Sarah Frank, Bsc

Data Analysis and Learning Behavior in Two Different Learning Settings

MASTER'S THESIS

to achieve the university degree of
Master of Science
Master's degree programme: Computer Science

submitted to

Graz University of Technology

Supervisor
Assoc.Prof. Dipl.-Ing. Dr.techn. Christian Gütl

Institute of Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Co-Supervisors
Assoc.Prof. Mar Pérez-Sangustín, Ph.D.
Dipl.-Ing. Dr.techn. Alexander Nussbaumer

Graz, December 2020

Sarah Frank, Bsc

Datenanalyse und Lernverhalten in zwei verschiedenen Lernumgebungen

MASTERARBEIT

zur Erlangung des akademischen Grades
Diplom-Ingeneurin
Masterstudium Computer Science

eingereicht an der

Technischen Universität Graz

Betreuer

Assoc.Prof. Dipl.-Ing. Dr.techn. Christian Gütl

Institute of Interactive Systems and Data Science
Leitung: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Mitbetreuer

Assoc.Prof. Mar Pérez-Sangustín, Ph.D.
Dipl.-Ing. Dr.techn. Alexander Nussbaumer

Graz, Dezember 2020

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date   Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Datum   Unterschrift

Abstract

E-learning systems have become a prevalent and almost unavoidable part of education over the past years, and can take many forms. From so-called Massive Open Online Courses (MOOCs) to smaller, locally-used training courses that take place online, the field of e-learning is diverse.

This thesis took two e-learning systems into account for its analyses. First, the data from six MOOCs was used to predict dropouts, and furthermore evaluate the influence of forum interactions on dropout rates. In this, the number of interactions, the user's active time, as well as the average time between clicks achieved high feature importance scores and showed promising results. However, while first statistics showed what seemed to be a positive correlation between forum posts and course completion, what could be found by using AdaBoosted decision trees was that at the average active forum users of 6.38%, the forum data was not sufficiently expressive to have much influence on the dropout prediction results, or to score highly on feature importance.

Secondly, a small, university-specific e-learning course, built on the principle of self-regulated learning, was investigated for patterns of user behavior. With only 31 students, the course showed a markedly lower user number than the previously considered MOOCs, which had an average of 5852.67 students each. Due to this, a combination of sequence extraction methods and heat maps was used to find user behavior patterns and make inferences regarding assessment outcomes. By splitting users according to their course performance, as well as by whether they exhibited organised or unorganised behavior in their course progression, it was possible to put those four user categories into relation. The findings showed that unordered students had a tendency towards repeating assessments, while ordered students most often reached full points on their first try. Furthermore, lower performing students tended to view less new content in a row than higher performance ones, and reviewed a higher number of older course content.

Overall, both sets of analyses showed limitations connected to a lack of data. For the MOOC data set from part one, this specifically applied to forum data due to the low number of forum participants. In part two, the small data set significantly limited the number of feasible machine learning methods that could be applied. For future work, this may be a useful point to consider when selecting data sets.

Kurzfassung

E-learning Systeme haben im Laufe der vergangenen Jahre zunehmend an Bedeutung gewonnen und stellen mittlerweile einen beinahe unersetzlichen Bestandteil im Bildungsbereich dar. Dabei sind die Formen die sie annehmen vielfältig, von sogenannten Massive Open Online Courses (MOOCs) bis zu kleineren, lokal eingesetzten Ausbildungskursen, die online stattfinden.

Diese Arbeit befasste sich für die Analysen mit zwei verschiedenen e-learning Systemen. Zunächst wurden mit den Daten von sechs verschiedenen MOOCs Abbrecher vorausgesagt, sowie der Einfluss von Forumaktivität auf die Abbrecherquote untersucht. Hierbei zeigte sich, dass die Anzahl der Interaktionen, die aktive Zeit des Users und die durchschnittliche Zeit zwischen zwei Klicks als Features starke Auswirkungen auf die Ergebnisse hatten. Obwohl erste Auswertungen einen positiven Zusammenhang zwischen Forumposts und Kursabschluss aufzuweisen schienen, konnte dies unter Nutzung von AdaBoosted Entscheidungsbäumen allerdings nicht gezeigt werden. Mit nur 6,38% aktiven Forumusern hatten die Forumdaten kaum Einfluss auf die Resultate und erreichten auch keinen hohen Wert bei der Auswertung der Feature Importances.

Darauffolgend wurde ein kleiner, universitäts-spezifischer e-learning Kurs untersucht, der auf dem Prinzip des selbstgesteuerten Lernens aufbaut. Insbesonderes Augenmerk wurde hierbei auf das Erkennen von Mustern im Userverhalten gelegt. Mit 31 Usern war der Kurs bedeutend kleiner als die zuvor untersuchten MOOCs mit durchschnittlich 5852,67 eingeschriebenen Usern. Aufgrund dessen wurde eine Kombination aus Sequence Extraction Methoden und Heatmap Diagrammen verwendet, um Muster im Userverhalten zu finden und Schlüsse auf die Testergebnisse der User ziehen zu können. Durch das Aufteilen der User in organisiert und unorganisiert Lernende, als auch nach ihren Kursergebnissen, konnten die vier Kategorien in Relation zueinander gesetzt und verglichen werden. Die Ergebnisse zeigten, dass unorganisiert Lernende eine Tendenz zum Wiederholen der Tests aufwiesen, während organisiert Lernende meist beim ersten Antritt die volle Punktezahl erreichten. Weiters riefen User mit schlechteren Testergebnissen weniger neue Inhalte nacheinander auf als jene die besser abschnitten und wiederholten mehr vergangene Kursinhalte.

Insgesamt zeigten sich in den Analysen beider Datensätze Schwächen, die mit dem Fehlen von Daten zusammenhingen. Für die MOOC Daten bezog sich dies insbesondere auf die Forumdaten aufgrund der niedrigen Anzahl an Forumsusern. In Teil zwei waren die anwendbaren Machine Learning Algorithmen aufgrund des kleinen Datensets eingeschränkt. Im Weiteren könnte es nützlich sein, dies bei der Auswahl der Datensets zu beachten.

Acknowledgements

I would like to thank my family and friends for their support during my studies, and for being there for me every step of the way.

I am also thankful to Christian Gütl for his insights, suggestions, and feedback, as well as his general support during my work on this thesis.

Furthermore, I thank Massimo Vitiello, as well as Alexander Nussbaumer, who both gave input and provided starting points for the work, as well as valuable feedback.

Finally, I would like to thank Maria Pérez-Sanagustín, Jorge Maldonado-Mahauad, and Josefina Hernández Correa for providing data access, ideas, input, and feedback along the way.

Contents

List of Tables

List of Figures

List of Abbreviations

BDT	Boosted Decision Tree
etc.	et cetera
i.e.	id est
LSA	Lag Sequential Analysis
MCC	Matthews Correlation Coefficient
MOOC	Massive Open Online Course
SMOTE	Synthetic Minority Over-sampling Technique
SRL	Self-Regulated Learning
SVM	Support Vector Machine
TPB	Theory of Planned Behavior

1. Introduction

This chapter focuses on the motivation for this thesis, as well as on giving a short overview of the general structure and topics of research discussed in it.

1.1. Motivation and Focus

With the fast-paced change of the job market and work environments in the current day and age, life-long learning has become a term that seems ever-present. However, this more recent requirement brings difficulties in addition to opportunities, when e-learning systems and MOOCs don't seem to perform as well as they were expected to back when they first started their rise to popularity.

While this lack of performance of students in MOOCs, especially, may not seem like a big problem when the learners are situated in a place where education is comparatively easy to obtain, there are many areas where this is not the case and e-learning systems present the only chance to continue learning. Conditions such as distance, lack of money or consistent time availability present difficulties to people who may want to, but are not able to seek further education on topics that interest them.

MOOCs aim to mitigate this disparity in education to give everyone the chance to learn. However, because there are high dropout rates in MOOC courses across the board, it is necessary to take a look at why this is the case, and how it is possible to predict and potentially mitigate dropouts.

In addition to, and in part due to the popularity of MOOCs, different e-learning methodologies have found their way into formerly traditional education institutions, so that self-directed and self-regulated learning have risen in importance once more. With this comes the necessity of investigating the effects that a self-directed or self-regulated learning environment has on its students' behaviors, as well as on their performance.

1.2. Contribution and Research Question

This work investigates the impact of the inclusion of forum data in classification tasks for dropout prediction. For this, boosted decision trees are used to classify users into "dropouts" and "completers". To measure the influence of the forum features and be able to track their development as time goes on, experiments with and without inclusion of forum data are run each at the beginning, middle, and end of each course.

With these analyses, this work aims to contribute to the research into the influence of social interaction such as forum activity in e-learning courses on dropout rates, as well as general student performance.

Furthermore, this work assesses the applicability of sequence extraction methods on small e-learning courses, and evaluates the meaningfulness of the results in combination with heat maps. In a further step, it also seeks to make connections between self-regulated learner behavior and assessment outcomes, with the intention of finding patterns, as well as differences between those students who performed more and less highly, respectively.

The hope is that by finding possible answers to these questions, the work can contribute to the research of how e-learning environments, especially self-regulated ones that typically require a high level of self-sufficiency from the student, can be constructed to best support various students' learning behaviors and needs.

1.3. Thesis Outline

This thesis is divided into six chapters.

Chapter 2 covers basic information about the history of e-learning in general, as well as that of MOOCs in particular. It describes previous research done in regards to learning behavior and dropout prediction, as well as underlying psychological theories that apply to user interaction with MOOCs and influence the completion rate, such as the theory of planned behavior. The chapter also discusses popular approaches of learning analytics that are commonly used in previous research to analyse user behavior. In addition, it explains some of the more popular machine learning algorithms used in research, such as decision trees, and Markov models, as well as concepts such as boosting and bagging, and model evaluation.

Chapter 3 first gives a brief explanation of the specific data set received from some of the institution's Coursera courses, its structure and an overview over the information that was considered, as well as the proposed strategies and concepts for the analyses. Subsequently, it then shows the results and goes on to discuss them and their meaning, describing how the new findings fit in with the previous research considered in the literature study of chapter 2.

Following this, chapter 4 similarly describes the received data, its contents and structure, and the significant strategies and concepts to be applied in the analyses of the information. Again, the chapter shows the results and discusses them in detail.

Chapter 5 briefly discusses the lessons learned throughout the work on this thesis, each for literature, development, and evaluation.

Finally, chapter 6 focuses on the results in relation to the research questions and proposes further research and possible avenues for future work, thus concluding the thesis.

2. Background and Related Work

This chapter briefly discusses the topics of interest relevant for this thesis, beginning with an overview of the history of e-learning, and an explanation of the important learning models that are present in technology-enhanced learning, in particular in view of the analysis this work will later be doing in chapter 4. Afterwards, the work elaborates on the history of MOOCs specifically. The work then takes a brief detour to pay additional attention to the popular MOOC platform Coursera, which is one of the sources of the course data that the work will later focus on. Following this, the chapter continues with an explanation of technology-enhanced learning and learning strategies that are particularly relevant for e-learning environments.

The second section focuses on data science and learning analytics, followed by the third section, which describes the machine learning approaches commonly used in learning analytics, and in particular in learner behaviour and dropout prediction. Here, there will be a focus on decision trees, Markov chains and Markov models, as well as sequence analysis, as those are relevant for the later analyses.

The last parts will discuss existing and related work in learning analytics and dropout prediction, respectively. They will give short descriptions of the ideas and approaches presented in the research.

2.1. E-Learning

E-learning in general and applications such as MOOCs in particular have grown in importance as computers and digital media increased in pervasiveness. The following sections will briefly explain the learning models that apply to e-learning, e-learning's history, give an overview over MOOCs in general, and the MOOC platform Coursera specifically, as well as give a short outline of the learning theories that are connected to technology-enhanced learning.

2.1.1. Learning Models

Many learning models, such as project-based learning, team-based learning, and problem-based learning, assume or even require students to be in physical proximity to one another, some more so than others. Due to the fact that many of them are based on the assumption of traditional modes of teaching, these learning techniques emphasise working together on a problem or project, often in the presence or with the help of the teacher (Bédard et al., 2012; Michaelsen and Sweet, 2008).

As such, they are only tangentially relevant for the field of e-learning, which is often used in cases where physical proximity is difficult or impossible to achieve, or the aim is to make the coursework time-independent. It follows that self-regulated learning and self-directed learning can, because of their emphasis on independence on the students' parts, in fact be considered the most significant learning theories in regards to e-learning.

With self-regulated learning focusing on the student's own motivation and drive to organize and complete their workload, Schunk (1989, p. 83) refers to it as "*learning that occurs from students' self-generated behaviors systematically oriented toward the attainment of their learning goals*". In practice, this means that self-regulated learners are assumed to have both motivation, and a functioning learning strategy that not only causes them to start the learning progress, but also to successfully navigate their way through the learning material without any outside regulation of their learning process (Perry & Winne, 2006).

Even though the exact definition of self-regulated learning depends on the specific theory, there are similarities and overlaps to be found. The typical characterization of a self-regulated learner expects activeness, and independent management of the learning process of the student (Greene & Azevedo, 2007). Some sources also consider goal setting a part of self-regulated learning (Pintrich, 2000). In this case, it is also necessary to consider motivation, as Locke and Latham (2004) explain, since the two influence each other.

Self-regulated learning itself can be further organised into phases. Once again, however, the specific terms and divisions differ between publications and the used model. In the time since the first publications during the late 1980s, multiple scholars have created self-regulated learning models and defined terms (Panadero, 2017). To name only one example of the differences between two such models and terminologies, Pintrich (2000) specified the phases as identification and planning, monitoring and control of learning strategies, and reaction and reflection, while Winne and Hadwin (1998) identified the phases as task definition, goal setting and planning, studying tactics, and adaptions to meta-cognition (Greene & Azevedo, 2007). While there are similarities, the terminology doesn't match entirely, though the underlying thought remains alike.

Self-regulated learning depends on the student controlling their own learning process—namely by using cognitive, motivational, and emotional functions (Lens & Vansteenkiste, 2012). This responsibility over their own behavior has the effect of increasing the feeling of ownership that the students feel in the process, which, in combination with feedback of their learning effectiveness, in turn has positive effects on their learning progress, as they feel more personal responsibility, but also more satisfaction for the outcome of the process (Zimmerman, 1990). Chapter 4 will be using data from a self-regulated learning system to analyse user behavior in later parts of the thesis.

Sharing several key-points with self-regulated learning, self-directed learning's most-used specification comes from Knowles (1975):

> *In its broadest meaning, self-directed learning describes a process in which individuals take the initiative, with or without the help of others, in diagnosing their learning needs, formulating learning goals, identifying human and material resources for learning, choosing and implementing appropriate learning strategies, and evaluating learning outcomes.* (p. 18)

While the four basic phases of self-regulated learning that were specified above closely match those of self-directed learning, and intrinsic motivation is seen as a crucial component of the learning process as well, differences can be pinpointed when looking at the two learning methods side by side.

Self-directed learning, sometimes also called self-guided learning or autonomous learning (Černochová & Selcuk, 2019), differs from self-regulated learning in that it has the tendency to focus more outside of traditional education environments, and *"involves planning trajectory"* (Saks & Leijen, 2014), which means that the student has total freedom in creating their learning goals and picking their learning material—thus, both the learning goal and the way there is more or less to be planned and, more crucially, decided by the student (alearningjourneyweb, 2017). This is unlike self-regulated learning, which tends to define specific learning tasks that have to be completed by the student, thus fixing a certain learning trajectory in place.

2.1.2. A Brief History of Technology-Enhanced Learning and E-Learning

Because e-learning wasn't created so much as it naturally developed, finding a specific "start" to its history is a difficult task, as it developed differently in various sectors (Nicholson, 2007). Generally it is fair to say that the concept of e-learning itself has existed for a long time, however.

Back in the 1960, Patrick Suppes and Don Bitzer already realised the possible use of computers almost like personalised teachers. While Patrick Suppes created a computer managed instruction system to be used in his courses which was already self-paced like a lot of MOOCs are nowadays (Molnar, 1997; Nicholson, 2007), Don Bitzer ended up creating PLATO, a system that intended to help develop and deliver computer-based education (Nicholson, 2007).

Contrary to the modern version of e-learning, computer-based training—generally considered to be the first form of electronic education—still required some sort of connection to a data medium, such as a CD-ROM. As this was before the rise of the web, computer-based training was still location-dependent, as well as time-dependent, unlike today's e-learning (Hubackova, 2015).

Following computer-based training and the emergence of the Web, Web-based training started to become popular. This, then new, form of training did not only provide a way of teaching, but also an opportunity for teachers and students to communicate (Hubackova, 2015). As this presented a range of new possibilities, new research and projects quickly followed.

It was around this time that the potential for the usage of technology in education started to truly be realised. Though there had been developments in the field before, the actual term *e-learning* has only been in use since 1999 (Hubackova, 2015). It was in this year too that John Chambers, then-CEO of Cisco Systems, predicted that education would ultimately take up a big portion of internet capacity, stating that schools that did not move with the times might end up seeing students pass them over in favor of more internet-based schools and courses (Friedman, 1999).

While not quite as drastic as Chambers predicted, interest in e-learning quickly began to emerge and evolve, which in turn caused a variety of platforms to be created in response.

One example of this heightened interest in using the Web for education and lifelong learning was the creation of MIT OpenCourseWare, which was one of the pioneers for sharing course materials across higher education. Starting out with only 50 courses in 2002, it has since only grown in content and participants, containing over 2500 courses in 2018 (OpenCourseWare, 2020).

Years later, Stanford University trialled taking some of its courses online and offering them to whoever wanted to join. Originally started with a selection of only three courses, they managed to draw in more than 100,000 students per course in their first semester in 2011 (Severance, 2012). One of the courses, "Introduction to Artificial Intelligence" by Peter Norvig and Sebastian Thrun, is generally credited with being the first open online course, whose platform eventually went on to create Udacity (Moe, 2015).

Reacting to the big success of the online courses, though especially that of "Introduction for Artificial Intelligence", various online platforms for e-learning were formed, Coursera being one of them (Severance, 2012).

2.1.3. Forms and Techniques of E-Learning

E-learning itself is a collective term that can be used to signify a number of different forms of learning. The most basic division is between synchronous and asynchronous e-learning. Synchronous e-learning uses technology such as video conferences and chats—communication methods that require the participants to be online at the same time (Hrastinski, 2008). On the other hand, asynchronous e-learning is time-independent, using technologies such as forums and e-mail, which makes it ideal for self-paced courses, such as most MOOCs nowadays (Soni, 2015).

Asynchronous e-learning can once again take different forms, depending on the learning method employed. Two of the more common learning methods were already mentioned in section 2.1.2, computer-based training and Web-based training. Nowadays, however, the use of other methods has grown, some of them being blended e-learning, social learning and game-based learning (Soni, 2015). In the following sections, this work will give a short description of these, comparatively newer forms of asynchronous e-learning.

Blended E-Learning

Blended (e-)learning can describe various teaching methods, depending on the context, as Driscoll (2002) explains. In the definition that will be used for this work, blended e-learning is defined as a combination of e-learning activities and instructional materials, and face-to-face teaching methods (Littlejohn and Pegler, 2007; Kahiigi et al., 2008). This way, traditional teaching methods that are typically location- and time-specific can be enriched with the opportunities newer technologies offer.

According to Driscoll (2002), some easy ways to employ blended e-learning are providing an online forum for students to exchange questions and answers regarding the learning subject, providing preparatory materials online before the face-to-face sessions, and putting the final assessments online—for example in the form of multiple-choice tests.

Social Learning

With social media taking up an increasingly large part of people's lives, it only makes sense that learning, too, has evolved to make use of this development. Consequently, social learning describes learning that utilizes social media of some kind to facilitate the learning process.

According to Wodzicki et al. (2012), "*social networking sites allow students to connect formal and informal learning settings*". As learning is, like Amry (2014) writes, "*the outcome of social interactions between students in collaborative learning activities*", this opportunity for knowledge exchange and student cooperation is typically expected to enrich the learning process.

Twitter, WordPress and YouTube are only a few examples for such technologies that can be used as education tools, with some, such as Google Documents, having become cornerstones of online collaboration between students (Dabbagh & Kitsantas, 2012). Facebook has become essential for networking between students (Dabbagh & Kitsantas, 2012), facilitating the creation of study groups or even online knowledge exchange in groups. Online blogs have become an increasingly more important source of supplementary learning materials or explanations, but are also used by educators to facilitate knowledge exchange by having students create blog posts about chosen or assigned topics to increase participation and the feeling of ownership in a course (Top, 2012).

Game-Based Learning

Game-based (e-)learning attempts to use what Garris et al. (2002) call "instructional computer games" for educational purposes, enabling the user to actively interact with and apply the knowledge to be learned (Pivec, 2007). According to Boeker et al. (2013), this has the effect of increasing students' motivation when interacting with the learning material. These interactions, and more generally game-based learning itself can take different forms, depending on the area in which it is used.

On the one hand there are, for example, mobile applications that allow players to compete in more or less trivia-based games, containing subjects from politics over geography to physics. On the other hand there are applications that are more direct in the skills they teach.

One example for this is the application Scratch, which provides a playful introduction into the basic structures of programming by giving the user puzzle pieces that represent certain programming blocks/actions (e.g. for-loops, if statements) (Scratch, 2020). The applications lets the user create their own games, change existing games by other users, and generally learn by applying the knowledge while also playing what essentially amounts to a game, thus making it reasonable to consider it an example for game-based learning.

Gamification

While gamification shares part of its name with game-based learning, there are significant differences in how gamification works. According to the most popular definition, gamification is considered to be *"the use of game design elements in non-game contexts"* (Deterding et al., 2011).

As opposed to game-based learning, which refers to games which were primarily built around the purpose of teaching the player, gamification works to extend (in some cases previously existing) study material by adding elements of playfulness and gaming. This can be done in various ways, with one of them being the addition of game-fiction, id est (i.e.) a fictional story-line or context (Sailer & Homner, 2020).

Persuasive Technologies and Design

Fogg (2003, p.15) defines persuasion as the attempt to change a person's behavior and/or attitude towards a topic or matter without the utilisation of outright manipulation, coercion or deception. Additionally, it is stipulated that true persuasion has to be intentional on the designer's part, not an unplanned side effect of the technology. Persuasive technologies are thus purposefully created technologies that strive to influence the user in some way.

In learning in particular, persuasive design is often used in order to increase learners' motivation, performance, and general engagement (Engelbertink et al., 2020). Personalized approaches have been found to be especially effective, with some of the most effective strategies being reward, competition, social comparison, and social learning (Orji et al., 2019). When looking at those strategies, it is noteworthy that all but reward are considered types of social support, with only reward being a type of dialogue support (Oinas-Kukkonen and Harjumaa, 2009; Engelbertink et al., 2020). The social aspects should thus not be underestimated, even when it comes to technologies.

Bite-Sized Learning

As life grows faster, sitting down to learn for long stretches of time, especially once the person is not learning full-time anymore, becomes difficult. Nonetheless, life-long learning is becoming more and more of a requirement in the fast-paced job market of today. Bite-sized learning is the answer to this often shortened time-frame the learners have available to concentrate on gaining new knowledge, which has at the same time become close to indispensable.

Kulkarni and Naik (2019) describe bite-sized learning as *"small, self-contained information nuggets with a focus to achieve defined objectives"*. Bite-sized learning modules are supposed to take little time to finish, so that one may complete such a "bite" quickly and not lose any of the underlying information value. This is possible by having only one single, very specific learning objective per module, as opposed to conventional e-learning courses, which often consist of multiple learning objectives (Kulkarni & Naik, 2019).

Mobile Learning

Mobile learning refers to learning that is intended to take place mobile, and considers this during the development of the lesson/application. Examples for this can range from productivity applications to flashcard learning applications, or even the use of the internet to find information. When including mobile learning in a more traditional learning environment, it is necessary to consider the available devices and constraints that follow because of them (Chen & Denoyelles, 2013).

2.1.4. State of the Art and Problems

Since e-learning's development, higher education institutions especially have started using increasing amounts of e-learning content in their teaching. Because of the technical advancements, both in general and in the learning sector specifically, e-learning now allows for multiple content presentations, personalization of the content and the learning tactics employed in the run of a course, and ubiquitous learning (Kahiigi et al., 2008).

While it has, in general, been accepted reasonably well, there are still roadblocks that stop the use of e-learning from being as far-reaching as it could be.

The most basic problem connected to e-learning is the possible lack of training in how to use the technological means to enhance the learning and/or teaching process (Riasati et al., 2012). Depending on the technology to be used, additional training may be necessary for the teachers, which is not always feasible. Furthermore, some teachers resist the changes in the teaching approach.

Another possible barrier that may play a role in hindering e-learning from being fully immersed in the teaching process is that it often requires teachers to revise their courses or create new course material instead of being able to reuse the content that they may have prepared in the past. In combination with what may be an entirely unfamiliar tool to the teachers, creating this new learning material often takes longer when taking e-learning technologies into consideration than it does to update and use traditional learning materials (Riasati et al., 2012).

In their work, Kulkarni and Naik (2019) mention the information overload that students are confronted with nowadays, and the shortened attention span many learners struggle with, partly due to that reason. They postulate that this is a major factor for why bite-sized learning is quickly gaining influence in the e-learning sector, especially in job training environments. Considering the vast amounts of information often conveyed in the course of a lecture session, bite-sized learning may not be a possibility in the education sector, however, or may at the very least have to be adapted to fit the environment.

Whether the answer is bite-sized learning or not, this shift in how information is processed needs to be addressed in the way e-learning presents its information and the way the learners interact with the learning material, not only in businesses but also schools.

In part, this increased shortness of the attention spans links to the shift in how learners use technology nowadays. Often, phones are faster and more easily available than ever before, so that shifting learning from computers to phones seems like a reasonable action and may be one way to deal with the changes in how people learn (Kulkarni & Naik, 2019). However, in many cases e-learning content is not optimised or even prepared to be mobile-ready, so that there is a need to review the presentation of the learning content and possibly even platform the learning content is made available through.

Another problem e-learning, especially as it pertains to courses that are entirely or mostly done online, has to deal with is that regularly updating big courses takes a lot of time and effort from the instructor, especially in fields that change rapidly. It may not be reasonable to expect the lecturer to create new video material, which is the preferred lecture medium in many MOOCs, every time the study field changes.

Depending on the subject matter covered in the course, the decision for how to design the MOOC may also present hardships. Some subjects lend themselves better to practically-oriented MOOCs, which has to be considered during the creation of the MOOC itself. To do this, it is necessary for the creator to have a certain familiarity with the different types of MOOCs that exist. Some of those types will be shortly described in the next section.

2.1.5. Introduction to MOOCs

Traditionally, the term MOOC is an overarching term for openly accessible online courses that are intended to educate a large number of people, irrespective of their place of residence, age, or living circumstances. The original meaning has since morphed to be less strict about how open the courses are, or about their massiveness, with the number of different MOOC types having multiplied since the initial creation of the nomenclature.

Originally, MOOCs were differentiated into two sub-types beyond this overarching term, with a look at the history of MOOCs revealing mostly cMOOCs (Kesim & Altınpulluk, 2015). In the following sections, this work will give a short description about both this type, and of xMOOCs.

Beside those two major categories, MOOCs have since also been categorized according to their various features, thus increasing the number of different types significantly. Pilli and Admiraal (2016), for example, proposed a new categorisation according to MOOC size and openness, which the work will give a short overview of. This taxonomy was chosen because it was one of the most recent ones, as well as one of the most accessible ones, with only two axes along which MOOCs can be placed. In the course of explaining this taxonomy, the chapter will also give an overview over various categorisations beyond the split into cMOOCs and xMOOCs, and how they fit into the categories proposed by Pilli and Admiraal.

cMOOCs

cMOOCs are based on the connectivist approach, hence their name (Kesim and Altınpulluk, 2015; Smith and Eng, 2013). The connectivist approach states that the student "*is responsible for their own learning*" (Kesim & Altınpulluk, 2015), a process that cMOOCs have adopted, emphasising the social connection across the network of learners (Conole, 2016). McLoughlin and Magnoni (2017) describe cMOOCs as being delivered "*through collaborative tools, blogs and discussion boards rather than content/learning management systems*", in a throwback to what has fallen out of the norm nowadays with platforms such as Coursera, Moodle and edX easily available.

Due to how they are presented, cMOOCs are quite free in their structure, especially regarding the learner's time management, resource use and the taken learning path (McLoughlin & Magnoni, 2017). Because of this immense freedom while navigating the cMOOC, users are especially dependent on their abilities to collaborate and communicate, and use critical thinking and information media skills (McLoughlin & Magnoni, 2017).

