DROPOUT PREDICTION WITH LEARNING ANALYTICS

Lappeenranta–Lahti University of Technology LUT

Master's Programme in Business Analytics, Master's thesis

2021

Juha Vehmas

Examiner(s): Professor Leonid Chechurin

Docent Jukka Korpela

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Science

Industrial Engineering and Management

Juha Vehmas

Dropout prediction with Learning Analytics

Master's thesis

2021

68 pages, 10 figures and 10 tables

Examiner(s): Professor Leonid Chechurin and Docent Jukka Korpela

Keywords: learning analytics, educational data mining, dropout prediction

Learning analytics is a growing field of research which focuses on analyzing data generated by digital learning methods to understand and optimize learning process. Digital learning has become more common as digitalization has moved forward and COVID-19 pandemic accelerated the move to online learning even further. The move to online learning has however raised the dropout rates.

In the thesis seven research areas of learning analytics are identified and presented. The research areas identified with literature review and confirmed with the use of LDA topic modelling. For dropout prediction in MOOC three machine learning models are built and evaluated. The models used are support vector machine, logistic regression, and random forest classifier. In addition, the prediction power of different data sources is evaluated with the help of literature and mutual information.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT Teknis-luonnontieteellinen

Tuotantotalous

Juha Vehmas

Opinnot keskeyttävien opiskelijoiden tunnistaminen oppimisanalytiikan avulla

Tuotantotalouden diplomityö

68 sivua, 10 kuvaa ja 10 taulukkoa

Tarkastajat: Professori Leonid Chechurin ja Dosentti Jukka Korpela

Avainsanat: oppimisanalytiikka, opetusdatan louhinta, opintojen keskeytysten ennustaminen

Oppimisanalytiikan on kasvava tutkimusalue, joka keskittyy analysoimaan dataa, jota digitaaliset opetusmenetelmät tuottavat. Analyysien tavoitteena on paremmin ymmärtää ja pyrkiä optimoimaan oppimisprosessia. Digitaaliset opetusmenetelmät ovat yleistyneet digitalisaation seurauksena. COVID-19-pandemia on pakottanut useat yliopistot ja koulut siirtymään etäopetukseen ja näin kiihdyttänyt verkon välityksellä tapahtuvan opetuksen kasvua. Internettiin siirtynyt opetus on kuitenkin kasvattanut opintonsa keskeyttäneiden määrää.

Työssä kirjallisuuskatsauksen avulla tunnistetaan seitsemän oppimisanalytiikan tutkimusaluetta ja esitellään ne. Kirjallisuuskatsauksessa löydettyjä tutkimusalueita verrataan LDA topic modelling koneoppimismenelmän avulla tunnistettuihin aiheisiin. Työn toisessa osassa pyritään ennustamaan massiivisen avoimen verkkokurssin keskeyttäviä opiskelijoita kolmen eri koneoppimismallin avulla. Käytetyt mallit ovat support vector machine, logistic regression ja random forest classifier. Lisäksi, eri datalähteiden hyödyllisyyttä arvioidaan kirjallisuuden ja kahden satunnaismuuttujan välisen informaation avulla.

Table of contents

Abstract

Figures

Tables

1. Introduction

Learning analytics is a growing field of research as different online learning opportunities keep emerging. It is influenced by many fields, most notably business intelligence, web analytics, educational data mining and recommender systems (Ferguson, 2012). Big data and online learning are major factors in the growth of learning analytics. The goal of learning analytics is utilizing the data generated by digital learning methods to achieve better results and to allocate resources effectively. Online learning has its pros and cons. Online courses are available to larger audiences and students have more freedom when and how they study. The freedom also means that instructors have fewer possibilities and less time to assist the students. The possibility to enroll with just a few clicks leads to high dropout rates. Another problem is that students can feel lonely without the connection to the teacher and other students. The limited connection also means that for teachers it is hard to notice students losing their motivation and to give the necessary support. Learning analytics aims to find solutions for these problems.

