it to the model that had the full 15 variables. The test proves that the RF model with only 10 variables performs almost as good as the one with 15 variables.

We originally selected six variables related to student features, six related to problem features and three related to submission features. The VSURF algorithm has removed the same ratio of each (1/3) leaving four, four and two features in each category respectively. From this finding, we can conclude that the three categories of features (student, problem and submission) have influence towards the prediction of CAMEO events. The first four variables in terms of VI are *average time correct answer*, *video time*, *performance first attempt* and *number attempts correct answer*, which correspond with the four student features that are kept in the selected model. We can conclude from these results that the four most important variables of the model are the student features, consequently student features are more important than submission and

Figure 6.5: Nested collection of random forest models.

Figure 6.6: Sequence of RF models constructed using a stepwise procedure.

problem features. The fifth and seventh variables in VI order are *time to deadline* and *attempt duration* are submission features, whereas the sixth, eight, ninth and tenth are *location*, *show answer*, *type response* and *type assignment* which are problem features. Thereupon, it seems that submission features have slightly more importance than problem features, but this hypothesis is not conclusive. Nevertheless, it is important to note, that although most of the problem features have the smallest VI towards the classification, these have been kept in the model because they provide valuable information regarding the predictability of CAMEO events.

In terms of the VI order of variables, we are not surprised to find out that *average time correct answer*, *performance first attempt* and *number attempts correct answer* have the highest VI values. Our previous findings in Subsection 4.5.2 already suggested that master accounts have the best performance in terms of solving questions in the first attempt and doing it very quickly, hence it makes sense that those are the most important features. Additionally, *time to deadline* and *attempt duration* features were also kept, and this is in line with our previous findings which indicated that CAMEO submissions are closer to the deadline and the attempt duration was much shorter (sometimes inhumanly shorter) than normal submissions. Finally, *location*, *show answer*, *type response* and *type assignment* are kept as problem features, which is also in line with our previous findings where we showed that students applied CAMEO more at the beginning and middle of the course (until they got the certificate), that they also CAMEO more when the show answer button was enabled before the assignments' deadline, when the type of response was multiple choice and it was a high stake question in terms of grade.

Related to the variables that are removed by the second step of VSURF algorithm, we agree that *max attempts* might not provide much information, and in the case of *attempts required* it probably has a high correlation with the student feature *number attempts correct answer*, which was kept as one of the top importance variables of the model. However, we are surprised to see that *random* feature is removed from the model, since we reported in Subsection 4.5.2.5 that CAMEO is found two times less in questions that contain random variables in the statement. We think that the variability provided by *random* feature might be in relationship with other features as well e.g., *type response* might provide part of this information as most question with random variables are 'formula' response types and this might be why *random* feature is removed. Finally, the two student features *page time* and *exercises accessed* are removed in the last step of the VSURF algorithm despite being variables with a medium importance. We believe that is due to the fact that they denote some indication of the amount of activity of the learner and this information might be provided already by other variables.

Part III

Recommendations to Improve Learning Processes and Summary of Conclusions

Chapter 7

Discussion and Recommendations to Improve Learning Processes

Contents

This chapter discuses some of the findings of previous Chapters 4, 5 and 6. More specifically, in Section 7.1 we talk over our findings in terms of the amount of activity while trying to delve into the motivations of students, in Section 7.2 we debate about the profiles of students that we found in Chapter 5 and the potential applications and outcomes for the learning process. Section 7.3 presents some recommendations to improve the learning process of students based on instructional design for optional activities, badges or to decrease CAMEO while Section 7.4 describes initial ideas to build automatic systems that can be used to improve the learning process of students based in our findings. Finally, Section 7.5.2 presents and discusses our results from evaluating the effectiveness, usefulness and usability of ANALYSE with 40 respondents.

7.1. Amount of Activity and Motivations

In Section 4.2, we analyzed the use of regular courseware activities and the effectiveness of students with those activities. We showed that some students focus their learning on solving problems while others prefer to watch videos. However, the majority of them combine both types of educational resources. This information can be used to detect the preferred learning method of each student. In Table 4.1 we showed correlations between effectiveness and other behavioral metrics. Our findings suggest that *exercise effectiveness* and *video effectiveness* are strongly correlated (0.63); a reasonable hypothesis is that part of the cause of *exercise effectiveness* is due to *video effectiveness*. We could also frame this relationship in motivational grounds, students that are highly motivated for this kind of learning experience might be more active with all educational resources. Other interesting correlation is between *exercise effectiveness* and *follow recommendations* (-0.115), which indicates that students following advice from Khan Academy recommender system might perform worse. This might sound surprising in first instance, the rationale can be that students whom are following those recommendations, might have a lower initial knowledge and that is why they are following the recommendations, whereas students with a better base knowledge might solve exercises following a more anarchic behavior. Another interesting association was found between *video effectiveness* and *video avoidance* (-0.234). This relationship indicates that students who avoid watching videos when they are not able to solve the correspondent exercise, are less effective when interacting with the videos they access. One could already expect this result, since they avoid watching videos even when they need to do it, their video effectiveness is reduced as they do not watch the correspondent videos in these situations. We can use these findings to analyze which behaviors might be negative for students' effectiveness to send recommendations.

Additionally, as part of Section 4.3, we analyzed the use of optional activities comparing it also with regular activities. We recall that the available optional activities were feedback, vote, setting up goals, using the badge display and changing the avatar profile, and that students were not advised regarding the availability of those activities. In general, our findings suggest that the use of these activities was low (Table 4.2). Only 23.2% of the users that logged into the platform used at least one of the optional activities, and the activities that were used more frequently were the profile avatar (10.8%) and badge display (12%), that are precisely those which are not related to learning activities. Furthermore, we presented in Table 4.3 a comparison between the percentage of regular learning activities and optional activities, showing that regular activities were used with a much higher frequency. We discuss more about the potential implications regarding these results in Section 7.3.

Moreover, in Table 4.4 we found that *optional activities*, and also the use of each one of the optional activities separately, was significantly correlated to *proficient exercises*. Nevertheless, after performing a partial correlation controlling for other variables, this correlation lowered and in some cases became non-significant. Our hypothesis here is that the use of any the optional activities does not necessarily produce a higher percentage of *proficient exercises*, since there might be a confounding effect with other activity variables, once controlling for them, the relationship is weakened a lot. Anyhow, we believe that the use of optional activities might engage students to use the platform more actively, thus leading to more learning. The optional activities that remained significantly correlated when applying partial correlations were goal (0.25), avatar (0.235) and display badges (0.229). The fact that students are personalizing their personal profile might motivate them to learn more, and in the case of goals, maybe students engage in the process of completing goals therefore learning more in the process of mastering those skills. Furthermore, Table 4.4 also showed the level of correlation between *learning gain* and *optional activities*. The Pearson correlation showed that *learning gain* was significantly correlated with *optional activities* (0.282), vote (0.333) and display badge (0.296). However, after removing the effect of third variables by applying a partial correlation, all relationships became non significant except for display badges (0.261). Similarly to our hypotheses for *proficient exercises*, we believe that despite the use of these activities might itself not lead to more learning achievement, it might bring motivation for students e.g., setting up badges or displaying an avatar might make students feel better, setting goals might motivate students to finish them, or making votes and giving feedback might make students feel as a part of the community. Consequently, another possible point of view is that the use of optional activities might be an indicator of students' engagement and motivation. Under this assumption, students who are motivated and engaged with the learning process might explore the platform more deeply, interacting with the different optional activities. Our results suggest that the use of optional activities can draw a positive influence regarding the learning outcomes of students.

We also compared the use of optional activities with other categorical variables by cross-tabulating using contingency tables. First, we compared the gender and the use of each optional activity separately as a 'yes' or 'no'. The results revealed that women use more often goals, avatar and badge display whereas men use more frequently vote and feedback. Nonetheless, the only relationship that was tested significant by the Pearson Chi-Square test was the use of feedback for men ($2.80, p = 0.045$). Other previous work in the literature have also found differences between men and women in the use of web-based learning environments e.g., (Caspi, Chajut & Saporta, 2008; Muñoz-Merino, Molina, Muñoz-Organero & Delgado Kloos, 2014). Additionally, in Table 4.5 we showed the log linear analysis of each optional activity as categorical variables. We found some associations between the use of avatar and display badges, between the use of feedback and votes, and also goal and avatar. We believe these results can be interesting for student profiling as well e.g., finding types of students very into profile personalization or social activity, which could be later on used for adaptation and personalization of their learning experience. Moreover, it might be interesting to check which specific associations might lead to more engagement and motivation.

During Subsection 4.4.1 we made an overview of the amount of badges delivered in Case Study 3.2.1.1. We showed how the amount of badges delivered is very unbalanced among the different students. We found that while many of those students earned many badges, others did not use them much. For example, we found more than 80 students who earned 100 badges or more in one course, and 15 that earned 500 badges or more. This could be indication that the student was interested in earning badges, but it might also mean just that the student earned those badges as part of the learning process. Badges are extrinsic motivators, therefore we should be careful in our gamification design (Deci et al., 2001). Indeed, we found some students who earned many badges by solving systematically easy exercises. We are concerned about that behavior since it would imply that the motivation of students lie in earning badges rather than in actually learning the available educational contents. Therefore, we decided to implement more complex metrics to measure the behavior of students with badges. During Subsection 4.4.3, we reported how the badge indicators *intentionality topic badges*, *intentionality repetitive badges*, *concentration badges* and *time efficiency badges* were distributed among the students, as we expected we found some students with very high intentionality, efficiency and concentration towards badges, which can help measure more accurately the interest of students for the badge system. We were specially concerned by those students with very high *intentionality repetitive badges*, which as we suspected, imply that they are earning the same *repetitive badges* systematically solving the same exercises for the mere purpose of having more badges. We believe this can be a troublesome issue and we discuss more about potential design in Section 7.3.

Furthermore, in Subsection 4.4.4 we analyzed the correlation between badge metrics and others, our first finding was that the different gamification indicators are highly correlated even when the definitions are different. As an example, we found a significant correlation (0.445) between *intentionality topic badges* and *intentionality repetitive badges* despite they address two completely different types of badges. We also found a strong correlation between badge metrics and those related to activity in the platform. It is only natural to think that students performing more activities will hence earn more badges as well. However, another potential line of thinking is that students that are motivated towards badges, will eventually interact more with the platform, and that increased motivation and interaction will lead them to finish more learning activities and potentially learn more. Additionally, we found a correlation between badge metrics with *videos accessed* and *completed videos*, despite the definitions of these badges do not take into account video activity at all. Even though these correlations were not very high, they exist because video activity might denote engagement with the platform, and this can be in relationship with their interest on gamification. We also found a very high correlation between *intentionality topic badges* and *concentration badges* (0.859), which can indicate that students focused on learning topics are concentrated on their task. We consider that as a good behavior that might be positive to encourage.

Another important part of our analysis was related to academic dishonestly and illicit collaborations in Section 4.5. Our first approach was to detect illicit collaborations by detecting *close submitters*. We presented our results in Subsection 4.5.1. We were able to detect 99 students in music course and 26 in the philosophy one, most of them grouped as simple pairs but also bigger communities submitting together their solutions. Additionally, based on Figure 4.12 and Table 4.8 we argued how these accounts that we categorize as *close submitters* are able to achieve a certificate doing a significant amount of less submissions, being active fewer days, watching fewer videos and viewing less discussion threads. Therefore, they are using some method that facilitates their way into a certificate of accomplishment. The motivation is probably connected to achieving a certificate. However, we cannot generalize since there are more complex relationships e.g., in some of the bigger communities showed in Figure 4.11 the motivation of some students might be receive a certificate without effort but others might be altruistically helping their friends by sharing their solutions. Besides, some of these communities might even be positive for the learning process of students. We delve into these different motivations and which of the associations can be more harmful for the learning process in next Section 7.2.

Second, we reported results and analyzed the cheating method CAMEO, which we believe to be very problematic. As a summary, our findings based on Figure 4.13 suggest that 12.9% of the certificate earners harvested at least a minimum of 1% of their correct answers, additionally 3.7% of the certificate earners cheated in more than 50% of their correct answers, what we can denominate as 'cheating through a certificate'. Moreover, the majority of CAMEO submissions (78% in the case of certificate earners), are premeditated i.e., students did not try to legitimately solve the question before cheating upon it. Furthermore, there is a correlation (0.359) between this premeditated behavior and the amount of CAMEO, which seems to convey that the more a student cheats, the less the student is trying to legitimately learn the contents. We believe that the main motivation for students to use CAMEO is in order to easily receive a certificate. This is supported by several of our findings. For example, we showed in Figure 4.15 that the amount of CAMEO dramatically decreased after students earn their certificate, and we also reported in Subsection 4.5.2.5 that it is more likely to find CAMEO on high-stake questions which provide more points towards the final grade. One interesting example is the case of non-certificate earners applying CAMEO, which initially can appear to weaken our hypothesis. Nonetheless, our findings in Figure 4.16 suggest that non-certificate earners tend to be more cynical, cheating a much higher percentage of their correct submissions than certificate earners. The most reasonable hypothesis from our point of view is that those non-certificate earners accessed the course with the intention of cheating through a certificate, however when they noticed that the course is not very CAMEO friendly i.e., more than 1000 questions and many of them randomized, they decided to drop out the course.

We believe CAMEO to be a serious issue due to several reasons. Pedagogy-wise, CAMEO is most likely related to poor learning, as any other cheating method students are achieving a passing grade without actually learning the required contents. Additionally, as research has suggested, students cheating might think that they have better skills that they actually have and damage long-term success in their career (Sparks, 2011). The second issue is that CAMEO represents a threat to the MOOC model based on certificate value, more than 10% of certificate earners used CAMEO and about 3.7% harvested the majority of their answers. Although MOOC providers are already aware of academic dishonesty problems in MOOCs i.e., trying to control identity and impersonation, there is still work in this direction, since if certificates are not trustworthy and valuable, then a MOOC model that uses certificates to accredit learning will not be sustainable in the following years. Finally, CAMEO also poses a serious problem with educational research in MOOCs. We showed in Figure 4.14a that master accounts were the faster ones and with the highest performance to solve problems. Additionally, Figure 4.14b showed that the density distribution of the performance of the master accounts was the higher while the harvester accounts had the worst performance, being normal accounts in the middle part. Therefore, any research that tries to identify variables correlated with student success in a MOOC is heavily biased by these accounts. For example, master accounts have a very high success rate, however these accounts do not use any of the learning contents since they are cheating, therefore these accounts would weaken the relationship between using instructional materials and certificate accomplishment. Thus, it is necessary to be able to detect and remove both master and harvester accounts in order to achieve reliable conclusions from MOOC research.