In addition, cMOOCs are harder to grade and award a certification for, as McLoughlin and Magnoni (2017) explain. This is because the learning path is free and completely open to the learner. It is possible for the direction of the course to change if the students' input and interests directs it that way, and as the students are welcome and even invited to present their own input and opinions into the course to influence its trajectory.

xMOOCs

According to McLoughlin and Magnoni (2017), xMOOCs are closer related to traditional learning environments than cMOOCs, focusing on content more than the cMOOC counterpart and adhering to the behaviorist approach. Due to the closer mirroring of traditional learning environments, xMOOCs are also easier to grade, as they set certain prerequisites for the user to be able to finish the course.

Current MOOC courses, especially those offered by universities as parts of a study program or on sites such as Coursera and edX, are most commonly xMOOCs, even though it is the younger type of the two (Kesim & Altınpulluk, 2015).

xMOOCs often present the possibility of being counted towards a degree at universities that accept their certifications, since they effectively act almost like traditional university courses, often with given timeframes in which the course has to be completed, graded assignments, and course contents that are clear and specified by the teacher and are not influenced by the students.

Pilli and Admiraal's Taxonomy

Pilli and Admiraal consider more than the two main types xMOOC and cMOOC for their taxonomy. They chose the two categories *Openness* and *Massiveness* to categorize a large number of newer MOOC types. As visible in Figure 2.1, they named four quadrants, each representing a MOOC category, where scale references the size/allowed number of participants in the MOOC and openness refers to how open the course is to participants in regards to, among other things, freedom of access, time constraints, place constraints, et cetera (etc.) (Pilli & Admiraal, 2016).

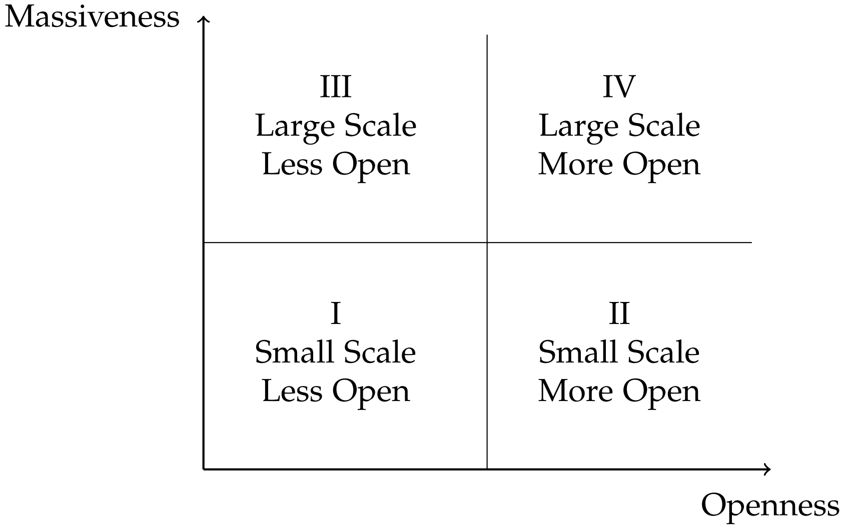


Figure 2.1.: Pilli and Admiraal's Taxonomy for MOOCs. Adapted from Pilli and Admiraal (2016).

Quadrant I contains MOOCs that are smaller in size and less open. An example for this would be a university-specific distance-learning course, as those are typically only available to the students of a particular university course. The MOOC types that Pilli and Admiraal (2016) consider part of this quadrant are described in short in the following paragraphs.

- **SPOCs** refer to *Small Private Online Courses*. In short, they could be described as a MOOC that is designed to be used as part of a traditional university course. These expect the collaboration between the students and the teachers, as well as between the students. These interactions can take place in person or online (Pilli & Admiraal, 2016).

- **groupMOOCs** focus on collaboration in smaller groups (Conole, 2016). Additionally, each student is assigned a mentor. Both of these things are done in an attempt to counteract the high dropout rates generally associated with MOOCs (Pilli & Admiraal, 2016).
- **Task-based MOOCs** set tasks for the learners. However, there are many ways that those tasks can be solved, and the decision is ultimately down to the learners themselves (Pilli & Admiraal, 2016).

Quadrant II considers MOOCs that are small and open—besides the previously described cMOOCs, this also encompasses a number of different types, such as *BOOCs, COOCs, DOCCs, POOCs, LOOCs, gMOOCs, pMOOCs, adaptive MOOCs and network-based MOOCs* (Pilli & Admiraal, 2016). In the following paragraphs, this work will describe a selection of the different types in short.

- **COOCs**—Community Open Online Courses—are typically courses that are provided by a corporation to train its employees and/or customers. According to Pilli and Admiraal (2016), these sorts of MOOCs are often also known as *Corporate MOOCs*.
- **gMOOCs** are game-based and game-based learning MOOCs and include serious games as part of the teaching material, or games that can otherwise be valuable to the learning experience (Pilli & Admiraal, 2016).
- **pMOOCs**, also known as *Project-Based MOOCs*, place special emphasis on the development of projects (Pilli & Admiraal, 2016).

Quadrant III categorises MOOCs that are large and less open. Example types of this category are VOOCs, SMOCs, HOOCs, miniMOOCs and POOCs, a selection of which this work will be describing in the following paragraphs.

- **VOOCs**, or Vocational Open Online Courses, are usually short (around an hour from start to completion) and show examples of practical procedures in addition to experts' advice (Pilli & Admiraal, 2016).
- **POOCs** are Personalized Open Online Courses and adjust their content to the learner, according to their preferences, knowledge and actions in the course (Pilli & Admiraal, 2016).

Finally, quadrant IV considers large and open MOOC types. The other one of the two originally considered MOOC types, xMOOC, is part of this category. Besides that, it also includes transferMOOCs, madeMOOCs, asynchMOOCs, SPOCs, Content-based MOOCs, Flex-MOOCs, iMOOCs, MOOC-Eds, and MOORs (Pilli & Admiraal, 2016).

- **TransferMOOCs** are MOOCs that are created with the transferred material from traditional, class-based courses (Pilli & Admiraal, 2016). The existing courses are simply moved over to the online version (Conole, 2016).

- **MadeMOOCs**, on the other hand, are specifically made with online knowledge transfer in mind. Because of this, they are more interactive than transferMOOCs, often having the learner doing exercise questions on the MOOC page itself (Pilli & Admiraal, 2016).
- **AsynchMOOCs** is a course type that can be found on Coursera and which has no specific start and/or end date. They do not have a fixed schedule, nor do they have deadlines, or if they do, they are more flexible than in other types of MOOCs (Pilli and Admiraal, 2016; Conole, 2016).
- **iMOOCs** place special focus on social learning and learner interaction, more so than other MOOC types. Completed assessments from the learners are uploaded and published for others to see (Pilli & Admiraal, 2016).

The list of MOOCs named in this section is not intended to be exhaustive—with the various taxonomies, there are also a large number of MOOC types, some of them very specific to only a small number of MOOCs, that simply cannot all be considered in this work in an efficient manner.

2.1.6. The Coursera Platform

As part of this thesis will be focusing on data retrieved from Coursera MOOCs, this section will be giving a short description of the platform.

Coursera was established as its own, independent company in 2012 by Daphne Koller and Andrew Ng, both of whom were professors for Computer Science at Stanford University, just one year after the use of online courses was trialled at Stanford (Severance, 2012).

With 3,800 active courses and 45 million learners by the end of 2019 (Shah, 2019), Coursera's aim to provide education and learning opportunity has shown to be well-accepted by both course creators as well as students. The large number of MOOCs across different fields of knowledge, from humanities over technological fields to languages, indicates that this is not just the case for technological fields, even if the first three courses that were originally offered in the pilot program at Stanford University were all part of their Computer Science program (Severance, 2012).

Nowadays, it is not just one university that supplies the teaching materials. Courses are supplied and maintained by organizations across the world, including universities and institutions such as IBM, Princeton and Yale, with participation itself being free of charge, though certification of completion involves payment.

Besides stand-alone courses, Coursera offers different modes of course organisations on its website. *Specializations* and *Professional Certificates* consist of sets of courses surrounding the more general topic topic, while *Degrees*, are set up by universities and function just as their conventional counterparts.

MasterTrack Certificates on Coursera can be used to count towards a master's degree. Similar to degrees, the courses of MasterTracks tend to be paid (Shah, 2019).

Courses on Coursera can have multiple sessions per year, which users can enroll in. Depending on the course in question, users may then be provided with video lectures, graded and/or peer reviewed assignments, programming assignments and quizzes. The courses also provide discussion forums where users can post and answer questions, and interact (Coursera, 2018a).

To be able deal with the amount of grading required to be able to assess the students of the MOOCs when there are only a limited amount of teachers responsible for the course, most of the courses nowadays make use of automated grading and peer assessments. In a similar vein, questions are commonly answered by other students and less often by the teacher. As MOOCs are often quite large, this also presents a positive aspect for the students, as the probability is high that there will be a student who is able to help, who is also available at about the time the question is asked (Severance, 2012). Because of this, the turnaround rate in the forums is significantly faster through student involvement than it would otherwise be.

2.1.7. Learning Strategies in E-Learning Environments

While one of the aims of MOOCs is to make education available to everyone and facilitate lifelong learning, taking part in them does present its own, very specific challenges. Due to their often close to teacher-less nature, MOOCs require a high level of self-efficacy, goal-setting and task interest from the student (Lung-Guang, 2019). In addition to that, self-organisation plays a significant role in MOOCs, especially for those that do not have a fixed schedule and where the student has to rely on themselves to continue on through the course.

However, a number of those traits—self-efficacy, goal-setting, and self-organisation in particular—are learned behavior. Because of this they need to be taken into account when considering different educational backgrounds of students. People who have finished higher education may have had the chance to develop those skills to a certain degree already, while those that did not attend higher education may have starting difficulties. Due to this, it is necessary to make considerations for how to mitigate possible negative effects of initially poor self-efficacy, etc., to retain students who may otherwise drop out.

To further explain user behavior and requirements for successful completion of MOOCs, Lung-Guang (2019) used both the Theory of Planned Behavior (TPB) and the model of Self-Regulated Learning (SRL) in his study, considering the students' decision-making processes as they were participating in the MOOCs.

According to Glanz et al. (2015), the TPB is a theoretical model that tries to connect motivational factors with the likelihood of specific human behavior in situations in which the person has or perceives to have a measure of control. The TPB also considers the influences of norms and perceived control of other parties, and the effects this may have on the performed actions (Lung-Guang, 2019). The TPB is widely accepted and has been used in multiple research into user behavior in MOOCs (Lung-Guang, 2019; Yang and Su, 2017).

The model of SRL contains strategies such as self-evaluation, goal setting, monitoring, etc. (Zimmerman & Martinez-Pons, 1990), making it relevant when considering MOOCs. As MOOCs do not, for the most part, offer a fixed schedule apart from a start and end date, it is necessary for the student to set their own goals and evaluate their performance throughout the course, as well as adapt accordingly. Additionally, student support is generally lower than in traditional learning environments, often only taking place through forum posts. According to Kizilcec et al. (2017), SRL is a teachable skill, and by identifying learner characteristics that point towards lower SRL skills, it may be possible to pinpoint students who are at risk of dropping out, and supporting them.

However, in a study by Kizilcec et al. (2016), it was found that simply recommending SRL strategies alone did not have a significant effect on student behavior. According to their research, the "*overall percentage of viewed lectures and passed assessments was similar*", both when students were shown SRL study tips and when they were not.

Looking at both TPB and SRL and the skills described in each, it becomes clear that both of these models place a high level of importance on self-interest and self-motivation as two significant factors to carry a person through a specific task (Lung-Guang, 2019). This does not differ when it comes to MOOCs, as students who take part are required to self-regulate their learning behavior.

In addition to TPB and SRL, it is also necessary to consider intrinsic and extrinsic motivation when it comes to MOOCs. Either or both can be the reason for a user to start a MOOC. In one study, Littlejohn et al. (2016) found that learners with higher SRL scores tended to be intrinsically motivated, while those with lower SRL scores tended to be extrinsically motivated. Motivation has also been found to increase when student interaction is encouraged (Richards, 2011), and in turn seems to increase student engagement with the course content, which furthermore has positive effects on completion rates (Xiong et al., 2015).

According to Yang et al. (2013), the growth of social connectedness during a MOOC's progression plays an important role in minimising student attrition. This is reflected in student behavior in both MOOCs, as well as distance universities, as both are faced with high dropout rates (Kotsiantis et al., 2003). Both of those education systems have little to no personal student interactions, which presents a challenge in retaining the students. According to Yang et al. (2013), increasing student interaction may have a positive effect on user retention. In our research, this work will also be taking a look if this held true for the data that was available to us.

In another study, Lung-Guang (2019) found that involving students in the planning of their actions, and encouraging them to do so, increases students' perception of planned behavior. This in turn improves decision-making skills, improving, furthermore, planned behavior. As planned behavior is a significant basis for successfully finishing a MOOC, this should be considered by the teacher. In their research, Kizilcec et al. (2017) found that both goal setting and strategic planning "*predicted attainment of personal course goals*".

2.1.8. Description of a Self-Regulated Learning System

In chapter 4, this work will be using data from a self-regulated learning system to model user behavior. This system was created by Alexander Nussbaumer to be used at an Austrian university, in context of a larger class.

The aim of the course is to supply learning material in a similar way as traditional learning systems, such as Moodle or MOOCs, in order to make them comparable. It leaves the choice of how and if to consume this material to the learner. The main page shows an index/table of contents, listing and linking to all course sections, as well as course content chapters.

The system forgoes more complicated navigation logic and supplies the user only with "Index", "Previous" and "Next" buttons on each page, where the "Next" and "Previous" buttons simply move along the index in the given direction.

An example structure of the index is shown in Listing 2.1

```
- Instruction
- Section 1:
  - Learning Goal
  - Content Chapter
  - Content Chapter
  - Assessment
- Section 2:
  - Learning Goal
  - Content Chapter
  - Assessment
- Learning Progress
```

Listing 2.1: Simplified course structure/index of the self-regulated learning system

Each section covers a singular topic, with content chapters focusing on a specific aspect or example.

2.2. Data Science and Learning Analytics

E-learning in general and MOOCs specifically produce a large amount of data that can be used to improve the learners' interactions and success with and in the course. When applying data science to click-stream, forum and interaction data, it is possible to draw important conclusions from it that can significantly help when improving the course experience. While some people argue that data science is just statistics with a new name, Blei and Smyth (2017) instead postulate that data science is "*the child of statistics and computer science*".

According to them, data science combines some features from statistics—the foundations of analyzing and reasoning about data—with some that are more typical for computer science, such as certain methods of computational thinking. While statistics provide the features to be able to deal with high dimensionality of data, as well as complexity and causality, computer science is responsible for optimisation, sampling and scaling (Blei & Smyth, 2017).

In the following section, this work will discuss further the contributions that computer science makes to data science, as well as shortly describe the field of data analytics. Then this work will provide a short overview over data and process mining, as well as give a brief summary of the field of learning analytics. Finally, selected data science methods will be introduced.

2.2.1. Data Science and Data Analytics

Data analytics is a broad field that consists of a number of different methods that are all considered part of it. There are, however, some basic, statistical concepts that are at the base of most or even all of them. Probability, sampling, and prediction are concepts that are particularly meaningful in this context (Berthold & Hand, 2003). They are heavily used across the different methods that can be considered to be part of data analytics.

Looking at data analytics, they can be further split into a number of categories, as listed by Berthold and Hand (2003). The most meaningful methods in context of the research covered in this thesis, however, are statistical methods such as linear models, regression modelling and multivariate analysis, Bayesian methods, and support vector and kernel methods, of which support vector machines are part. Furthermore, there are rule inductions, and neural networks. Lastly there are stochastic search methods, which can consist of genetic algorithms, simulated annealing and evolution (Berthold & Hand, 2003).

In section 2.2.4, this work will give short explanations on the methods that are most influential in the research direction that this thesis will be covering.

2.2.2. Data and Process Mining

The line between learning analytics and data mining—especially data mining in the educational sector—can seem blurry and vague. Even in literature, (educational) data mining and learning analytics seem to go hand in hand and are often only used together.

According to Baker and Inventado (2014), data mining differs from learning analytics in that it focuses more on automated methods than learning analytics do, though they stated that ultimately the difference comes down to the use of models more than the methods used to get the results. Educational data mining especially is most often used in conjunction with learning analytics, with each using methods of the other field, which makes a clear delineation of where one ends and the other one begins difficult.

Data mining does its best to gain useful information from the huge amounts of data that are available nowadays. Hand et al. (2001, p.1) define data mining as *"the analysis of (often large) observational data sets to find unsuspected relationships and* *to summarize the data in novel ways that are both understandable and useful to the data owner*".

Cherkassky and Mulier (2007, p.15) explain that "*in a more narrow sense, many data mining algorithms attempt to extract a subset of data samples (from a given large data set) with useful (or interesting) properties*". To do this, some of the common methods and algorithms are Decision Trees, Bayesian networks, and Support Vector Machines (Grigorova et al., 2017).

To show how these and other methods fit into the wider field of data mining paradigms, this work references Maimon and Rokach (2014) and their description of how various types of data mining paradigms are in relation to each other. The graphical representation of this relationship can be found in Figure 2.2.

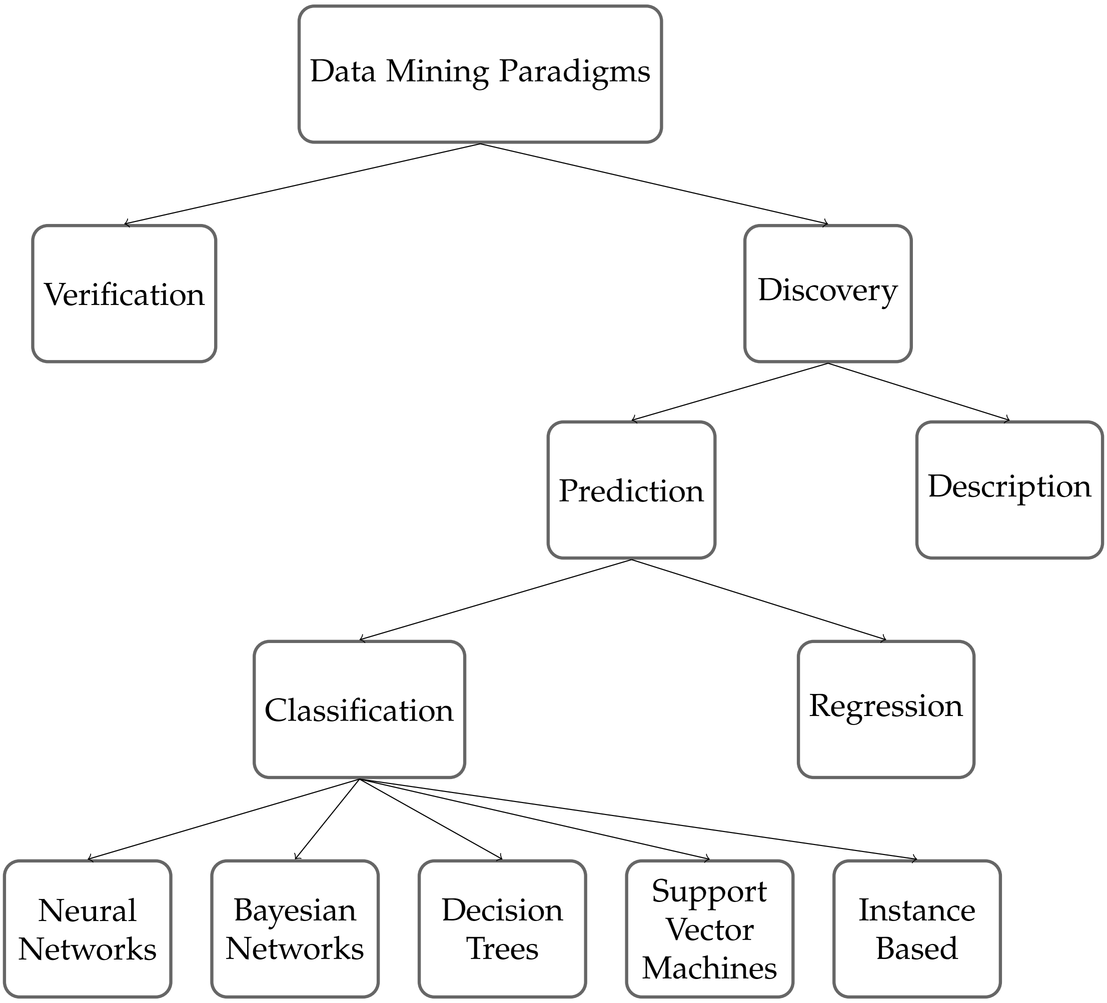


Figure 2.2.: Taxonomy of data mining paradigms. Adapted from Maimon and Rokach (2014, p. 166).

This work will mostly be focusing on classification, and furthermore on decision trees, as these are what will be used for the analyses in later sections.

As opposed to data mining, process mining is more concerned with discovering patterns of and insights into the process itself, rather than into singular actions and data points (Mukala, Buijs, & Van Der Aalst, 2015). In the learning environment, process mining can be used to analyse student learning behavior by using process models built from click data collected from course interactions (Mukala, Buijs, Leemans, et al., 2015). One such example would be the fuzzy miner, as used by Mukala, Buijs, Leemans, et al. (2015), to find process models.

Similarly to process mining, relationship mining focuses on the relationship between features in a given data set. This can take various forms, depending on both the type of data available and the knowledge the researcher is aiming for. According to Baker and Inventado (2014), there are four kinds of relationship mining that are particularly meaningful in the educational sector, those being association rule mining, sequential pattern mining, correlation mining, and casual data mining.

2.2.3. Learning Analytics

The U.S. Department of Education (2012) defines learning analytics as "*incorporating concepts and techniques from information science and sociology, in addition to computer science, statistics, psychology, and the learning sciences*". With the overall aim to, basically, understand the students' learning processes, learning analytics place a strong emphasis on human reasoning, as opposed to data mining, which attempts to automate the discovery process (Viberg et al., 2018). Often used side by side with (educational) data mining, learning analytics by themselves focus on making use of business intelligence, and various types of analytics, such as academic, action and predictive (Papamitsiou & Economides, 2014).

Independent fields of their own, *academic analytics* specifically focus on analysing educational data at a regional or institutional level rather than across the whole field, while *action analytics* focus more on the teacher than the student—they function as a sort of self-reflection and promote reflection on teaching methods and materials employed by the teacher (Steiner et al., 2014).

Furthermore, other than social network analysis, which can be used to, amongst other things, find students who are not sufficiently connected to the group, and/or students with high influence (U.S. Department of Education, 2012), learning analytics also use technologies such as data visualisation, prediction and relationship mining to analyse the available learning data (Viberg et al., 2018).

In much of the recent literature, learning analytics were used not after a course finished, but while it was still ongoing, as this gives the teacher the chance to intervene and counteract negative student performance or negative developments in the learning progress (Steiner et al., 2014). Chatti et al. (2012) described the learning analytics process as circular, as visualised in Figure 2.3, with the effect that it lends itself well to being used for constant evaluation and adjustment.

Learning analytics have been used to detect students who are considered to be at risk of failing a course to allow intervention and thus an increase in the chance of success (U.S. Department of Education, 2012). However, while they have been found to achieve reasonable results in raising learning support and teaching in

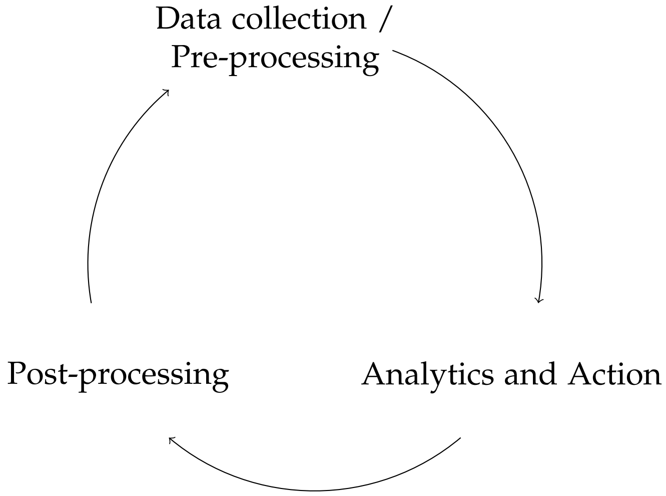


Figure 2.3.: Learning Analytics Process. Adapted from Chatti et al. (2012, p. 6).

35% of the cases, they have not brought a significant increase in student learning outcomes, with only 9% of the research results Viberg et al. (2018) considered in their literature review showing improvements. Though Viberg et al. (2020) later found that this was the case with 20% instead, even this higher number does not do the possibilities of learning analytics justice.

On the other hand, when Schumacher and Ifenthaler (2018) asked students what they expected of learning analytic features, they found that planning and organization support, self-assessments, and adaptive recommendations were frequent answers, as well as personalised analyses of learning activities. This makes learning analytics not just interesting from the teachers' views, but also from the students'. This is especially the case in self-regulated learning, as the high amount of independence and autonomy can lead to a higher need for external encouragement and support.

2.2.4. Selected Data Science Methods and Models

In the past years of research, a broad variety of different methods to analyse and model data were used to give insight into the in part massive amounts of data that exist nowadays. Some of them have become especially prevalent in the area of e-learning, such as decision trees, which are often used in an attempt to predict dropouts.

Additionally, this chapter will pay attention to splitting criteria, boosting and bagging, and the problem of model evaluation, all of which become relevant in chapter 3. Finally, there will be a short description and explanation of Markov chains, as well as sequential analysis, the last of which will be used to examine the user behavior in the self-directed course analysed in chapter 4.

Decision Trees

The decision tree algorithm is a supervised data mining technique that is used to build a predictive model which can be used for classification tasks (Li, 2019; Rajeshkanna and Arunesh, 2018). This tree represents the decision rules on the way from an initial situation to an ultimate result/class label (Alharan et al., 2017). Due to the way it is constructed it is always a directed graph, starting at a root with no incoming edges, and ending at the leaves, which have no outgoing nodes and represent the most likely class affiliation according to the edges traversed from the root (Maimon & Rokach, 2006).

Decision trees (see Figure 2.4) can either be built from the top down or from the bottom up, depending on the algorithm used, with literature typically preferring the top-down approach (Maimon & Rokach, 2006). In this case, the building process starts at the root, with the first attribute chosen through whichever method was used to pick it. Then, another node is drawn for each of the options that follow—these can be boolean values, numerical ranges, or even strings—and the second attribute is added for each. The attributes do not necessarily have to match on the two branches, since the information gain may change depending on the previously given true/false answer.

Each non-leaf node represents a feature that the data is split on, while each edge constitutes as a value, or range of values in the case of numeric attributes, that the feature can take on (Maimon & Rokach, 2006). An example of a simple decision tree using a small subset of features that will be used in the analyses in chapter 3 is shown in Figure 2.4.

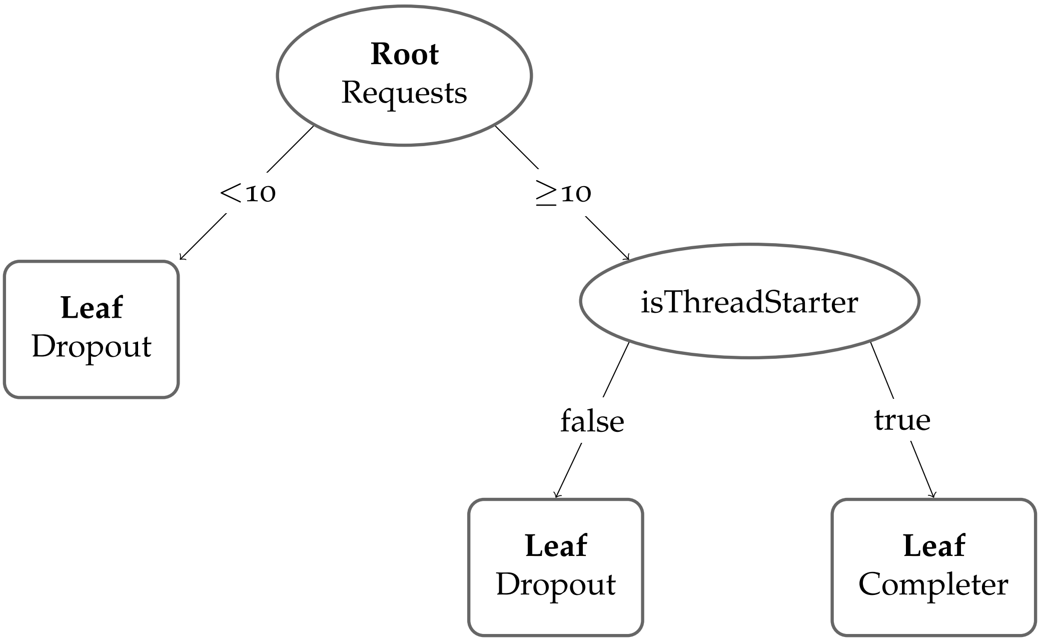


Figure 2.4.: Structure of a decision tree, exemplified with features used in this work. Adapted from Maimon and Rokach (2006, p. 166).