COVID-19 pandemic has affected schools and other educational institutions around the world forcing face-to-face classes to move online or to be cancelled. UNICEF (2020) reported that 90 percent of ministries of education put into practice some form of remote learning. It depended on the country how well the transition to remote methods went. For example, in Italy COVID-19 pandemic acted as a point of acceleration for digitalization of education as the education system was tightly built around "bricks-and-mortar" classrooms (Taglietti et al, 2021). Maity et al. (2021) found out in their study that during the COVID-19 pandemic in India the accessibility and the quality of teaching was higher in universities than in colleges and for school students it was even lower. In another study the possibilities and challenges of transforming courses to online teaching format during COVID-19 pandemic are discussed. Overall, the students gave positive feedback on the courses. The resulting use of digital tools can be seen as the new normal of future learning. Specific events will still be held face-to-face to increase learning success. (Voigt et al. 2021)

Online learning produces huge amounts of data that can be used to follow the learning process and give useful insights for both teachers and students. Finding the best methods to utilize this data is one of the main goals of learning analytics. There are multiple factors affecting the learning outcome and for learning analytics to cover all these there are a wide variety of methods researched. In this thesis the goal is to present these methods and specially to have a closer look into dropout prediction.

1.1 Objective of the thesis

The thesis consists of two main parts: literature review and practical dropout prediction task. In the literature review 50 articles on learning analytics are selected to find out the main topics of learning analytics. The topics are first formed as a result of subjective analysis and then compared to topics found with Latent Dirichlet allocation topic modelling. In the practical part the objective is to present the process of building machine learning models for dropout prediction. The process starts from the raw data and ends to the evaluation of the models. Only log data from a single MOOC course was utilized to build the models.

1.2 Research questions

The aim of this research is to give the reader an overview of the field of research called learning analytics, and to find a way to predict which students are at risk of dropping out of an online course. For the prediction to be as beneficial as possible for the course organizer it should happen in an early stage of the course, and it should be easy to execute and understand. There are three research question for this research:

1) What are the main topics of learning analytics?

Learning analytics is a broad field with many objectives and methods. The main goal is to improve both teaching and learning but there are many ways to achieve that. To answer this question learning analytics is divided into topics.

2) What data should be collected for dropout prediction?

Machine learning methods need data to first train the model and then to predict the outcome. To have good results the wanted outcomes should be separable with the features in the data. The goal is to find which features have the strongest predictive power when it comes to dropout prediction.

3) How to use the learning management system log data to predict dropouts?

The goal is to test which machine learning algorithms have the best performance utilizing only the log data. This can be measured by prediction accuracy and by how early after the start of the course the algorithm is able to predict accurately.

1.3 Structure of the thesis

The thesis consists of six chapters. The first chapter is the introduction in which the subject of the thesis is introduced as well as the objective and research questions.

In chapter 2 the background for the later chapters is outlined. The most important terms and topics related to learning analytics are explained to give the reader an idea of the landscape of the thesis. The aim is to make the rest of the thesis easier to understand for the reader.

Chapter 3 is the literature review of learning analytics. Aim of the chapter is to conduct literature review to find the research areas of learning analytics. Each of the found research areas are explained and the most common methods used in each are presented. In the end LDA topic modelling is used for validating the research areas found by subjective analysis.

Chapter 4 focuses to explain the goal and methods of dropout prediction in more detail. The chapter works as a background for chapter 5 in which the practical part is presented step by step. The classification models and the performance metrics for them are also explained in this chapter.

Chapter 5 goes through the building process of the three machine learning models. The process is divided into dataset introduction, data preprocessing, the model building, evaluating the results and comparing the models. The features are also compared to understand what data should be collected.

In chapter 6 the research questions are answered. The first research question’s answer is mainly based on chapter 3 while the second and the third question are answered by the findings in chapters 4 and 5.

2. Background

In this chapter the necessary background for learning analytics and the literature review is presented. The aim is to define the main terms and subjects related to learning analytics as well as give learning analytics a definition.