7.2. Student Profiling Outcomes

Chapter 5 presented how to cluster and profile students in terms of their different behaviors: first regarding their activity with regular and optional activities, second regarding their behavior with badges and third in terms of their roles and behavior in academic dishonestly and illicit collaboration. The overall idea is to present some ideas regarding the potential outcomes of such clustering and student profiling, such as analyzing the different student profiles and how these can be useful for adaptation purposes.

In the first instance, Section 5.1 presented the clustering results when we used as input the variables *exercise effectiveness*, *video effectiveness* and *optional activities*. More specifically, Figure 5.1 showed a boxplot visualization and Figure 5.2 a parallel coordinates visualization of the indicators of the resulting four clusters. Cluster 4 was composed by students who did not show interest in either regular or optional activities. Generally speaking, these students invested very little time in the platform and were therefore not motivated by the learning experience. We can argue that these students might not perform well in self-regulated online environments where they need to self-manage their time and learning process, or maybe they just were not motivated by the specific contents and they left the platform just after checking them out. Maybe a potential recommendation for this cohort of students would be to consult with them if they enjoy these kind of online environments in order to offer them an alternative possibility closer to more traditional learning environments like face-to-face classes. Then, students from cluster 2 had a moderate interest in the platform with an average time of 270 minutes, and though most of the students belonging to this cluster did not show interest in optional activities, some of them used several optional activities. There is not a clear recommendation for students in this cluster, as we can see there is a high variability also in *exercise effectiveness* and *video effectiveness*, specially the latter were we can see some students very close to a 100%. Due to this high variance within the cluster, we cannot establish a clear recommendation for all students, and it would be advisable to customize recommendations for each particular individual. We discuss next students that belong to clusters 1 and 3, which shall be analyzed jointly. First, students from cluster 3 were quite active with the platform spending an average of 631 minutes, and effectively progressing in both exercises and videos. Nevertheless, most of these students did not used optional activities at all or just used one of them. Second, students from cluster 1 also did a nice effort in the platform with an average of 697 minutes, they also did a good effective progress in exercises and videos. Nonetheless, the variance of both *exercise effectiveness* and *video effectiveness* is much higher than the one in cluster 3. The key distinction of cluster 1 and cluster 3, is that all students of cluster 1 used optional activities, and it is a cluster that is fundamentally composed by students that used optional activities. Although the progress of students in cluster 1 is a bit lower than the one of students in cluster 3, students in cluster 1 invested more time on average (697 min compared to 631 min) than the ones in cluster 3. One hypothesis is the idea that we discussed in previous Section 7.1, which is that the engagement with the learning environment led them to spend more time in the platform and at the same time to use optional activities. These two student profiles are very interesting, both of them showed a motivation for the learning environment as their high effectivenesses reveal, with one key distinction which is that one cluster showed interest in optional activities and the other did not. We believe that students from both clusters can benefit from blended or online pedagogies where the use of self-regulated online learning environments is possible. Students from cluster 1 showed a natural motivation for the environment which they explored using optional activities, despite they were not advised about the existence of these activities, spending a very high amount of time and clearly showing engagement with the learning process; they probably do not need clear guidelines as they enjoy the exploration of the contents. Students from cluster 3 also showed a clear interest in the learning contents achieving the highest effectiveness in the course, however they did not make use of optional activities. Therefore, it might be the case that in order to take advantage of all the features of the online learning environment they should receive more formal advise and recommendation to use the different available tools and optional features. Overall, we think that these findings can be useful to find appropriate pedagogies for different profiles of students.

A similar approach was followed to analyze the behavior of students with badges in Section 5.2. We showed the clustering results in Figures 5.4 and 5.5. The general idea is that we can use these gamification metrics to try to assess which students are interested in gamification and which are not, in order to personalize their learning experiences or for group formation purposes i.e., to make working groups or classes setting up together students that might feel motivated by the same features. We detected three different clusters that we analyze as follows: cluster 2 was composed by those students that did not do much effort in the platform, analogously with the findings of previous student profiling analysis, we can infer that this is the cohort of students that were not motivated by this self-regulated approach since they invested little time and did not show interest in badges or educational resources. Consequently, they would probably perform better in more traditional settings as we already concluded in previous analysis. Students in cluster 1 showed interested in the Khan Academy environment by actively using videos and exercises, and also the badge indicators are high. It looks like these students enjoyed both the self-regulated and the gamification environment, they have the highest time investment and effectiveness, and showed interested in both *topic badges* and *repetitive badges*. As a result, we can conclude that this is an appropriate learning environment for this cluster of students. Finally, cluster 3 represents a very interesting group of students, which have spent a moderate time and accomplished a moderately low progress with educational resources, but has shown a lot of interest in badges. More specifically, their *intentionality repetitive badges* indicator is the highest of all, showing that despite their progress and time invested is lower than the one of those students belonging to cluster 1, their interest for *repetitive badges* is higher. Therefore, it is clear that these students were interested in achieving those badges and that they can feel motivated by gamification features if provided during their learning experience. However, we can also see that their progress has been lower than students in cluster 1. This might be due to repetitively solving the same exercises to gain more *repetitive badges*, instead of advancing with different exercises. In any case, this behavior will not lead them to learn more or progress in their learning experience, so we should be careful with this kind of behavior as it can be counterproductive for the learning process. We discuss more about this aspect in next Section 7.3.

As part of Section 5.3 we applied clustering to those accounts detected as *close submitters* after applying the algorithm as described in Subsection 3.3.7.1, the input variables were *exercise effectiveness*, *number active days*, *videos accessed* and *number submissions*. We note out again that the global idea is that these students always submit their assignments very close in time, therefore there is something suspicious and probably illicit happening here. We applied this analysis to both music and philosophy courses separately and obtained the same cluster types. Overall the global idea is that accounts from cluster 1 did a great effort by watching most videos and being active many days and managed to achieve a certificate. Accounts from cluster 2 did a small effort, they almost did not watch videos, they were active a moderate amount of days and attempted few submissions, still they were able to achieve a high score and get a certificate. Finally, accounts in cluster 3 were active very few days, did not watch videos but still performed plenty of submissions. Accounts in cluster 3 are most likely harvesting accounts as described as part of CAMEO strategy in Subsection 3.3.7.2 that are used to harvest quiz solutions by applying exhaustive try and error, that explains the low success and very high number of submissions.

Carrell, Malmstrom & West (2008) reported that peer cheating influences students to also cheat, so it is only natural that we found these illicit associations. An association between two accounts that belong to cluster 1 can potentially represent two students that made an effort in the course by watching videos and trying to learn and understand the contents and meeting once per week to submit their assignments together, potentially sharing their answers in a reciprocal relationship. Motivation here can be ambition to improve the grades, and we might argue that this relationship does not represent a severe problem for the learning process of these students. An association between a student from cluster 1 and other from cluster 2, might represent a less balanced interaction where student from cluster 1 is potentially having a passive attitude and passing the answers to student from cluster 2 (potentially a friend or acquaintance), so that this latter user can obtain a certificate without investing effort in the course. Indeed, literature has reported that one typical attitude toward cheating is that, one copies from other (‘active’) and the other lets others to copy ('passive') (Eisenberg, 2004), which resembles quite well this situation where student of cluster 1 usually submits the assignment before the one in cluster 2. Additionally, letting others to copy from you is regarded as less severe than actually copying from others (Yardley, Rodríguez, Bates & Nelson, 2009). In the case of this specific association, the impact on the learning process of students from cluster 2 is obviously more severe. Then, we also found associations between two accounts from cluster 2, which implies that two students are trying to get a certificate without investing much effort in the course. As found in previous research students do not usually cheat alone but with friends that are close and they feel they can trust (Scrimpshire, Stone, Kisamore & Jawahar, 2016), therefore we are not surprised to find couples of students cheating together to accomplish a passing grade without effort. We believe that this can be severe for the learning process and future beliefs of these students, since they come to think that they are able to accomplish goals putting effort on them. Finally, we have the association of one account from cluster 1 and one from cluster 3, and also one account from cluster 2 and one from cluster 3. As there is a presumably harvesting account (cluster 3) in both associations, most probably these are CAMEO associations as were described in Subsection 3.3.7.2. Therefore, we note that under these associations, both accounts are run by the same student. First the association between an account from cluster 1 and one from cluster 3 might represent a slightly less severe situation since, the account from cluster 1 invested an effort in the platform and maybe is using the harvesting account to be on the safe side and get a passing grade without problems. We can see this association closer to the idea of applying CAMEO as a 'helper-mode' that we reported (Alexandron et al., 2017). The second scenario is one account from cluster 2 and one from cluster 3, which might represent a more severe situation since the student is managing to get a certificate without any effort at all and seems to be closer to the 'premeditated-mode' that we reported (Alexandron et al., 2017). In the case of bigger communities, each one shall be analyze separately, but we already showed in Table 5.2, how some communities might combine CAMEO and peer cheating, and some of them might be more ethic and real working groups that can actually be fruitful for their learning process. It would be interesting to analyze the role of different students (Boud, Cohen & Sampson, 1999) e.g., leaders or followers and other behaviors. Although we think that in most cases students know each other prior to starting the course, they can also meet online in study groups and then decide to perform an 'unethical collaboration' (Lampe, Wohn, Vitak, Ellison & Wash, 2011).

7.3. Instructional Design

Section 7.1 discussed about the relationship between optional activities and learning outcomes. However, this correlation was weakened once removing influence of third variables. Nonetheless, we hypothesize then that the use of optional activities can be used as an indicator of engagement and motivation, and at the same time, these activities can keep students engaged through their learning experience and motivated them to spend more time in the platform and use more activities, hence leading to improve their learning outcomes. Due to these reasons, we think that the inclusion of optional activities can bring benefits to the learning process of students. However, one major problem that we found in our experiment is that under our conditions (the use of the platform was not mandatory and instructors did not inform their student about the availability of these optional activities), the use of optional activities was very low compared to the use of regular activities like exercises and videos, and for most students optional activities appeared to be unnoticed during their learning process. Consequently, our recommendation is to encourage or at least to share information regarding the availability of different optional activities and tools, so that students that might not have a very exploratory behavior know about them as well. One interesting finding was that more than half of the goals that students started, were also finished, which gives an indication that students are resilient and are motivated to achieve their learning goals. Moreover, we also found that most of the votes that students gave to their peers were positive, and this can be good to reinforce other peers. However, there might be some cases in which more neutral or even negative votes can have a good impact on the overall learning process since it can help distinguish good from bad learning content. So students should not be afraid of emitting neutral or negative votes when adequate. One problem that we detect is that the activities that were used more (avatar and display badges) are actually not related to learning. The setting of goal, votes and feedback were used less frequently, therefore we think that an additional reinforcement for these activities might be necessary, specially in the case of goals since we already presented that students were motivated to complete most of the started goals. One possibility, would be to encourage the use of these optional activities related to the learning process by providing additional points towards the final grade, or a small percentage of the grade based on these activities.

In Section 4.4 we also presented an overview of which badges were delivered more frequently, finding that the top-10 ranking was very similar for the three courses. Some of these badges are easy to acquire and were received by most of the students just after doing some interaction with the platform. Therefore, we believe that these easy participatory badges will not play an important role in the motivation of students. Nevertheless, instructors might be able to set up more specific and motivating badges in those exercises or topics that are tough to master for students, this way students might feel an extra motivation to learn a difficult topic. Additionally, if we want to boost some specific aspect or tools, then we can provide specific badges for the desired behavior that we want to encourage e.g., let us say that we want to increment the use of goals, we can provide new badges when students start and finish goals; this can apply to many other objectives. During Subsection 4.4.2, we explored which factors might have an effect on the amount of badges that were delivered. Although we are not able to be conclusive about our results, our findings seem to indicate that easier exercises and longer videos will trigger more badges, although this would be obviously dependent of the criteria to deliver those badges. Therefore, we mention again the previous idea that we might want to use some easier badges on those exercises that are difficult to resolve. Furthermore, the position of the item within the course structure had a significant effect in the number of badges triggered. As one could expect items at the beginning of the course triggered more badges than those at the end. To overcome that obstacle and keep students engaged, we would recommend to spread badges during the course duration with special emphasis on the contents in the middle and end of the course, to improve the motivation through the whole duration. As many other studies, our findings suggest that gamification was helpful to engage many students, and this can be useful for example in engineering education, where students have severe issues in terms engagement and motivation. Indeed a report by the American Society for Quality informed than 80% of the children were not interested in engineering careers (Henke, 2009). The use of game-thinking and game elements have been proven to useful in engineering education e.g., in a course for computer science engineers (Mas-Sanso & Manresa-Yee, 2016). Nonetheless, creating a good technology-enhanced learning experience is quite a challenging task that requires a well fitted pedagogy design for each course (Riojas, Lysecky & Rozenblit, 2012). As an example, we remark again the cluster of students that we discussed in previous Section 7.2, which tended to solve the same easy exercises in a systematic way without no additional learning involved, in order to earn more *repetitive badges*. In that case, we can clearly see the effect of badges as an extrinsic motivator, since students stopped solving exercises for the purpose of learning. Therefore, the final gamification design need to have the focus on learning and trying to boost the intrinsic motivation of students (Zirk, 2014).

Finally, in Subsection 4.5.2.5 we analyzed which factors were associated with CAMEO. We believe that this is already a serious problem and will likely increase if the value of MOOCs certificates increases over the next years, therefore we provide some instructional advise that can be helpful for instructors and course designers. Although this is not a final solution to the problem, it can help to decrease the prevalence and severeness of CAMEO in future courses. One of our findings was that the use of randomized questions reduces CAMEO by half, and this finding was also in the same line that the work of Northcutt et al. (2016). Therefore, we recommend the use of randomization as frequently as possible, specially for high stake questions. More specifically, randomization allows to select variables of a statement as a range, or select the specific variables from a pool of values, therefore as these different values lead to different potential solutions, students are unable to use the solution of the harvester account in their main account. However, as randomization difficulties CAMEO, it does not remove it entirely as students can find other ways. A more general approach could be to implement very big pools of questions, for each topic. However, this could be hardly implemented by just a course staff, but more a bigger scale with different stakeholders in order to create large problem pools for different topics, levels of complexity and with the appropriate technology to be used by different courses. We also found that delaying the feedback of the question decreased the prevalence of CAMEO by half as well. The main feature that facilitates CAMEO to students is the use of 'show answer' functionality that can be available in edX. However, even when 'show answer' is completely disabled, students can also try exhaustive search and get correct/incorrect feedback that it is always given; this latter approach is specially effective in multiple choice quizzes. Indeed, the main caveat of delaying feedback is that is a design pattern that is not beneficial pedagogy-wise. Also, delaying feedback would be complicated in self-paced courses which do not have hard deadlines. Overall, this solution improves security regarding CAMEO at the expense of deteriorating the learning experience of students. As a result, one could argue that we should focus on those students who actually want to learn instead than on those who do not. Our recommendation is finding a balance between reducing CAMEO and maintaining quality learning process e.g., to use randomized questions as frequently as possible and delaying feedback only on really high stake questions like exams, while maintaining the feedback available for formative assessment activities that do not have any weight towards the final grade. Moreover, we advise to modify high-stake final exams in future editions of the course.