In this example, the feature *Requests* is numerical, thus the outgoing edges give ranges of values to split the sample set on. The feature *isThreadStarter*, however, is binary with only True and False being valid values. Numeric attributes, especially if they do not split in a binary way (unlike in the example mentioned in this section), increase the complexity of the decision tree, and anything but binary splits should be used with caution, as increased complexity affects model accuracy (Breiman et al., 1984).

Once a certain stopping criterion is hit—e.g. a maximum tree depth is reached—the algorithm terminates, ending each branch with a final node—the leaf—with the corresponding class label of the combination of attribute values along the branch. The depth of a decision tree is measured by the number of "levels" it has from the root node to the last/deepest leaf. The *Root* node would be level 0, with the one below that being level 1, and so forth.

After the tree has been created, it can then be used to predict class affiliation.

While the basic decision trees are only rarely used nowadays, there are newer variations that have found broad application in the area of dropout prediction, e.g. Boosted Decision Trees (BDTs), which will be used for some of the analyses to come.

Splitting Criteria

To be able to choose the best split at each level, decision trees need some measure to differentiate between a good and a bad split, and how to decide which feature to split on next. These measures are referred to as *splitting criteria*. There are many different criteria that can be used when building decision trees, often depending on the type of data to be analysed. Two of the most commonly used criteria and the ones this section will be focusing on are *Information Gain* and *Gini Index*.

At its core, *Information Gain* is used to express the impurity of a set of samples (Raileanu & Stoffel, 2004). To do so, it basically compares the entropy before a split with that after a split and calculates the difference. The formula for this method is shown in Equation 2.1, as given by Maimon and Rokach (2006).

$$
\begin{aligned}
InformationGain(a_i, S) &= Entropy(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{|\sigma_{a_i = v_{i,j}} S|}{|S|} \cdot Entropy(y, \sigma_{a_i = v_{i,j}} S) \\
Entropy(y, S) &= - \sum_{c_j \in dom(y)} \frac{|\sigma_{y = c_j} S|}{|S|} \cdot log_2 \frac{|\sigma_{y = c_j} S|}{|S|}
\end{aligned}
\tag{2.1}
$$

Raileanu and Stoffel (2004) derived the formula by first giving the probability for a random sample to be assigned to a random class c_i. This so-called prior probability, denoted as $p(c_i)$, gives the likelihood that a sample belongs to some class c_i. The information content of this probability can be calculated by simply taking the negative binary logarithm of this probability, which can then further be used to formulate the average necessary amount of information to classify a sample as belonging to the class, as shown in Equation 2.2.

$$info(S) = -\sum_{i=1}^{k} p(c_i) log_2(p(c_i)) \tag{2.2}$$

After a split, where n is the number of different outcomes, and T is the attribute test, this equation would then look as in Equation 2.3

$$info_T(S) = \sum_{i=1}^{n} p(t_i) info(T_i) \tag{2.3}$$

The information gain by conducting a split would thus be the difference between $info(S)$ and $info_T(S)$, as in Equation 2.1. Or, as Raileanu and Stoffel (2004) describe it, Equation 2.4.

$$InformationGain(T) = -\sum_{i=1}^{k} p(c_i) log_2(p(c_i)) + \sum_{i=1}^{n} p(t_i) \sum_{j=1}^{k} p(c_j|t_i) log_2(p(c_j|t_i)) \tag{2.4}$$

According to Raileanu and Stoffel (2004), for a node/test T with n different outcomes and k different classes, the *Gini Index* is defined as in Equation 2.5, and estimates the probability of misclassification.

$$Gini(T) = 1 - \sum_{i=1}^{k} (p(c_i)^2 - \sum_{i=1}^{n} p(t_i) \sum_{j=1}^{k} p(c_j|t_i)(1 - p(c_j|t_i) \tag{2.5}$$

Boosting and Bagging

Boosting and bagging are meta-algorithms that can be used with both classification and regression models (Ghojogh & Crowley, 2019). Because boosting and bagging train multiple individual classifiers that are created with varying training sets, which are then combined into one, combined classifier, they are considered examples of ensemble methods (Opitz & Maclin, 1999).

According to the research conducted by Freund and Schapire (1996), *boosting* can improve classifier performance in two ways: It generates many hypotheses that may or may not only be marginally better than randomly guessing, with a large error, which are then combined into one hypothesis with small error. This is done by having the algorithm find hypotheses that minimize the training errors for subsets of the data set. Secondly, the algorithm responds to the training samples, so hypotheses for partial sample splits may vary from each other strongly. In fact, boosting actively encourages this effect by changing the distribution over the training set in reference to the previous hypothesis' performance.

Bagging—short for "bootstrap aggregating"—differs from boosting in that it does not choose features according to previous results, they are picked uniformly at random with replacement (Breiman, 1996). In fact, the different weak hypotheses do not have any influence on each other during the creation process at all (Freund & Schapire, 1996). It is only after the hypotheses were created that they are combined by simple voting—this means that the samples get assigned those classes that the partial hypotheses most often assigned to them (Breiman, 1996).

While boosting is a technique that is effective when learning from imbalanced data, bagging is still considered to outperform boosting when the data contains noise (Khoshgoftaar et al., 2010). Freund and Schapire (1996) also found that boosting works particularly better with simpler algorithms, while the improvement on bagging is smaller—though still present—when it comes to more involved ones.

Model Evaluation

A common method used for the evaluation of decision tree classifiers is the calculation of the accuracy score, which puts into relation true positives, false positives, true negatives, and false negatives. As visible in Equation 2.6, accuracy is calculated by dividing the true classifications by the total number of classified samples.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.6}$$

Because of the way it is calculated, accuracy score should not be used when the data set is imbalanced—that is, when one class has a lot more data samples than the other one. This throws off the accuracy score, as the class with the majority of samples overwhelms the few of the other class, which leads to an accuracy score that does not represent the performance of the model correctly. It is because of this that a model might reach a 99% accuracy score on an unbalanced data set, if one of the classes contains 99% of the samples, by simply guessing the bigger class every time (Khoshgoftaar et al., 2010).

Due to accuracy being considered an insufficient measurement for model performance with unbalanced classes, the data in this work was artificially balanced. However, this does marginally skew the results with duplicated samples for the smaller class, which was an accepted risk in this case.

If the aim is to get a meaningful evaluation result for the original, unbalanced data set, a different evaluation method is necessary. As a consequence of the shortfalls of accuracy, some research prefers the utilization of metrics such as the F_1 score, which is considered more appropriate for use with imbalanced data (Koyejo et al., 2014).

Another approach is the Matthews Correlation Coefficient (MCC), which can also be used for the evaluation of model performance on unbalanced data sets, and should in fact be preferred over accuracy and F_1 score according to Chicco and Jurman (2020). The calculation, once again, considers true positives, true negatives, false positives, and false negatives. Unlike accuracy, the MCC considers false positives and negatives in the numerator as well, however. Equation 2.7 shows the equation in its entirety, as given by Chicco and Jurman (2020).

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{2.7}$$

In Chapter 3, this work will compare the results of the model performance calculated with balanced data and using the accuracy score, with that of the MCC calculated with the original, unbalanced data set.

Markov Chain and Markov Models

Markov models are an broad category of models that can vary widely in their complexity. The least complex form could probably be considered to be Markov chains, which are used to illustrate event sequences and their probability (Fink, 2014).

At the base, Markov chains are made up of states and transitions between the states. Each of the transitions has a probability, meaning for an optimal creation of a Markov chain, a transition matrix is indispensable. A transition matrix gives probabilities for moves between each of the states (see Table 2.1).

To name an example, the table contains the value 0 at S_0 / S_0. This means that there is no possibility for the transition from the first to the second of these states. Because the states are the same for S_i and S_j, this means that there is no circle back to the same state. S_1, on the other hand, forms a circle because it has a non-zero possibility of circling back to itself.

With the information from Table 2.1, it is then possible to calculate the total probability for the occurrence of a sequence/chain in a specific context in a further step (Fink, 2014).

	S_0	S_1	S_2
S_0	0	.5	.5
S_1	.5	.1	.4
S_2	.3	.4	.3

Table 2.1.: An example of a transition matrix

With simple Markov chains, the next state is only influenced by the current state, not any of the previous ones. It is not necessary for every node in the Markov chain to have a transition to every other node. In the case of a non-existent connection between two nodes, it is simply not possible for the one state to follow after the other. Markov chains can be depicted as flow charts, where each connection between nodes has a certain probability (see Figure 2.5).

Hidden Markov models present the additional difficulty of their underlying state sequence being unknown ("hidden"), unless a specific state is active, in which case a generated output of the state is transmitted (Fink, 2014).

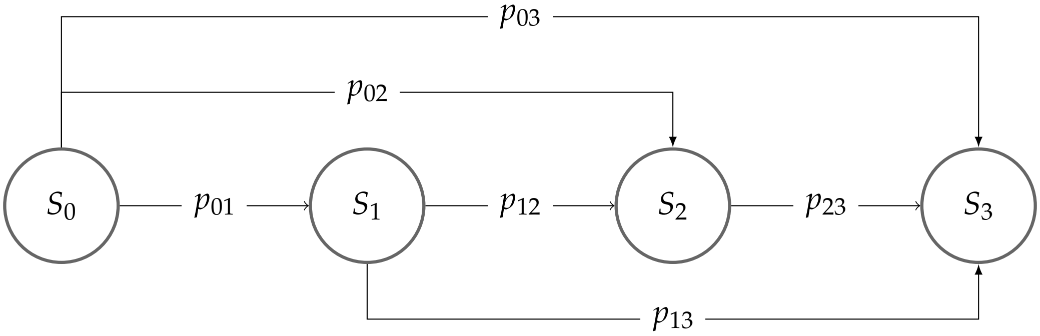


Figure 2.5.: Structure of a Markov Chain

Sequential Analysis

Sequential analysis methods look for patterns of behavior in actions that take place over time (Pohl et al., 2016). *Lag sequential analysis* is one variety of method that can be applied to this effect (Lloyd et al., 2016).

At its base, (lag) sequential analysis assigns codes to certain behaviors and then models sequences as they appear over time. As Kenny et al. (2006) explains, lag sequential analysis is then *"used to study interpersonal behavior by measuring the number of times certain behaviors precede or follow a selected behavior"* using conditional probability, and is a method often used in interaction analysis in human sciences.

In this, "lag" refers to the number of events that are considered for the sequence. Furthermore, it is important to note that in Lag Sequential Analysis (LSA), two codes are chosen as start and end nodes, and that it does not necessitate a specific order of states between them to consider the start and end nodes sequential—it specifies the number of transitions between the states, not the actual sequences (Pitkänen, 2017).

Sequence: ABCABACBABBCADC
Lag: [-2, 2]
Start Event: A
End Event: C

Lag	-2	-1	1	2
Occurrences	1	2	1	2

Table 2.2.: Example of lag sequential analysis parameters and output. Adapted from Pitkänen (2017, p.34)

For the example in Table 2.2, the range of the lag means that with lag -2, the algorithm starts from the occurrence of the chosen start event, and moves two places to the left in the sequence to find the end event. If found, it counts as an occurrence. For lag 2, the algorithm looks to the right of the occurrence of the start event.

Faraone and Dorfman (1987) explain that lag sequential analysis can be used to find and express cross-dependencies in sequences of interactions; however, the assumption that there are cross-dependencies must hold, or the results will be incorrect. Furthermore, Gottman et al. (1990) noted that lag sequential analysis should only be used as a fallback if it is not possible to conduct a full analysis with a Markov model of second or third order due to a lack of data, as it functions as a *"trick to get around the problem"*. Pitkänen (2017) mentions that noise presents a significant problem to the method, as well as high lags, as they may find irrelevant sequences.

2.3. Related Work in the Area of Learning Analytics

In previous research, both models based on logistic regression and support vectors, as well as those based on clustering have been used to analyse data and predict user behavior in MOOCs (AL-Shabandar et al., 2017). For clustering, self-organised map clustering has been found to be a viable approach to visualizing user behavior patterns (Alias et al., 2015).

Wong et al. (2015) used keyword taxonomy on MOOC forum data to analyse learning processes in MOOC courses. They classified the messages into "Remembering", "Understanding", "Applying", "Analyzing", "Evaluating", and "Creating", as proposed by Anderson and Krathwohl's taxonomy (Anderson et al., 2000).

With a focus on teach-outs—learning sessions that focus more on the creation of a discussion around a topic of *"pressing social urgency"* without formal assessments—Yan et al. (2019) used content analysis and topic modeling on forum data to evaluate learner engagement features.

Focusing more on the students' study approaches, Akçapinar et al. (2020) analysed logs collected by an eBook system, and furthermore investigated the connection between students' reading behavior and academic performance by using association rule mining analysis. In a previous attempt at a similar analysis, Cheng and Chu (2019) employed LSA to find behavior patterns and model student learning achievements. This was then followed by research by Yang and Chen (2020), which focused further on the online learning behavior of students, and used LSA to model the learning behavior of different cognitive styles as they corresponded to learning behavior.

Researching a similar topic, Saint et al. (2020) used simple frequency analysis, epistemic network analysis, temporal process mining and stochastic process mining to analyse patterns of SRL. They extracted the top and bottom decile and compared their behavior in the course, finding that the top decile students put a higher focus on summative tasks, goal setting, and reviewing answers to formative quizzes. The bottom decile students, on the other hand, were found to be only half as active as top decile ones, on average.

Malekian et al. (2020) researched student readiness for assessment tasks in MOOCs, and found that by using sequential pattern mining with neural network models, they were able to investigate the differences in sequences of behavior for higher- and lower-performing students. The results showed that higher-performing students tended to view and review lecture materials more, while lower-performing students had more sequences that related to repeated failed assessment submissions.

2.4. Related Work in the Area of Dropout Prediction

In the past years, there has been much interest in the area of dropout prediction using various machine learning algorithms and models. This section will be focusing on looking at some of the recent research, the methods used, and the eventual results that were achieved by applying them.

Lee and Chung (2019) discussed the class imbalance problem as it applies to dropout prediction in face-to-face education. Because the number of dropouts tends to be very small in this case—in contradiction to MOOCs, where the dropout rate can make up about 90% of the students who originally signed up for the course—classifiers may treat the dropouts as "noise", leading to inaccurate predictions in later stages. When comparing random forest, boosted decision tree, random forest with Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), and boosted decision tree with SMOTE, they found that the classifier that performed best when it came to precision-recall the was boosted decision tree classifier.

AL-Shabandar et al. (2017) found that there was "*strong correlation between click stream actions and successful learner outcomes*" in their research, using a variety of different supervised machine learning algorithms for the data classification, in combination with learning analytics, which have already been used to predict student dropouts in the past, not only in MOOCs but traditional learning environments, as well (Tseng et al., 2014).

Likewise, Support Vector Machines (SVMs) and least mean square algorithms have been found to be useful means to estimate the dropout rates with only click stream data (He et al., 2015). In their research, AL-Shabandar et al. (2017) used self-organised maps and feed-forward neural networks to predict whether a user would complete a course or not.

Hagedoorn and Spanakis (2017) considered the temporal aspect in their approach to predict user attrition. They found that while it was possible to get good results with all of their analyses, which compared logistic regression, random forest and AdaBoost results, it was significantly more difficult to predict the week in which the dropout would occur. They also found that a profile that was left empty by the user was a high indication that they would later drop out.

As profiles may be seen as a way for MOOC students to form connections with each other and become more involved with the MOOC and/or its platform, this finding also leads back to those presented by Yang et al. (2013), which mentioned that higher social connectedness had positive influence on MOOC completion rates. This may indicate that it is beneficial to include tasks that urge students to somehow interact with each other, e.g. through a forum.

In their research paper, He et al. (2018) used multiple linear regression to investigate the connection between forum posts and study outcome, and found that those that participated in the MOOC forums had a tendency to get better results than those that did not.

Wen et al. (2014) used collective sentiment analysis combined with survival analysis to investigate correlation between user sentiment in MOOC forum posts and user dropouts. Previously, Ramesh et al. (2013) used probabilistic soft logic to predict (dis-)engagement through the use of sentiment and subjectivity, neither of which brought the desired results. Building on this knowledge, Wen et al. (2014) acknowledge that using sentiment polarity analysis techniques to analyse single forum posts is prone to erroneous results due to the lack of context. They postulate, however, that with a large enough number of analyses, those errors may diminish in severity.

Meanwhile, Qiu et al. (2016) found that higher effort and asking more questions did not necessarily lead to a higher probability for passing the course. They did, however, note that forum activity seemed to be a positive indicator for learning performance, as a whole, even though the level of student engagement in the forums lay at only 6% active users. This echoed the findings by Ramesh et al. (2014), who also observed that forum activity was a high-ranking feature when predicting course completion, second to only the number of lecture views, especially in the middle and end phases of the course.

Xing et al. (2019) investigated the achievement emotions in connection with dropout rates, using a classifier to predict achievement emotion in forum posts, which was then in turn used in connection with a survival modeling technique to further study the effects of those emotions.

A case study by Pigeau et al. (2019) compared eight classification models in how they performed in dropout prediction of 12 different courses from the MOOC platform OpenClassrooms, focusing on random forest, AdaBoost, SVM, logistic regression, and neural networks, as well as LSTM neural network, a process mining method, and a method based on sequence mining. They subsequently found that AdaBoost and logistic regression gave the most promising results, with AdaBoost performing more reliably when predicting failing learners, and logistic regression being better with passing ones.

From a different point of view, Henderikx et al. (2018) explain that it is necessary to reconsider the meaning of failure in relation to MOOCs. Their paper on how intentions play into user dropouts in MOOCs proposes that the intentions of the user need to be taken into account when looking at dropout rates, as some students may only wish to view a specific lecture and have no intention to finish the entire course from the start.

Because of this, dropout prediction may have to learn to account for these people who only ever intend to “sit in” on some classes, or tracking of intentions and their changes might be useful.

Hew et al. (2020) also mention this when they write about how to measure the success of a MOOC. Furthermore, they argue that using completion as a measure for student success may not be useful, "*due to the considerable difference between MOOC and traditional university courses*". They state that a large number of students start a MOOC because of reasons other than finishing the course, different from how it would be at a traditional university. Due to this, using metrics that one would use at a traditional university—like dropout rate—will most likely not be representative of the MOOC's benefit to the students. Instead, the proposed approach is to use indicators such as "learner satisfaction" (Hew et al., 2020; Rabin et al. (2019)).

Similarly, Brochenin et al. (2017) do not consider completion of the course for their analysis, but instead choose to focus on resource usage behavior of the users. By finding out which resources were accessed particularly often or tended to be skipped, they aim to draw conclusion from the behavior data, and furthermore figure out how to improve the course content.

2.5. Summary

E-learning has continually gained significance over the past few years, as continuing education has become more and more important, and e-learning has become more present in traditional teaching environments. With life-long learning almost becoming a necessity in the fast-changing job market, and e-learning swiftly becoming ever-present, some e-learning types such as mobile, as well as bite-sized learning have crucially shot up in importance as attention spans shorten.

With this surge in relevance, learning theories such as self-directed and self-regulated learning have also gained a central role, as researchers and teachers are working to get them more integrated and more strongly considered in the area of e-learning.

Much of the research focuses on MOOCs, which are by now well-established e-learning systems. Nevertheless, they still have the seeming issue of high dropout rates and thus appear not to live up to their potential. However, because the definition for what can be considered "success" for the student is dependant on the student's goals when signing up for the course in the first place, research seems to be rapidly moving in two directions with this research question.

Some researchers are working to find ways to predict and mitigate these low completion rates, while others are attempting to find different ways of classifying whether a student was successful in attaining their specific goal in a MOOC, even if this goal was not course-completion by itself. Furthermore, some research wholly focuses on how students use and interact with e-learning courses and their material. In either case, student behavior and the analysis thereof is an interesting factor that has quickly become a central point of research in both areas. In one such example, for instance, research has found that social connectedness, as well as social support rewards seem particularly relevant to e-learning, as both appear to influence student outcomes, motivation and engagement positively.

Due to the similarity and overlap of the fields, they often work with similar methods to analyse data, with a heavy focus on learning analytics—data mining algorithms such as decision trees and support vector machines especially find regular use, as well as Markov chains and models.

3. Analysing Coursera Data from a South-American University

This work aims to perform drop-out and completion prediction by using data from a selection of Coursera MOOCs, as well as their connected course forums. The original data is processed to create a set of features using the user actions with and without the forum data, which is then used with a boosted decision tree classifier to get the results.

3.1. Dataset

The six MOOCs that will be analyzed in the course of this work were offered on the MOOC platform *Coursera*. While most of the courses were STEM or STEM-adjacent, their specific subject focus was varied. To give an overview over the topics included, a list of the MOOC titles, as well as a short description of the general course topic can be reviewed in Table 3.1. From this point onward, this work will be using abbreviations to refer to the MOOCs throughout the rest of the text. These abbreviations, functioning as codes to identify the MOOC in question, are also listed in the table.

To enable students to assess time requirements of a course, Coursera allows course creators to specify course period/duration lengths and estimated workloads for the course. The considered MOOCs had listed course periods that ranged from 4 to 6 months, and an estimated workload of 2 to 7 hours per month. However, MOOCs on Coursera are often self-paced, with this information functioning as a suggestion more than strict deadline. Furthermore, Coursera does not usually formally require previous knowledge when accessing a course.

Code	MOOC	Content
AST	Análisis de Sistemas de Transporte	Fundamentals of management of passenger transport systems
DSV	Decodificando Silicon Valley	Interacting with Silicon Valley as a technological entrepeneur
EA	Electrones en Acción	Introduction to electronics and Arduinos
GOE	Gestión de organizaciones efectivas	Building successful organisations
PCA	Hacia una práctica constructivista en el aula	Designing classes for a constructive environment
WS	La Web Semántica	Fundamentals about the Semantic Web

Table 3.1.: This table lists the MOOC titles with a short content description. Code refers to the abbreviation that will be used throughout the work to reference specific courses.

MOOC Title	Launch	Duration	Estimated workload
Análisis de Sistemas de Transporte (AST)	October 2015	6 months	2-3 h/month
Decodificando Silicon Valley (DSV)	October 2015	4 months	5-7 h/month
Electrones en Acción (EA)	October 2015	4 months	4-6 h/month
Gestión de organizaciones efectivas (GOE)	October 2015	6 months	2-3 h/month
Hacia una práctica constructivista en el aula (PCA)	April 2015	N/A	N/A
La Web Semántica (WS)	October 2015	7 months	2-4 h/month

Table 3.2.: The table shows the MOOCs considered in the course of this work, as well as basic data about the course iterations.

Due to this, the given workload and suggested course duration may not represent the reality for each student, and can in fact vary widely according to the users' pre-existing knowledge levels and/or the length of their learning sessions.

At the time of the course iterations, course sessions on Coursera had an official start date that functioned as a kick-off day for the students enrolled in the course. While it was possible to join after a course has already started, it was only possible to pre-enroll for the course if the start date was in the future. Once the start date had passed, the pre-enrollment then automatically transformed into a normal enrollment, and the user could access the course forum and assessments. From then, courses were self-paced with suggested deadlines for the students (Coursera, 2018a). Their material was available to the users without additional time constraints. This means that students could view all course material over the span of several months, or even just all in one day, if they wished.

Each of the specific course iterations analysed in this work took place and were active between the years 2015 and 2016, with sessions having started in either April or October 2015. Detailed information about launch month, duration, and estimated workload for each course (if available) is specified in Table 3.2, where the *Launch* column shows the original launch date of the course, and *Duration* is the time in months that the session was considered active. *Estimated workload* describes the time in hours that was estimated by the creators of the course to be necessary for a user to pass the course. In one case, no course period or suggested workload was given. This is denoted with N/A.

To successfully complete a course, all its graded assignments need to be passed. However, while course access on Coursera is, for the most part, freely available, assessments cannot always be completed or even viewed if the user does not pay the course fee.

To join courses without paying the fee, it is possible to enroll as *auditor*, in which case most of the course material is freely available to the user. It is likely that they will not be able to submit certain assignments or get grades, however, and they will not get a certificate after finishing the course. If a course is joined as

ID	Mode
1	In video
2	Course quick
3	Open single page
4	Quick questions
5	Survey
6	Formative
7	Summative

Table 3.3.: Coursera's classification for assessment types

an auditor and the student decides they want to receive a certificate, they can choose to pay for the course after having already started it. At this point, any previously inaccessible course material and assessments will be made available and the student becomes eligible to receive a certificate after having successfully completed the course (Coursera, 2018c). Students who paid the course fee are also referred to as being verified, which is the term that will be used in this work in any analyses/statistics using this particular information going forward.

Coursera supports a number of different assessment types that teachers can utilize in their courses. The six examples that were analyzed in this work included assessments that can roughly be divided into quizzes, peer review assignments and programming assignments. In this specific context, quizzes are automatically-graded and can be multiple-choice or in-video, whereas peer review assignments are graded by fellow students of the course.

In the available Coursera data, these rough assessment types are more finely associated with IDs between 1 and 7, the array of which can be found in Table 3.3, together with their corresponding assessment type names as found in Coursera's data export.

Not all of the different modes found application in the six courses that were analyzed for this work. In fact, only types 1 (in video), 6 (formative) and 7 (summative) were used. In the context, *in video* refers to quizzes that were integrated in the lecture videos. From the data available, this seemed to be the most popular option in the analysed courses. Over 54% of the 172 assessments from six courses were in video assessments.

The second most popular assessment type was summative, which still made up over 27% of assessments. It is typically used at the end of a learning section/period to assess the student's knowledge (Garrison & Ehringhaus, 2007).

On the other hand, formative assessment refers to assessments that are not, in and of themselves, supposed to be used by the teacher to grade the students or for final assessments; instead the goal is for the student to learn from the assessment

			Assessment Type Frequency		
MOOC	# Modules	# Assessments	1: In video	6: Formative	7: Summative
AST	6	32	20	6	6
DSV	4	31	1	26	4
EA	4	27	11	0	16
GOE	7	17	11	0	6
PCA	10	58	50	0	8
WS	7	7	0	0	7
Total	38	172	93	32	47

Table 3.4.: The table shows the number of modules and assessments for each MOOC, as well as the frequency of assessment types across the different MOOCs. For the considered courses, types 2 to 5 were not used.

and have it function as a self-assessment of knowledge, and to be able to adapt the further learning process according to the students' needs (Ventista, 2018; Garrison and Ehringhaus, 2007). Assessments of this type made up the remaining 19% in the studied courses.

The complete breakdown of the number of modules in each MOOC, as well as the number of assessments and distribution of assessment types can be found in Table 3.4.

In addition to general facts about course structure and content, Coursera provides a wealth of further, in-depth information about its courses and students. Some of this data which are relevant for this work are the records of course sign-ups and assignment actions, the data about the assignments, and the information whether the student ultimately passed the course.

Furthermore, one of the relevant points of interest for this work is the log containing the initial interactions of the user with the course and/or course description, those of importance being users with the membership role *Learner* and *Pre_enrolled_learner*. In the context of the log, *Learners* are those who enroll in a course where the launch date has already passed. Pre-enrolled learners (pre-)enrolled in a course before the start date, and are automatically enrolled—creating a further data entry with the *Learner* membership role—once a course's start date has arrived. Further states include *Browser*, which is a user who viewed the course description but did not enroll, and *Not_enrolled*, which describes a user who un-enrolls after having enrolled for the course in the past.

The log may contain multiple entries for the same user, for example one each with the roles *Browser*, *Learner* and *Not_enrolled* for someone who first looked at the course description, then enrolled in it and finally un-enrolled at a later date (Coursera, 2018b). An overview over these user states in tabular form was compiled

Tag	Description
Browser	User viewed the course description but did not enroll
Pre_enrolled_learner	User enrolled in the course before the start date
Learner	User enrolled in the course after the start date
Not_enrolled	User un-enrolled from a course after previously enrolling

Table 3.5.: User states in relation to a MOOC as seen in the data sets, with explanation of terms

in Table 3.5.

Upon evaluation of each course's enrollment data, it was found that the student numbers of the analyzed MOOCs varied widely, with between 1579 and 14630 enrollments for each of the courses. Altogether, the sign-ups encompassed over 36,000 students, though the same student may have been counted twice if they enrolled in more than one of the six courses that were part of the data set considered in this work.