2.1 Digital education

Digital education is the use of digital tools and technologies to support the teaching process. Another term often used for digital education is Technology Enhanced Learning (TEL). Term digital learning is used when the use of digital tools is studied from the learning perspective (Kumar Basak et al., 2018). As the use of digital learning has been growing, many new research communities have emerged or shifted their interest to it. These include Artificial Intelligence in Education (AIED), Intelligent Tutoring Systems, Computer-Supported Collaborative Learning (CSCL), Learning Sciences, Learning analytics, Educational Data Mining and various MOOC-related communities (Dillenbourg, 2016). This can be perceived as a natural continuation to digitalization of businesses.

Dillenbourg (2016) identified the following six trends in digital education:

1. **More physical**: this is a bidirectional trend where physical objects or events enter the digital realm and digital objects are brought to the physical environment. For example, robotics can blend digital and physical worlds and augmented reality can bring digital objects to the classroom table**.**
2. **Less semantic**: instead of measuring correct/wrong answers the behavior patterns can be studied. Semantic information does not need to be excluded but it can be integrated with multiple levels of abstraction
3. **More social**: at the start of digital education learning was mostly thought of as an individual activity. It has become clear that social learning processes must be integrated with individual learning.

4. **Less design**: digital education allows teachers and students more freedom than before. There is no need to design strict predefined paths; learners can explore the learning environment freely. The challenge is to find the balance between freedom and design.
5. **More Open**: learning technologies have become more open in many ways. There are free to access courses, open-source platforms, everyone is open to contribute material and open architecture solutions.
6. **More teachers**: in the field of learning technologies formal learning has lost interest. Because of this the teachers' needs have not been addressed and focus has been on improving learning without an active instructor. To change this teacher must be listened to.

One of the challenges of digital education is the fact that designing courses and completing the courses digitally requires a certain level of technical proficiency. Especially in the transition phase from traditional classrooms to digital learning the teacher's technical abilities can limit the course structure. Having easy to use tools and systems for teaching is vital because any time used for setting up the learning environment and learning to use the environment is not used for learning the actual subject. (Nielsen, Miller & Hoban, 2014) It should be noted that present day students have grown in the digital age and often learn new systems fast. The use of digital tools prepares students for the future work environment which is more and more digitalized.

2.2 Online learning

Online learning is gaining popularity all over the world. Factors driving the online learning adoption are improving the access to learning, higher quality of learning and reducing costs. (Panigrahi, Srivastava & Sharma, 2018) As mentioned earlier COVID-19 pandemic forced many education institutes to adopt online learning but it was growing already before that. There are multiple benefits in online learning for all stakeholders. One of the biggest for the education providers is the easy scalability. Use of online education platforms offers education institutions an opportunity to reach new students in a cost-effective way

(Yashalova & Vasiltsov, 2020). Online learning made worldwide distance learning possible as now there are no limitations from where students can access the courses (Kaplan & Haenlein, 2016). For students online learning offers freedom but also demands stronger self-discipline. Shen et al. (2013) found strong correlation between self-efficacy and learner satisfaction in online learning.

The most popular format of online learning is Massive Open Online Courses (MOOCs). Another common format of online learning is Small Private Online Courses (SPOCs) which provide students with better instructions and support but lack the accessibility of MOOCs. The rise of these new educational formats is expected to reform business schools and other higher education institutions. (Kaplan & Haenlein, 2016) Most MOOCs consist of pre-recorded videos, quizzes with automatic checking, and discussion forums to create social interactions. The quality of education in MOOCs has high variation and there is not an established MOOC business model. MOOCs are easy to enroll which contributes to the freedom that the students have with online learning. The dropout rates in MOOCs are often as high as 80-90 %. The reasons for dropping out include a lack of incentive for completion, failure to comprehend the content material and lack of support. (Hew & Cheung, 2014) MOOCs can be made more engaging by having game-based elements, interactive content, immediate feedback, guiding students to pick courses with correct difficulty level, links to advanced material, and real-world challenges and use cases (de Freitas, Morgan & Gibson, 2015). The challenges of online learning are further discussed later in the thesis, specifically the dropout problem in MOOCs.