7.4. Towards Actuators

In this section, we provide some guidelines regarding how to use the lessons learned to potentially design and implement automatic systems that can have different tasks, such as sending warnings or recommendations or to detect behaviors. One potential use of these metrics is to automatically detect problems in educational resources or students. For example, the system can analyze effectiveness in exercises and send a warning to instructors for those exercises which have very low correctness ratio which might imply that the exercise is poorly designed or that the difficulty is too high. Similarly, the system can analyze the repetition ratio of videos and send warnings for those videos that have a very high repetition ratio, which could imply that students are confused by the video and they are watching it repetitively. Instructors can use these warnings to revise the educational contents with problems, this is specially useful in MOOC environments since instructors are unable to analyze students or educational resources individually, therefore they need data-driven approaches that might facilitate the detection of outliers and issues.

During this dissertation we analyzed the relationship of different variables with learning outcomes. In Section 6.1 we built a model to predict learning gains of students and we found that some behaviors had a negative effect towards learning gains. For example, we learned that being unreflective (*unreflective user*) was found to be bad for learning achievement. As a result, a recommender system could send a warning to students when this behavior is detected advising them to reflect more on their previous attempts. Another potential example is that if a student is detected to be avoiding videos (*video avoidance*) and still failing to solve a certain exercise, the recommender system can advise him to watch the video which actually explains the concepts associated with the exercise he is struggling with. Additionally, during the last sections we discussed about those students that are solving the same easy exercises systematically to earn more *repetitive badges*, thus showing very high values of *intentionality repetitive badges*. It could also be in our best interest to send warnings to students whenever they are detected carrying out this noxious behavior. Furthermore, this system could also encourage good behaviors, such as for example the use of goals or social activity functionality. As part of Section 6.2, we approached the implementation of machine learning models with the objective of early predicting certificate accomplishment with the idea of providing tools, such as a warning system, that can intervene and help the student get back on track before it is too late. We tried several machine learning models and found that in our case, the GBM model was overall the best fit, however since we saw differences in these models as more data becomes available, maybe it would be interesting to consider using different models for predictions at different time frames of the course. We also found that the importance of variables changed greatly at the different time frames of the course i.e., at the beginning of the course participatory variables (like *number active sessions* or *number events*) were very important, while as the course advances their importance decreases in favor of *exercise effectiveness*. We can use these insights for the implementation of a early warning system. In this direction, even though we followed a hard class prediction approached in Section 6.2 (we selected the class with the probability above 0.5), it is possible to follow a soft prediction approach based on probabilities and we can adapt the different warnings according these probabilities e.g., a very high probability of not getting a certificate might receive a strong warning whereas a threshold probability can receive a more moderate warning, and those students who are doing really good with a very low probability will not receive a warning at all (or maybe an encouraging message).

Finally, in Section 6.3 we implemented a model to predict submission events as CAMEO or not, which achieved a sensitivity and specificity of 0.96 and 0.99 respectively. We used VSURF package for feature selection and to analyze the importance of variables. Despite all variables provided useful information regarding CAMEO, student features had the highest importance, followed by submission and problem features. Using these findings we can work towards implementing a run-time detector without the use of IP. Since our model in Section 6.3 used all data available as a retrospective analysis we would need to develop a proof on concept to analyze its viability by adapting student features to take into account only the data available until the very instant of time of each submission, which could potentially lead to some deterioration of the quality metrics of the model. Finally, we could run the model in real time whenever a submission is made. For example, sending a warning after 3 submissions detected as CAMEO in a row, since the sensitivity of the model is very high the probability of this event happening by mistake is very low.

7.5. Evaluation of Visual Analytics Dashboard

In order to gain insight regarding the effectiveness and usability of ANALYSE (Subsection 3.1.2.2), we performed an evaluation of the tool. The respondents were 40 students taking ‘Design of Telematics Applications’ class, which is part of the Telecommunications Engineering master’s degree at UC3M. We expect that ANALYSE can be used by non-technical users with no additional training and that its visualizations are effective and usable. More information regarding the specific visualizations of ANALYSE can be consulted in our original research papers about the tool (Ruipérez-Valiente et al., 2016, 2017). The complete questionnaire that the respondents received is in Appendix A.

7.5.1. Overview of the Survey

The survey objective is to evaluate the usability, usefulness and effectiveness of ANALYSE. The intervention took place for around 60 minutes with the following phases:

1. Initial interaction with a typical course using Open edX (about 8 minutes).

2. Initial interaction with ANALYSE (about 7 minutes).

3. Respondents interacted with the 12 visualizations of ANALYSE with the role of instructor. Students had to respond the questions from 'Block 1' (see Appendix A for all specific questions) of the survey by interacting with each visualization to complete a task, and therefore obtaining the proper conclusions to respond the question (about 25 minutes).

4. Respondents had to rate in a 5-point Likert scale the degree of usefulness of each one of the 12 visualizations and also three general questions about the usefulness of ANALYSE ('Block 2' and 'Block 3', about 5 minutes).

5. Respondents were asked the 10 questions of the SUS survey (Brooke, 1996) to evaluate the general usability of the tool ('Block 4', about 5 minutes).

6. Finally, the respondents received two open questions regarding the best features and potential improvements for ANALYSE ('Block 5', about 5 minutes).

As a summary the participants solved the 39 questions survey within Appendix A, which included 12 questions to measure the effectiveness of the visualizations, 3 questions about the general usefulness of ANALYSE, 12 questions regarding the usefulness of each visualization, 10 questions regarding the general usability and 2 qualitative open questions about the best features and potential improvements.

7.5.2. Discussion about the Survey

This subsection discusses the results of the survey. Table 7.1 contains the results for all questions of the survey in Appendix A divided by blocks as we described in previous Subsection 7.5.1. For the questions of 'Block 1' where respondents had to carry out a task we report the percentage of correct answers, for the rest of 5-Likert scale questions involving 'Blocks 2, 3 and 4', we report the mean and standard deviation of the answers.

First in terms of the effectiveness of the tool we analyze questions belonging to 'Block 1' in Table 7.1. As we can see, most of the questions are above a 90% of effectiveness except for question 11. We can infer that most of the visualizations are effective and easy to use, even for users that interact for the first time with ANALYSE. The problem with question 11 is that students had to interact with 'Chapter Time' visualization by explicitly clicking to switch from 'total time' in the chapter to the 'graded time' which was asked by the question, most students did not perform that action and directly responded the value of the 'total time'. Therefore, our lesson learned here is that we should rethink how to make clear to users the available interaction with the different visualizations. Actually, this issue was also addressed in some of the open questions as part on 'Block 5'.

In terms of the visualization usefulness based on 'Block 2', we believe that the results are positive since the average value is 3.69/5. All visualizations are close or above a mean value of 3, thus we can argue that the respondents think that all visualizations have a potential use. More particularly, the most useful visualizations were 'Course Summary' and 'Students Grades' (4.43 and 4.45 respectively), we think this matches our initial ideas since these are the most straightforward and meaningful visualizations related to the learning process. The least useful visualization was 'Video Event Distribution' (2.95), which we believe that this is due to a bad visualization design that we actually improved in the next version of ANALYSE developed after this survey. In the case of the global usefulness based on 'Block 3' the results are good being all average values above 4.2, thus respondents considered ANALYSE to be globally helpful to track students' progress during their learning progress as well as to detect problems and issues in educational resources.

The 'Block 4' implemented the SUS questionnaire which has been extensively used in the literature to evaluate the usability. We obtained a total score of 78.4, which according a study developed by Sauro (2011) it is within the 15% better percentile. These results are really good and seem to indicate that ANALYSE is usable. There are other studies that have applied SUS to their learning analytics platforms such as LARAe system (Charleer et al., 2014) with an SUS score of 76 or the SAM tool (Govaerts et al., 2012) with a score of 71.36.

Finally, 'Block 5' contained the two open questions. The first one regarding the most useful features threw many positive statements such as "I knew everything had happened", "It permits knowing where students fail more generally in order to detect problems", "I was able to see the evolution of students from the very beginning of the course" or "I am able to see my progress and compare it with the rest of the students". We think that the respondents were able to see the potential use of ANALYSE, and benefits for awareness and self-reflection, which have been reported as one of the main benefits from learning analytics tool in the literature (Govaerts et al., 2012; Govaerts, Verbert, Klerkx & Duval, 2010). We also found several positive comments regarding the usability of the tool such as "The app was very intuitive and I did not need any previous knowledge to make use of it" or "The best features are that is really easy to use and the interactive visualizations makes it very intuitive", which is also in line with the high SUS score that we have reported. The second question regarding which features could be improved provided interesting insight from the respondents, such as the implementation of social activity visualizations or adapting language to the user (note that ANALYSE is in English while the respondents were Spanish). Several respondents also mentioned 'Repetitions of Video Intervals' visualization, some of them

Table 7.1: Results of the evaluation of ANALYSE based on the questions in Appendix A.

BLOCK 1					
#	**Percentage correct**		**#**	**Percentage correct**	
1	100 %		**2**	100 %	
3	92.5 %		**4**	100 %	
5	95 %		**6**	100 %	
7	97.5 %		**8**	97.5 %	
9	95 %		**10**	97.5 %	
11	30 %		**12**	95 %	
BLOCK 2					
#	**Mean**	**Std. deviation**	**#**	**Mean**	**Std. deviation**
13	4.42	0.75	**14**	4.45	0.68
15	3.8	0.88	**16**	3.38	1.13
17	3.48	1.26	**18**	3.43	1.15
19	2.95	1.22	**20**	3.45	0.99
21	3.73	1.11	**22**	3.8	0.94
23	3.8	0.91	**24**	3.5	1.3
BLOCK 3					
#	**Mean**	**Std. deviation**	**#**	**Mean**	**Std. deviation**
25	4.2	0.76	**26**	4.38	0.77
27	4.35	0.89			
BLOCK 4					
#	**Mean**	**Std. deviation**	**#**	**Mean**	**Std. deviation**
28	3.98	0.69	**29**	2	0.92
30	3.98	0.88	**31**	1.75	1.02
32	4.05	0.59	**33**	1.63	0.62
34	3.95	0.84	**35**	1.58	0.89
36	3.83	0.77	**37**	1.45	0.84

regarding positive feedback and others indicating that it was hard to interpret. Therefore, it looks like this visualization generated a bit of controversy, which it is also in line with the high standard deviation of the usefulness rating of this visualization (1.26), hence we might want to revise it in the future.

Chapter 8

Conclusions and Future Work

Contents

This last chapter of the thesis comprises final conclusions (Section 8.1) and some ideas regarding future work (Section 8.2).

8.1. Conclusions

This section is divided first in Subsection 8.1.1 which presents some final remarks and limitations of our work, Subsection 8.1.2 mentions the research projects that have used part of our findings, and Subsection 8.1.3 describes the research stays that the author has performed.

8.1.1. Final Remarks and Limitations

During this dissertation, we contributed to the area of learning analytics visualization dashboards. We analyzed the effectiveness, usefulness and usability of ANALYSE with 40 respondents and the results showed high average values for the three areas. The free responses of the respondents were also positive and with good ideas for future work. One specific issue was that respondents needed to interact and click in a visualization to obtain advanced information receive a low correctness ratio. They felt confused and only 30% of them solved this question correctly. Therefore, our advice would be that for those visualizations that enable interactivity, users should receive clear information regarding the availability of that functionality.

We explored the activity and behavior of students with regular and optional activities. Our findings suggest that there are some students with a more visual (only videos) or active (only exercises) learning profiles, however most of them had a more balanced learning using both educational resources. We also found some interesting correlations between effectiveness and some behavioral patterns. In terms of optional activities, the use was much lower in general than for regular activities. However, we did find some students that showed a nice interest for these activities. The optional activities that were used with more frequency were those not related to learning, avatar and display badges, a reasonable hypothesis is that these activities require less effort than others such as commenting, thus we would recommend instructors to encourage the use of optional activities related to learning. Although the correlation between optional activities and learning outcomes was weakened once controlling for third variables such as the time, we believe that there might exist an indirect effect, and optional activities can help generate motivation, which might imply better learning as supported in previous works. As limitations, we can argue that we do not have final conclusions but hypothesis and more work would be required to confirm some of the hypotheses established as part of these findings. Furthermore, some of the correlations might be spurious correlations i.e., they might be caused by one (or more) additional factors. Finally, the studies were observational, thus, we cannot conclude on any causal relationships.

We used the different variables related to the activity and behavior of students with the platform to build two prediction models of learning outcomes, one of learning gains and the other for certificate accomplishment. For the first one to predict learning gains we used a multivariate linear regression that were able to predict 68% of the learning gains variability. We were able to increment the prediction power of previous work by adding behavioral variables that we found might be negative for the learning achievement of students e.g., *forgetful user* and *unreflective user*. One of the main limitations of our model was the heavy importance of the prior knowledge of students (*pre test score*) in the prediction model, which might not always be feasible to have available. We showed how if we remove *pre test score* variable, the prediction power would decrease greatly, thus a challenge would be to maintain a good prediction power without the use of this variable. Additionally, we would need to replicate this research using other MOOC platforms such as Coursera or Open edX in order to be able to evaluate the generalization of these findings. We also analyzed the problem of high attrition rates in MOOCs by developing early prediction models of certificate accomplishment. We used several MLs models and found that GBM was the best suited in terms of performance and stability as more data became available. Nevertheless, different models performed better at different time frames, which might raise the idea that it can be interesting to use different models at different time frames. From our perspective the most interesting finding was that in the first weeks participatory variables were even more important than effectiveness variables, which we believe it is very important if we want to develop an early warning system. One of the main limitations is that both training and test data sets rely on data from a course which was already finished in a retrospective analysis and we have not been able to compare with other courses. We believe that the most immediate next step would be to test the model on a second re-run from the same course, to determine whether the prediction model can at least extrapolate to courses with similar contents and course structure. Finally, we could use these findings with the purpose of including a behavioral recommender system or a warning system in our LA dashboards ALAS-KA or ANALYSE.