To ascertain the MOOCs' dropout rates, further data was gained from *course_grades*, which contains information about the users' course grades. New entries are created each time a grading event takes place. Each of the entries contains the course id and a learner id. Furthermore, they include information on the time of the grading event, as well as the number of course assessments that the user has passed at that specific point in time, the number of course assessments that the user passed and the user's identity was verified for, and the grade according to the passed items, as well as the grade according to the passed items for which the user identity was verified. Besides this, they also contain the overall passing state, which can take on one of four values, described in Table 3.6.

State	Description
0	Course was started, not passed
1	Course was finished, not verified
2	Course was finished, verified
3	Course is not passable

Table 3.6.: The four states describe the user's state in regards to the course. A course can be passed without verifying the graded assignments, though no course certificate will be handed out in that case.

However, while Coursera specifies four different states, the analyses in this work going forward only consider two. Since all courses were passable by design, state 3 becomes redundant for this purpose. Furthermore, while both state 1 and state 2 occur in the data, they are both combined in the "Completers" class that this work

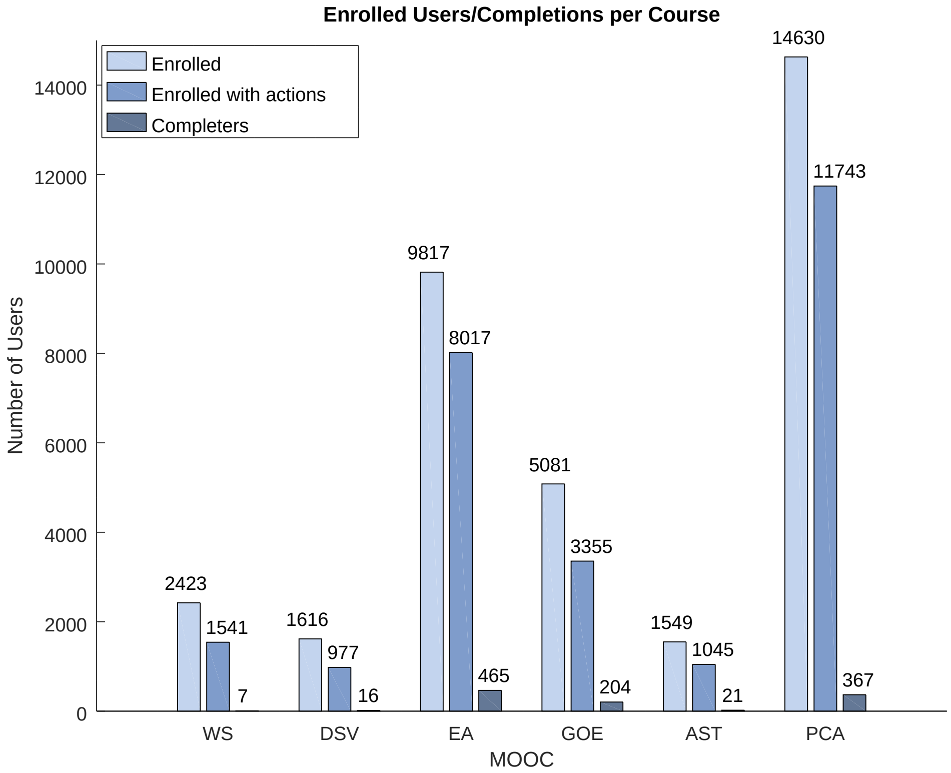


Figure 3.1.: Number of Users per Course with Final Completion State

will be using to classify users. From here on, anytime "Completers" are mentioned, they are intended to refer to both students who did and did not get verified by paying the course fee. State 0 is mirrored in the "Dropouts" class that is used henceforth.

Like most MOOCs, the six courses analysed in this work were also found to have high dropout rates. To statistically evaluate the dropouts, this work first put into relation the number of enrolled students with those who enrolled and further interacted with the course after their initial sign-up, and finally those who successfully finished the course. As visible from Figure 3.1, there is a sharp drop in the numbers from enrolled students to only those who accessed the course after, with another, even sharper drop to the number of students who completed the course, both verified and not.

To differentiate between users who only signed up, and those who participated after their enrollment, Table 3.7 gives an overview over the enrolled students of each course, both those who were active in the course after the initial sign-up, and overall, as well as the dropouts, again with the numbers of only those students who had additional activities besides the sign-up, and the total number. As visible from the dropout rates, they show similarly high percentages in either case, with 2.1% difference in the best case for GOE, and 0.2% difference in the worst case for WS,

	Enrollments		Dropouts		Dropout Rate				
MOOC	All	Active	All	Active	All	Active	Completers	Verified	Survey
AST	1549	1045	1528	1024	98.6%	98.0%	21	18	39
DSV	1616	977	1600	961	99.0%	98.4%	16	14	41
EA	9817	8017	9352	7552	95.3%	94.2%	465	244	189
GOE	5081	3355	4877	3151	96.0%	93.9%	204	154	74
PCA	14630	11743	14263	11376	97.5%	96.9%	367	272	388
WS	2423	1541	2416	1534	99.7%	99.5%	7	4	84
Overall	35116	26678	34036	25598	96.9%	96.0%	1080	706	815

Table 3.7.: MOOC enrollment, and dropout statistics

and a median difference of 0.6% between all and active dropout rates.

Dropouts comprised between 95.3% and 99.7% of the enrolled students when considering all students who enrolled, and between 93.9% and 99.5% when only considering students who had at least one additional action after the enrollment. Both those who formally un-enrolled and those who did not finish the course without un-enrolling were counted among dropouts. The majority of users who passed a MOOC also chose to verify their result by paying the course fee (before or after qualifying for completion), thus becoming eligible for a certificate from the course.

The "Survey" column corresponds to the number of users enrolled in that course who gave information about their demographic data.

Coursera allows users to specify certain information about themselves, such as age and place of residence. Through this data, provided by the users themselves in their profiles and through surveys, it is possible to gain knowledge about demographics. Due to the voluntary status of this data, only a small sample of students is represented from each course. Interestingly, while the total number of enrolled students showed big differences from course to course, it was in this data that similarities surfaced.

The majority of users who made their birth years available were born in the 1980s, something that matched across all six MOOCs. At the point in time that the courses were available, this would have made this portion of users about 26 to 35 years old. Similarly, 1970s and 1990s were strongly represented as well in each of the courses. The complete statistical analysis can be found in Table 3.8, with age groups as percentages in Figure 3.2.

The courses themselves were taught in Spanish, a fact that is also represented in the demographic data—the majority of users were living in Spanish speaking countries at the time of the course sessions. From the demographics data collected in the user survey, students from Mexico, Spain, Chile and Colombia constituted a large part of each course's user base, as visible in Table 3.9.

	Births by Decade							
MOOC	<1950	1950s	1960s	1970s	1980s	1990s	N/A	Total
AST	0	1	3	7	17	11	0	39
DSV	0	1	2	8	27	3	0	41
EA	4	11	30	40	64	38	2	189
GOE	0	4	9	18	31	12	0	74
PCA	4	28	74	95	147	38	2	388
WS	0	5	11	25	34	9	0	84
Total	8	50	129	193	320	111	4	815

Table 3.8.: Number of users by their decade of birth. Of those who answered the question, the largest fraction of students of each course were born in the 1980s, making them between 26 and 35 at the time of taking the courses.

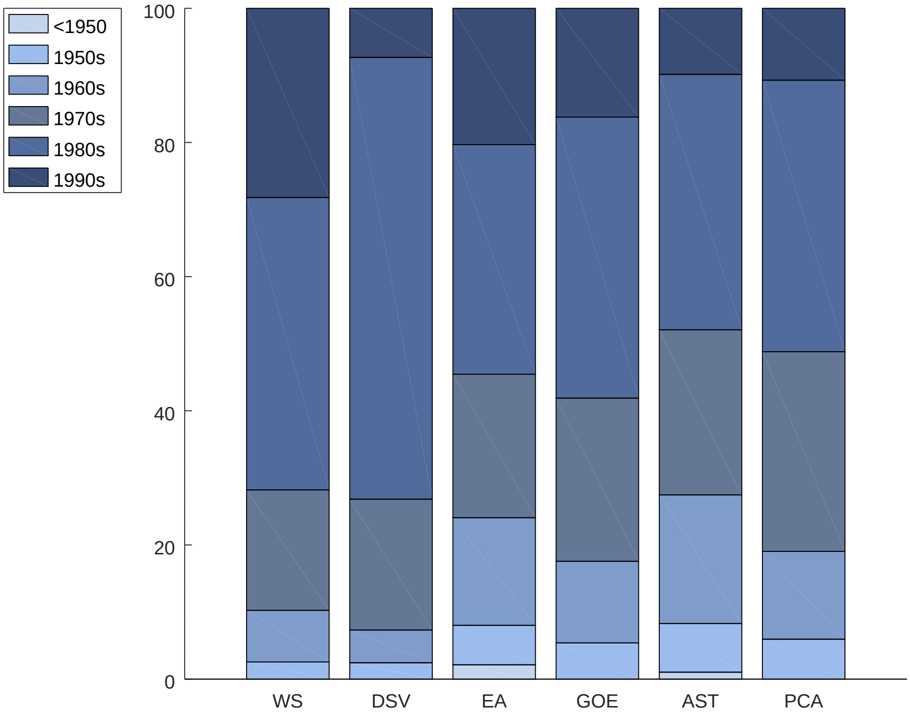


Figure 3.2.: Age group percentages as represented in the survey.

	Country of Residence							
MOOC	Argentina	Chile	Colombia	Mexico	Peru	Spain	Other	Total
AST	2	4	4	3	3	5	18	39
DSV	3	5	4	2	2	2	23	41
EA	7	23	20	26	14	45	60	187
GOE	0	10	7	16	8	8	25	74
PCA	15	45	43	106	14	33	127	383
WS	7	11	8	13	2	11	32	84
Total	34	95	91	166	43	104	285	808

Table 3.9.: Number of users by their country of residency. As the classes were in Spanish language, the core audience were those from Spanish speaking countries.

Finally, the last data file of particular interest to this work contains information on the learners' course progress. Each course item (e.g. lectures, assignments, quizzes) and the learners' interactions with them are listed, together with information on whether the course item was started (1) or completed (2), and a timestamp for each action. Additionally, they contain the user and course IDs.

A course item may be restarted multiple times, with only the last completion counting for the grade, if it is indeed graded. For the purpose of this work, however, repeated accesses of course items are counted as interactions like any other; there is no differentiation in how they are treated.

The data about the interactions with the MOOCs and their content were further supplemented with forum data. Coursera provides each course with its own dedicated discussion forum section, where the students are able to swap ideas and ask, as well as answer each other's questions.

General participation in the forums was mostly shown to be low across the board for all courses. To compare the numbers for completers and dropouts, this work put into relation the number of students who participated in the forums and those who passed or did not pass the course. Both are also shown in percentages, for better comparability. It shows that, in general, the percentage of forum post writers among completers is significantly higher than the same percentage among dropouts. The full statistical evaluation for each course can be found in Table 3.10.

Using both the course interaction data and the forum data, Table 3.11 compiled the interaction information to show how course activity varied between the "Completers" and "Dropouts" classes, and the influence of the forum data on the calculated averages. It also compares these average number of interactions with the average number of interactions that took place overall, or "globally". This is done both including and excluding forum data.

	Completers		Dropouts		
MOOC	No Posts	Posts	No Posts	Posts	Total
AST	6 (28.57%)	15 (71.43%)	940 (91.80%)	84 (8.20%)	1045
DSV	7 (43.75%)	9 (56.25%)	916 (95.32%)	45 (4.68%)	977
EA	344 (73.98%)	121 (26.02%)	7194 (95.26%)	358 (4.74%)	8017
GOE	166 (81.37%)	38 (18.63%)	2946 (93.49%)	205 (6.51%)	3355
PCA	269 (73.30%)	98 (26.70%)	11015 (96.83%)	361 (3.17%)	11743
WS	1 (14.29%)	6 (85.71%)	1445 (94.20%)	89 (5.80%)	1541

Table 3.10.: Statistics of forum posts in relation to completers/dropouts

	Avg. Interactions with Forum			Avg. Interactions without Forum		
MOOC	Global	Completers	Dropouts	Global	Completers	Dropouts
AST	44.09	46.43	42.91	42.91	46.19	42.84
DSV	66.57	58.50	66.23	66.04	58.38	66.17
EA	71.15	154.30	65.62	70.78	154.07	65.65
GOE	39.03	57.69	37.20	38.44	57.60	37.20
PCA	42.62	75.50	41.18	42.22	75.30	41.15
WS	36.47	130.86	35.47	35.79	129.29	35.36
Overall	49.99	87.21	48.10	49.36	86.81	48.06

Table 3.11.: Comparison between global average interactions in the forum per user with those of completers, and dropouts, done both with and without considering forum data.

In an attempt to get a more in-depth understanding about users' utilisation of these forums, and to find whether there were patterns in how the forums were used, this work used this data to look at the general interaction behavior of the users. The statistics created from this data additionally differentiated the type of forum interaction, namely between thread questions—posts that started a new thread—and thread replies—posts that were written in response to an existing thread—in the statistics created from the available data.

Implementing this separation, Table 3.12 attempts to investigate how forum untilisation differs when it comes to users asking questions, as opposed to users answering them. It shows the total number of unique users who were active in the course forum, as well as the total number of posted question and answers, each. Additionally, the table shows the average number of posts per user who was active

MOOC	Active Users	Active User %	Questions	Replies	Posts/User	Replies/Thread
AST	99	9.47%	74	163	2.39	4.29
DSV	54	5.53%	56	70	2.33	1.94
EA	479	5.97%	476	667	2.39	2.59
GOE	243	7.24%	220	146	1.51	2.75
PCA	459	3.91%	345	556	1.96	4.03
WS	95	6.16%	92	149	2.54	4.38
Total	1429	6.38%	1263	1751	2.19	3.33

Table 3.12.: Forum statistics for each MOOC

in the forum. The percentage shown in the "Active Users" column represents the percentage of users who participated in the forums of those who signed up for the course and had at least one course content interaction. Lastly, the table gives the mean number of replies per thread per course to investigate general engagement with the forum threads.

	Completers			Dropouts		
MOOC	Only Q	Only R	Both	Only Q	Only R	Both
AST	2 (13.33%)	6 (40.00%)	7 (46.67%)	21 (25.00%)	43 (51.19%)	20 (23.81%)
DSV	2 (22.22%)	2 (22.22%)	5 (55.56%)	27 (60.00%)	11 (24.44%)	7 (15.56%)
EA	30 (24.79%)	37 (30.58%)	54 (44.63%)	167 (46.65%)	104 (29.05%)	87 (24.30%)
GOE	18 (47.37%)	13 (34.21%)	7 (18.42%)	137 (66.18%)	40 (20.29%)	28 (13.53%)
PCA	29 (29.59%)	30 (30.61%)	39 (39.80%)	113 (31.30%)	194 (53.74%)	54 (14.96%)
WS	0 (0.00%)	1 (16.67%)	5 (83.33%)	46 (51.69%)	25 (28.09%)	18 (20.22%)
Total	81 (22.88%)	89 (29.05%)	117 (48.07%)	511 (44.67%)	419 (36.63%)	214 (18.71%)

Table 3.13.: Statistics regarding the user activity in the forums in each MOOC, where 'Only Q' are students who only created new threads, and 'Only R' corresponds to students who only replied to existing threads.

Further statistics can be found in Table 3.13, which shows how learners used the forum, and how their utilisation plays into and/or differs with their completion status. "Only Q" gives the number of students who only started a new thread in the forum but did not write any answers in existing threads, even their own. "Only R" is the opposite—students who only replied to existing threads but did not start a new one. Students who both started a new thread and replied to at least one thread are counted in the "Both" column. In addition to absolute numbers, each column also gives percentages, wherein the percentages were calculated per class.

Following this, Table 3.14 takes the differentiation of forum posts into question

	Started Threads/User				Thread Replies/User				
MOOC	0	1-5	6-10	>10	0	1-5	6-10	11-15	>15
AST	995	50	0	0	969	70	6	0	0
DSV	936	41	0	0	952	23	1	0	1
EA	7679	331	6	1	7735	258	20	2	2
GOE	3165	189	1	0	3267	86	2	0	0
PCA	11508	232	3	0	11426	309	7	0	1
WS	1472	69	0	0	1492	41	6	0	2

Table 3.14.: The number of posts that a user typically made in the forum, and whether this post was the start of a new thread or a reply. The table shows the number of users who made a number of posts that is included in the given range.

and reply posts, and creates categories going by frequency to show a more finely granulated distribution across the ratio of questions to replies in the forums.

3.2. Framework and Concept

The following section expands on the pre-processing and analysis processes, as well as explaining the selected features used in the analyses. It further explains the training and testing process, as well as how some of the model parameters were chosen.

3.2.1. Pre-processing and Feature Extraction

In the first step, the available data was cleaned of incomplete and faulty rows. During this process, user interaction data without timestamps or invalid data in the timestamp field were disregarded. As were users without any actions after the initial enrollment.

Following this, the user interactions were compiled and used to create user-specific interaction data. This data consisted of both activity with the course content and, optionally, the course discussion forums. Each user's interaction data was sorted by timestamps, after which they were split into sessions. A new session was started if there were more than 30 minutes of downtime between two actions, as previously done by Vitiello (2017). These sessions were then used to create the features shown in Table 3.15, which were afterwards used in the further analyses.

For the runs where forum interactions were included in the feature analysis, the additional features from Table 3.16 were added.

Due to the large difference between the number of completers and dropouts in

Feature	Description
sessions	The number of sessions the user started
requests	The total number of actions the user made in the course
days	The number of days where the user made at least one action
activetime	The total time that the user spent active
sessionlen	The average length of the user's sessions
clicktime	The average length between two clicks
sessionrequests	The average number of requests per session
dayrequests	The average number of requests per active day

Table 3.15.: The table gives short descriptions of the features used for the dropout prediction

Feature	Description
forumposts	The number of forum posts the user posted in the considered timeframe
threadstarter	Boolean value, whether the student started a new thread in the considered timeframe
threadresponder	Boolean value, whether the student posted a reply to an existing thread in the considered timeframe

Table 3.16.: Additional forum features created for the training and testing stages

the courses, it is necessary to perform class balancing to avoid improperly taught classifiers. This can be done in one of two basic ways.

Firstly, the class sizes of the training data can be equalized by removing data from the larger class. This approach, however, is often less desirable, as it shrinks the data the algorithm can learn from.

The other possibility is to add data to the smaller class. Generally, this is the preferred way of dealing with class imbalances. Depending on the used approach, existing data from the smaller class can simply be copied and thus multiplied, or the data can be changed in a minuscule manner and added thus. For this work, the copying approach was chosen, with the smaller, completers class being sampled until the class members matched the larger number of dropouts.

In reference to Vitiello (2017), analyses were done splitting the data into *Absolute* partitions, *Days*, *Percentage* wise partitions, and *Scaled* portions of the data.

For the *Absolute* approach, only an absolute number of initial user actions are considered when creating the features. The feature creation process uses from 1 to 100 initial actions per user, with an increase of 10 initial actions for each experiment. So the experiments will take place with the features using the first 10 interactions the user has with the course according to the timestamps, go on to 20, 30, etc. and finally end with the first 100 interactions.

Days. In total, this approach considers the first seven days of the user's course actions for the features. This does not just refer to active days, but *does* start from the first time the user interacts with the course, even if this is after the official course start. Considering the timestamps, each user's actions were considered in 24-hour steps for each of the runs. Actions within the first 24 hours were considered in the features for one day, within the first 48 hours the features for two, etc.

Percentage uses certain percentages of user actions for the feature creation, depending on each user's total amount of actions. Starting at just 1% of the user's action, the algorithm continues with 10, 20, etc. At 100%, all of the user's actions are considered in the feature creation.

Scaled, similarly to *Percentage*, considers percentages of the user's data for the feature creation, though in this case it also considers the timing. Taking the total amount of time that the user spent interacting with the course, the *Scaled* setting considers the actions taken in the first 1% of the time the user spent interacting with the course material, then going on to 10, 20, etc.

This work uses BDTs to model and predict student dropouts, as Vitiello (2017) used in his experiment considering the data from two different MOOC platforms at once. Gradient BDTs also showed the best results when Liang et al. (2016) compared the results from SVMs, logistic regression, random forest and gradient BDT for dropout predictions. Together with the great results presented by Vitiello (2017), this made BDTs the most promising algorithm for this analysis. The method was previously described in Chapter 2.2.4.

3.2.2. Training and Testing

For this analysis, both the AdaBoost and the decision tree classifiers were used to fit the data. With the combination of those two algorithms, there are several parameters that can be set. One of those parameters is *max_depth*, which describes the maximum depth the decision tree can reach before the algorithm terminates. Because the choice of this maximum depth plays a significant role in the quality of the final model, this value has to be considered carefully. There is, however, no universal "best way" for choosing this parameter. The official documentation of scikit-learn suggests starting with a maximum depth of 3, and evaluating and potentially increasing from there (scikit-learn, 2020a). This approach was adopted for this work.

Once the maximum depth value becomes too high, the train and test scores start to differ increasingly, which is a sign of overfitting that can be recognised from the accuracy plots. For this work, only one course (EA) and one mode (Percentage) was considered in the plot of each maximum depth value experiment to avoid overwhelming it. However, the results do represent the general trend, and are thus meaningful by themselves. Furthermore, to be able to compare accuracies across runs with varying max depth, the accuracy mean across all splits of the Percentage mode was taken for each training and test run with the data from the MOOC.

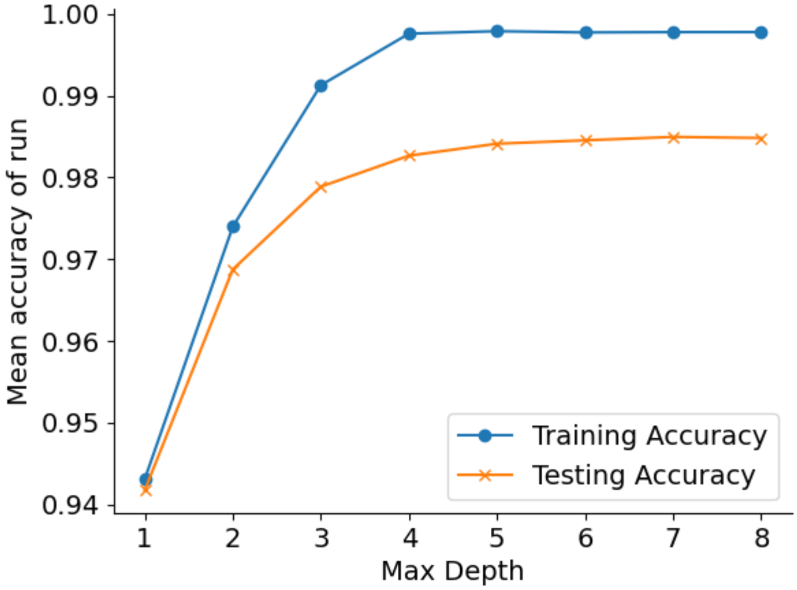


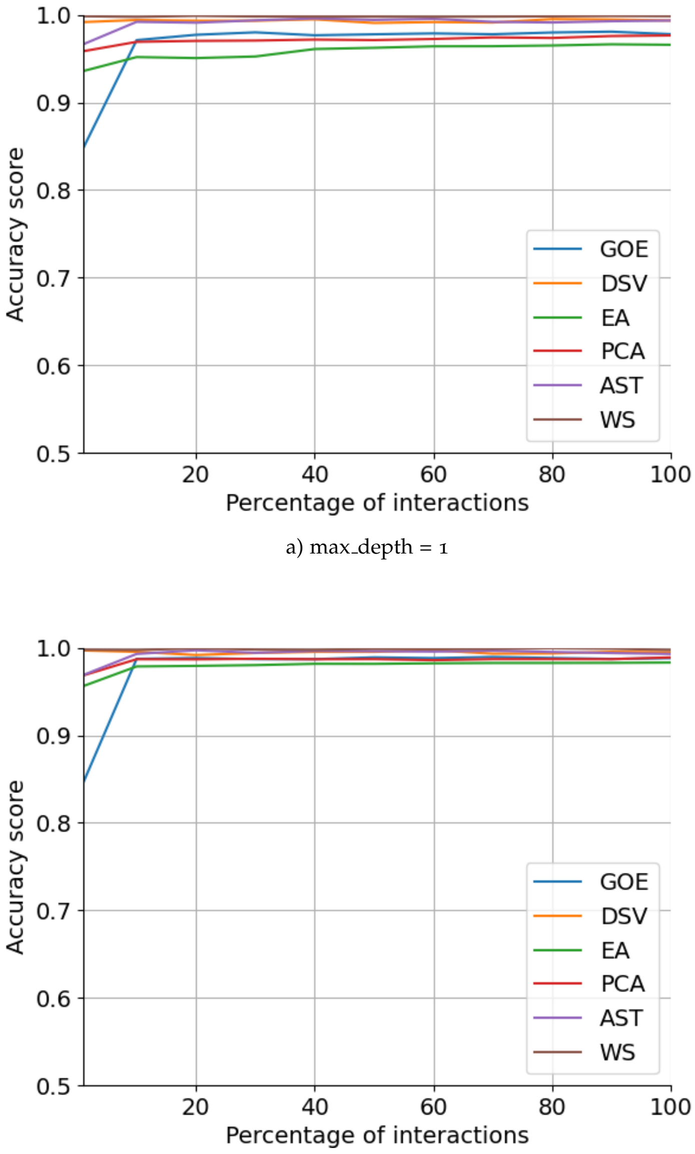
Figure 3.3.: Comparison of training and testing accuracies over varying max depths during decision tree creation.

As visible from the results to be found in Figure 3.3, experiments were done with maximum depths of 1 to 8. The difference between training and testing accuracy is minimal for a maximum tree depth of 1, and only slightly larger at 2, but increases noticeably starting at maximum depth 3 before reaching a peak at 4. After this point, the difference stays mostly constant for the rest of the experiments that were conducted.

Due to this constancy at the higher maximum depth values, it is reasonable to assume that further increasing the maximum depth of the decision tree will lead to similarly consistent differences between training and testing accuracy going forward, or even increase more. As overfitting already has a noticeable effect, it is unlikely that the effect of overfitting will become less pronounced with higher maximum depth, once more.

Taking into consideration those insights, the pronounced effect of the overfitting present starting at maximum depth 3 is also visible in the results when comparing the accuracy graphs from maximum depth 1 and maximum depth 3 in Figure 3.4. While the increase in accuracy could be legitimate under other circumstances, the previous comparison of training and test scores suggests that it is most likely overfitting that led to the increased accuracy, instead.

From the multiple trials with varying settings for the maximum decision tree depth, it was possible to decide on the parameter value for this particular data set, as explained above. While maximum depth 2 showed marginally higher difference



a) max_depth = 1

b) max_depth = 3

Figure 3.4.: Prediction accuracies for two different max_depth settings, showing the effects of overfitting.

between training and test accuracy than maximum depth 1, it also increased accuracy significantly. Because of this, a maximum depth value of 2 was used going forward.

The number of weak learners (*n_estimators*) was set to 50 for the AdaBoostClassifier, with *learning_rate* set to 1.0, as is the default.

Due to the previously performed class balancing and in combination with other validation measures, prediction accuracy is a reasonably good measurement for the goodness of fit or, in other words, the final classifier performance. Because overfitting can, even with balanced classes, lead to a misleadingly good accuracy score, it is still a good idea to check for overfitting with balanced classes, as was done with the comparison of train and test scores.