2.3 Blended Learning

The common definition for blended learning is the combination of traditional face-to-face teaching and online learning methods (Graham, 2013). In higher education blended learning is often implanted with the goal of offering flexibility in time and place. To achieve the full potential of blended learning the teachers must understand the needs of the students. Using differentiated instructions for student groups provide best learning outcomes. (Boehlens, Voet & De Wever, 2018) In a study by Dziuban et al. (2018) students ranked the blended learning environment as the most effective way of learning. Blended learning is nowadays the normal method in many universities and schools. It provides students with better support than online learning which is important especially with younger students who lack self-discipline.

2.4 Learning analytics

The 1st International Conference on Learning Analytics and Knowledge was held in 2011 in Banff, Alberta, Canada. The motivation to establish a dedicated forum for learning analytics is described by three indicators:

1. The ability of organizations to utilize data does not keep up with the growth of data. This is especially pronounced in relation to knowledge, teaching and learning.
2. Learning institutions and corporations ignore most of the data the learners leave behind in the process of accessing learning materials, interacting with teachers and other learners, and creating new content.
3. Educational institutions are under growing pressure to lower costs and increase efficiency. Analytics can provide important insights on how to view and plan for change at course and institutional level.

The organizers of the conference presented the following definition of learning analytics: "Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs." (LAK '11, 2011) This definition is widely accepted and used in literature (Siemens, 2013). It is noteworthy that this definition does not limit learning analytics to automatically conducted data analysis (Chatti, Dyckhoff, Schroeder & Thüs, 2012). Both the definition and the motivation of learning analytics allows a wide variety of methods to be used to achieve the wanted goals. Learning analytics is a field of research where the expertise of professionals from many different fields can be combined.

Learning analytics focuses on Technology-Enhanced Learning (TEL) research. It is closely connected to and inspired by many fields of research. The most influential of these are business intelligence, web analytics, educational data mining and recommender systems. Having strong connections to these fields means that researchers have approached learning analytics from different perspectives which have led to learning to have multiple goals and ways to achieve these goals. (Ferguson, 2012) Generally, the goal of learning analytics is to utilize educational data to improve and support the learning process by developing suitable methods. The process of learning analytics is often an iterative cycle. The cycle can be divided in three major parts:

1. **Data collection and preprocessing**: the basis of the learning analytics process is educational data. The data is collected from various learning environments and systems. Most of the time the data includes attributes which are not useful, or it can be too large to be directly analyzed. In these cases, the data must be pre-processed to a suitable format which can then be used as an input for a learning analytics method.

2. **Analytics and action**: Different learning analytics techniques can be applied to the pre-processed data to understand hidden patterns in the data. These patterns can help to provide better a learning experience. The choice of method depends on the data and on the objective of the analysis. The main point of this step is to take action which is justified by the analysis.

3. **Post-processing**: this an important step for the continuity of the improvement. It includes collecting new data from new sources, refining the dataset, selecting new attributes for a new iteration, identifying new metrics, and possibly choosing a new analytics method. (Chatti et al., 2012)

2.5 Educational data mining

Education data mining (EDM) is closely related to learning analytics and the objectives and methods of the two overlap on many occasions. Both fields of research require similar data and researchers share similar skill sets. However, the communities of these fields have grown separately to answer the same questions. One key difference between the communities is that EDM focuses more on automated discovery while learning analytics takes an approach which leverages human judgement through visualization and other methods. (Siemens & Baker, 2012)

EDM utilizes statistical, machine learning, and data mining algorithms over educational data. The main goal of EDM is to analyze the data to answer education research questions. The EDM process converts the raw data from learning management systems into information that can potentially have help to improve educational research or teaching methods. The process is similar to data mining processes in other application areas (i.e., business, genetics, medicine, etc.,) (Romero & Ventura, 2010) The process is also similar to the learning analytics process explained earlier. The main difference is in the methods used even though there is overlap in them. EDM is more focused on automating the process from data to information, and in learning analytics it is more common to have human input in the analysis.