We analyzed the use and behavior of students with badges. We found students earning many badges e.g., 80 students who earned 100 badges or more, which can be an indication of interest. In order to delve into their behavior and real motivations, we applied algorithms that try to infer their interest, concentration and efficiency towards badges. We explored these results finding students that indeed showed a very high interest and concentration, while others did not, even when they invested similar amount of effort in the platform. Additionally, we analyzed the correlations between these indicators and others, founding that badge metrics are correlated to other activity and effectiveness metrics. Similar to our hypotheses with optional activities, the use of badges is not going to increment students' learning. However, the use of badges might lead students to be more engaged and participative. Our results indicate that those students that were interested in badges and gamification, manage to spend more time in the platform and progressed more with educational resources. Additionally, we also recommended the use of badges to encourage certain behaviors and actions that might be carried out with less frequency. As an example, we found that social activity and goals were underused in our experiments, hence we could target this specific positive behaviors by adding badges that would stimulate students to use these optional activities. Nonetheless, we also did special emphasis on the fact that badges are external motivators and as such, we found some students that were solving the same exercises consistently to earn more *repetitive badges* and these actions would not lead to additional learning. Therefore, the gamification approach should focus on promoting intrinsic motivation so that students do not miss the real purpose and goal, which is learning. Additionally, we note out two important characteristics, the first one is that most students were around 17-18 years old and the second is that the use of Khan Academy was non-mandatory. We believe that these two characteristics can lead to a high participation with gamification elements, since they are young (thus familiar with video games) and voluntarily joined this experiment. If we change some of these characteristics we might find different results e.g., we could expect that older and busier people might not have the interest to spend an extra time in the platform because of badges.

In terms of our contribution on online academic dishonesty and illicit collaborations, we investigated CAMEO cheating method, in which students are found to create fake accounts to harvest correct solutions that are later on used with their master account to gain credit, and ultimately a certificate of accomplishment in many cases. We implemented an algorithm based on heuristics to detect CAMEO and applied it to a MOOC in introductory physics, finding that 12.9% (65 accounts) of the certificate earners used this methodology to copy at least 10 questions. We found profiles of students applying CAMEO more as a back up plan and help mode, but others using it systematically and very deliberately to obtain a certificate. We believe that CAMEO is already a significant issue that can really threaten the value of certificates in the long run. Regarding these results we applied our algorithm only to one MOOC, therefore, in order to get a better estimation and improve generalization we plan to extend the research across a larger sample of courses. In any case, other independent studies (Northcutt et al., 2016) have found the prevalence of CAMEO across a bigger MOOC portfolio, therefore we do not hold any doubts and this is most likely a global problem. Additionally, we also designed a RF classification model based on student, submission and problem features. This model has a high performance achieving a sensitivity and specificity of 0.96 and 0.99 respectively. Additionally, the model does not rely on the IP of the submissions, which was one important requirement of our original algorithm based on heuristics. Since we achieved such high quality metrics, we used the model to predict on 'suspicious' events, which were those submissions that did not pass all filtering criteria established by our detection algorithm (see Subsection 3.3.7.2) but were still suspicious, and we found that 9.9% of those events were classified as CAMEO, indicating that our previous estimate regarding CAMEO prevalence might need to increment in approximately a 37%. Finally, we used VSURF package for feature selection and ranking the variable importance of the different features based on RF. Our findings suggest that student features are more important than those related to the problem and the submission, which seems to indicate that the student has more influence to detect cheating as was found also in the context of gaming the system (Muldner et al., 2011). As part of the limitations of this model is that we used both for the training and evaluation the tagged sample by our previous algorithm (Subsection 3.3.7.2), therefore we rely on its initial effectiveness to detect CAMEO. Additionally, we have trained and evaluated the model using data from the same MOOC that, although is a perfectly valid approach to evaluate the model, impedes us to generalize these results to other MOOCs.

In this same direction of academic dishonesty we also developed an algorithm and method to detect accounts that submit their assignments close in time in online learning environments (see Subsection 3.3.7.1). We used this algorithm on two MOOCs and label the accounts detected by the algorithm as *close submitters*. We presented that most *close submitters* are grouped in couples of accounts, but other bigger communities were found as well. We showed that it was statistically significant that those accounts labeled as *close submitters* were able to achieve a certificate of accomplishment being active less days, with less submission attempts, watching fewer videos and accessing less forum threads, which we hypothesized that can be a clear indication that those students were collaborating or engaged in some academically dishonest behavior. As next step we used clustering to detect different types of accounts and using these cluster assignments we discussed the pedagogical associations in the different couples and communities that we detected. For example, we found some couples that resembled a CAMEO association with two accounts run by the same student. In other cases, we found a more hard-working student with a leading role investing a lot of time in the course and sharing the solutions with a peer friend that did not interacted with the course at all. We also discussed other potential associations in Section 5.3.2. We reported some examples of communities, for example one noticeable community of 5 accounts that in 68 minutes were able to solve the 5 quizzes in a MOOC and achieve certification. We discussed how harmful the different associations might be for the learning process, despite we argue that some real collaborations might even be positive for their learning. As main limitation we point out that we are conveying reasonable hypotheses but we have not been able to confirm our findings with more objective criteria like contacting those students.

Overall, we think that if not addressed properly the problem of academic dishonesty can endanger the future MOOC sustainability. First, we showed that CAMEO but also other illicit collaborations are already a significant issue in MOOCs and that it could jeopardize the validity of the MOOC certificate system. Second, since these accounts have outlier behaviors, and usually either very high or very poor performance, their actions and behavior can heavily and systematically interfere with MOOC research if not removed from the data sample. Third, cheating is usually associated with poor learning, and we found that in most cases the motivation of those students is to obtain a certificate without actually investing any effort and learning the contents. Finally, despite the research was conducted in MOOCs, our findings and conclusions might be similar for other learning environments that allow users to register additional accounts and/or that provide some kind of feedback before graded assignment deadlines.

8.1.2. Research Projects

Some of the results achieved during this dissertation have been transferred to the following projects:

- **Educational Reflected Spaces (EEE)**[1]:
    - Funding organization: Spanish Ministry of Science and Innovation, National I+D+I plan under grant TIN2011-28308-C03-01.
    - Partners: Universidad Carlos III de Madrid, Universitat Pompeu Fabra and Universidad de Valladolid.
    - Duration: January, 2012 – December, 2014.
- **eMadrid**[2]:
    - Funding organization: Regional Government of Madrid under grant S2013/ICE-2715.
    - Partners: Universidad Carlos III is the network coordinator and works jointly with the Autónoma de Madrid, Complutense de Madrid, Politécnica de Madrid, Rey Juan Carlos and UNED universities
    - Duration: October, 2014 – October, 2018.
- **Reformulate Scalable Educational Ecosystems Offering Technological Innovations (RESET)**[3]:

[1] http://eee.gast.it.uc3m.es/
[2] http://www.emadridnet.org/
[3] http://reset.gast.it.uc3m.es/

• Funding organization: Spanish Ministry of Economy and Competitiveness project RESET under grant TIN2014-53199-C3-1-R.

• Partners: Universidad Carlos III de Madrid, Universitat Pompeu Fabra and Universidad de Valladolid.

• Duration: January, 2014 – December, 2017.

- **Supporting Higher Education to Integrate Learning Analytics (SHEILA)**[4]:

• Funding organization: Erasmus+ Programme of the European Union under grant 562080-EPP-1-2015-BE-EPPKA3-PI-FORWARD.

• Partners: The University of Edinburgh, Brussels Educational Services, Open University of the Netherlands, Tallinn University, Universidad Carlos III de Madrid, European Association for Quality Assurance in Higher Education and Erasmus Student Network.

• Duration: January, 2016 – July, 2018.

- **Spanish Network of Learning Analytics (SNOLA)**[5]:

• Funding organization: Spanish Ministry of Economy and Competitiveness project RESET under grant TIN2015-71669-REDT.

• Partners: Universidad de Salamanca, DeustoTech Learning, Universidad de León, Universidad del País Vasco, Universidad Politécnica de Madrid, Universidad de Valladolid, Universidad Carlos III de Madrid, Universidad de Vigo and Universidad Nacional de Educación a Distancia.

• Duration: January, 2016 – December, 2017.

8.1.3. Research Stays

As part of the research carried out during this dissertation the author performed two research stays at the following high quality universities:

1. The first research stay was from the 2nd of June (2015) to the 4th of September (2015) at the Physics department of MIT where the author joined Research in Learning, Assessing and Tutoring Effectively (RELATE) group[6] and was supervised by Prof. Dr. David E. Pritchard[7] which holds a position as Cecil and Ida Green Professor of Physics at MIT. As a result one workshop presentation (Alexandron et al., 2015) and two articles (Ruipérez-Valiente et al., 2016; Alexandron et al., 2017) have been published already. Additionally,

[4] http://sheilaproject.eu/
[5] https://snola.es
[6] http://relate.mit.edu/
[7] http://web.mit.edu/physics/people/faculty/pritchard_david.html

this is still an ongoing collaboration and we have an additional article which is currently under review (Ruipérez-Valiente et al., 2017).

2. The second research stay was from the 19th of September (2016) to the 21st of December (2016) at the School of Informatics at the University of Edinburgh where the author joined the Institute for Adaptive and Neural Computation[8] and was supervised by Prof. Dr. Dragan Gašević[9] which is Professor and Chair in Learning Analytics and Informatics in the Moray House School of Education and the School of Informatics at the University of Edinburgh. We have already one paper published (Ruipérez-Valiente et al., 2017a). Additionally, we are working on one additional journal paper that will be submitted soon (Ruipérez-Valiente, Joksimović, Kovanović, Gašević, Muñoz-Merino & Delgado Kloos, 2017b).

8.2. Future Work

There are many potential areas of future work as a follow up of the results achieved during this dissertation. In the direction of visual analytics with our platforms ALAS-KA and ANALYSE, one potential future action is the design and implementation of new and novel indicators as well as new visualizations e.g., a more in depth analysis of the social activity in these environments including for example social network analysis visualizations of their activity in the discussion forums using tools such as qgraph[10]. We would also like to integrate other tools to perform more advanced statistical analysis. Generally speaking, we would also want to implement several of the recommendations received by the respondents of the evaluation survey such as enabling different languages and clearing up the available interactivity of each visualization. A very challenging project would be to develop a modular and common framework of indicators and visualization that can be easily adapted to different courses and platforms, this would increment interoperability greatly. Another interesting line of work would be to compare the usability of these platforms between 'digital natives' and 'forced digital immigrants'.

Based on our findings in prediction models and detectors we would like to integrate our lessons learned within ALAS-KA and ANALYSE. For example, we can introduce a recommender system that when negative behaviors are detected, it can provide feedback to the student regarding how to improve his past negative action or an early warning that when a student is detected to be in risk of not achieving a certificate send a warning, so that the student can turnover this situation before it is too late. As part of our prediction model of certificate accomplishment, we analyzed the evolution of this model per week, but we would also like to carry out a more granular approach, maybe a day-to-day analysis which would also help into analyzing the effect of other variables such as deadlines or days of the week. We would also like to try out forecasting algorithms that can be useful to find trends in the evolution of data as time goes by. One of the

[8] http://www.anc.ed.ac.uk/
[9] http://www.ed.ac.uk/profile/dragan-gasevic
[10] http://sachaepskamp.com/qgraph

most challenging projects would be to delve into the development of prediction models that can extrapolate to different topics, courses and even platforms. One idea to accomplish that would be the use of ratios instead of absolute variables, and use those as the input variables of the prediction model. The final stage of these ideas would be to perform A/B experiments in which a treatment group would receive some sort of recommendation or warning (e.g., positive or negative behaviors, warning regarding risk of not getting a certificate) and the control group would not; then we can analyze if the treatment group improved their learning outcomes compared to the control group (higher learning gains or higher certificate accomplishment ratio).

In terms of the interaction of students with regular activities we would like to work towards defining new metrics that are able to efficiently characterize the effectiveness of students. For that purpose new experiments including pre-test and post-test to estimate students' learning might be good, since we would be able to analyze which effectiveness metrics are more correlated with the actual learning achievement of students. In terms of optional activities, we plan to formulate a common framework that can be applicable to other VLEs and also replicate our experiments in other platforms. We also want to conduct A/B testing experiments with optional activities where some students will have them available and others will not, so that we can measure with more reliability the actual impact of optional activities in the learning process of students.

Related to the behavior with badges, we would like to extend our models to be applicable to more badge types. Additionally, we would like to replicate our experiment in other environments to see if we obtain similar findings that can help us generalize. Another interesting direction would be to survey students regarding their interest about badges, and check for correlations between our metrics and the reported interest. This can be helpful to validate the effectiveness of our badge metrics. Finally, we would also like to delve into assessing the influence of interest in badges on learning achievement, specially in those cases of students that had really high *intentionality repetitive badges*.

Our research on online academic dishonesty also opens many potential future work lines. The most straightforward would be to examine both in terms of CAMEO and *close submitter* algorithms a larger portfolio of MOOCs, but also maybe other types of online learning courses, such as on-campus blended learning or corporate training. We can gain insight in terms of generalization and to find trends across courses depending on other factors such as university, topic or platform delivering the course. A broader analysis might enable also the implementation of ML models that can work well across courses, this would permit developing run-time detectors that could be embedded within VLEs or learning analytics systems such as ANALYSE. We would also like to corroborate our findings and the effectiveness of our detectors by interviewing some of the detected students, although this would problematic for several reasons such as students might feel that their rights are violated, or they might also lie due to embarrassment and fear to potential repercussions due to their illicit behavior. On a more general way, there are probably many other methods that students might be using to facilitate their way into certificates, thus more work towards developing detectors would be useful. Finally, pedagogy-wise we would like to quantify the effect that each one of the different associations and different illicit behaviors have on learning achievement. Although generally speaking cheating is often related to poor learning, we might be surprised to find that some behaviors, such as using CAMEO as support, might not be harmful for learning, or that some associations, such as sharing answers with friends but taking the course seriously, might even be beneficial for learning.

Appendix A

Evaluation Questionnaire for ANALYSE

- **BLOCK 1**. Specific questions that require that you interact and carry out a certain task with ANALYSE in order to be able to give an adequate response. Each question is associated with an specific visualization:

1. Course Summary. What percentage of students have acquired a proficiency grade in ‘Homework’ assignments.
2. Students’ Grades. What is the grade of student ‘Verified’ in the midterm exam?
3. Problem Time Distribution. In which problems have the student ‘Audit’ spent the most time? How much time?
4. Video Time Watched. Which video has highest difference between different video watched and total video watched by all students?
5. Repetitions of Video Intervals. In the video ‘Radioactive’, which approximate interval of seconds has been watched more times by all the class?
6. Video Time Distribution. In which video did the student ‘Staff’ spend more time?
7. Video Events Distribution. In the video ‘Passenger - Let her go’, what is the approximate range of seconds where we can find more ‘Change Speed’ events by ‘Audit’ student?
8. Exercise and Video Progression. Check the progression for the student ‘jruipere’. Which score was higher on the date ‘15/02/2015’, video progress or exercise grades?
9. Daily Time on Exercises and Videos. Which day did users engage the largest amount of time on problems with the platform? How much time?
10. Course Accesses. Which section of the course has the highest number of accesses by student ‘Honor’?

11. Chapter Time. How much ‘Graded time’ time has been spent by all students in section 2?
12. Students Time Schedule. Which is the time interval in which the student ‘JoseRuiperez’ has us28ed the platform the most? How many minutes?