3.3. Results

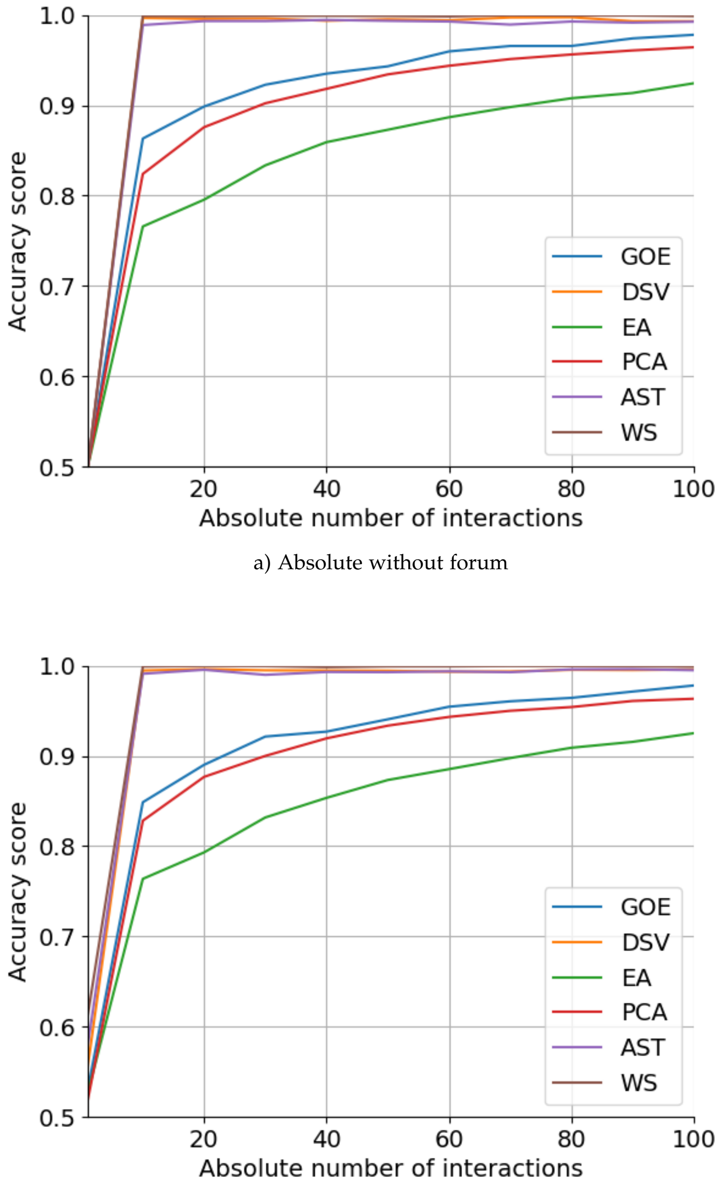
For comparison purposes, the dropout prediction analyses were run both for the data that contained forum activity in addition to the basic course interaction data, and for the data that did not. Thus, the analysis was run twice in its entirety. Furthermore, the four modes *Absolute*, *Days*, *Percentage*, and *Scaled* were implemented to use an increasing amount of data for the feature creation for each experiment. The modes were previously explained in Section 3.2.1. Each analysis was run over ten folds and the results were quantified using the average of the normalized accuracy scores of each fold for each of the analyzed courses. The results of each mode's experiments—Absolute, Days, Percentage and Scaled—are represented in one plot, with the respective x-axis representing the dimension used by the approach.

As visible in Figures 3.5, 3.6, 3.7, and 3.8, each showing the results excluding forum features at the top and those including forum features at the bottom, the difference in accuracy is minimal, and in some cases worse for the experiment including forum features than for the one excluding them.

As visible from the accuracy scores in Figure 3.5, accuracy for the mode *Absolute* is close to 0.5 when using only the first course interaction to create features. This makes sense, as this would be only marginally better than the algorithm simply guessing the class affiliation for each sample. At 10 interactions, however, the result is already significantly better, and the scores continue to increase for EA, PCA, and GOE. DSV, WS, and AST are all clustered close to 1.0. Both experiments, with and without forum features, show very similar results. The main difference seems to be that there is a small shift towards better accuracy at one singular data point, where the accuracy is closer to 0.55.

Figure 3.6 analyses the results with the actions during the first 1 to 7 days of the users' enrollments in the MOOCs. Once again, DSV, AST, and WS are closely clustered together at almost 1.0 for both the experiment with forum features, and that without. In this case, however, Figure 3.6b shows some lower scores at various points for AST, GOE, and EA, in particular. In general, accuracy still rises with time overall for each of the courses in both cases, though not as smoothly as was the case for the *Absolute* mode. It also does not reach the same heights for the courses GOE, EA, and PCA as it did in *Absolute*.

When considering percentages from 1 to 100, the accuracy scores became very stable for each MOOC, and showed a smaller range than the other solutions across all MOOCs. This is shown in Figure 3.7. GOE starts with the lowest accuracy for both the experiment with the forum features and the one without, though the one including forum features is marginally higher.

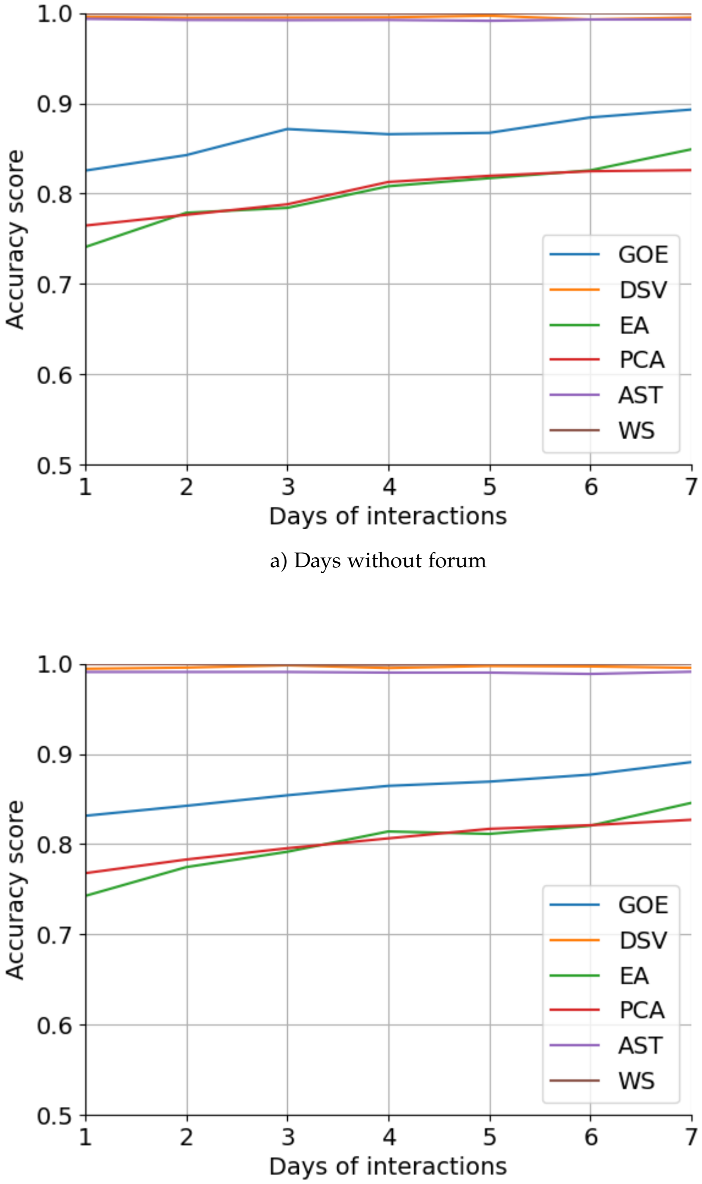


a) Absolute without forum

b) Absolute with forum

Figure 3.5.: Absolute: Accuracy of dropout prediction using BTDs as averaged over 10 folds, with and without forum features.

The results of the scaled approach, shown in Figure 3.8 seem almost like a combination of the results of *Absolute* and *Days*. Once again, DSV, AST, and WS are clustered at the top. The accuracies of GOE, PCA, and EA, however, start at similar values as they did in *Days*, and rise with increasing number of actions similarly to how they did in *Absolute*. Once more, the results with forum features swing between being better and worse than the accuracy scores for the experiment disregarding



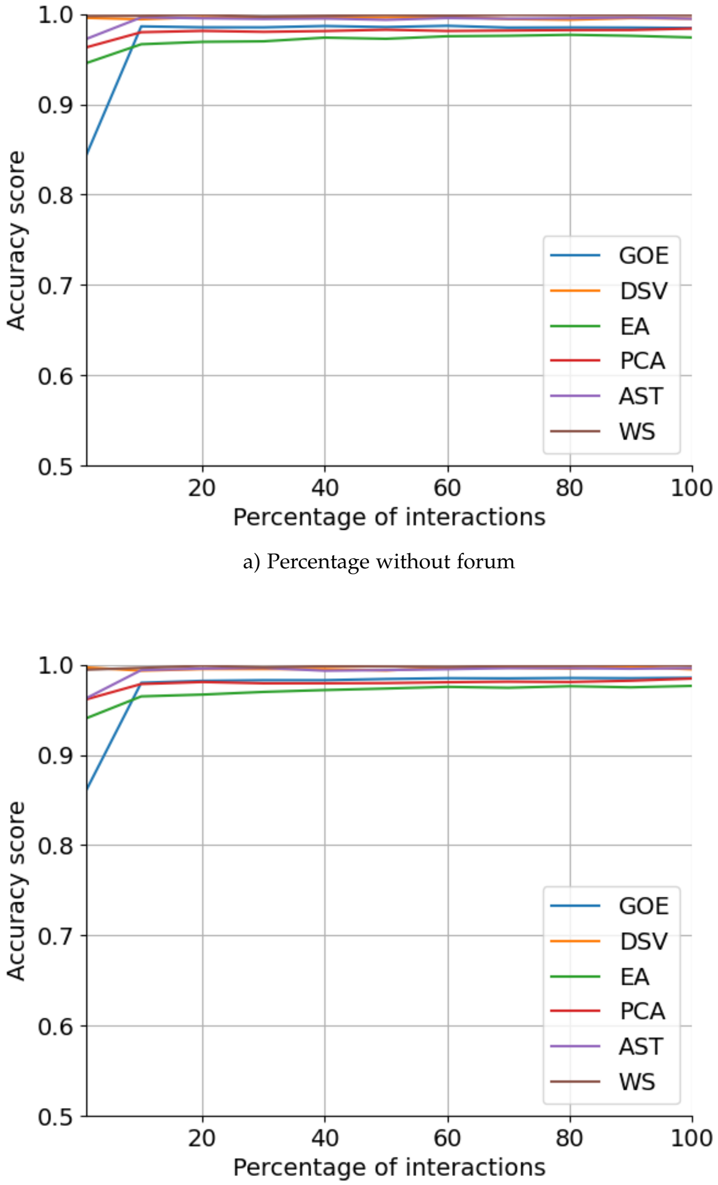
a) Days without forum

b) Days with forum

Figure 3.6.: Accuracy of dropout prediction as averaged over 10 folds, with and without forum features

the forum features.

To compare the results of the analyses using class balancing and accuracy scores with analyses made with the MCC and no class balancing, another run took place, once again with included forum data. However, while MCC is generally expected to work well for imbalanced data sets, this did not appear to be the case with the given data. The results, found in Figures 3.9 and 3.10 for Percentage and Absolute,



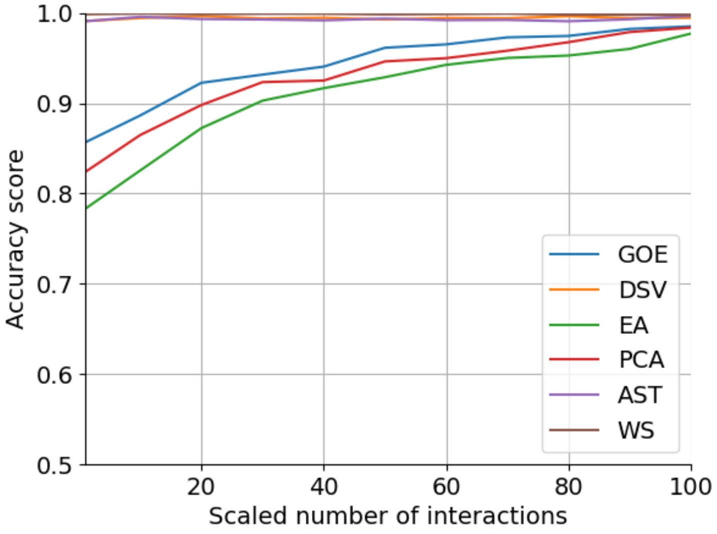
a) Percentage without forum

b) Percentage with forum

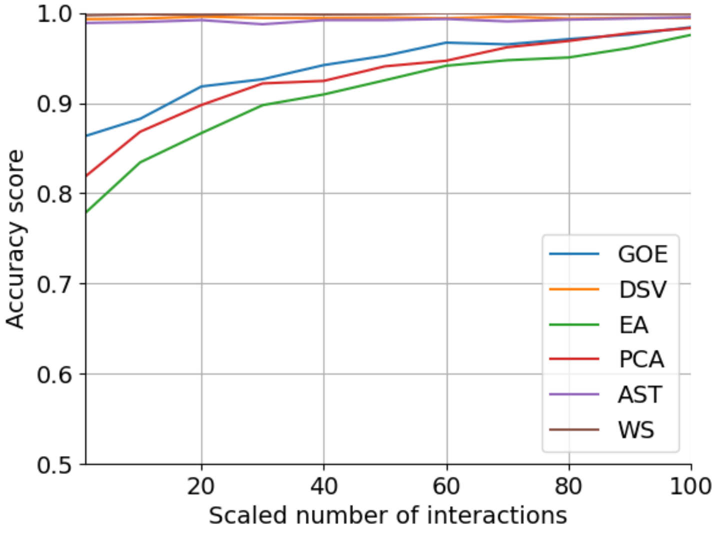
Figure 3.7.: Accuracy of dropout prediction as averaged over 10 folds, with and without forum features

respectively, showed high fluctuations in their scores when not using class balancing. This is especially the case for the courses with the three highest dropout rates, AST, DSV and WS.

However, the three courses also had significantly fewer enrollments, with between 977 and 1541 active students, than the other three, which show better results and had between 3355 and 11743 active students.



a) Scaled without forum

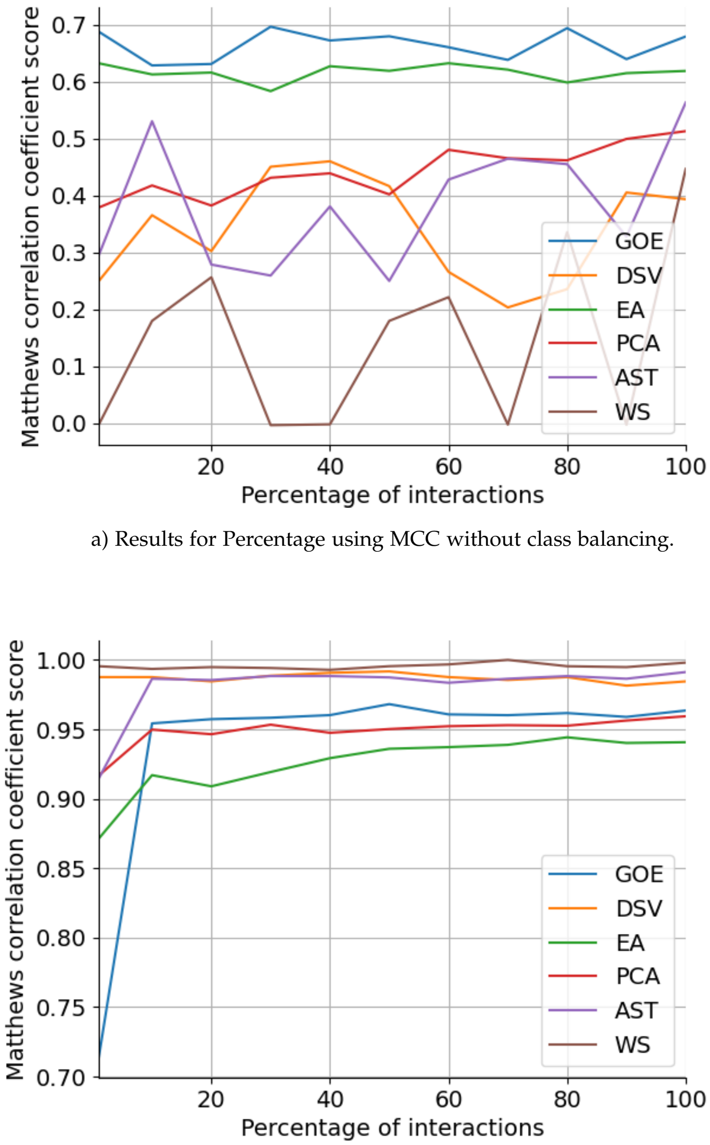


b) Scaled with forum

Figure 3.8.: Accuracy of Dropout Prediction Using Boosted Decision Trees as averaged over 10 folds

Since class balancing as used for this analysis also increases the amount of data, it is possible that the results follow from a lack of data for the smaller courses when class balancing does not take place.

Thus, after another try with class balancing, the results once again become more reasonable and closely match the accuracy scores, when taking into account that the MCC signifies the score 0 as matching the average random prediction (scikit-learn, 2020b).



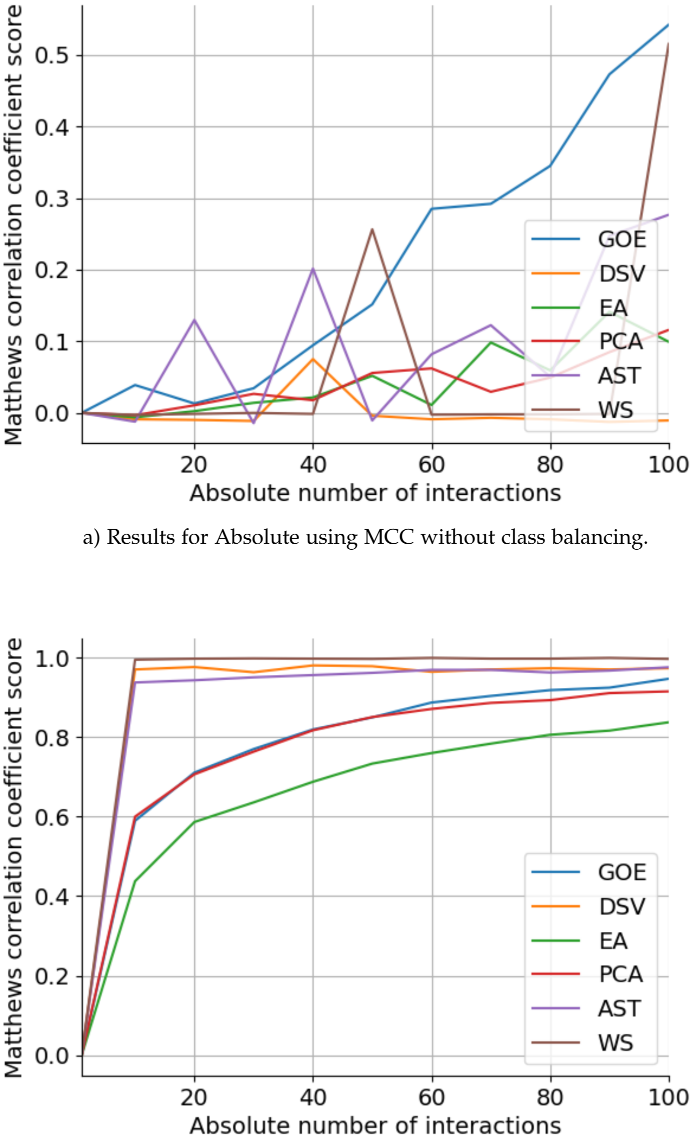
a) Results for Percentage using MCC without class balancing.

b) Results for Percentage using MCC with class balancing.

Figure 3.9.: Comparing MCC scores for Percentage with balanced and unbalanced data.

Due to this, class balancing seems necessary for the given analyses, irrespective of the selected evaluation method.

Following this, the most relevant features for the boosted decision tree classifier when used in conjunction with AdaBoost were analysed. This was done at the beginning, middle and end of the course, to be able to see differences in the progression of each feature's significance. Once again, this analysis took place for both the feature set with and without features gained through incorporating forum data to be able to compare the results, as seen in Table 3.17. The most significant



a) Results for Absolute using MCC without class balancing.

b) Results for Absolute using MCC with class balancing.

Figure 3.10.: Comparing accuracy score for Absolute with balanced data with MCC score without balancing data.

features for the experiment including forum features are shaded in blue, while the ones for the experiment excluding forum features are colored in red.

Absolute	10		50		100	
	F	N	F	N	F	N
sessions	0.024	0.004	0.044	0.056	0.060	0.024

requests	0.224	0.020	0.228	0.020	0.280	0.196
days	0.020	0.000	0.036	0.028	0.048	0.036
activetime	0.152	0.256	0.244	0.232	0.188	0.244
forumposts	0.064	-	0.132	-	0.108	-
threadstarter	0.024	-	0.016	-	0.016	-
threadresponder	0.024	-	0.000	-	0.000	-
sessionlen	0.084	0.224	0.064	0.316	0.112	0.164
clicktime	0.304	0.436	0.084	0.216	0.088	0.256
sessionrequests	0.024	0.004	0.060	0.060	0.028	0.044
dayrequests	0.056	0.056	0.092	0.072	0.072	0.036

Days	1		3		5		7	
	F	N	F	N	F	N	F	N
sessions	0.036	0.040	0.064	0.048	0.024	0.032	0.008	0.020
requests	0.144	0.100	0.184	0.216	0.228	0.228	0.224	0.264
days	0.004	0.004	0.040	0.040	0.040	0.048	0.040	0.040
activetime	0.316	0.360	0.248	0.288	0.288	0.300	0.268	0.236
forumposts	0.020		0.036		0.064		0.056	
threadstarter	0.008		0.020		0.004		0.008	
threadresponder	0.000		0.000		0.000		0.000	
sessionlen	0.132	0.184	0.108	0.108	0.080	0.108	0.064	0.116
clicktime	0.168	0.160	0.104	0.152	0.108	0.104	0.112	0.128
sessionrequests	0.112	0.132	0.068	0.052	0.100	0.120	0.096	0.084
dayrequests	0.060	0.020	0.128	0.096	0.064	0.060	0.124	0.112

Percentage	10%		50%		100%	
	F	N	F	N	F	N
sessions	0.024	0.048	0.032	0.068	0.080	0.096
requests	0.508	0.312	0.296	0.248	0.296	0.308
days	0.004	0.016	0.036	0.048	0.032	0.048
activetime	0.104	0.212	0.136	0.188	0.180	0.160
forumposts	0.084		0.124		0.124	
threadstarter	0.000		0.020		0.000	
threadresponder	0.020		0.000		0.000	
sessionlen	0.048	0.164	0.044	0.084	0.052	0.084
clicktime	0.064	0.096	0.152	0.168	0.072	0.128
sessionrequests	0.064	0.068	0.120	0.096	0.080	0.092
dayrequests	0.080	0.084	0.040	0.100	0.088	0.104

Scaled	10		50		100	

	F	N	F	N	F	N
sessions	0.044	0.032	0.040	0.020	0.064	0.076
requests	0.268	0.304	0.248	0.252	0.308	0.312
days	0.056	0.060	0.096	0.092	0.028	0.056
activetime	0.136	0.168	0.144	0.208	0.184	0.176
forumposts	0.064		0.052		0.136	
threadstarter	0.000		0.000		0.000	
threadresponder	0.020		0.008		0.000	
sessionlen	0.084	0.088	0.064	0.092	0.032	0.076
clicktime	0.144	0.152	0.164	0.192	0.088	0.128
sessionrequests	0.084	0.072	0.132	0.088	0.080	0.088
dayrequests	0.100	0.124	0.052	0.056	0.080	0.088

Table 3.17.: The feature importances at the beginning, middle and end of the course when considering forum-derived features (F) in comparison to not including forum data (N).

Both result sets showed that without fail *activetime* was amongst the three highest ranked features. When inspecting the results containing the forum features, *requests* becomes more important overall. While it appears in the ranking features 10 out of 13 times when forum data is not considered, it reaches one of the highest three scores every single time when also incorporating forum features.

Clicktime, on the other hand, seemed to get lower scores overall when forum features were present. Even when *clicktime* reached a remarkably high score both with and without forum feature consideration in the *Absolute* category at the 10 action split, the score without forum feature consideration ranked over 0.1 points above that with forum features.

The feature scores of *sessions* and *days* were low in all analysed use cases. They should not be dismissed out of hand, however, as they may be contributing to the model quality only when regarded in combination with another feature.

For the analysis that was run with the addition of forum features, the results show that *threadresponder* reached particularly low importances throughout the entire prediction process. While *threadstarter* performed better, it still did not score high enough to be considered among the three highest ranked features even once. In contrast, the feature *forumposts* was among the three highest ranking features in 4 out of 13 trials.

3.4. Discussion

One of the interests of this chapter was to examine the significance that forum interactions/forum features in dropout prediction. As Table 3.10 showed, there is a marked difference between the percentage of forum users among completers and that among dropouts. In fact, for the courses AST, DSV, and WS, more than 50% of completers wrote at least one forum post, whether it was a question or reply. This is not the case when looking at the dropouts, where participation percentages were below 9% for every course.

However, when looking at the classification results, there was no marked difference between the result with and without forum features for the most part. In fact, the scores become worse in some cases. This leads to the possible conclusion that the inclusion of the forum features may lead to overfitting. Feature selection may be necessary to counteract this. As the features *threadstarter* and *threadresponder* didn't score particularly high in the feature importance analysis found at 3.17, they may be possible candidates for exclusion, with *days* and *sessions* being two more candidates for removal for analyses with this particular data set.

When reviewing the classification results themselves, it quickly becomes apparent that the experiment where percentage amounts of actions were used to create the features performed best, even at just 10% of the actions. In comparison to the experiment where an absolute number of interactions was picked, the difference in performance is indisputable. Absolute numbers performed the worst of all experiment modes in the beginning, then jumped up sharply at 10 interactions, after which it improved steadily for GOE, EA, and PCA. DSV, AST, and WS jump from about 0.5 to almost 1.0 at 10 considered interactions. This might be because dropouts stop interacting very early on in the course, leading to drastically different features between completers and dropouts.

Overall, however, accuracy scores are only worse for the experiment considering day-wise interactions. Because that experiment considers a fixed number of days, it falls into a similar category as the one considering an absolute number of actions. Fixed numbers of days/interactions do not take into account the length of the course, or when the user actually starts interacting with the course. Longer courses will usually generate more user interactions, simply due to the higher number of course items that the users will access on average, and some students may sign up for the course only to come back to it at a later time to start and finish it, while technically only falling partially or not at all into the window of time that the day-wise experiment takes into consideration. Thus the experiment using absolute numbers will in most cases be using much less actions to fit the model than an experiment using percentages. This matched the findings in the work of Vitiello (2017), who observed that the results for Percentage were the best of the four experiments.

In closer examination of the results for the Days approach, what stands out is that there is a significant difference in the performance depending on the MOOC. AST, DSV, and WS all perform extremely well, while GOE, PCA, and EA perform in the range between 0.75 at worst and 0.89 at best. Furthermore, the accuracies are mostly stable for AC and GOE. When looking at the average of active days of users according to classes, as in Table 3.18, the difference between completer and dropouts are among the lowest for GOE, PCA, and EA. However, AST shows a smaller difference than PCA, and still performs significantly better than the other three courses. In general, the numbers show why importance of the *days*

	Average Active Days		
MOOC	Completers	Dropouts	Difference
AST	13.67	3.65	10.02
DSV	24.00	3.50	20.50
EA	13.62	4.16	9.46
GOE	9.46	2.81	6.65
PCA	16.95	3.42	13.53
WS	18.33	3.04	15.29

Table 3.18.: The average number of active days of users per class.

feature in 3.17 is generally low. Except for DSV and WS, the difference is generally comparatively low, especially considering the courses were generally without a hard deadline/closing date, and does not make for a good indicator whether someone will drop out or complete a course.

In a further attempt to explain the comparatively low performances in the results, the average interactions for each day were calculated for both completers, and dropouts, respectively. Here, the relatively low performances of EA, GOE, and PCA become clearer. As Figure 3.11 shows, the average of interactions during the first 10 days are relatively low, with similar results to be found in Figure 3.12 for both GOE and PCA. Assuming the majority of students signed up close to the start of the course, this would mean that the three courses work with only a small amount of data for the Days approach.

In any case, the average number of interactions for EA, GOE, and PCA is in general relatively low at a maximum of 8 to around 10, depending on the course. The courses AST, DSV, and WS, in comparison, all show interaction averages of 30 to 55 at the maximum, offering a lot more interactions to work with. Since the *requests* feature is consistently high in importance in Table 3.17, this explains why AST, DSV, and WS perform so well in the classification task.

When comparing the feature importance results without forum features from this work with the one done in the thesis written by Vitiello (2017), some notable overlaps in the highest scoring features can be observed. Just as was found there, this work also found *days* to be one of, if not the, lowest scoring feature. In fact, there are only six instances in Table 3.17 where *days* did not get the lowest score, and in five of those cases it was only *sessions* that scored lower—another low scorer for Vitiello (2017), as well.