Learning analytics and EDM share many objectives, methods, and processes. In this thesis term learning analytics is used from this point onward to cover both of them.

2.6 Learning analytics stakeholders

Who benefits from the results of the learning analytics? The easy answer is everyone who wants to learn or teach something. However, it is important to identify different stakeholders as the target audience affects how the research problem is constructed. The evident groups are the students and the teachers but there are many others as well. Ferguson (2012) mentions three groups which benefit from learning analytics: governments, educational institutions, and teachers/learners. Chatti et al. (2012) divides the stakeholders into students, teachers, tutors/mentors, educational institutions, researchers, and system designers. Similar division is done by Romero & Ventura (2010) who identify students, teachers, educational researchers, learning providers, and administrators as the stakeholders. The stakeholders and the objectives related to them are summarized in Table 1.

Table 1 Learning analytics stakeholders

Stakeholders	Objective of learning analytics
Students	• Recommendations on learning activities • Different learning paths • Adaptive hints • Interesting discussions • Visualization of the learning process
Teachers	• Feedback about instructions • Analysis of the student's progress and behavior • Identifying students who need support • Predicting student performance • Clustering students by different attributes • Finding common mistakes • Improving courses with more effective activities and customization
Researchers	• Evaluation of learning management system • Evaluation of course content • Automatic construction of student models • Comparison of data mining tools
Education provider	• Assist decision making • Finding cost-effective ways to improve courses • Lower the dropout rates • Help in student selection
Administrator/government	• Help in resource allocation (money, human and material) • Improving educational programs • Determining the efficiency of online (distance) learning • Evaluation of teachers and programs

It can be said that the main stakeholders of learning analytics are students and teachers. Most of the objectives of learning analytics are aimed towards assisting the teachers. Students also closely benefit from many of the objectives aimed for teachers. Researchers are a special group as they enable learning analytics methods for the other groups. For education providers and administrators/government the objectives share similarities as they focused on decision making and resource allocation. There is also a separate field of research which focuses on these two stakeholders called academic analytics.

3. Literature review of learning analytics

The chapter consists of literature review conducted on learning analytics. At the start of the chapter the process of selecting the articles to include in the review is explained. Then the selected pool of articles is analyzed with Scopus statistics. Seven topics are identified from the articles and each article is clustered in one of the topics. The main objectives and methods of each topic are then presented. In the end the identified topics are compared to topics found with LDA topic modelling.

3.1 The selection of articles

To find relevant articles for the literature review Scopus database was used. The selection process is illustrated in Figure 1. As discussed earlier, learning analytics and educational data mining are fields of research with similar objectives and stakeholders and there is significant amount of overlap in the research. For this reason, the keywords selected for the search are “learning analytics” and “educational data mining”. The publication years were searched from 2010 to 2020. To include only studies with strong academic background the search was limited to articles. Finally, the language of the article was filtered to English. With these search parameters (Table 2) 1,904 articles were found from the Scopus database. To narrow down the list of articles they were ordered by citations, and literature reviews and articles defining learning analytics framework were discarded from the pool. After this the 50 most cited articles were picked for a closer analysis which results are presented later in this chapter. The cut point in the articles was at 75 citations while the most cited by paper in the selected pool has 313 citations.