- **BLOCK 2**. Now that you have used all visualizations, rate from 1 (not useful at all) to 5 (very useful) the usefulness of each one of the visualizations from the perspective of an instructor:

13. Course Summary.
14. Students’ Grades.
15. Problem Time Distribution.
16. Video Time Watched.
17. Repetitions of Video Intervals.
18. Video Time Distribution.
19. Video Events Distribution.
20. Exercise and Video Progression.
21. Daily Time on Exercises and Videos.
22. Course Accesses.
23. Chapter Time.
24. Students Time Schedule.

- **BLOCK 3**. Specific questions about the usability of the learning analytics extension. Mark from 1 (strongly disagree) to 5 (strongly agree) your degree of agreement with the following statements from the perspective of an instructor:

25. I think that the use of this application would help to evaluate students that are taking an online course more easily.
26. I think these visualizations are useful and help understand the learning process of students.
27. I think that these visualizations can be used to detect problems in learning resources of an online course (exercises, videos, etc).

- **BLOCK 4**. System Usability Scale questionnaire. Mark from 1 (strongly disagree) to 5 (strongly agree) your degree of agreement with the following statements from the perspective of an instructor:

28. I think that I would like to use this web application frequently.
29. I found the web application unnecessarily complex.

30. I thought the web application was easy to use.
31. I think that I would need the support of a technical person to be able to use this web application.
32. I found the various functions in this web application were well integrated.
33. I thought there was too much inconsistency in this web application.
34. I would imagine that most people would learn to use this web application very quickly.
35. I found the web application very cumbersome to use.
36. I felt very confident using the web application.
37. I needed to learn a lot of things before I could get going with this web application.

- **BLOCK 5**. Questions regarding best features and potential improvements allowing an open text response:

38. Which features and/or visualizations do you think are the most useful/important after interacting with ANALYSE?
39. Which features could be improved and what new functionality could be implemented to improve ANALYSE?

Appendix B

Publications by the Author

B.1. JCR-Indexed

The journal publications indexed in the Journal Citations Report (JCR) are the following:

1. **Ruipérez-Valiente, J.A.**, Muñoz-Merino, P.J., Leony, D. & Delgado Kloos, C. 2015. ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. *Computers in Human Behavior*. 47, 139–148 (Ruipérez-Valiente et al., 2015). Impact factor as of 2015: 2.88, JCR(21/129), Q1, category: psychology, multidisciplinary.

2. Leony, D., Muñoz-Merino, P. J., **Ruipérez-Valiente, J. A.**, Pardo, A., Martín-Caro, D. A., & Delgado Kloos, C. (2015). Detection and evaluation of emotions in Massive Open Online Courses. *Journal of Universal Computer Science*, 21(5), 638–655 (Derick Leony et al., 2015). Impact factor as of 2015: 0.546, JCR(89/106), Q4, category: computer science, software engineering.

3. Muñoz-Merino, P. J., **Ruipérez-Valiente, J. A.**, Alario-Hoyos, C., Pérez-Sanagustín, M., & Delgado Kloos, C. (2015). Precise Effectiveness Strategy for analyzing the effectiveness of students with educational resources and activities in MOOCs. *Computers in Human Behavior*, 47, 108–118 (Muñoz-Merino et al., 2015). Impact factor as of 2015: 2.88, JCR(21/129), Q1, category: psychology, multidisciplinary.

4. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., Delgado Kloos, C., Niemann, K., Scheffel, M., & Wolpers, M. (2016). Analyzing the Impact of Using Optional Activities in Self-Regulated Learning. *IEEE Transactions on Learning Technologies*, 9(3), 231–243 (Ruipérez-Valiente et al., 2016). Impact factor as of 2015: 1.129, JCR(78/231), Q2, category: education & educational research.

5. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., Gascón-Pinedo, J. A., & Delgado Kloos, C. (2016). Scaling to Massiveness with ANALYSE: A Learning Analytics Tool for Open edX.

IEEE Transactions on Human-Machine Systems (Ruipérez-Valiente et al., 2016). Impact factor as of 2015: 1.8, JCR(5/22), Q1, category: computer science, cybernetics.

6. Muñoz-Merino, P. J., **Ruipérez-Valiente, J. A.**, Delgado Kloos, C., Auger, M. A., Briz, S., Castro, V. De, & Santalla, S. N. (2016). Flipping the Classroom to Improve Learning with MOOCs Technology. *Computer Applications in Engineering Education*, 25(1), 15–25 (Muñoz-Merino et al., 2016). Impact factor as of 2015: 0.935, JCR(45/85), Q3, category: engineering, multidisciplinary.

7. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., Pijeira Díaz, H. J., Santofimia Ruiz, J., & Delgado Kloos, C. (2017). Evaluation of a Learning Analytics Application for Open edX Platform. *Computer Science and Information Systems*, 14(1), 51–73 (Ruipérez-Valiente et al., 2017). Impact factor as of 2015: 0.623, JCR(83/106), Q4, category: computer science, software engineering.

8. Alexandron, G., **Ruipérez-Valiente, J. A.**, Chen, Z., Pedro J. Muñoz-Merino, & Pritchard, D. E. (2017). Copying@Scale: Using Harvesting Accounts for Collecting Correct Answers in a MOOC. *Computers & Education*, 108, 96–114 (Alexandron et al., 2017). Impact factor as of 2015: 2.881, JCR(14/104), Q1, category: computer science, interdisciplinary applications.

9. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., & Delgado Kloos, C. (2017). Detecting and Clustering Students by their Gamification Behavior with Badges: A Case Study in Engineering Education (In press). *The International Journal of Engineering Education*, (Engineering Behind Technology-Based Educational Innovations) (Ruipérez-Valiente et al., 2017). Impact factor as of 2015: 0.559, JCR(61/85), Q3, category: engineering, multidisciplinary.

10. Muñoz-Merino, P. J., Méndez Rodríguez, E., Delgado Kloos, C., & **Ruipérez-Valiente, J. A.** (2017). Design, Implementation and Evaluation of SPOCs at the Universidad Carlos III de Madrid. *Journal of Universal Computer Science*, 23(2), 167–186 (Muñoz-Merino et al., 2017). Impact factor as of 2015: 0.546, JCR(89/106), Q4, category: computer science, software engineering.

B.2. International Conferences

The following papers have been published in international conferences:

1. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., Delgado Kloos, C., Niemann, K., & Scheffel, M. (2014). Do Optional Activities Matter in Virtual Learning Environments? In *Ninth European Conference on Technology Enhanced Learning* (pp. 331–344) (Ruipérez-Valiente et al., 2014). Springer International Publishing. Google Scholar h5-index factor: 16.

2. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., & Delgado Kloos, C. (2014). A demonstration of ALAS-KA: a learning analytics tool for the khan academy platform. In *Ninth European Conference on Technology Enhanced Learning* (pp. 518–521) (Ruipérez-Valiente et al., 2014). Springer International Publishing. Google Scholar h5-index factor: 16.

3. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., & Delgado Kloos, C. (2015). A Predictive Model of Learning Gains for a Video and Exercise Intensive Learning Environment. In *International Conference on Artificial Intelligence in Education* (pp. 760–763) (Ruipérez-Valiente et al., 2015a). Springer International Publishing. CORE Rank: A.

4. Pijeira, H. J., Santofimia, J., **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., & Delgado Kloos, C. (2015). Using Video Visualizations in Open edX to Understand Learning Interactions of Students. In *10th European Conference on Technology Enhanced Learning* (pp. 522–525) (Pijeira Díaz et al., 2015). Springer International Publishing. Google Scholar h5-index factor: 16.

5. Redondo, D., Muñoz-Merino, P. J., **Ruipérez-Valiente, J. A.**, Delgado Kloos, C., Pijeira Díaz, H. J., & Santofimia Ruiz, J. (2015). Combining Learning Analytics and the Flipped Classroom in a MOOC of maths. In *International Workshop on Applied and Practical Learning Analytics* (pp. 71–79) (Redondo et al., 2015). CEUR Workshop Proceedings.

6. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., & Delgado Kloos, C. (2016). An analysis of the use of badges in an educational experiment. In *Frontiers in Education Conference* (pp. 1–8) (Ruipérez-Valiente et al., 2016a). IEEE. CORE Rank: B.

7. Pijeira Díaz, H. J., Santofimia Ruiz, J., **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., & Delgado Kloos, C. (2016). A Demonstration of ANALYSE: A Learning Analytics Tool for Open edX. In *Proceedings of the Third ACM Conference on Learning@Scale* (pp. 329–330) (Pijeira Díaz et al., 2016). ACM. Google Scholar h5-index factor: 14.

8. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., & Delgado Kloos, C. (2016). Analyzing students' intentionality towards badges within a case study using Khan academy. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 536–537) (Ruipérez-Valiente et al., 2016b). ACM. Google Scholar h5-index factor: 32.

9. **Ruipérez-Valiente, J. A.**, Alexandron, G., Chen, Z., & Pritchard, D. E. (2016, April). Using multiple accounts for harvesting solutions in MOOCs. In *Proceedings of the Third ACM Conference on Learning@Scale* (pp. 63–70) (Ruipérez-Valiente et al., 2016). ACM. Google Scholar h5-index factor: 14.

10. **Ruipérez-Valiente, J. A.**, Joksimović, S., Kovanović, V., Gašević, D., Muñoz-Merino, P. J., & Delgado Kloos, C. (2017). A Data-driven Method for the Detection of Close Submitters in Online Learning Environments. In *Proceedings of the 26th International Conference*

on World Wide Web Companion (pp. 361–368) (Ruipérez-Valiente et al., 2017a). ACM. CORE Rank: A*.

11. **Ruipérez-Valiente, J. A.**, Cobos, R., Muñoz-Merino, P. J., Andújar, A., & Delgado Kloos, C. (2017). Early Prediction and Variable Importance of Certificate Accomplishment in a MOOC (Ruipérez-Valiente et al., 2017). In *Proceedings of the Fifth European MOOCs Stakeholders Summit*. Springer International Publishing.

B.3. Under Review or Work in Progress

Several publications are still under review or as work in progress to be submitted:

1. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., and Delgado Kloos, C. Improving the Prediction of Learning Outcomes in Educational Platforms including Higher Level Interaction Indicators (Under review). *Expert Systems*.

2. **Ruipérez-Valiente, J. A.**, Muñoz-Merino, P. J., Alexandron, G., and Pritchard, D. E. Using Machine Learning to Detect "Multiple-Account" Cheating and Analyze the Influence of Student and Problem Features (Major revision). *IEEE Transactions on Learning Technologies*.

3. **Ruipérez-Valiente, J. A.**, Joksimović, S., Kovanović, V., Gašević, D., Muñoz-Merino, P. J., and Delgado Kloos, C. (2017). Collaboration or "Collaboration": Detecting and Characterizing Communities of Accounts in MOOCs (Work in Progress). *IEEE Transactions on Learning Technologies*.

Abbreviations

ALAS-KA Add-on of the Learning Analytics Support of the Khan Academy - See Subsection 3.1.2.1 for details.

ANALYSE Add-on of the Learning Analytics Support for Open edX - See Subsection 3.1.2.2 for details.

ANOVA One-way Analyses of Variance - Read more https://en.wikipedia.org/wiki/Analysis_of_variance.

API Application Programming Interface.

AUC Area Under the ROC Curve - Read more http://gim.unmc.edu/dxtests/roc3.htm.

CAMEO Copying Answers using Multiple Existences Online - See Subsection 4.5.2 for details.

CSCL Computer-Supported Collaborative Learning.

CSV Comma-Separated Values.

EDM Educational Data Mining - See Section 2.2 for details.

GAE Google App Engine - Platform as a Service infrastructure. Read more https://cloud.google.com/appengine/.

GBM Gradient Boosting Machine - Read more https://en.wikipedia.org/wiki/Gradient_boosting.

ITS Intelligent Tutoring System - See Section 2.2 for details.

JCR Journal Citations Report.

JSON JavaScript Object Notation.

kNN k-Nearest Neighbours - Read more https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

LA Learning Analytics - See Section 2.2 for details.

LMS Learning Management System - See Section 2.2 for details.

MAD Mean Absolute Deviation - See Equation 3.2.

MANOVA Multivariate Analysis of Variance - Read more https://en.wikipedia.org/wiki/Multivariate_analysis_of_variance.

MIT Massachusetts Institute of Technology - http://www.mit.edu/.

ML Machine Learning - Read more https://en.wikipedia.org/wiki/Machine_learning.

MOOC Massive Open Online Course - See Section 2.2 for details.

MSD Mean Squared Deviation - See Equation 3.3.

PLE Personal Learning Environment.

RELATE Research in Learning, Assessing and Tutoring Effectively group at MIT – http://relate.mit.edu/.

RF Random Forests - Read more https://en.wikipedia.org/wiki/Random_forest.

ROC Receiver Operating Characteristic - Read more https://en.wikipedia.org/wiki/Receiver_operating_characteristic.

SPOC Small Private Online Course - See Section 2.2 for details.

SUS System Usability Scale (Brooke, 1996).

SVM Support Vector Machine - Read more https://en.wikipedia.org/wiki/Support_vector_machine.

UAM Universidad Autónoma of Madrid - https://www.uam.es/ss/Satellite/es/home/.

UC3M Universidad Carlos III of Madrid - http://www.uc3m.es/Home.

VLE Virtual Learning Environment - See Section 2.2 for details.

VSURF Algorithm for variable importance and selection using RF - Read more https://cran.r-project.org/web/packages/VSURF/VSURF.pdf.

XML eXtensible Markup Language.