The list of features that are among the highest-ranked at least twice, *requests*, *activetime*, *sessionlen*, and *clicktime* also exactly matches the ones identified as most important by Vitiello (2017). AL-Shabandar et al. (2017) also found the number of user events/actions to be the highest weighted feature in their analysis, which

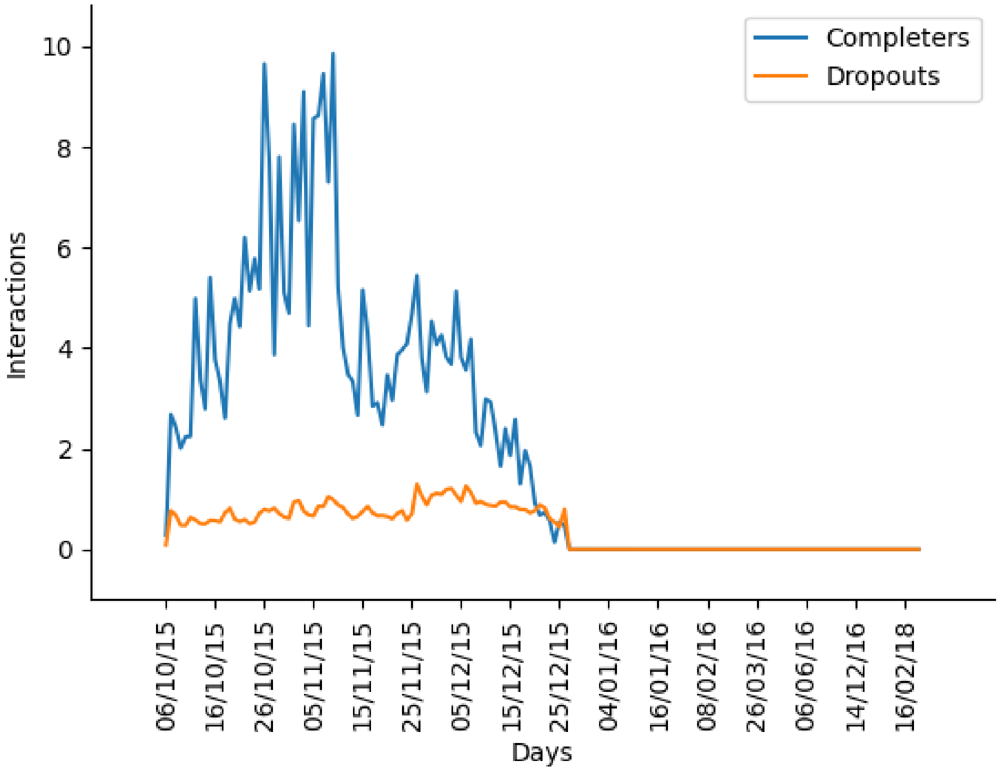


Figure 3.11.: The average number of course interactions in EA, calculated per class.

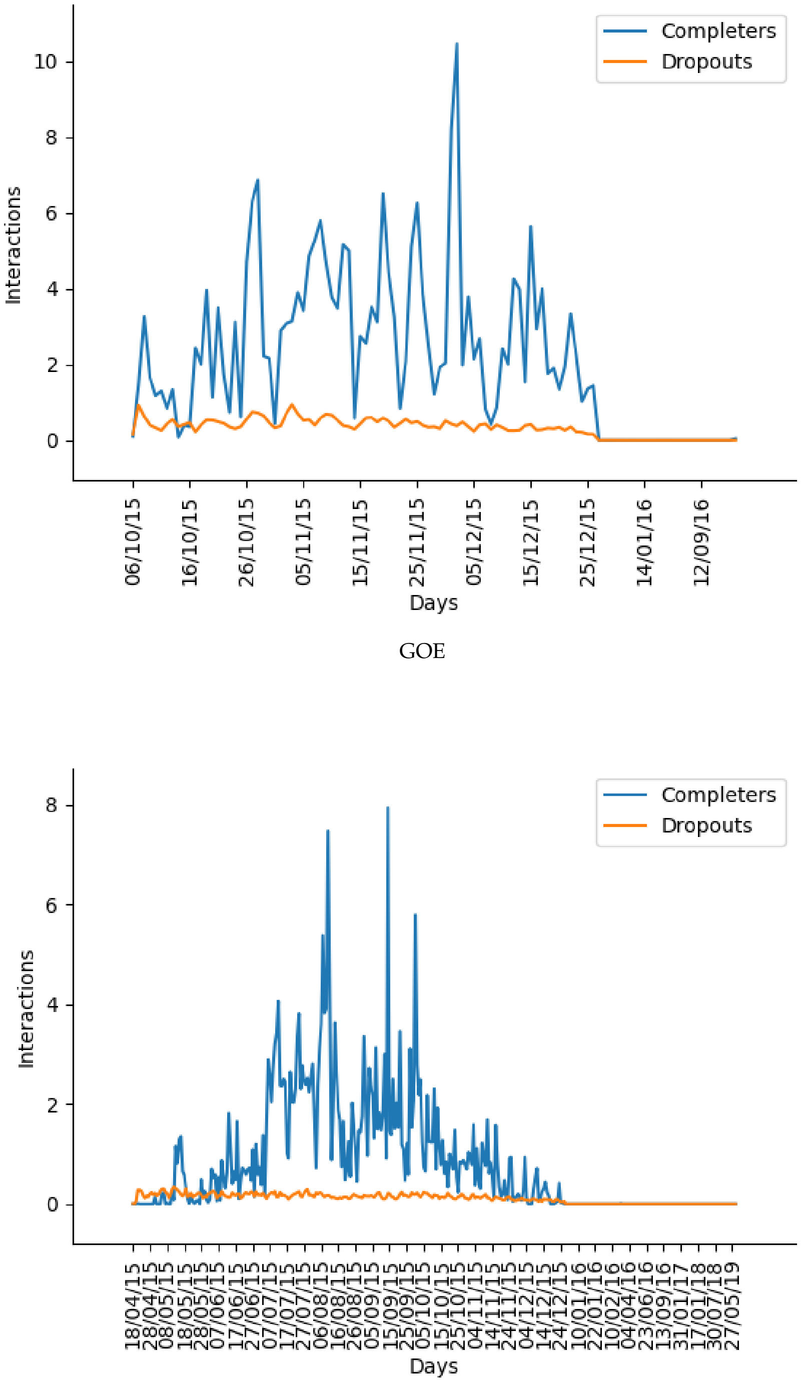
matches the results of most of the experiments in Table 3.17.

Table 3.19 summarizes the highest and lowest importance features by counting their occurrences in the top three over all experiments, in order of decreasing frequency for highest, and of increasing frequency for lowest. As it shows, the results exhibit considerable overlap in both section, with the lowest three matching in order exactly. The highest scoring features then show some differences, with active time and requests swapping places between the two works.

Furthermore, while it still is among the highest scoring features in half of the experiments, clicktime doesn't appear among the highest frequency top-scorers for Vitiello (2017) often enough to be in the top three, but does for this work. Even

	Features as found in this work	Features as found by Vitiello (2017)
Highest	Active Time, Requests, Clicktime	Requests, Active Time, Session Length
Lowest	Days, Sessions, Active Day Requests	Days, Sessions, Active Day Requests

Table 3.19.: Comparison of highest scoring features with the results found by Vitiello (2017).



GOE

PCA

Figure 3.12.: The average number of course interactions in GOE and PCA over time, calculated per class.

more notably, Session Length, which was the third highest frequency result for top scoring features in the reference work, is far from being in the top three influential features in this work, with only 19.23% of experiments where it was among the three features with the highest feature importances. In the results presented by Vitiello (2017), this was the case in 66.67% of the experiments.

What further stands out in this work are the mostly low feature importances of the forum-derived features, which does not match the results of Ramesh et al. (2014), who found forum activity to be the second-most important feature in their analysis. In Section 3.1, it was shown in Table 3.10 that the percentage of forum users was consistently higher among completers than it was among dropouts. While participation percentages ranged between 3.17% and 8.20% for dropouts, they were between 18.63% to 85.71% for completers. In fact, the average participation percentage over all courses for completers was 47.46%, but only 5.52% for dropouts.

In both cases, however, the participation rates were higher than the 3% participation rate over all students that Breslow et al. (2013) found in their analyzed MOOC. This seems to stem from more active Dropouts, as his numbers for certificate earners—52% of them were active in the forum—closely matches the 47.46% of active completers in the forums in this work. From these statistical analyses in regards to forums, it is possible to postulate that there seems to be a correlation between course completion and forum activity in most evaluated courses.

Further looking into this possible connection, Table 3.13 shows that in terms of behavior, completers were generally more likely to interact with the forums by both creating new threads, and responding to existing ones, with percentages between 18.42% and 83.33%. Dropouts usually either leaned towards creating only new threads, or only answering to existing ones, though which one their behavior is skewed to differs between courses. For DSV, EA, GOE, and WS, students who only created new threads outweigh the others, while in AST and PCA, dropouts showed a preference for only responding to existing posts.

This is also reflected when looking at the average question and response posts per person in Table 3.20, where completers and dropouts were treated separately. While there are 0.04 question posts per enrolled dropout in PCA, there are 0.03 response posts. Because the number of students who are active in the forum is so much smaller than the number of enrolled students altogether, however, the numbers are generally very low, and differences small.

This does not change significantly when looking at the average number of total forum interactions per user, where the numbers don't rise over 5.43 posts at the maximum, for the course WS. This can be seen in Table 3.21. Even when considering only active forum posters, the highest average is 6.33, once again for WS. Because of these low numbers, it is reasonable to expect low importance of features derived from the number of forum posts—the differences between completers and dropouts is ultimately too small to have much impact.

	Avg. Q		Active Avg. Q		Avg. R		Active Avg. R	
MOOC	C	D	C	D	C	D	C	D
AST	0.81	0.06	1.13	0.68	1.76	0.12	2.47	1.50
DSV	0.81	0.04	1.44	0.96	2.06	0.04	3.67	0.82
EA	0.31	0.04	1.21	0.92	0.65	0.05	2.48	1.03
GOE	0.19	0.06	1.03	0.88	0.19	0.03	1.00	0.53
PCA	0.36	0.02	1.36	0.59	0.47	0.03	1.77	1.06
WS	1.43	0.05	1.67	0.92	4.00	0.08	4.67	1.36
Overall	0.65	0.05	1.31	0.83	1.52	0.06	2.68	6.30

Table 3.20.: Average number of forum questions and responses per class, averaged over all students, and only those who were active in the forum. "Q" refers to new threads being posted, while "R" is a reply to an existing thread. "C" is short for completers, and "D" for dropouts.

	Avg. Posts		Active Avg. Posts	
MOOC	Completers	Dropouts	Completers	Dropouts
AST	2.57	0.18	3.60	2.18
DSV	2.88	0.08	5.11	1.78
EA	0.96	0.09	3.69	1.95
GOE	0.38	0.09	2.02	1.41
PCA	0.83	0.05	3.12	1.65
WS	5.43	0.13	6.33	2.28
Overall	2.18	0.10	3.98	1.88

Table 3.21.: Average number of forum posts per class, both averaged over all students, and only over those who were active in the forum.

3.5. Summary

High dropout rates are a pervasive problem in the area of MOOCs. This chapter looked at user course interaction data and forum data in an attempt to find correlations between dropouts and user actions. With feature creation processes being run at different points throughout the course/with varying amounts of data, the analysis aimed to investigate the progression of the results from earlier in the course to later on. Using decision trees that were boosted with AdaBoost and with a maximum three depth of 2, it found a high degree of accuracy was able to be reached with this approach.

Over all courses, feature importance results showed some differences to previous research, in that session length did not seem to be as highly ranked in this case. What was found in this case was that the three most important features were the amount of active time in the course, the number of requests, and the average time between two clicks.

When investigating the effects of inclusion of forum interaction data, a larger number of forum interactions, especially if the user engaged both in existing threads and created new ones, seemed to be positively correlated with course completion in the statistical analyses at the start of the chapter. Upon inclusion of forum data in the classification learning data, however, comparisons of the results showed that features that were derived from forums did not necessarily lead to improved performance of the used classifier, most likely due to the mostly small difference between the number of forum posts of the two classes, and feature importances for forum features were mostly low.

Such being the case, it might be of interest to try similar analyses with courses that have a higher percentage of forum interactions among the students—perhaps by having one or more assessments directly include interactions with the forum in some shape or form.

4. Analysing E-Learning Data from an Austrian University

This work also considers a data set collected throughout a class at an Austrian university to investigate learning behavior in connection with student performance. The online system is built to support self-regulated learning, and represented a mandatory section of the coursework, which the students were required to pass in itself to be able to pass the course.

As previously described in Section 2.1.8, the e-learning course consists of multiple sections, which are themselves made up of chapters or subsections that the students work through. It is left up to them whether they do this in the given order or not, or even whether they study any of the section chapters at all. Each section contains an associated assessment that the students can take whenever they feel ready to do so. If they are not satisfied with the result, they can retake the assessment immediately, or go back and revise the section contents once more before another attempt of the assessment. There is no maximum limit on the number of allowed assessment attempts, and only the best result is counted towards the student's final grade.

4.1. Dataset

The analysed course is part of a class at an Austrian university and was only made available to students who were signed up for it during the winter term in 2018/19. The class itself enrolled 31 students, none of which dropped out during its progression. It was not openly available on the web. Because of this, the target group of the course was quite small, and the data set gained from it is limited. With the trial group being only one class of students, the resulting data set is made up of the same 31 users who were originally enrolled in the parent-class, with most of them having a background in technology, particularly computer science (see Table 4.1).

At the beginning and end of the course, a survey was sent out to the participants where they were asked to supply demographic data as well as information about expectations and motivations, and assessments of the course structure and execution, respectively. The introductory survey was filled out by 19 of the 31 students. The summarised information from the first survey can be found in Table 4.1, as well.

Besides the information collected through the surveys, two data files contained the information collected throughout the progression of the course. One of the available files for this course consists of user traces, showing interactions of the

Demographic	Response distribution		
	Response	Frequency	Percentage
Gender	Female	8	25.80%
	Male	11	35.48%
	Other	0	0.00%
	No Answer	12	38.71%
Age	18-20	1	3.23%
	21-23	6	19.35%
	24-26	10	32.26%
	>26	2	6.45%
	No Answer	12	38.71%
Field of Study	Computer Science	13	41.94%
	Software Engineering and Management	5	16.13%
	Information and Computer Engineering	1	3.23%
	No Answer	12	38.71%

Table 4.1.: Demographic data of the students who filled out the survey at the beginning of the course. Of the 31 students taking the course, 12 chose not to answer the survey and are recorded in the "No Answer" options of the table.

users with the course material in the form of visits and assessments, as well as login and logout actions. When applicable, the log entries show the section that an action was part of, as well as the visited material's id, in addition to a timestamp, among other information.

Additionally, a second log file shows the users' assessment results, as well as the number of times they were repeated, giving the information listed in Table 4.2.

The course website is built to show an overview page with links to each of the different content pages after the successful login. From there, it is possible to get an overview over the course topics, and access every content chapter.

The course itself is structured in three main topics/sections, with additional instructional and learning progress pages, which are also accessible from the overview page. Each of the sections consists of a number of content chapters (see Table 4.3) and has learning goal information and a corresponding assessment that can be retaken an infinite amount of times, with the best try counting towards the final result. There are no prerequisites to be able to see any piece of learning material, and though there is an order to the content chapters and sections, it is not a requirement for the student to study them in order, or even at all, to be able to

Tag	Description
id	Interaction ID
user	User ID
assessment	List of assessment sections with interactions
section	Section ID
completedsections	Number of successfully completed sections
status	0 if not successfully completed, 1 if successful
timestamp_start	timestamp of assessment start
timestamp_finish	timestamp of assessment end

Table 4.2.: Description of the data received through a course at an Austrian university.

		# of Chapters per Section		
# of Students	# of Graded Assessments	Section 1	Section 2	Section 3
31	3	7	5	5

Table 4.3.: The course in numbers—student count, assessment numbers and content chapters per section

take the section exam. The content chapters themselves primarily contain textual descriptions and information about the section topic.

Besides using the overview page to navigate the content, it is also possible to change to the previous or next chapter directly from the currently viewed one.

4.2. Framework and Concept

The small size of the data set presented a challenge, in that typical machine learning algorithms require more information to be able to create an appropriate model. Because of this factor, originally the decision was made to build this analysis on the concept of sequential analysis, which was explained in Section 2.2.4. However, due to the fact that no assumptions of cross-dependencies can be safely made in this case, the analysis was ultimately further simplified the process by applying a sequence mining approach. This method was previously presented by Malekian et al. (2020) in their attempt to find whether there were specific patterns to be found in students who failed assessments, as opposed to those who passed.

For this, the data was first cleaned of any test and/or invalid users. Then, the actions were used to create user profiles, where each profile contained the actions with timestamps sorted by time. Similarly to the preparation done in Chapter 3, actions that took place within 30 minutes from each other were considered to be part of one "session". If the break time between two consecutive actions exceeded 30 minutes, a logout/login action pair was added in between the two actions if it was not already present, as previously done by Saint et al. (2020). This served to make the switch between two sessions visible in the heat map visualising the user behavior.

Code	Description
IN	Login
OUT	Logout
IND	Index page
INS	Instruction page
PRO	Progress page
U1	Section 1 content
U2	Section 2 content
U3	Section 3 content
G1	Goal page of section 1
G2	Goal page of section 2
G3	Goal page of section 3
A1	Assessment of section 1
A2	Assessment of section 2
A3	Assessment of section 3

Table 4.4.: Action codes with short descriptions.

In the next step, so-called "codes" were defined for specific actions and/or action types of students. The actions were substituted with these codes, defined in Table 4.4, to reduce complexity. Any accesses that involved a content chapter of a particular section were coded with that section—e.g. an access of an item belonging to section 2 would simply be referred to as "U2". The one exception to this case was the overview page for the section's exam, which was dropped as it functions only as an information page before the assessment effectively starts. The outcome was a list of codes for each student which described the student's path through the online course.

Following this, the students were split into two classes—those who repeated at least one of the assessments, and those who did not, denoted from here on with "O" for the students who did not repeat, and "R" for those who did. This data was then used to create the transition matrices/heat maps depicted in Appendix A.1 for the O and Appendix A.2 for R. Due to the fact that at the end of the course, each

Code	Description
v	The content was new to the user and was viewed for the first time
rv	The content was known to the user and being reviewed
a-p	The assessment was passed with 100% of the points
a-f	The assessment was passed with less than 100% of the points

Table 4.5.: Codes used in sequence mining, adapted from those used by Malekian et al. (2020).

student had reached the full 100% of each assessment, these codes may effectively be used interchangeably with passing and failing in coming sections, if failing is considered to be any result below the 100% mark, in this specific case.

Based on the work of Malekian et al. (2020), the data was then further simplified by using only content accesses, and assessment attempts with their result. In this way, content was broken down to extract whether a student was viewing content that was new to them, or reviewing content that they had already viewed in the past during the course. This process also removed the information how students moved from content to content, as this information can also be extracted from the heat maps in Appendices A.1 and A.2 and does not influence how the students interacted with the actual learning content.

Similarly, assessments were simplified into passing and failing tries, or as applicable for this paper: attempts where 100% of the points were reached and attempts where less than 100% were reached. Furthermore, while the original work considered attempts of the same assessment as the previous one, and attempts of a different assessment to the previous one as two separate cases each with their own label, it was decided due to the size of the available data in this case not to make the distinction.

One further reason for this was that due to the structure of the course, the support for sequences including assessments were significantly lower than those including only views and reviews, as the number of course content pages far outnumbered the assessment attempts. Overall, assessment repetitions did not take place often enough for each individual assessment, or for other assessments than the most recent one, to be able to use this approach with the available data. Even with this change, sequences containing assessment attempts were ultimately far outnumbered by those that contained only view and review actions.

The final codes used in this process and going forward are documented in Table 4.5, together with short descriptions of each.

Following this, sequence extraction was used on the combined data of O and R, and for each class' data, individually. Referring once more to Malekian et al. (2020), this was done with different lags, which were previously described in Section 2.2.4, and selected to be in the range of 1 to 9. For each lag, the occurring sequences were extracted and their support was calculated to be able to evaluate importance.

Sequences below a specific minimum support value were dropped. This minimum support value was selected to be 20% in the original work. However, in this thesis, the lack of data caused results to be sparse for the 20% minimum, leading for the limit to be dropped to 10%.

4.3. Results

The data consisted of 31 students with an average of 65.53 actions per student. There were a total of three sections, each with one final, graded assessment that the students were supposed to pass to complete the course. 46 of the 93 assessments in total were completed with the maximum reachable points at first try. Generally, students re-attempted the assessments until they got full points. At maximum, this took four tries. The distribution of the number of assessment attempts per user can be viewed in Figure 4.1.

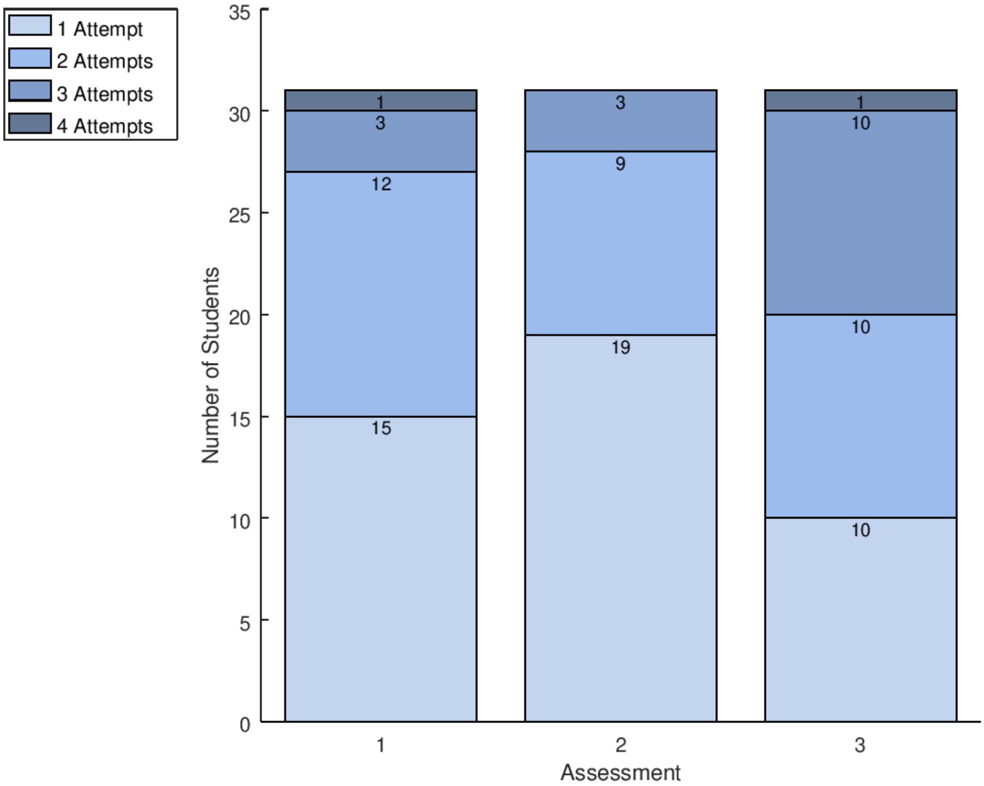


Figure 4.1.: Number of assessment attempts per student, per assessment

As shown, the majority of students completed assessments 1 and 2 only once, while assessment 3 had an even distribution of students who attempted it one, two and three times. Assessment 2 seems to have a lower number of attempts in

Assessment	# of Attempts	Average # of Attempts
1	52	1.68
2	46	1.48
3	64	2.06

Table 4.6.: The table shows the number of attempts for the assessments, average number of attempts per student per assessment

comparison to the others altogether, with more than half of the students making only one attempt. One student in both assessment 1 and assessment 3 attempted the assessment four times, while the maximum for assessment 2 was three attempts. An overview of the total number of attempts per assessment, as well as the average number of assessment attempts for each of the three assessments can be found in Table 4.6.

With the multiple assessment attempts, there was a steady increase in the point percentage, overall and per student. The increase in the overall achieved percentage across each assessment is shown in Figure 4.2.

During the second attempt, there was an average of 75% increase in the total points for the students who did not finish at full points on their first try, and another average of 41.6% total point increase on the third try, if there was one.

The view numbers for each section can be found in Table 4.7. As the table shows, the instruction page was visited a total of 55 times. Most of the views of this section took place before any of the sections and their contents were viewed. Only 15 of the 55 checks took place during an active session, which for this analysis is considered the time between the first interaction with a section's content—including the section goal page—and the first visit of the section's assessment page.

The number of accesses for each of the learning goal pages was very even, as can be seen in Table 4.7. The learning goal page of the relevant section was generally checked before the section's learning content chapters were visited, as shown in the column "Before LC".

Content	Accesses	Before LC	Average per Student
Instruction Page	55		1.77
Section 1: Learning Goal	48	38	1.54
Section 2: Learning Goal	46	35	1.48
Section 3: Learning Goal	48	35	1.54
Progress Page	107		3.45

Table 4.7.: Number of accesses per course section, where LC is short for "Learning Content"

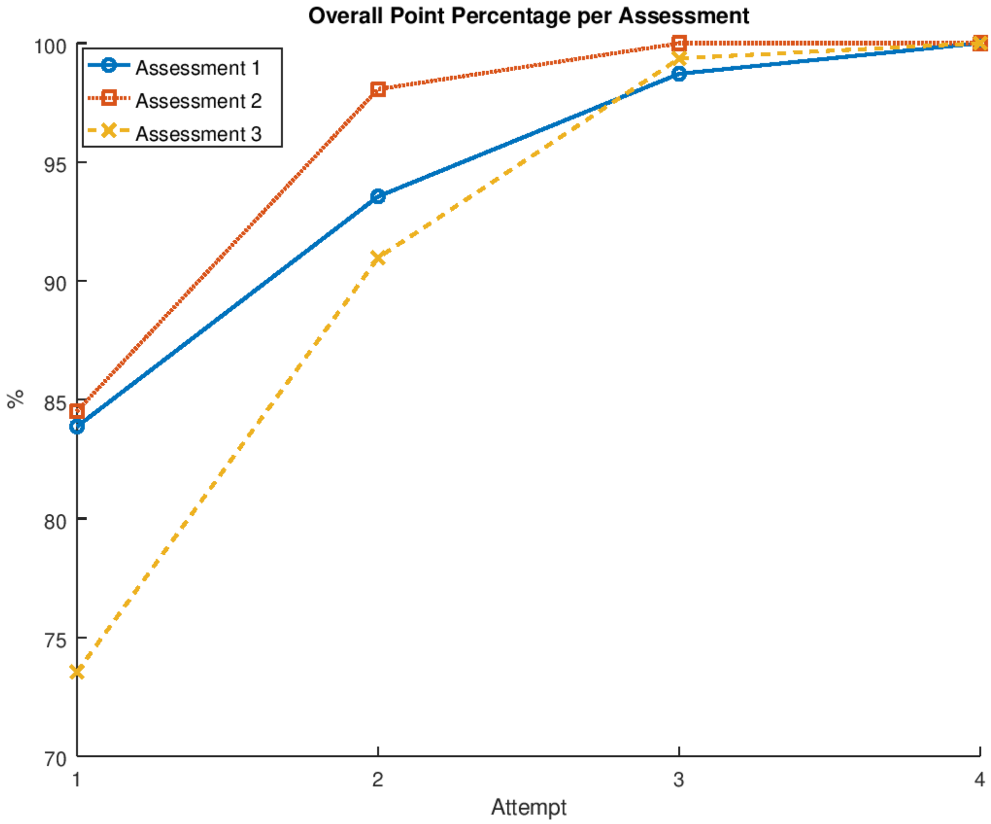


Figure 4.2.: Improvement in the overall point percentage across all students, per assessment.

Learners checked their progress on the progress page a total of 107 times. Only 11 of those checks took place during an active session, and in fact most often the learning progress was only checked after the student had finished all three assessments, though 12 students also viewed their progress before retaking previously finished assessments.

About half of the students viewed the content chapters in a section in the order they were intended to be viewed, while a small subset of 5 students did not view the learning units at all or only a select few before taking the assessment. The numbers were calculated by section and show some variation in the relation between ordered and unordered learners, as visible in Figure 4.3.

About half—15 of the 31—of the students switched between sections during an active sessions.

Of 31 students, 29 took the assessments in order. Of the others, one student worked through the assessments in order, then started over to do them once more, ending with a third try of assessment 1. The last student started with 2, went on to finish section 1 and ended with 3.