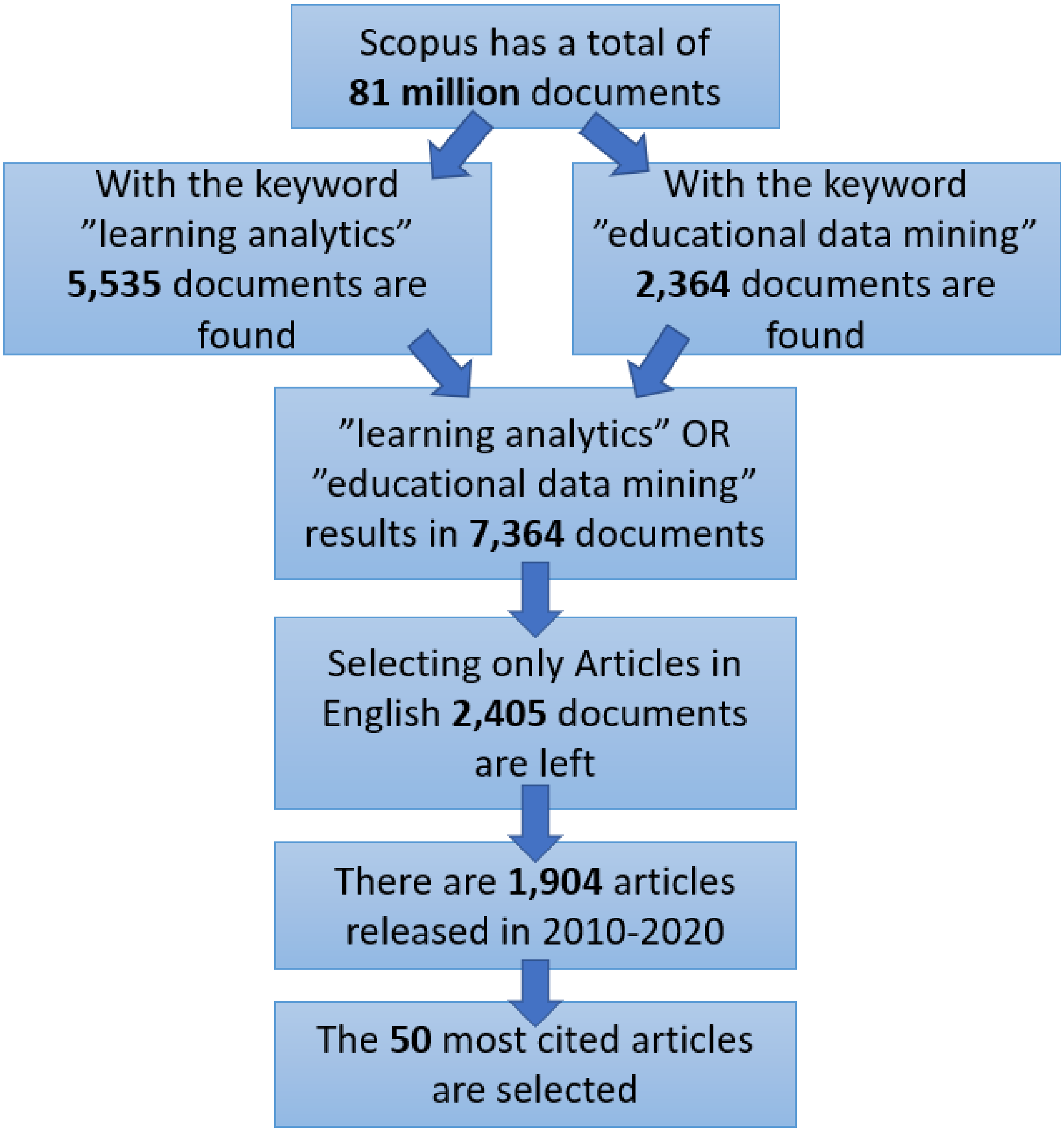

Scopus has a total of 81 million documents
With the keyword ”learning analytics” 5,535 documents are found
With the keyword ”educational data mining” 2,364 documents are found
”learning analytics” OR ”educational data mining” results in 7,364 documents
Selecting only Articles in English 2,405 documents are left
There are 1,904 articles released in 2010-2020
The 50 most cited articles are selected