Nomenclature

attempt duration See Subsection 3.3.4 for details.

attempts required See Subsection 3.3.4 for details.

average number attempts See Subsection 3.3.1 for details.

average number hints See Subsection 3.3.1 for details.

average time correct answer See Subsection 3.3.2 for details.

average time per day See Subsection 3.3.3 for details.

certificate See Subsection 3.3.2 for details.

close submitter See Subsection 3.3.7.1 for details.

completed videos See Subsection 3.3.2 for details.

concentration badges See Subsection 3.3.6 for details.

dispersion time per day See Subsection 3.3.3 for details.

dispersion time per exercise See Subsection 3.3.3 for details.

dispersion time per video See Subsection 3.3.3 for details.

exercise abandonment See Subsection 3.3.5 for details.

exercise effectiveness no help See Subsection 3.3.2 for details.

exercise effectiveness See Subsection 3.3.2 for details.

exercise time See Subsection 3.3.3 for details.

exercises accessed See Subsection 3.3.1 for details.

follow recommendations See Subsection 3.3.5 for details.

forgetful user See Subsection 3.3.5 for details.

harvested See Subsection 3.3.7.2 for details.

hint abuse See Subsection 3.3.5 for details.

hint avoidance See Subsection 3.3.5 for details.

intentionality repetitive badges See Subsection 3.3.6 for details.

intentionality topic badges See Subsection 3.3.6 for details.

learning gain See Case Study 3.2.1.2 for details.

location See Subsection 3.3.4 for details.

max attempts See Subsection 3.3.4 for details.

negative behaviors See Subsection 6.1.1 for details.

number active days See Subsection 3.3.1 for details.

number active sessions See Subsection 3.3.1 for details.

number attempts correct answer See Subsection 3.3.2 for details.

number events See Subsection 3.3.1 for details.

number submissions See Subsection 3.3.1 for details.

number threads viewed See Subsection 3.3.1 for details.

optional activities See Subsection 3.3.1 for details.

order See Subsection 3.3.7.1 for details.

page time See Subsection 3.3.3 for details.

percentage correct exercise See Subsection 4.4.2 for details.

performance first attempt See Subsection 3.3.2 for details.

pre test score Pre Test Score - See Case Study 3.2.1.2 for details.

pre test time Pre Test Time - See Case Study 3.2.1.2 for details.

proficient exercises See Subsection 3.3.2 for details.

random See Subsection 3.3.4 for details.

repetitive badge See Subsection 3.1.1.1 for details.

show answer See Subsection 3.3.4 for details.

time efficiency badges See Subsection 3.3.6 for details.

time to deadline See Subsection 3.3.4 for details.

topic badge See Subsection 3.1.1.1 for details.

total abandonment See Subsection 6.1.1 for details.

total time See Subsection 3.3.3 for details.

type assignment See Subsection 3.3.4 for details.

type response See Subsection 3.3.4 for details.

unreflective user See Subsection 3.3.5 for details.

video abandonment See Subsection 3.3.5 for details.

video avoidance See Subsection 3.3.5 for details.

video duration See Subsection 4.4.2 for details.

video effectiveness See Subsection 3.3.2 for details.

video time See Subsection 3.3.3 for details.

videos accessed See Subsection 3.3.1 for details.

References

Abramovich, S. & Schunn, C. (2011). Achievement Systems' to Boost Achievement Motivation. In *Games Learning Society Conference*, (pp. 3–10). Citeseer.

Abramovich, S., Schunn, C., & Higashi, R. M. (2013). Are badges useful in education?: It depends upon the type of badge and expertise of learner. *Educational Technology Research and Development*, *61*(2), 217–232.

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., & Addison, K. L. (2015). Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, (pp. 93–102). ACM.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking. In *International Conference on Intelligent Tutoring Systems*, (pp. 227–239). Springer.

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), 101–128.

Alexandron, G., Ruipérez-Valiente, J. A., Chen, Z., Muñoz-Merino, P. J., & Pritchard, D. E. (2017). Copying@Scale: Using harvesting accounts for collecting correct answers in a MOOC. *Computers & Education*, *108*, 96–114.

Alexandron, G., Ruipérez-Valiente, J. A., & Pritchard, D. E. (2015). Evidence of MOOC Students Using Multiple Accounts to Harvest Correct Answers. In *Learning with MOOCs II: A Workshop for Practitioners: New Approaches to Teaching & Learning*.

Aljohani, N. R. & Davis, H. C. (2012). Learning analytics in mobile and ubiquitous learning environments. In *11th World Conference on Mobile and Contextual Learning: mLearn 2012*.

Anderman, E. M., Cupp, P. K., & Lane, D. (2009). Impulsivity and academic cheating. *The Journal of Experimental Education*, *78*(1), 135–150.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2013). Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*, (pp. 95–106). ACM.

Anozie, N. & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In *American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06)*, (pp. 1–6)., Boston, EEUU.

Arnab, S., Berta, R., Earp, J., De Freitas, S., Popescu, M., Romero, M., Stanescu, I., & Usart, M. (2012). Framing the adoption of serious games in formal education. *Electronic Journal of e-Learning*, *10*(2), 159–171.

Atapattu, T., Falkner, K., & Tarmazdi, H. (2016). Topic-wise classification of MOOC discussions: A visual analytics approach. In *Proceedings of the 9th International conference on Educational Data Mining (EDM)*.

Baker, R. S., Corbett, A. T., Gowda, S. M., Wagner, A. Z., MacLaren, B. A., Kauffman, L. R., Mitchell, A. P., & Giguere, S. (2010). Contextual slip and prediction of student performance after use of an intelligent tutor. In *International Conference on User Modeling, Adaptation, and Personalization*, (pp. 52–63). Springer.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, (pp. 531–540). Springer.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, (pp. 383–390). ACM.

Baker, R. S., Gowda, S. M., & Corbett, A. T. (2011). Towards predicting future transfer of learning. In *International Conference on Artificial Intelligence in Education*, (pp. 23–30). Springer.

Baker, R. S., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185.

Balakrishnan, G. & Coetzee, D. (2013). Predicting student retention in massive open online courses using hidden markov models. *Electrical Engineering and Computer Sciences University of California at Berkeley*.

Balbuena, S. E. & Lamela, R. A. (2015). Prevalence, motives, and views of academic dishonesty in higher education. *Asia Pacific Journal of Multidisciplinary Research*, *3*(2).

Bansal, S. & Singh, P. (2015). Blending active learning in a modified SPOC based classroom. In *3rd International Conference on Innovation and Technology in Education*, (pp. 251–256). IEEE.

Barata, G., Gama, S., Jorge, J., & Gonçalves, D. (2013a). Engaging engineering students with gamification. In *5th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, (pp. 1–8). IEEE.

Barata, G., Gama, S., Jorge, J., & Gonçalves, D. (2013b). So fun it hurts – gamifying an engineering course. In *International Conference on Augmented Cognition*, (pp. 639–648). Springer.

Berkling, K. & Thomas, C. (2013). Gamification of a Software Engineering course and a detailed analysis of the factors that lead to it's failure. In *International Conference on Interactive Collaborative Learning (ICL)*, (pp. 525–530). IEEE.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, *19*(1-2), 205–220.

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the 1st international conference on learning analytics and knowledge*, (pp. 110–116). ACM.

Bogle, K. D. (2000). Effect of perspective, type of student, and gender on the attribution of cheating. *Proceedings of Oclahoma Academic Science. Oclahoma City*, *80*, 91–97.

Bolton, G. (2010). *Reflective practice: Writing and professional development*. Sage publications.

Borras-Gene, O., Martinez-Nunez, M., & Fidalgo-Blanco, A. (2016). New Challenges for the motivation and learning in engineering education using gamification in MOOC. *International Journal of Engineering Education*, *32*(1), 501–512.

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & evaluation in higher education*, *24*(4), 413–426.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, *8*.

Brinton, C. G. & Chiang, M. (2015). MOOC performance prediction via clickstream data and social learning networks. In *IEEE Conference on Computer Communications (INFOCOM)*, (pp. 2299–2307). IEEE.

Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z., & Wong, F. M. F. (2014). Learning about social learning in MOOCs: From statistical analysis to generative model. *IEEE transactions on Learning Technologies*, *7*(4), 346–359.

Broeckelman-Post, M. A. (2008). Faculty and student classroom influences on academic dishonesty. *IEEE Transactions on Education*, *51*(2), 206–211.

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4–7.

Brooks, C., Stalburg, C., Dillahunt, T., & Robert, L. (2015). Learn With Friends: The Effects of Student Face-to-Face Collaborations on Massive Open Online Course Activities. In *Proceedings of the Second ACM Conference on Learning@Scale*, (pp. 241–244). ACM.

Burge, J., Fox, A., Grossman, D., Roth, G., & Warren, J. (2015). Spocs: what, why, and how. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, (pp. 595–596). ACM.

Calvo-Flores, M. D., Galindo, E. G., Jiménez, M. P., & Piñeiro, O. P. (2006). Predicting students' marks from moodle logs using neural network models. *Current Developments in Technology-Assisted Education*, *1*, 586–590.

Carneiro, R., Lefrere, P., Steffens, K., & Underwood, J. (2012). *Self-regulated learning in technology enhanced learning environments*, volume 5. Springer Science & Business Media.

Carrell, S. E., Malmstrom, F. V., & West, J. E. (2008). Peer effects in academic cheating. *Journal of human resources*, *43*(1), 173–207.

Caspi, A., Chajut, E., & Saporta, K. (2008). Participation in class and in online discussions: Gender differences. *Computers & Education*, *50*(3), 718–724.

Cela, K. L., Sicilia, M. Á., & Sánchez, S. (2015). Social network analysis in e-learning environments: A preliminary systematic review. *Educational Psychology Review*, *27*(1), 219–246.

Chandramouli, B., Goldstein, J., & Duan, S. (2012). Temporal analytics on big data for web advertising. In *28th International Conference on Data Engineering (ICDE)*, (pp. 90–101). IEEE.

Chaplot, D. S., Rhim, E., & Kim, J. (2015). Predicting student attrition in moocs using sentiment analysis and neural networks. In *Proceedings of AIED 2015 Fourth Workshop on Intelligent Support for Learning in Groups*.

Chapman, K. J., Davis, R., Toy, D., & Wright, L. (2004). Academic integrity in the business school environment: I'll get by with a little help from my friends. *Journal of Marketing Education*, *26*(3), 236–249.

Charleer, S., Santos, J. L., Klerkx, J., & Duval, E. (2014). Improving teacher awareness through activity, badge and content visualizations. In *International Conference on Web-Based Learning*, (pp. 143–152). Springer.

Chen, Q., Chen, Y., Liu, D., Shi, C., Wu, Y., & Qu, H. (2016). Peakvizor: Visual analytics of peaks in video clickstreams from massive open online courses. *IEEE transactions on visualization and computer graphics*, *22*(10), 2315–2330.

Cheng, C. K., Paré, D. E., Collimore, L.-M., & Joordens, S. (2011). Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers & Education*, *56*(1), 253–261.

Chitsaz, M., Vigentini, L., & Clayphan, A. (2016). Toward the development of a dynamic dashboard for FutureLearn MOOCs : insights and directions. In *Proceedings of Australasian Society for Computers in Learning in Tertiary Education Conference 2016*, (pp. 116–121).

Cobos, R., Gil, S., Lareo, A., & Vargas, F. A. (2016). Open-DLAs: An Open Dashboard for Learning Analytics. In *Proceedings of the Third (2016) ACM Conference on Learning@Scale*, (pp. 265–268). ACM.

Coetzee, D., Fox, A., Hearst, M. A., & Hartmann, B. (2014). Should your MOOC forum use a reputation system? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, (pp. 1176–1187). ACM.

Coffrin, C., Corrin, L., de Barba, P., & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in MOOCs. In *Proceedings of the fourth international conference on learning analytics and knowledge*, (pp. 83–92). ACM.

Conde, M. Á., García-Peñalvo, F. J., Gómez-Aguilar, D.-A., & Therón, R. (2015). Exploring Software Engineering Subjects by Using Visual Learning Analytics Techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, *10*(4), 242–252.

Corbett, A., Kauffman, L., Maclaren, B., Wagner, A., & Jones, E. (2010). A cognitive tutor for genetics problem solving: Learning gains and student modeling. *Journal of Educational Computing Research*, *42*(2), 219–239.

CSC (2012). The rapid growth of global data.

Curtis, D. D. & Lawson, M. J. (2001). Exploring collaborative online learning. *Journal of Asynchronous learning networks*, *5*(1), 21–34.

Dabbagh, N. & Kitsantas, A. (2012). Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. *The Internet and higher education*, *15*(1), 3–8.

Daniel, B. (2015). Big data and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, *46*(5), 904–920.

Davies, J. & Graff, M. (2005). Performance in e-learning: online participation and student grades. *British Journal of Educational Technology*, *36*(4), 657–663.

De Bruin, G. P. & Rudnick, H. (2007). Examining the cheats: The role of conscientiousness and excitement seeking in academic dishonesty. *South African Journal of Psychology*, *37*(1), 153–164.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of educational research*, *71*(1), 1–27.

Delgado Kloos, C., Muñoz-Merino, P. J., Muñoz-Organero, M., Alario-Hoyos, C., Pérez-Sanagustin, M., Parada G., H. A., Ruipérez-Valiente, J. A., & Sanz, J. L. (2014). Experiences of running MOOCs and SPOCs at UC3M. In *IEEE Global Engineering Education Conference, EDUCON*, (pp. 884–891).

Derick Leony, Pedro J. Muñoz-Merino, José A. Ruipérez-Valiente, Abelardo Pardo, David Arellano Martín-Caro, & Carlos Delgado Kloos (2015). Detection and evaluation of emotions in Massive Open Online Courses. *Journal of Universal Computer Science*, *21*(5), 638–655.

Desmarais, M. C. & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1-2), 9–38.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, (pp. 9–15). ACM.

Domínguez, A., Saenz-de Navarrete, J., de Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education*, *63*, 380–392.

Douglas, J., Iversen, E., & Kalyandurg, C. (2004). Engineering in the k-12 classroom: An analysis of current practices and guidelines for the future. *ASEE Engineering K12 Center*.

Duval, E. (2011). Attention please!: learning analytics for visualization and recommendation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, (pp. 9–17). ACM.

Dwivedi, P. & Bharadwaj, K. K. (2015). e-learning recommender system for a group of learners based on the unified learner profile approach. *Expert Systems*, *32*(2), 264–276.

Dyckhoff, A. L., Lukarov, V., Muslim, A., Chatti, M. A., & Schroeder, U. (2013). Supporting action research with learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, (pp. 220–229). ACM.

Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and implementation of a learning analytics toolkit for teachers. *Educational Technology & Society*, *15*(3), 58–76.

Eaton, C. D. & Wade, S. (2014). Collaborative Learning Through Formative Peer Review With Technology. *Journal of Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *24*(6), 529–543.

Eisenberg, J. (2004). To cheat or not to cheat: effects of moral perspective and situational variables on students' attitudes. *Journal of moral education*, *33*(2), 163–178.

Elbadrawy, A., Studham, R. S., & Karypis, G. (2015). Collaborative multi-regression models for predicting students' performance in course activities. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, (pp. 103–107). ACM.

Essa, A. & Ayad, H. (2012). Improving student success using predictive models and data visualisations. *Research in Learning Technology*, *20*.

Fancsali, S. E. (2013). Data-Driven Causal Modeling of "Gaming the System" and Off-Task Behavior in Cognitive Tutor Algebra. In *NIPS Workshop on Data Driven Education*.

Feng, M., Beck, J., Heffernan, N., & Koedinger, K. (2008). Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test? In *Educational Data Mining 2008*.

Feng, M., Heffernan, N., & Koedinger, K. (2006). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In *Proceedings of 8th International Conference, Intelligent Tutoring Systems Jhongli, Taiwan*, (pp. 31–40)., Berlin. Springer-Verlag.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304–317.

Fox, A. (2013). From MOOCs to SPOCs. *Communications of the ACM*, *56*(12), 38–40.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Galbraith, M. W. & Jones, M. S. (2010). Understanding incivility in online teaching. *Journal of Adult Education*, *39*(2), 1.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71.

Gašević, D., Kovanović, V., & Joksimović, S. (2017). Piecing the Learning Analytics Puzzle: A Consolidated Model of a Field of Research and Practice.