In the case that a student repeated an assessment attempt, the majority of students did not view the goal or learning sections again. Of the students who repeated the

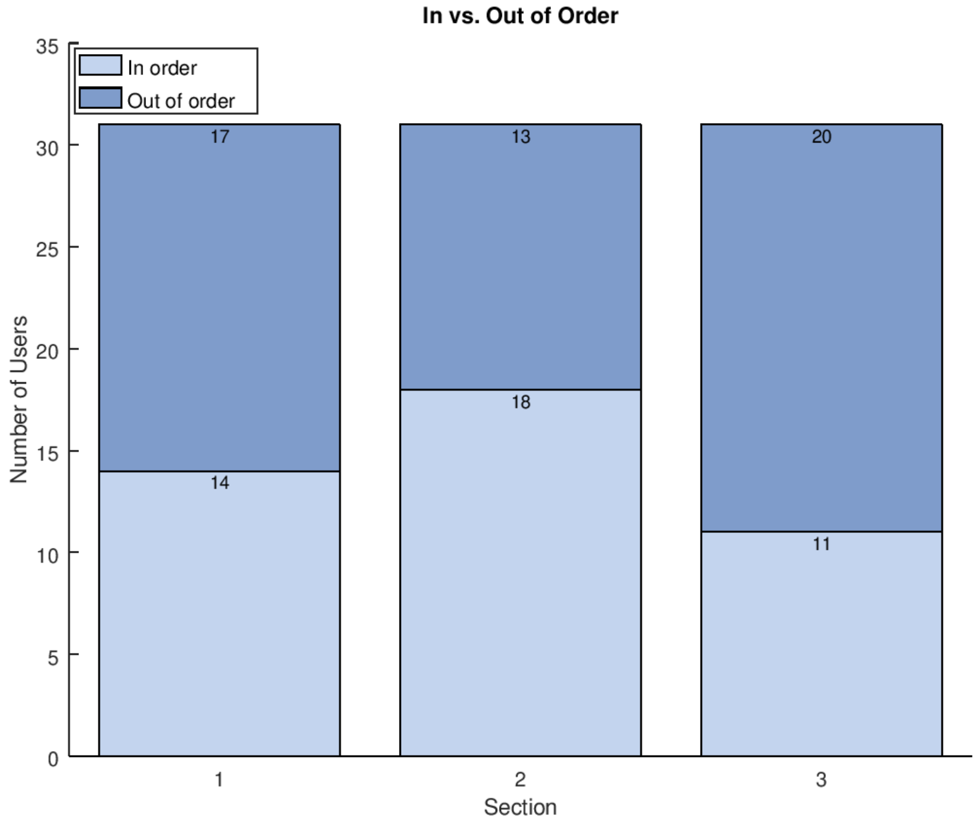
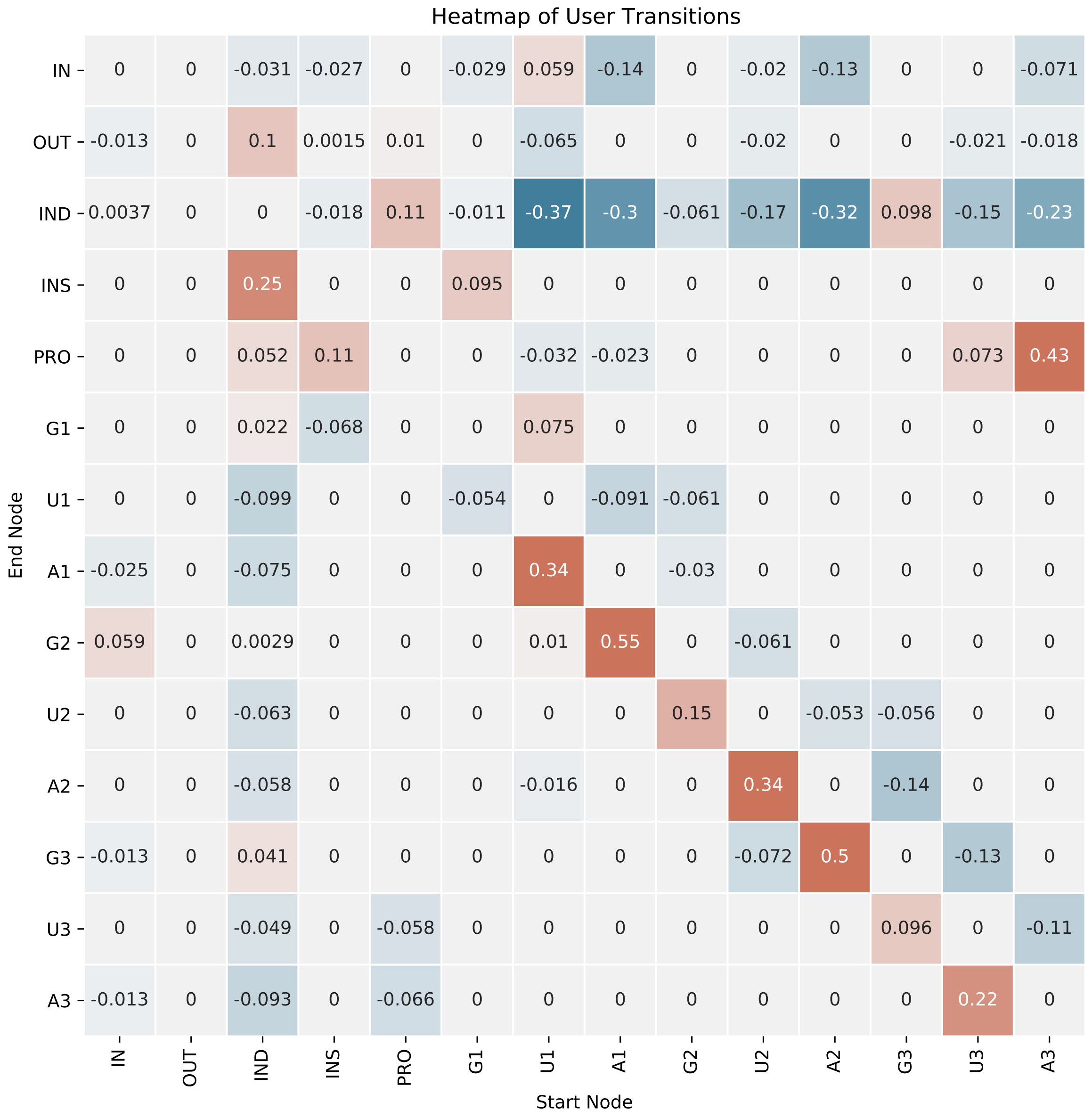


Figure 4.3.: Number of students who viewed the sections in order/out of order

assessments one or multiple times, half reviewed one or more learning contents for at least one retry in section 1, 5 of 7 students reviewed learning contents in section 2, and only 6 of 15 students reviewed contents in section 3.

To be able to assess differences in student behavior in more depth, this work calculated the transition probabilities between each pair of states for both the students who repeated assessments, and those who did not. The probabilities were then visualised in heat maps, where higher probabilities are higher saturated and lower probabilities have less saturation. Figure A.1 shows the results for students who did not repeat assessments, and Figure A.2 shows those of students who did repeat at least one assessment once or more times. Both of these heat maps can be found in the Appendix.

Figure 4.4 shows the difference in behavior between students who repeated assessment and those who did not. For this, the probability values for transitions found for students who repeated were subtracted from those found for students who did not repeat. Because of this, negative, or blue-shaded numbers are transitions that had higher probability for repeating students, while positive, or red-shaded numbers had higher probability for students who did not repeat. The darker the shade, the higher the difference between the results for the two classes.



	IN	OUT	IND	INS	PRO	G1	U1	A1	G2	U2	A2	G3	U3	A3
IN	0	0	-0.031	-0.027	0	-0.029	0.059	-0.14	0	-0.02	-0.13	0	0	-0.071
OUT	-0.013	0	0.1	0.0015	0.01	0	-0.065	0	0	-0.02	0	0	-0.021	-0.018
IND	0.0037	0	0	-0.018	0.11	-0.011	-0.37	-0.3	-0.061	-0.17	-0.32	0.098	-0.15	-0.23
INS	0	0	0.25	0	0	0.095	0	0	0	0	0	0	0	0
PRO	0	0	0.052	0.11	0	0	-0.032	-0.023	0	0	0	0	0.073	0.43
G1	0	0	0.022	-0.068	0	0	0.075	0	0	0	0	0	0	0
U1	0	0	-0.099	0	0	-0.054	0	-0.091	-0.061	0	0	0	0	0
A1	-0.025	0	-0.075	0	0	0	0.34	0	-0.03	0	0	0	0	0
G2	0.059	0	0.0029	0	0	0	0.01	0.55	0	-0.061	0	0	0	0
U2	0	0	-0.063	0	0	0	0	0	0.15	0	-0.053	-0.056	0	0
A2	0	0	-0.058	0	0	0	-0.016	0	0	0.34	0	-0.14	0	0
G3	-0.013	0	0.041	0	0	0	0	0	0	-0.072	0.5	0	-0.13	0
U3	0	0	-0.049	0	-0.058	0	0	0	0	0	0	0.096	0	-0.11
A3	-0.013	0	-0.093	0	-0.066	0	0	0	0	0	0	0	0.22	0

Figure 4.4.: Heat map showing the difference between transition probabilities for students repeated assessments and those who did not.

To further investigate the connection between learner behavior and outcome, the work specified ordered/unordered behavior as additional categories, to be combined with the previously defined classes O (for students who did each assessment only once), and R (for students who repeated at least one assessment). Class affiliation for O/R was calculated for each of the three assessments separately, which means that this specific class could change for a student, depending on whether

	unordered			ordered		
	A1	A2	A3	A1	A2	A3
O	29.41%	46.15%	25.00%	71.43%	72.22%	45.45%
R	70.59%	53.85%	75.00%	28.57%	27.78%	54.55%

Table 4.8.: Percentages of users who took an assessment once/multiple times per assessment number, for ordered learners, and unordered learners.

the student repeated that particular assessment. Ordered/unordered classes were calculated over the whole course, as it considered the overall learning behavior the student displayed. Using these numbers, Table 4.8 shows the calculated percentages of unordered learners who took the specific assessment one/multiple times, as well as the same for ordered learners.

The sequence extraction approach was originally run with a set minimum support of 20%, where support is calculated by dividing the number of occurrences of a specific sequence by the total number of extracted sequences. Any sequences that did not reach this number were dropped from the results. Like in the reference work of Malekian et al. (2020), the sequences were then modeled in sequence diagrams, where arrows leading back to the state itself signify an additional repeat of the state, and the numbers in the square brackets above the arrows denote the number of repetitions to be found among the set of extracted sequences. For example, the annotation "[1...3]" means that the repetition of the state takes place between (and including) 1 and 3 times. The state names used in the sequence diagrams were previously described in Table 4.5. This led to the results visualised in Figure 4.5, collecting all sequences with the lags in the range of 1 to 9. The term "lags", in this case, is adopted from sequential analysis and used as described in Chapter 2.2.4.

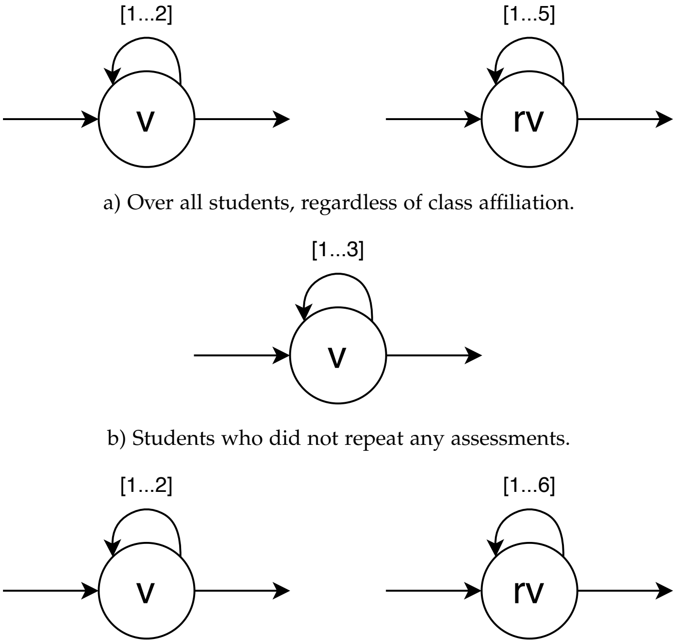
However, the result set was quite sparse, with no extracted sequences that reached the minimum support score of 20% for lag 6 and up for the data set with all students. For the class O, there were no valid results from lag 4 upwards, while the results for R only showed a lack of results starting at lag 7.

In an effort to gain more insight into the behavioral patterns and improve the results, the minimum support percentage was subsequently lowered to 10. For the combined-class data set, as well as class R, all nine lag sizes returned results in this case, though the data set associated with the class O still did not show results for lags larger than 4. The results for this experiment can be found in Figure 4.6.

4.4. Discussion

Generally, the number of students who viewed a section's content in order varies by section.

Section one was viewed in order by 45.16% of the students, while section two

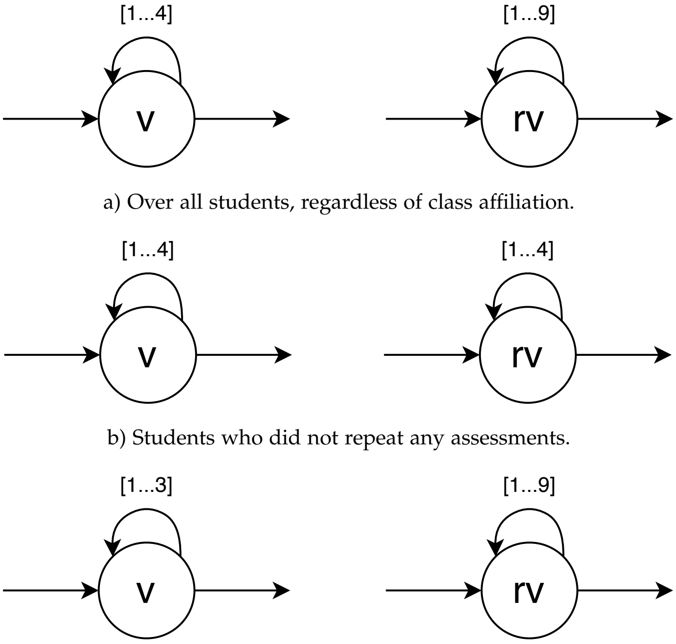


a) Over all students, regardless of class affiliation.

b) Students who did not repeat any assessments.

c) Students who repeated at least one assessment.

Figure 4.5.: State transitions with support 20, overall, for students with one assessment attempt, and for students with repeated assessment attempts.



a) Over all students, regardless of class affiliation.

b) Students who did not repeat any assessments.

c) Students who repeated at least one assessment.

Figure 4.6.: State transitions with support 10, overall, for students with one assessment attempt, and for students with repeated assessment attempts.

even increased this percentage to 58.06%. After section two, the percentage falls once more, with section three only showing 35.48% of students viewing its content in order. Interestingly, this behavior seems to negatively correlate with the number of assessment attempts, as assessment 1 was repeated only 1.48 times per student on average, but assessment 2 had an average of 2.06 attempts per student, as seen in Table 4.6.

This apparent connection between organised learning behavior and higher average results seems to be further confirmed by the results in Table 4.8. Learners that displayed unordered learning behavior appear to have a higher probability of repeating an assessment, while people that viewed all learning content in order have a high percentage of users who used only one assessment attempt to reach the full points. This seems related to, and matches previous findings from Mukala, Buijs, and Van Der Aalst (2015), who discovered that students who performed better usually showed a higher degree of structural learning behavior than those who did not perform as well.

While a seemingly minor difference, the unorganised learning behavior also shows itself in disjointed view actions, which causes the lower number of repetitions of them in Figures 4.5 and 4.6. As opposed to students of the class O, who have fewer review actions in a row—so much so, in fact, that it did not pass the 20% support minimum in Figure 4.5—members of the class R show a tendency to review a larger number of past content chapters.

When comparing the heat maps for the classes O (Figure A.1) and R (Figure A.2), this general mode-changing behavior shows as well in the number of connections. Several aspects stand out in the heat map, one of them being that O learners were apparently more likely to check the instructions. For O students, the probability of a transition from the index—the entry point into the system after the log in, so a point everyone would pass—to the instruction page was at 40%, while it was at 15% for students of the class R. In fact, the class R seemed to use the index in a different way than the class O, in that they jumped to different parts of the course with it more often, and, noticeably so, returned to it more often from the various course content sections in the course, as well.

Users from class R also appeared to directly return to previous material more often, further showing why the 'rv' action was so highly supported in Figures 4.5 and 4.6. On the other hand, users from class O did not generally return to previous content directly, with the only exception being the return from content of the section 1 back to the goal of the same section.

From the numbers in Table 4.7, it is reasonable to conclude that students paid a large amount of attention to their progress page. For a course that contained three assessments, the average progress page checks per student were 3.45. For students who did not repeat their assessments, 50% of transitions were from assessment 3 to the progress page. For repeaters, the transition probability from assessment 3 to progress page was still at 45%. This can also be found mirrored in Figures A.1 and A.2. Class O especially checked the progress page after finishing assessment 3 in

88% of the cases, while for class R it was still 45%. From the progress page, most users returned to the index or logged out, but a small number of transitions—5% for class O, and 12% for class R—accessed assessment 3, either to take it for the first time, or to retake it. Because of this, it appears that students generally re-evaluated their results after the last assessment, possibly to go back and improve a result if they were not satisfied with their final grade.

Generally, the acceptance of the course structure and according learning behavior showed higher student assessment results initially, showing that the course's order and organisation is well defined. In any case, however, assessment results invariably improved with additional attempts, and all students finished with 100% of the points. As the course represented part of a longer class, and took place partway through the semester, it is likely that the students had motivation to finish the course with good results because they were already invested in the outcome. Due to the fact that they had likely also already put work into the class previous to this segment, this probably functioned as further impetus.

This is also reflected in the fact that the course was finished in a single session by the majority (61.30%) of students. On the average, the course was finished in 1.55 sessions, overall.

4.5. Summary

This chapter focused on investigating the link between learning behavior in a self-regulated learning environment and assessment performance. Machine learning algorithms do not represent the best solution when the number of samples is small, even when the number of classes and/or features is small, as well. Due to this limitation, the data set in this chapter, with 31 students and 3 assessments, was analysed by way of sequence extraction and calculating their support.

In combination with heat maps and various statistical analyses, this revealed that students who showed better self-organisation skills had a tendency to need fewer attempts in reaching 100% of the points on all assessments. Students who repeated one or more assessments reviewed a larger number of content chapters than those who completed all assessments with full points on the first try. Additionally, it was found that students who did not repeat any assessments tended to view more new content chapters without interruptions than the other class.

General behavior differed significantly in the use of the index, where users who repeated assessments were found to use it significantly more to navigate the course than users who did not, who mostly jumped directly from one content chapter to the following one. The inclusion of a progress page, in particular, was found to be well-accepted and often-used by the students, and the acceptance of the suggested structure showed some connection to initial higher student performance.

However, because the students all completed the course with full points, the learning method of the students who did not keep as closely to the suggested order of the learning material should ultimately still be considered as being successful.

5. Lessons Learned

This chapter briefly examines the insights gained throughout the work on this thesis.

5.1. Literature

Over the course of working on chapter 2, what became increasingly evident was the wide scope of e-learning technologies, as well as the underlying learning and behavioral theories. It quickly became clear that it was not enough to merely concentrate on technical aspects when the focus of much of the research is on the user behavior in relation to the respective e-learning system. This was also often represented in background literature that mentioned links of user behavior to different psychological theories such as the model of self-regulated learning and the theory of planned behavior. Additionally, factors such as motivation, goal-setting, etc. play a big role for learning systems when it comes to successfully engaging users.

Similarly diverse as the technologies used for e-learning are the methods used to analyse e-learning data. Depending on the amount of data available and the research objective, anything from statistical analysis over decision trees to support vector machines finds application. Picking the right method(s) for the available data plays a big role in getting useful results.

Much of the research over past years focused on analysing user behavior in an attempt to predict dropouts. However, more recent research has started moving towards a more differentiated view on what should be considered successful participation in MOOCs, which may not always correspond to completing the entire MOOC. As such, literature showed that the understanding on what succeeding in an e-learning course means may be changing, and that it is not only high final scores that equate to a successful course participation.

5.2. Development

Throughout development of the analyses for chapter 3, the main difficulty was finding the right parameters for the decision tree classifier. Due to setting the maximum depth of the decision tree too high in the beginning, overfitted results led to unrealistic numbers. Additionally, class balancing was utterly crucial. A first attempt at the analyses was made without class balancing and showed a near-constant almost-100% accuracy without regard for the number of samples used in the machine learning process. After class balancing, the numbers became more realistic, and the accuracy score more meaningful. Additionally, what should not be disregarded is the side-effect of gaining more data through class balancing, when it is done by generating new samples. This was especially notable in the results of the analyses using MCC as opposed to accuracy.

In contrast, the main difficulty with the data set analysed in chapter 4 was that it was small. The first idea was to use a (hidden) Markov model to analyse the user behavior. However, throughout the preparation of the data and statistical analyses it quickly became apparent that this would not be possible due to lack of data. The transition matrix was then used to create the heat maps to visualise user transitions. Since one of the goals was to see whether the students visited the sections in order or jumped between them, it quickly became apparent that it would not make sense to merge the separate learning sections into one bigger 'learning content' section for the heat maps. However, the same did not make sense for the sequence diagrams, which show the view vs. review behavior. Because of this, the section chapters were ultimately considered regardless of the section they belonged to for this analysis, only in view of whether they were being viewed for the first time or reviewed.

5.3. Evaluation

If there is one thing to be learned from the evaluation in chapter 3, it is that even when one expects there to be enough data, there probably is not. In this specific case, the forum data was not significant enough to show the results previous research was able to find. Furthermore, while feature importance can give an indication for which features should be included and which can be dropped, it is not the be-all and end-all, as it can be specific to the analysed data set, as found when comparing the results from Vitiello (2017) to the ones from this thesis. A feature that performs highly for one data set may not do so for a different one. The feature importance scores do not represent this. Thus, they need to be used with caution when it comes to feature selection.

In regards to the analyses in chapter 4, structured learning seems to either be a reason for initial higher performance, or a side-effect of a particular learning type. While it is not possible to tell whether students who display structured learning behavior show a tendency to perform highly because of their learning behavior, or if they display structured learning behavior and perform highly because of a different factor altogether, there does seem to be a connection in some way. Finally, students who are already part-way into a course will likely finish an e-learning course and not drop out. If there has been a certain amount of work put into the course before the e-learning part starts, it seems to be likely that the students will attempt to perform highly.

6. Conclusion and Future Work

This chapter focuses on summarizing the findings gained in this thesis, and considers possible future improvements and directions for further research.

6.1. Conclusion

With digitisation and digital learning becoming prevalent, the varieties of systems and underlying learning theories have become numerous. By picking two different systems, this thesis aimed to get a wider view over the various forms of learner behavior and their patterns, and how the data that is accumulated in online learning systems can be used to draw conclusions about student performance, and dropout prediction. Specifically, this thesis focused on the MOOC courses of a South-American university, hosted on the platform Coursera, and the self-hosted, self-regulated learning based e-learning course of an Austrian university.

The general theory behind the two systems was similar, in that they both assumed a measure of self-organising skill from the students to be able to complete the course, and the content could be viewed in whichever order the student preferred. However, while the Coursera MOOCs all had a minimum of 1.5 thousand students, and were openly available on the web, the other learning system was part of a university-specific class and allowed only the students enrolled in the class access.

Due to the different sizes, data scopes, functionalities, and general research questions involved with each of the data sources, two different approaches were chosen to analyse the courses.

For the more traditional MOOC courses from the South-American university, this thesis aimed to conduct dropout prediction using boosted decision trees, using steadily increasing amounts of the available data. The influence of forum data on the outcome of the dropout prediction was evaluated, to see if the inclusion had positive effects on accuracy. Furthermore, the feature importances were calculated for both the case including forum data, and the one excluding it.

Findings showed that features such as the number of user actions, the user's active time, and the average time between clicks generally scored highly in the results. The inclusion of forum features did not have a notable effect on the results, with the features threadstarter and threadresponder scoring especially low feature importances. This represents the notable limitation of the analysis, namely the lack of forum participation of the users. Only an average of 6.38% of users participated in the forums, which are too few to lead to a meaningful result when it comes to the influence of forum activity on user retention.

For the self-directed learning system from the Austrian university, the aim was to evaluate the general user behavior and find possible patterns, as well as links to learner performance. Due to the considerably smaller size than the Coursera courses, which were open to the general public and thus drew in a much larger group of learners, this presented a challenge for the usage of machine learning algorithms. To avoid the problems of over- and underfitting due to missing data, traditional sequence extraction was used in combination with heat maps and statistical analysis to find patterns in behavior and outcomes.

The results showed a possible connection between ordered learning behavior and an initial higher performance in the assessments. Students who viewed the learning contents in order tended to score full points on their first assessment attempts more often, though the end results after sometimes multiple assessment attempts had every student at full points. However, due to the small size of the data set and the fact that the course took place partway through a (face-to-face) university course, the students would have already been motivated to finish, so the results need to be viewed in its context. In a stand-alone e-learning course, the students may not have reattempted the assessments until they reached full points, or they might have dropped out entirely.

6.2. Future Work

For MOOCs, there are many possible directions where further research can take place. Because dropout rates are so high, what may be of interest in the future is what motivations and factors make a learner decide *not* drop out. If we assume dropping out to be the rule, what differs for the students who do not do so? Investigating which factors completers may have in common to the exclusion of dropouts may be a possible direction to take this in. It may also be of interest to take the learners' original motivations into account, since it is a possibility that not everyone who enrolls in a course enrolls with the specific goal of completing it.

Furthermore, as briefly mentioned in the summary of Chapter 3, focusing on courses that place a stronger focus on utilisation of forums may be of interest when including forum features in classification tasks. Results may show more significant difference before and after inclusion of forum interactions if, for example, one of the initial tasks involves interacting with the forum, even if this consists of a semi-formal task such as a short introduction post. It may be of interest if this initial, mandatory interaction with the forum functions as a factor to convince users who may otherwise not have become aware of or active in the forum to consider participation even after the task.

Similarly, it would be interesting to evaluate whether semi-formal assignments such as the task to write an introduction post in an existing thread have positive effects on forum participation, or function as a block that causes more people to drop out before doing the forum task. And if this is the case, do dropout rates fall later in the course or do they behave as usual?

In the case of the self-directed learning system, it may be of interest to review the analysis with more data. Due to the lack of data, Markov models could not be applied in this specific case, but may be a possible direction for further research.

It may also be interesting to consider whether the behavior of the students for this course matches the behavior in courses that are not part of a larger course. The same thing is true with a course that is the first task of a longer class. Do students feel higher motivation to repeat assessments until they have reached the full points even if the self-directed learning course is unattached to/not partway through a university class?

In that case, it would also be possible to study the effect the self-directedness has on the dropout rates. Due to the specific nature of how the self-directed course was included in the university course, no students dropped out after starting, but would this still be the case if the course were longer and not, for the most part, able to be finished in one session? Do shorter courses increase student motivation in this case? Do multiple short, self-directed courses with very specific topics fare better than one long, self-directed course that covers multiple topics?