Figure 1 The article selection process for the literature review

Table 2 Search parameters

Keywords	“learning analytics” OR “educational data mining”
Search in	Article title, Abstract, Keywords
Years	2010 - 2020
Document type	Article
Language	English

Figure 2 shows that the interest in learning analytics is growing fast. The first document found in Scopus by just searching "learning analytics" is from 2004. Berk (2004) defined learning analytics as "the set of activities an organization does that helps it understand how to better train and develop employees and customers". The definition is narrower than the one presented in chapter 2.4 but the core idea is the same. However, the research started to gain traction in 2010 and the first paper in this period is from Bach (2010) in which a conceptual framework for the development of learning analytics is outlined. Educational data mining emerged a couple of years earlier than learning analytics but similarly the growth started in 2010. The number of publications is led by the United States with 401 articles followed by Spain (239 articles), Australia (179 articles) and the United Kingdom (176 articles). The subject areas for the articles are dominated by Social Sciences and Computer Science with 1,243 and 1,129 articles respectively. In other words, 62.8 % of the articles are categorized in these subjects. The other subjects with over two percentage shares in the field are in order: Engineering, Psychology, Mathematics, Arts and Humanities, and Business, Management and Accounting. The journal with the most publications is *Computers In Human Behavior* as they have 72 articles published. On the second place *British Journal Of Education Technology* has 50 articles and the third place is divided by three journals with 44 articles: *Computers And Education*, *Interactive Learning Environments*, and *International Journal of Emerging Technologies In Learning*.

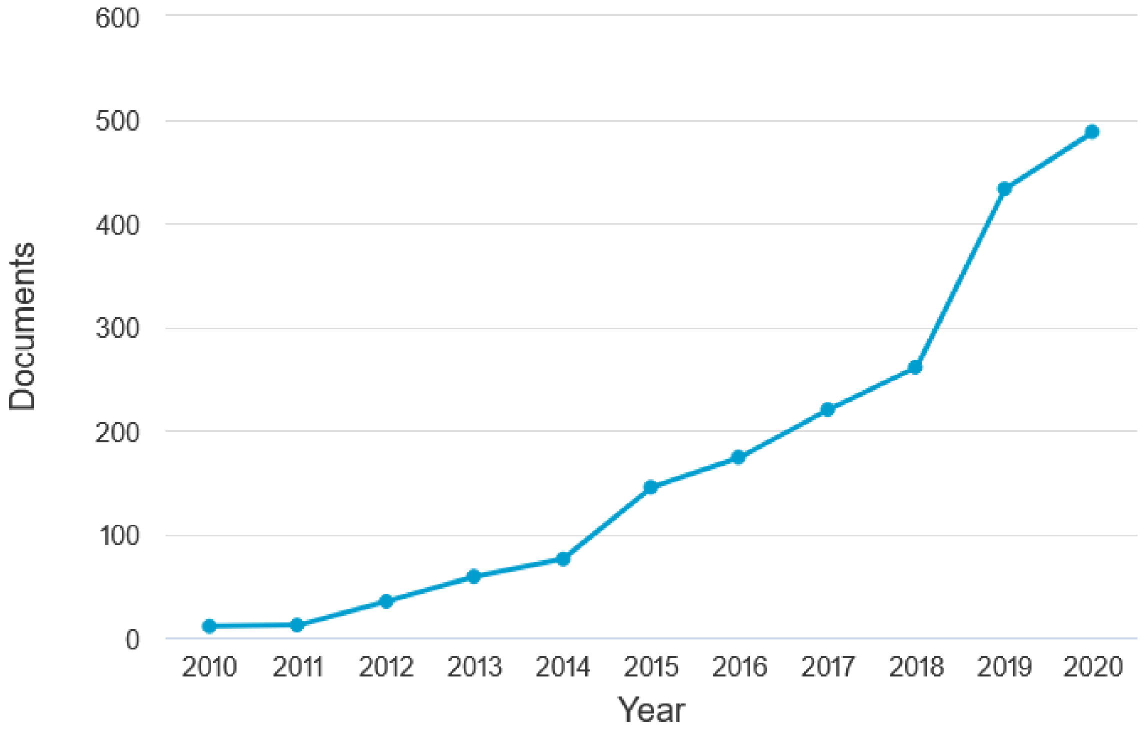


Figure 2 Number of articles found with the search parameters in years 2010 to 2020