Gašević, D., Mirriahi, N., & Dawson, S. (2014). Analytics of the effects of video use and instruction to support reflective learning. In *Proceedings of the fourth international conference on learning analytics and Knowledge*, (pp. 123–132). ACM.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2015). VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal*, *7*(2), 19–33.

Gibson, D., Ostashewski, N., Flintoff, K., Grant, S., & Knight, E. (2015). Digital badges in education. *Education and Information Technologies*, *20*(2), 403–410.

Giluk, T. L. & Postlethwaite, B. E. (2015). Big five personality and academic dishonesty: A meta-analytic review. *Personality and Individual Differences*, *72*, 59–67.

Goligoski, E. (2012). Motivating the learner: Mozilla's open badges program. *Access to Knowledge: A Course Journal*, *4*(1).

Govaerts, S., Verbert, K., Duval, E., & Pardo, A. (2012). The student activity meter for awareness and self-reflection. In *Human Factors in Computing Systems*, (pp. 869–884). ACM.

Govaerts, S., Verbert, K., Klerkx, J., & Duval, E. (2010). Visualizing activities for self-reflection and awareness. In *International Conference on Web-Based Learning*, (pp. 91–100). Springer.

Grafsgaard, J., Wiggins, J., & Boyer, K. (2014). Predicting Learning and Affect from Multimodal Data Streams in Task-Oriented Tutorial Dialogue. In Stamper, J., Pardos, Z., Mavrikis, M., & McLaren, B. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining*, (pp. 122–129).

Grant, S. & Betts, B. (2013). Encouraging user behaviour with achievements: an empirical study. In *Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on*, (pp. 65–68). IEEE.

Groves, P., Kayyali, B., Knott, D., & Kuiken, S. V. (2016). The 'big data' revolution in healthcare: Accelerating value and innovation.

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@Scale conference*, (pp. 41–50). ACM.

Guo, S. & Wu, W. (2015). Modeling Student Learning Outcomes in MOOCs. *Department of Computer Science and Engineering, Beihang University Beijing, China*.

Hakulinen, L., Auvinen, T., & Korhonen, A. (2013). Empirical study on the effect of achievement badges in TRAKLA2 online learning environment. In *Learning and Teaching in Computing and Engineering (LaTiCE)*, (pp. 47–54). IEEE.

Hamari, J. & Eranti, V. (2011). Framework for designing and evaluating game achievements. *Proceedings Digital Games Research Association Workshop*, *115*(115), 122–134.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? – a literature review of empirical studies on gamification. In *47th International Conference on System Sciences (HICSS)*, (pp. 3025–3034). IEEE.

Harding, T. S., Mayhew, M. J., Finelli, C. J., & Carpenter, D. D. (2007). The theory of planned behavior as a model of academic dishonesty in engineering and humanities undergraduates. *Ethics & Behavior*, *17*(3), 255–279.

Harmon, O. R. & Lambrinos, J. (2008). Are online exams an invitation to cheat? *The Journal of Economic Education*, *39*(2), 116–125.

Hart, C. (2012). Factors associated with student persistence in an online program of study: A review of the literature. *Journal of Interactive Online Learning*, *11*(1), 19–42.

Henke, C. (2009). Engineering Image Problem Could Fuel Shortage. ASQ Survey: Career Not on Radar for Kids or Parents.

Hiltz, S. & Arbaugh, J. (2003). Improving quantitative research methods in studies of asynchronous learning networks (ALN). *Elements of Quality Online Education, Practice and Direction. Needham, MA: Sloan Center for Online Education*, 59–72.

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469–478.

Inselberg, A. & Dimsdale, B. (1990). Parallel coordinates: a tool for visualizing multidimensional geometry. In *Proceedings of the 1st conference on Visualization*, (pp. 361–378). IEEE Computer Society Press.

Intel (2013). Getting started with big data.

Jing, L. (2004). Data Mining Applications in Higher Education. *Executive report, SPSS Inc*, 22–24.

Joksimović, S., Dowell, N., Skrypnyk, O., Kovanović, V., Gašević, D., Dawson, S., & Graesser, A. C. (2015). How do you connect?: analysis of social capital accumulation in connectivist MOOCs. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, (pp. 64–68). ACM.

Jordan, A. E. (2001). College student cheating: The role of motivation, perceived norms, attitudes, and knowledge of institutional policy. *Ethics & Behavior*, *11*(3), 233–247.

Judd, T., Kennedy, G., & Cropper, S. (2010). Using wikis for collaborative learning: Assessing collaboration through contribution. *Australasian Journal of Educational Technology*, *26*(3).

Kauffman, Y. & Young, M. F. (2015). Digital plagiarism: An experimental study of the effect of instructional goals and copy-and-paste affordance. *Computers & Education*, *83*, 44–56.

Kay, J. (2001). Learner control. *User modeling and user-adapted interaction*, *11*(1-2), 111–127.

Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & education*, *55*(2), 427–443.

Kelly, K., Arroyo, I., & Heffernan, N. (2013). Using ITS Generated Data to Predict Standardized Test Scores. In *Educational Data Mining*, (pp. 3–4).

Keppell, M., Souter, K., & Riddle, M. (2011). *Physical and virtual learning spaces in higher education: Concepts for the modern learning environment*. IGI Publishing (IGI Global).

Khalil, H. & Ebner, M. (2014). MOOCs completion rates and possible methods to improve retention-a literature review. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, (pp. 1305–1313).

Khalil, M., Kastl, C., & Ebner, M. (2016). Portraying MOOCs learners: a clustering experience using learning analytics. *The European Stakeholder Summit on experiences and best practices in and around MOOCs*, 265–278.

Kim, S. (2013). Effects of the gamified class in engineering education environments. *Journal of Convergence Information Technology*, *8*(13), 253.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the Workshop on Analysis of Large Scale Social Interaction in MOOCs*, (pp. 60–65).

Koedinger, K. R., Kim, J., Jia, J. Z., McLaughlin, E. A., & Bier, N. L. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, (pp. 111–120). ACM.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, *37*(4), 331–344.

Koutina, M. & Kermanidis, K. L. (2011). Predicting postgraduate students' performance using machine learning techniques. In *Artificial Intelligence Applications and Innovations* (pp. 159–168). Springer.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143–152.

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education*, *31*(1), 30–43.

Lambert, E. G., Hogan, N. L., & Barton, S. M. (2003). Collegiate academic dishonesty revisited: What have they done, how often have they done it, who does it, and why did they do it. *Electronic Journal of Sociology*, *7*(4), 1–27.

Lampe, C., Wohn, D. Y., Vitak, J., Ellison, N. B., & Wash, R. (2011). Student use of facebook for organizing collaborative classroom activities. *International Journal of Computer-Supported Collaborative Learning*, *6*(3), 329–347.

Lee, J. E., Recker, M. M., Choi, H., Hong, W. J., Kim, N., Lee, K., Lefler, M., Louviere, J., & Walker, A. (2016). Applying Data Mining Methods to Understand User Interactions within Learning Management Systems: Approaches and Lessons Learned. *Journal of Educational Technology Development & Exchange*, *8*(2).

Lee, L.-H., Yu, L.-C., Lee, K.-C., Tseng, Y.-H., Chang, L.-P., & Chen, H.-H. (2014). A Sentence Judgment System for Grammatical Error Detection. In *Proceedings of the 25th International Conference on Computational Linguistics*, (pp. 67–70).

Leony, D., Pardo, A., de la Fuente Valentín, L., de Castro, D. S., & Delgado Kloos, C. (2012). GLASS: a learning analytics visualization tool. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, (pp. 162–163). ACM.

Li, N., Verma, H., Skevi, A., Zufferey, G., Blom, J., & Dillenbourg, P. (2014). Watching MOOCs together: Investigating co-located MOOC study groups. *Distance Education*, *35*(2), 217–233.

Li, W., Grossman, T., & Fitzmaurice, G. (2012). GamiCAD: a gamified tutorial system for first time autocad users. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, (pp. 103–112). ACM.

Littlejohn, A., Hood, N., Milligan, C., & Mustain, P. (2016). Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *The Internet and Higher Education*, *29*, 40–48.

Liyanagunawardena, T. R., Williams, S., & Adams, A. (2013). The Impact and Reach of MOOCs: A Developing Countries' Perspective. *eLearning Papers*, 1–8.

LoSchiavo, F. M. & Shatz, M. A. (2011). The impact of an honor code on cheating in online courses. *Journal of Online Learning and Teaching*, *7*(2), 179.

Macfadyen, L. P. & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & education*, *54*(2), 588–599.

Magaña, V. C. & Organero, M. M. (2014). The impact of using gamification on the eco-driving learning. In *Ambient Intelligence-Software and Applications* (pp. 45–52). Springer.

Maggs, L. A. (2014). A case study of staff and student satisfaction with assessment feedback at a small specialised higher education institution. *Journal of further and higher education*, *38*(1), 1–18.

Mahajan, P., Gaba, G., & Chauhan, N. S. (2016). Big data security. *Journal of Management and IT*, *7*(1), 89–94.

Mas-Sanso, R. & Manresa-Yee, C. (2016). Gamifying an artificial intelligence course in engineering education. *International Journal of Engineering Education*, *32*(1), 513–520.

Mashey, J. R. (1997). Big data and the next wave of infras-tress. In *Computer Science Division Seminar, University of California, Berkeley*.

Masters, K. (2011). A Brief Guide To Understanding MOOCs. *The Internet Journal of Medical Education*, *1*(2), 1–5.

May, M., George, S., & Prévôt, P. (2011). TrAVis to enhance students' self-monitoring in online learning supported by computer-mediated communication tools. *Computer Information Systems and Industrial Management Applications*, *3*, 623–634.

Mazza, R. & Dimitrova, V. (2004). Visualising student tracking data to support instructors in web-based distance education. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, (pp. 154–161). ACM.

Mazza, R. & Milani, C. (2005). Exploring usage analysis in learning systems: Gaining insights from visualisations. In *Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education*, (pp. 65–72).

McCabe, D. L. & Trevino, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education*, *64*(5), 522–538.

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics &Behavior*, *11*(3), 219–232.

McLoughlin, C. & Lee, M. J. (2010). Personalised and self regulated learning in the web 2.0 era: International exemplars of innovative pedagogy using social software. *Australasian Journal of Educational Technology*, *26*(1).

Mladenović, S., Krpan, D., & Mladenović, M. (2016). Using Games to Help Novices Embrace Programming: From Elementary to Higher Education. *International journal of engineering education*, *32*(1), 521–531.

Młynarska, E., Greene, D., & Cunningham, P. (2016). Time Series Analysis of VLE Activity Data. In *9th International Conference on Educational Data Mining: EDM 2016, Raleigh, North Carolina, 29 June-2 July 2016 United States*.

Montola, M., Nummenmaa, T., Lucero, A., Boberg, M., & Korhonen, H. (2009). Applying game achievement systems to enhance user experience in a photo sharing service. In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*, (pp. 94–97). ACM.

Moody, D. L. & Sindre, G. (2003). Evaluating the effectiveness of learning interventions: an information systems case study. *ECIS 2003 Proceedings*, 80.

Muldner, K., Burleson, W., Van de Sande, B., & VanLehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts. *User modeling and user-adapted interaction*, *21*(1-2), 99–135.

Muñoz-Merino, P. J., Delgado Kloos, C., Seepold, R., & García, R. M. C. (2006). Rating the importance of different LMS functionalities. In *Frontiers in Education Conference, 36th Annual*, (pp. 13–18). IEEE.

Muñoz-Merino, P. J., Méndez Rodríguez, E., Delgado Kloos, C., & Ruiperez-Valiente, J. A. (2017). Design, Implementation and Evaluation of SPOCs at the Universidad Carlos III de Madrid. *Journal of Universal Computer Science*, *23*(2), 167–186.

Muñoz-Merino, P. J., Molina, M. F., Muñoz-Organero, M., & Delgado Kloos, C. (2014). Motivation and emotions in competition systems for education: An empirical study. *IEEE Transactions on Education*, *57*(3), 182–187.

Muñoz-Merino, P. J., Ruipérez-Valiente, J. A., Alario-Hoyos, C., Pérez-Sanagustín, M., & Delgado Kloos, C. (2015). Precise Effectiveness Strategy for analyzing the effectiveness of students with educational resources and activities in MOOCs. *Computers in Human Behavior*, *47*, 108–118.

Muñoz-Merino, P. J., Ruipérez-Valiente, J. A., Alario-Hoyos, C., Pérez-Sanagustín, M., & Kloos, C. D. (2014). Analítica del Aprendizaje para la Evaluación Precisa de la Efectividad del Alumno con Actividades y Recursos Educativos. In *Proceedings of the 9th Iberian Conference on Information Systems and Technologies (CISTI)*, (pp. 1–6).

Muñoz-Merino, P. J., Ruipérez-Valiente, J. A., Delgado Kloos, C., Auger, M. A., Briz, S., Castro, V. D., & Santalla, S. N. (2016). Flipping the Classroom to Improve Learning with MOOCs Technology. *25*(1), 15–25.

Muñoz-Merino, P. J., Ruipérez Valiente, J. A., & Kloos, C. D. (2013). Inferring higher level learning information from low level data for the Khan Academy platform. In *Proceedings* *of the Third International Conference on Learning Analytics and Knowledge - LAK '13*, (pp. 112–116)., New York, New York, USA. ACM Press.

Muñoz-Organero, M., Muñoz-Merino, P. J., & Kloos, C. D. (2010). Student behavior and interaction patterns with an lms as motivation predictors in e-learning settings. *IEEE Transactions on Education*, *53*(3), 463–470.

Muntean, C. I. (2011). Raising engagement in e-learning through gamification. In *Proceedings of the 6th International Conference on Virtual Learning*, (pp. 323–329).

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *Proceedings of the 37th Frontiers In Education Conference*, (pp. T2G–7). IEEE.

Ni, A. Y. (2013). Comparing the effectiveness of classroom and online learning: Teaching research methods. *Journal of Public Affairs Education*, 199–215.

Nicoară, E. S. (2013). The impact of massive online open courses in academic environments. In *Conference Proceedings of eLearning and Software for Education (eLSE)*, (pp. 644–649). Universitatea Nationala de Aparare Carol I.

Northcutt, C. G., Ho, A. D., & Chuang, I. L. (2016). Detecting and preventing "multiple-account" cheating in massive open online courses. *Computers & Education*, *100*, 71–80.

Ochoa, X., Worsley, M., Weibel, N., & Oviatt, S. (2016). Multimodal learning analytics data challenges. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, (pp. 498–499). ACM.

O'Toole, R. (2013). Pedagogical strategies and technologies for peer assessment in Massively Open Online Courses (MOOCs).

Palazzo, D. J. (2006). *Detection, patterns, consequences, and remediation of electronic homework copying*. PhD thesis, Massachusetts Institute of Technology.