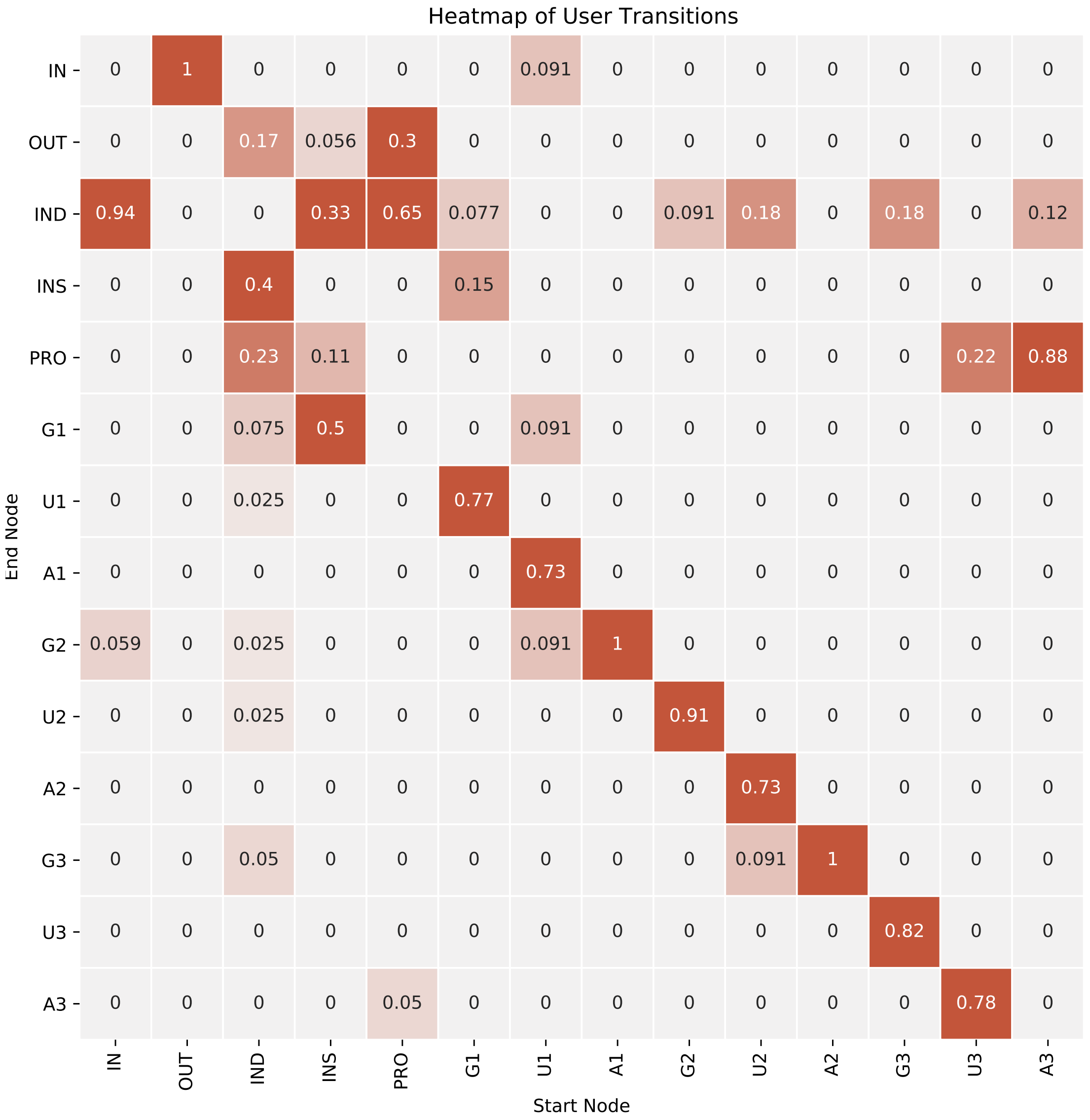
Finally, the difference in behavior, and possible connection between ordered/ unordered learning and performance could also be used to predict student outcomes during active course sessions, and allow for the teachers to intervene if a student is in danger of dropping out or failing the course.

Appendix

Appendix A.

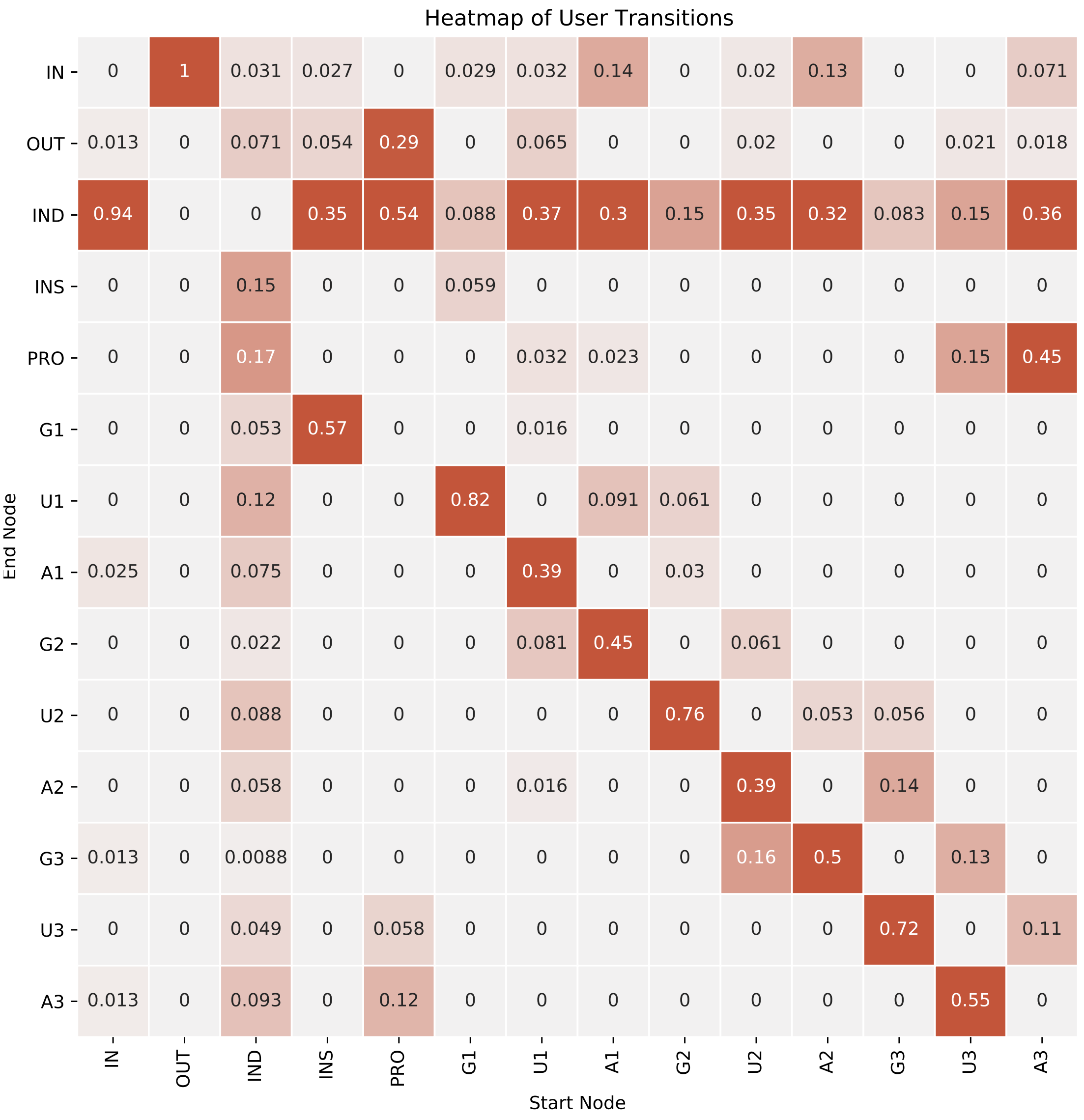
Transition Heat Maps

A.1. Students who did not repeat assessments



	IN	OUT	IND	INS	PRO	G1	U1	A1	G2	U2	A2	G3	U3	A3
IN	0	1	0	0	0	0	0.091	0	0	0	0	0	0	0
OUT	0	0	0.17	0.056	0.3	0	0	0	0	0	0	0	0	0
IND	0.94	0	0	0.33	0.65	0.077	0	0	0.091	0.18	0	0.18	0	0.12
INS	0	0	0.4	0	0	0.15	0	0	0	0	0	0	0	0
PRO	0	0	0.23	0.11	0	0	0	0	0	0	0	0	0.22	0.88
G1	0	0	0.075	0.5	0	0	0.091	0	0	0	0	0	0	0
U1	0	0	0.025	0	0	0.77	0	0	0	0	0	0	0	0
A1	0	0	0	0	0	0	0.73	0	0	0	0	0	0	0
G2	0.059	0	0.025	0	0	0	0.091	1	0	0	0	0	0	0
U2	0	0	0.025	0	0	0	0	0	0.91	0	0	0	0	0
A2	0	0	0	0	0	0	0	0	0	0.73	0	0	0	0
G3	0	0	0.05	0	0	0	0	0	0	0.091	1	0	0	0
U3	0	0	0	0	0	0	0	0	0	0	0	0.82	0	0
A3	0	0	0	0	0.05	0	0	0	0	0	0	0	0.78	0

A.2. Students who repeated at least one assessment



End Node \ Start Node	IN	OUT	IND	INS	PRO	G1	U1	A1	G2	U2	A2	G3	U3	A3
IN	0	1	0.031	0.027	0	0.029	0.032	0.14	0	0.02	0.13	0	0	0.071
OUT	0.013	0	0.071	0.054	0.29	0	0.065	0	0	0.02	0	0	0.021	0.018
IND	0.94	0	0	0.35	0.54	0.088	0.37	0.3	0.15	0.35	0.32	0.083	0.15	0.36
INS	0	0	0.15	0	0	0.059	0	0	0	0	0	0	0	0
PRO	0	0	0.17	0	0	0	0.032	0.023	0	0	0	0	0.15	0.45
G1	0	0	0.053	0.57	0	0	0.016	0	0	0	0	0	0	0
U1	0	0	0.12	0	0	0.82	0	0.091	0.061	0	0	0	0	0
A1	0.025	0	0.075	0	0	0	0.39	0	0.03	0	0	0	0	0
G2	0	0	0.022	0	0	0	0.081	0.45	0	0.061	0	0	0	0
U2	0	0	0.088	0	0	0	0	0	0.76	0	0.053	0.056	0	0
A2	0	0	0.058	0	0	0	0.016	0	0	0.39	0	0.14	0	0
G3	0.013	0	0.0088	0	0	0	0	0	0	0.16	0.5	0	0.13	0
U3	0	0	0.049	0	0.058	0	0	0	0	0	0	0.72	0	0.11
A3	0.013	0	0.093	0	0.12	0	0	0	0	0	0	0	0.55	0

Bibliography

Akçapinar, G., Chen, M.-R. A., Majumdar, R., Flanagan, B., & Ogata, H. (2020). Exploring student approaches to learning through sequence analysis of reading logs. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 106–111. https://doi.org/10.1145/3375462.3375492

alearningjourneyweb. (2017). *Self-directed vs self-regulated learning*. Retrieved February 22, 2020, from https://alearningjourneyweb.wordpress.com/2017/03/13/self-directed-vs-self-regulated-learning/

Alharan, A., Alsagheer, R., & Al-Haboobi, A. (2017). Popular decision tree algorithms of data mining techniques: A review. *International Journal of Computer Science and Mobile Computing*, *6*, 133–142.

Alias, U. F., Ahmad, N. B., & Hasan, S. (2015). Student behavior analysis using self-organizing map clustering technique. *ARPN Journal of Engineering and Applied Sciences*, *10*, 17987–17995.

AL-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., & Radi, N. (2017). Machine learning approaches to predict learning outcomes in massive open online courses. *2017 International Joint Conference on Neural Networks (IJCNN)*, 713–720. https://doi.org/10.1109/IJCNN.2017.7965922

Amry, A. B. (2014). The impact of whatsapp mobile social learning on the achievement and attitudes of female students compared with face to face learning in the classroom. *European Scientific Journal*, *10*(22).

Anderson, L. W., Krathwohl, D. R., & Airasian, P. W. (2000). A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Bédard, D., Lison, C., Dalle, D., Côté, D., & Boutin, N. (2012). Problem-based and project-based learning in engineering and medicine: Determinants of students' engagement and persistance. *Interdisciplinary Journal of Problem-based Learning*, *6*(2), 7–30.

Berthold, M., & Hand, D. J. (2003). *Intelligent data analysis* (Vol. 2). Springer.

Blei, D. M., & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences of the United States of America*.

Boeker, M., Andel, P., Vack, W., & Frankenschmidt, A. (2013). Game-based e-learning is more effective than a conventional instructional method: A randomized controlled trial with third-year medical students. *PLoS One*, *8*(12).

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. *Wadsworth Int. Group*.

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*, 13–25.

Brochenin, R., Buijs, J. C. A. M., Vahdat, M., & van der Aalst, W. M. P. (2017). Resource usage analysis from a different perspective on mooc dropout. *ArXiv*, *abs/1710.05917*.

Černochová, M., & Selcuk, H. (2019). Digital literacy, creativity, and autonomous learning. *Encyclopedia of Education and Information Technologies*, *10*.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5-6), 318–331.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, B., & Denoyelles, A. (2013). Exploring students' mobile learning practices in higher education. *Educause Review*, *7*(1), 36–43.

Cheng, L.-C., & Chu, H.-C. (2019). An innovative consensus map-embedded collaborative learning system for er diagram learning: Sequential analysis of students' learning achievements. *Interactive Learning Environments*, *27*(3), 410–425.

Cherkassky, V., & Mulier, F. (2007). *Learning from data: Concepts, theory, and methods* (Second). John Wiley & Sons, Ibc.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6.

Conole, G. (2016). Moocs as disruptive technologies: Strategies for enhancing the learner experience and quality of moocs. *Revista de Educacion a Distancia*, (50).

Coursera. (2018a, July 20). *Coursera*. https://about.coursera.org/

Coursera. (2018b, July 20). *Coursera data exports guide*. https://www.gitbook.com/book/coursera/data-exports/

Coursera. (2018c, July 21). *Enrollment options*. https://learner.coursera.help/hc/en-us/articles/209818613-Enrollment-options

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. https://doi.org/https://doi.org/10.1016/j.iheduc.2011.06.002

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 9–15. https://doi.org/10.1145/2181037.2181040

Driscoll, M. (2002). *Blended learning: Let's get beyond the hype*. Retrieved February 4, 2020, from http://www-07.ibm.com/services/pdf/blended_learning.pdf

Engelbertink, M. M., Kelders, S. M., Woudt-Mittendorff, K. M., & Westerhof, G. J. (2020). Participatory design of persuasive technology in a blended learning course: A qualitative study. *Education and Information Technologies*, 1–24.

Faraone, S. V., & Dorfman, D. D. (1987). Lag sequential analysis: Robust statistical methods. *Psychological bulletin*, *101*(2), 312.

Fink, G. (2014). *Markov models for pattern recognition: From theory to applications*. Springer London.

Fogg, B. (2003). *Persuasive technology: Using computers to change what we think and do*. Elsevier Science. https://books.google.at/books?id=gfiYh%5C_BHj94C

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*, *96*, 148–156.

Friedman, T. (1999). Next, it's e-ducation. *New York Times*, *17*, A29.

Garris, R., Ahlers, R., & Driskell, J. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, *33*, 441–467. https://doi.org/10.1177/1046878102238607

Garrison, C., & Ehringhaus, M. (2007). *Formative and summative assessments in the classroom*. http://www.amle.org/Publications/WebExclusive/Assessment/tabid%20/1120/Default.aspx

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial.

Glanz, K., Rimer, B., & Viswanath, K. (2015). *Health behavior: Theory, research, and practice*. Wiley. https://books.google.at/books?id=9BQWCgAAQBAJ

Gottman, J., Gottman, J., Roy, A., Press, C. U., & Roy, I. (1990). *Sequential analysis: A guide for behavioral researchers*. Cambridge University Press. https://books.google.at/books?id=TIC1gMlgXGsC

Greene, J. A., & Azevedo, R. (2007). A theoretical review of winne and hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, *77*(3), 334–372.

Grigorova, K., Malysheva, E., & Bobrovskiy, S. A. (2017). Application of data mining and process mining approaches for improving e-learning processes. *Proceedings of the International conference Information Technology and Nanotechnology*, 1960–1966.

Hagedoorn, T. R., & Spanakis, G. (2017). Massive open online courses temporal profiling for dropout prediction. *2017 International Conference on Tools with Artificial Intelligence*, 231–238. https://doi.org/10.1109/ICTAI.2017.00045

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Massachusetts Institute of Technology.

He, C., Ma, P., Zhou, L., & Wu, J. (2018). Is participating in mooc forums important for students? a data-driven study from the perspective of the supernetwork. *Journal of Data and Information Science*, *3*(2), 62–77.

He, J., Bailey, J., Rubinstein, B. I. P., & Zhang, R. (2015). Identifying at-risk students in massive open online courses. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 1749–1755.

Henderikx, M., Kreijns, K., & Kalz, M. (2018). Intention-behavior dynamics in moocs learning; what happens to good intentions along the way? *2018 Learning with MOOCs (LWMOOCS)*, 110–112.

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with moocs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, *145*. https://doi.org/https://doi.org/10.1016/j.compedu.2019.103724

Hrastinski, S. (2008). Asynchronous and synchronous e-learning. *EDUCAUSE Quarterly*, *31*(4).

Hubackova, S. (2015). History and perspectives of elearning. *Procedia - Social and Behavioral Sciences*, *191*, 1187–1190. https://doi.org/10.1016/j.sbspro.2015.04.594

Kahiigi, E. K., Hansson, H., Ekenberg, L., & Tusubira, F. F. (2008). Explorting the e-learning state of art. *Electronic Journal of e-Learning*.

Kenny, D., Kashy, D., & Cook, W. (2006). *Dyadic data analysis*. Guilford Publications.

Kesim, M., & Altınpulluk, H. (2015). A theoretical analysis of moocs types from a perspective of learning theories [The Proceedings of 5th World Conference on Learning, Teaching and Educational Leadership]. *Procedia - Social and Behavioral Sciences*, *186*, 15–19. https://doi.org/https://doi.org/10.1016/j.sbspro.2015.04.056

Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *41*(3), 552–568.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & Education*, *104*, 18–33. https://doi.org/https://doi.org/10.1016/j.compedu.2016.10.001

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2016). Recommending self-regulated learning strategies does not improve performance in a mooc. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, 101–104. https://doi.org/10.1145/2876034.2893378

Knowles, M. (1975). *Self-directed learning: A guide for learners and teachers*. Association Press.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2003). Preventing student dropout in distance learning using machine learning techniques. *Knowledge-Based Intelligent Information and Engineering Systems*, 267–274. https://doi.org/10.1007/978-3-540-45226-3_37

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., & Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. *Advances in Neural Information Processing Systems*, 2744–2752.

Kulkarni, A. S., & Naik, K. R. (2019). Bite sized learning: Transforming global e-learning (16 bold). *Journal of Emerging Technologies and Innovative Research (JETIR)*, *6*(3).

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, *9*, 3093. https://doi.org/10.3390/app9153093

Lens, W., & Vansteenkiste, M. (2012). Promoting self-regulated learning: A motivational analysis. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications*. Taylor & Francis. https://books.google.at/books?id=MDQLfOg0jXoC

Li, L. (2019). *Classification and regression analysis with decision trees*. Retrieved November 20, 2020, from https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054

Liang, J., Li, C., & Zheng, L. (2016). Machine learning application in moocs: Dropout prediction. *2016 11th International Conference on Computer Science & Education (ICCSE)*, 52–57.

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. https://doi.org/https://doi.org/10.1016/j.iheduc.2015.12.003

Littlejohn, A., & Pegler, C. (2007). *Preparing for blended e-learning*. Routledge.

Lloyd, B. P., Yoder, P. J., Tapp, J., & Staubitz, J. L. (2016). The relative accuracy and interpretability of five sequential analysis methods: A simulation study. *Behavior research methods*, *48*(4), 1482–1491.

Locke, E. A., & Latham, G. P. (2004). What should we do about motivation theory? six recommendations for the twenty-first century. *Academy of Management Review*, *29*(3), 388–403.

Lung-Guang, N. (2019). Decision-making determinants of students participating in moocs: Merging the theory of planned behavior and self-regulated learning model. *Computers & Education*, *134*, 50–62. https://doi.org/https://doi.org/10.1016/j.compedu.2019.02.004

Maimon, O., & Rokach, L. (2006). Decision trees. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 165–192). Springer US.

Maimon, O., & Rokach, L. (2014). *Data mining with decision trees: Theory and applications (2nd edition)*. World Scientific Publishing Company. https://books.google.at/books?id=OVYCCwAAQBAJ

Malekian, D., Bailey, J., & Kennedy, G. (2020). Prediction of students' assessment readiness in online learning environments: The sequence matters. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 382–391. https://doi.org/10.1145/3375462.3375468

McLoughlin, L., & Magnoni, F. (2017). The move-me project: Reflecting on xmooc and cmooc structure and pedagogical implementation. In K. Quian & S. Bax

(Eds.), *Beyond the language classroom: Researching moocs and other innovations* (pp. 59–69). Research-publishing.net.

Michaelsen, L. K., & Sweet, M. (2008). The essential elements of team-based learning. *New directions for teaching and learning*, *2008*(116), 7–27.

Moe, R. (2015). The brief & expansive history (and future) of the mooc: Why two divergent models share the same name. *Current issues in emerging elearning*, *2*(1), 2.

Molnar, A. (1997). Computers in education: A brief history. *The journal*, *24*(11), 63–68.

Mukala, P., Buijs, J., & Van Der Aalst, W. (2015). Exploring students' learning behaviour in moocs using process mining techniques. *Department of Mathematics and Computer Science, University of Technology, Eindhoven, The Netherlands*, 179–196.

Mukala, P., Buijs, J. C. A. M., Leemans, M., & van der Aalst, W. M. P. (2015). Learning analytics on coursera event data: A process mining approach. *SIMPDA*.

Nicholson, P. (2007). A history of e-learning: Echoes of the pioneers. *Computers and Education: E-Learning, from Theory to Practice*, 1–11.

Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, *24*, 485–500. https://doi.org/10.17705/1CAIS.02428

OpenCourseWare, M. (2020). *Milestones*. Retrieved January 3, 2020, from https://ocw.mit.edu/about/milestones/

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, *11*, 169–198.

Orji, F. A., Oyibo, K., Orji, R., Greer, J., & Vassileva, J. (2019). Personalization of persuasive technology in higher education. *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 336–340. https://doi.org/10.1145/3320435.3320478

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

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

Perry, N., & Winne, P. (2006). Learning from learning kits: Gstudy traces of students' self-regulated engagements with computerized content. https://doi.org/10.1007/s10648-006-9014-3

Pigeau, A., Aubert, O., & Prié, Y. (2019). Success prediction in moocs: A case study. *Educational Data Mining 2019*.

Pilli, O., & Admiraal, W. (2016). A taxonomy of massive open online courses. *Contemporary Educational Technology*, *7*(3), 223–240.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Academic Press. https://doi.org/https://doi.org/10.1016/B978-012109890-2/50043-3

Pitkänen, H. (2017). *Exploratory sequential data analysis of user interaction in contemporary bim applications* (Master's thesis). Aalto University.

Pivec, M. (2007). Editorial: Play and learn: Potentials of game-based learning. *British Journal of Educational Technology*, *38*(3), 387–393. https://doi.org/10.1111/j.1467-8535.2007.00722.x

Pohl, M., Wallner, G., & Kriglstein, S. (2016). Using lag-sequential analysis for understanding interaction sequences in visualizations. *International Journal of Human-Computer Studies*, *96*, 54–66.

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., & Xue, Y. (2016). Modeling and predicting learning behavior in moocs. *Proceedings of the ninth ACM international conference on web search and data mining*, 93–102.

Rabin, E., Kalman, Y. M., & Kalz, M. (2019). An empirical investigation of the antecedents of learner-centered outcome measures in moocs. *International Journal of Educational Technology in Higher Education*, *16*, 1–20.

Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77–93.

Rajeshkanna, A., & Arunesh, K. (2018). Role of decision tree classification in data mining. *International Journal of Pure and Applied Mathematics*, *119*(15), 2533–2543.

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013). Modeling learner engagement in moocs using probabilistic soft logic. *NIPS workshop on data driven education*, *21*, 62.

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2014). Learning latent engagement patterns of students in online courses. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1272–1278.

Riasati, M. J., Allahyar, N., & Tan, K.-E. (2012). Technology in language education: Benefits and barriers. *Journal of education and practice*, *3*(5), 25–30.

Richards, G. (2011). *Measuring engagement: Learning analytics in online learning*. Retrieved March 10, 2020, from https://www.academia.edu/779650/Measuring_Engagement_Learning_Analytics_in_Online_Learning

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, *32*, 77–112.

Saint, J., Gašević, D., Matcha, W., Uzir, N. A., & Pardo, A. (2020). Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 402–411. https://doi.org/10.1145/3375462.3375487

Saks, K., & Leijen, Ä. (2014). Distinguishing self-directed and self-regulated learning and measuring them in the e-learning context [International Conference on Education & Educational Psychology 2013 (ICEEPSY 2013)]. *Procedia - Social* *and Behavioral Sciences*, *112*, 190–198. https://doi.org/https://doi.org/10.1016/j.sbspro.2014.01.1155

Schumacher, C., & Ifenthaler, D. (2018). Features students really expect from learning analytics. *Computers in Human Behavior*, *78*, 397–407.

Schunk, D. H. (1989). Social cognitive theory and self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement* (pp. 83–110). Springer.

scikit-learn. (2020a). *1.10.5. tips on practical use*. Retrieved September 21, 2020, from https://scikit-learn.org/stable/modules/tree.html#tips-on-practical-use

scikit-learn. (2020b). *Sklearn.metrics.matthews_corrcoef*. Retrieved November 13, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html

Scratch. (2020). *About scratch*. Retrieved February 5, 2020, from https://scratch.mit.edu/about

Severance, C. (2012). Teaching the world: Daphne koller and coursera. *Computer*, *45*(8), 8–9.

Shah, D. (2019). *Coursera's 2019: Year in review*. Retrieved January 7, 2020, from https://www.classcentral.com/report/coursera-2019-year-review/

Smith, B., & Eng, M. (2013). Moocs: A learning journey two continuing education practitioners investigate and compare cmooc and xmooc learning models and experiences. In S. K. S. Cheung, J. Fong, W. Fing, F. L. Wang, & L. F. Kwok (Eds.), *Hybrid learning and continuing education*.

Soni, A. (2015). *eLearning Industry choosing the right elearning methods: Factors and elements*. Retrieved February 4, 2020, from https://elearningindustry.com/choosing-right-elearning-methods-factors-elements

Steiner, C., Kickmeier-Rust, M., & Albert, D. (2014). Learning analytics and educational data mining: An overview of recent techniques. *Learning analytics for and in serious games*, 6–15.

Top, E. (2012). Blogging as a social medium in undergraduate courses: Sense of community best predictor of perceived learning [Social Media in Higher Education]. *The Internet and Higher Education*, *15*(1), 24–28. https://doi.org/https://doi.org/10.1016/j.iheduc.2011.02.001

Tseng, S.-F., Chou, C.-Y., Chen, Z.-H., & Chao, P.-Y. (2014). Learning analytics: An enabler for dropout prediction. *Proceedings of the 22nd International Conference on Computers in Education*, 286–288.

U.S. Department of Education, O. o. E. T. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Retrieved February 25, 2020, from https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Ventista, O. M. (2018). Self-assessment in massive open online courses. *E-Learning and Digital Media*, *15*(4), 165–175. https://doi.org/10.1177/2042753018784950

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110. https://doi.org/https://doi.org/10.1016/j.chb.2018.07.027

Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online learning environments: A review of empirical research. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 524–533. https://doi.org/10.1145/3375462.3375483

Vitiello, M. (2017). *User prediction in moocs* (Master's thesis). Graz University of Technology.

Wen, M., Yang, D., & Rosé, C. (2014). Sentiment analysis in mooc discussion forums: What does it tell us?

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated engagement in learning. *Metacognition in educational theory and practice*, 277–304.

Wodzicki, K., Schwämmlein, E., & Moskaliuk, J. (2012). "actually, i wanted to learn": Study-related knowledge exchange on social networking sites [Social Media in Higher Education]. *The Internet and Higher Education*, *15*(1), 9–14. https://doi.org/https://doi.org/10.1016/j.iheduc.2011.05.008

Wong, J.-S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015). An analysis of mooc discussion forum interactions from the most active users. In N. Agarwal, K. Xu, & N. Osgood (Eds.), *Social computing, behavioral-cultural modeling, and prediction* (pp. 452–457). Springer International Publishing.

Xing, W., Tang, H., & Pei, B. (2019). Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of moocs. *The Internet and Higher Education*, *43*. https://doi.org/https://doi.org/10.1016/j.iheduc.2019.100690

Xiong, Y., Li, H., Kornhaber, M. L., Suen, H. K., Pursel, B., & Goins, D. D. (2015). Examining the relations among student motivation, engagement, and retention in a mooc: A structural equation modeling approach. *Global Education Review*, *2*(3), 23–33.

Yan, W., Dowell, N., Holman, C., Welsh, S. S., Choi, H., & Brooks, C. (2019). Exploring learner engagement patterns in teach-outs using topic, sentiment and on-topicness to reflect on pedagogy. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 180–184. https://doi.org/10.1145/3303772.3303836

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. *Proceedings of the 2013 NIPS Data-driven education workshop*, *11*, 14.

Yang, H.-H., & Su, C.-H. (2017). Learner behaviour in a mooc practice-oriented course: In empirical study integrating tam and tpb. *The International Review of Research in Open and Distributed Learning*, *18*(5), 36–63. https://doi.org/10.19173/irrodl.v18i5.2991

Yang, T.-C., & Chen, S. Y. (2020). Investigating students' online learning behavior with a learning analytic approach: Field dependence/independence vs. holism/serialism. *Interactive Learning Environments*, 1–19.

Zimmerman, B. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist - EDUC PSYCHOL*, *25*, 3–17. https://doi.org/10.1207/s15326985ep2501_2

Zimmerman, B., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, *82*, 51–59. https://doi.org/10.1037/0022-0663.82.1.51