Palazzo, D. J., Lee, Y. J., Warnakulasooriya, R., & Pritchard, D. E. (2010). Patterns, correlates, and reduction of homework copying. *Physical Review Special Topics - Physics Education Research*, *6*(1), 1–11.

Papamitsiou, Z. & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society*, *17*(4), 49–64.

Pardo, A., Mirriahi, N., Martinez-Maldonado, R., Jovanovic, J., Dawson, S., & Gašević, D. (2016). Generating actionable predictive models of academic performance. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, (pp. 474–478). ACM.

Pardos, Z., Gowda, S., Baker, R. S., & Heffernan, N. (2010). Ensembling predictions of student post-test scores for an intelligent tutoring system. In *Educational Data Mining 2011*.

Park, Y. & Jo, I.-h. (2015). Development of the Learning Analytics Dashboard to Support Students' Learning Performance Learning Analytics Dashboards (LADs). *Journal of Universal Computer Science*, *21*(1), 110–133.

Payan, J., Reardon, J., & McCorkle, D. E. (2010). The effect of culture on the academic honesty of marketing and business students. *Journal of Marketing Education*, *32*(3), 275–291.

Pedro, M. O., Ocumpaugh, J., Baker, R. S., & Heffernan, N. (2014). Predicting stem and non-stem college major enrollment from middle school interaction with mathematics educational software. In *Educational Data Mining 2014*.

Peñalvo, F. J. G., Conde, M. A., Bravo, S., Gómez, D. A., & Therón, R. (2011). Visual Analysis of a Moodle-based Object Oriented Programming Course. *International Journal of Computers Applications Proceedings on Design and Evaluation of Digital Content for Education*, *1*(7), 8–14.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 759–772.

Pijeira Díaz, H. J., Santofimia, J., Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2015). Using Video Visualizations in Open edX to Understand Learning Interactions of Students. In *10th European Conference on Technology Enhanced Learning, EC-TEL 2015, Toledo, Spain*, (pp. 522–525).

Pijeira Díaz, H. J., Santofimia Ruiz, J., Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2016). A Demonstration of ANALYSE: A Learning Analytics Tool for Open edX. In *Proceedings of the Third (2016) ACM Conference on Learning@Scale*, (pp. 329–330). ACM.

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., & Xue, Y. (2016). Modeling and predicting learning behavior in MOOCs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, (pp. 93–102). ACM.

Qu, H. & Chen, Q. (2015). Visual analytics for MOOC data. *IEEE computer graphics and applications*, *35*(6), 69–75.

Rabbany, R., Elatia, S., Takaffoli, M., & Zaïane, O. R. (2014). Collaborative learning of students in online discussion forums: A social network analysis perspective. In *Educational data mining* (pp. 441–466). Springer.

Redondo, D., Muñoz-Merino, P. J., Ruipérez-Valiente, J. A., Delgado Kloos, C., Pijeira Díaz, H. J., & Santofimia Ruiz, J. (2015). Combining Learning Analytics and the Flipped Classroom in a MOOC of maths. In *International Workshop on Applied and Practical Learning Analytics*, (pp. 71–79).

Resta, P. & Laferrière, T. (2007). Technology in support of collaborative learning. *Educational Psychology Review*, *19*(1), 65–83.

Riojas, M., Lysecky, S., & Rozenblit, J. (2012). Educational technologies for precollege engineering education. *IEEE transactions on learning technologies*, *5*(1), 20–37.

Romero, C. & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135–146.

Romero, C. & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics*, *40*(6), 601–618.

Romero, C. & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27.

Rosé, C. P. & Siemens, G. (2014). Shared task on prediction of dropout over time in massively open online courses. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing Workshop*, volume 14, (pp. 39–41).

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2007). Assessing social presence in asynchronous text-based computer conferencing. *International Journal of E-Learning & Distance Education*, *14*(2), 50–71.

Ruipérez-Valiente, J. A., Alexandron, G., Chen, Z., & Pritchard, D. E. (2016). Using Multiple Accounts for Harvesting Solutions in MOOCs. In *Proceedings of the Third (2016) ACM Conference on Learning@Scale*, (pp. 63–70). ACM.

Ruipérez-Valiente, J. A., Cobos, R., Muñoz-Merino, P. J., Andújar, Á., & Delgado Kloos, C. (2017). Early Prediction and Variable Importance of Certificate Accomplishment in a MOOC. In *Fifth European MOOCs Stakeholders Summit*.

Ruipérez-Valiente, J. A., Joksimović, S., Kovanović, V., Gašević, D., Muñoz-Merino, P. J., & Delgado Kloos, C. (2017a). A Data-driven Method for the Detection of Close Submitters in Online Learning Environments. In *Proceedings of the 26th International Conference on World Wide Web Companion*, (pp. 361–368).

Ruipérez-Valiente, J. A., Joksimović, S., Kovanović, V., Gašević, D., Muñoz-Merino, P. J., & Delgado Kloos, C. (2017b). Collaboration or "Collaboration": Detecting and Characterizing Communities of Accounts in MOOCs (Work in Progress – To Be Submitted). *IEEE Transactions on Learning Technologies*.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Alexandron, G., & Pritchard, D. E. (2017). Using Machine Learning to Detect 'Multiple-Account' Cheating and Analyze the Influence of Student and Problem Features (Major revision). *IEEE Transactions on Learning Technologies*.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. Improving the Prediction of Learning Outcomes in Educational Platforms including Higher Level Interaction Indicators (Under review). *Expert Systems*.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2013). An architecture for extending the learning analytics support in the Khan Academy framework. In *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality*, (pp. 277–284).

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2014). A demonstration of ALAS-KA: A learning analytics tool for the Khan Academy platform. In *9th European Conference on Technology Enhanced Learning (EC-TEL 2014)*, (pp. 518–521). Springer International Publishing.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2015a). A Predictive Model of Learning Gains for a Video and Exercise Intensive Learning Environment. In *17th International Conference on Artificial Intelligence in Education*, (pp. 760–763).

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2015b). Diseño e Implementación de un Módulo de Analítica de Aprendizaje, y su Aplicación para la Evaluación de Experiencias Educativas. *TICs para el Aprendizaje de la Ingeniería (TICAI)*, 77–84.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2016a). An Analysis of the Use of Badges in an Educational Experiment. In *Frontiers in Education Conference (FIE)*, (pp. 1–8).

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2016b). Analyzing students' intentionality towards badges within a case study using Khan academy. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, (pp. 536–537)., New York, New York, USA. ACM Press.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2017). Detecting and Clustering Students by their Gamification Behavior with Badges: A Case Study in Engineering Education. *The International Journal of Engineering Education*, *33*(2-B), 816–830.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Delgado Kloos, C., Niemann, K., & Scheffel, M. (2014). Do Optional Activities Matter in Virtual Learning Environments? In *Ninth European Conference on Technology Enhanced Learning*, (pp. 331–344)., Graz, Austria. Springer International Publishing.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Delgado Kloos, C., Niemann, K., Scheffel, M., & Wolpers, M. (2016). Analyzing the Impact of Using Optional Activities in Self-Regulated Learning. *IEEE Transactions on Learning Technologies*, *9*(3), 231 – 243.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Gascón-Pinedo, J. A., & Delgado Kloos, C. (2016). Scaling to Massiveness with ANALYSE: A Learning Analytics Tool for Open edX (In press). *IEEE Transactions on Human-Machine Systems*, 1–6.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Leony, D., & Delgado Kloos, C. (2015). ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. *Computers in Human Behavior*, *47*, 139–148.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Pijeira Díaz, H. J., Santofimia Ruiz, J., & Delgado Kloos, C. (2017). Evaluation of a Learning Analytics Application for Open edX Platform. *Computer Science and Information Systems*, *14*(1), 51–73.

Salehi, M. & Kamalabadi, I. N. (2013). Hybrid recommendation approach for learning material based on sequential pattern of the accessed material and the learner's preference tree. *Knowledge-Based Systems*, *48*, 57–69.

Sanecka, E. & Baran, L. (2015). Explicit and implicit attitudes toward academic cheating and its frequency among university students. *Polish Journal of Applied Psychology*, *13*(2), 69–92.

Santofimia Ruiz, J., Pijeira Díaz, H. J., Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2014). Towards the Development of a Learning Analytics Extension in Open edX. In *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality*, (pp. 299–306).

Santos, J. L., Klerkx, J., Duval, E., Gago, D., & Rodríguez, L. (2014). Success, activity and drop-outs in MOOCs an exploratory study on the UNED COMA courses. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, (pp. 98–102). ACM.

Sauro, J. (2011). SUSTisfied? little-known System Usability Scale facts. *UX Magazine*, *10*(3), 2011–3.

Schmitz, H.-C., Scheffel, M., Friedrich, M., Jahn, M., Niemann, K., & Wolpers, M. (2009). CAMera for PLE. In *European Conference on Technology Enhanced Learning*, (pp. 507–520). Springer.

Schon, D. A. (1984). *The reflective practitioner: How professionals think in action*, volume 5126. Basic books.

Scrimpshire, A. J., Stone, T. H., Kisamore, J. L., & Jawahar, I. (2016). Do Birds of a Feather Cheat Together? How Personality and Relationships Affect Student Cheating. *Journal of Academic Ethics*, 1–22.

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2012). Tracing a little for big improvements: Application of learning analytics and videogames for student assessment. *Procedia Computer Science*, *15*, 203–209.

Sharma, K., Alavi, H. S., Jermann, P., & Dillenbourg, P. (2016). A gaze-based learning analytics model: in-video visual feedback to improve learner's attention in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, (pp. 417–421). ACM.

Sharma, K., Jermann, P., & Dillenbourg, P. (2014). How students learn using MOOCs: An eye-tracking insight. In *Second MOOC European Stakeholders Summit, EMOOCs*.

Sheridan, J., Alany, R., & Brake, D.-J. (2005). Pharmacy students' views and experiences of turnitinw - an online tool for detecting academic dishonest. *Pharmacy Education*, *5*.

Shi, C., Fu, S., Chen, Q., & Qu, H. (2015). VisMOOC: Visualizing video clickstream data from massive open online courses. In *Visualization Symposium (PacificVis)*, (pp. 159–166). IEEE.

Shu, L. L. & Gino, F. (2012). Sweeping dishonesty under the rug: how unethical actions lead to forgetting of moral rules. *Journal of personality and social psychology*, *102*(6), 1164.

Siemens, G. (2013). Massive Open Online Courses: Innovation in Education? *Open educational resources: innovation, research and practice*, *1833*, 5–16.

Siemens, G. & Baker, R. S. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, (pp. 252–254). ACM.

Singhal, A. C. (1982). Factors in students' dishonesty. *Psychological Reports*, *51*(3), 775–780.

Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014). Your click decides your fate: Inferring information processing and attrition behavior from MOOC video clickstream interactions. *arXiv preprint arXiv:1407.7131*.

Sparks, S. D. (2011). Studies Shed Light on How Cheating Impedes Learning. *Education Week*, *30*(26), 1–16.

Strohbach, M., Ziekow, H., Gazis, V., & Akiva, N. (2015). Towards a big data analytics framework for IoT and smart city applications. In *Modeling and processing for next-generation big-data technologies* (pp. 257–282). Springer.

Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning*, *15*(3).

Swan, K. (2003). Learning effectiveness online: What the research tells us. *Elements of quality online education, practice and direction*, *4*, 13–47.

Tabuenca, B., Kalz, M., Drachsler, H., & Specht, M. (2015). Time will tell: The role of mobile learning analytics in self-regulated learning. *Computers & Education*, *89*, 53–74.

Tervakari, A.-M., Marttila, J., Kailanto, M., Huhtamäki, J., Koro, J., & Silius, K. (2013). Developing learning analytics for TUT circle. In *Open and Social Technologies for Networked Learning* (pp. 101–110). Springer.

Tsui, J., Lau, J., & Shieh, L. (2014). Septris and SICKO: implementing and using learning analytics and gamification in medical education. *Educause*.

Tucker, C., Pursel, B., & Divinsky, A. (2014). Mining student-generated textual data in MOOCs and quantifying their effects on student performance and learning outcomes. In *2014 ASEE Annual Conference*.

Urdan, T. A. & Weggen, C. C. (2000). Corporate elearning: Exploring a new frontier.

Vaibhav, A. & Gupta, P. (2014). Gamification of moocs for increasing user engagement. In *International Conference on Innovation and Technology in Education*, (pp. 290–295). IEEE.

Vatrapu, R., Teplovs, C., Fujita, N., & Bull, S. (2011). Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, (pp. 93–98). ACM.

Voss, B. D. (2013). Massive open online courses (MOOCs): A primer for university and college board members. *AGB Association of Governing Boards of Universities and Colleges*.

Walonoski, J. A. & Heffernan, N. T. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, (pp. 382–391). Springer.

Wankel, C. & Blessinger, P. (2012). *Increasing student engagement and retention using online learning activities*. Emerald Group Publishing Limited.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International journal of Educational research*, *13*(1), 21–39.

Wen, M., Yang, D., & Rose, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In *Educational Data Mining 2014*.

Whitehill, J., Williams, J., Lopez, G., Coleman, C., & Reich, J. (2015). Beyond prediction: First steps toward automatic intervention in MOOC student stopout. *International Educational Data Mining Society*.

Witherspoon, M., Maldonado, N., & Lacey, C. H. (2012). Undergraduates and academic dishonesty. *International Journal of Business and Social Science*, *3*(1).

Witmer, H. & Johansson, J. (2015). Disciplinary action for academic dishonesty: does the student's gender matter? *International Journal for Educational Integrity*, *11*(1), 6.

Wood, H. & Wood, D. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, *33*(2), 153–169.

Worsley, M. & Blikstein, P. (2015). Leveraging multimodal learning analytics to differentiate student learning strategies. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, (pp. 360–367). ACM.

Xie, T., Tillmann, N., De Halleux, J., & Bishop, J. (2015). Educational software engineering: Where software engineering, education, and gaming meet. In *Computer Games and Software Engineering* (pp. 113–132). Chapman and Hall/CRC.

Xu, Z., Goldwasser, D., Bederson, B. B., & Lin, J. (2014). Visual analytics of MOOCs at Maryland. In *Proceedings of the first ACM conference on Learning@Scale conference*, (pp. 195–196). ACM.

Yardley, J., Rodríguez, M. D., Bates, S. C., & Nelson, J. (2009). True confessions?: Alumni's retrospective reports on undergraduate cheating behaviors. *Ethics & Behavior*, *19*(1), 1–14.

Zhang, H., Almeroth, K., Knight, A., Bulger, M., & Mayer, R. (2007). Moodog: Tracking students' online learning activities. In *Proceedings of World conference on educational multimedia, hypermedia and telecommunications*, volume 25, (pp. 4415–4422).

Zirk, M. (2014). *Gamification for Software Engineering Education*. PhD thesis, University of Tartu.