When analyzing the pool of selected 50 articles the same countries are at the top and the shares of subject areas are similar. Two journals, *Computers in Human Behavior* and *Computers and education*, have a major number of articles in the selection as they have ten and eight articles respectively. The selection is spread between years from 2012 to 2017 and one article from 2010. The most common year of publication is 2013 with 13 articles. Six authors (Figure 3) have three or more articles in the selection. In total there are 143 authors presented in the selected pool of articles.

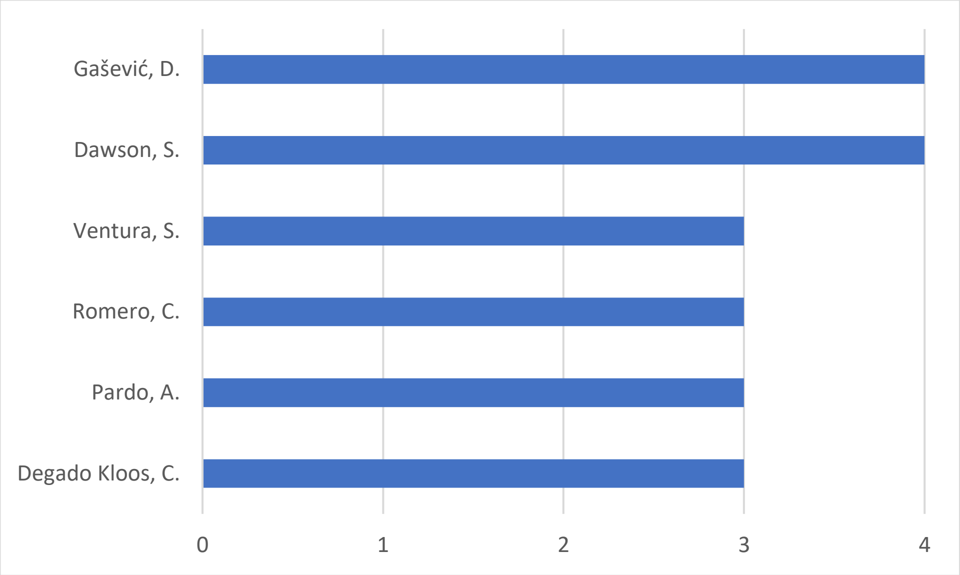


Figure 3 Authors who most publications in the selected pool

The analysis of search results shows that learning analytics is a growing field with many researchers around the globe. The shares of different subject areas confirm the nature of learning analytics as a mix of technical analysis and social sciences. In the next section the selected pool of articles is analyzed more closely.

3.2 Learning analytics research areas

To achieve a better understanding of the learning analytics landscape the selected pool of articles was divided into topics. The basis for these topics was adopted from the background presented in Chapter 2. During the clustering process the topics were finalized to seven distinct research areas. There is inevitable overlap in these topics and almost every article has some attributes from many topics. Despite this every article is only categorized to one topic to give a clear picture of the popularity of the research trends. The topics and the descriptions of them from the most common to least common one are:

1) **Performance prediction**: the goal is to predict student's grade and find out which variables have the biggest impact on the grade

2) **Dropout prediction**: the goal is to predict if a student will pass or fail the course and identify as early as possible students that need extra support

3) **Course design**: creation and development of a course to better support the needs of the students and teachers by utilizing data
4) **Learning strategy**: finding different strategies and identifying which strategies lead to best results
5) **Learning visualization**: group of methods to visualize learning strategies and progression for both students and teachers
6) **Social network analysis**: analyzing the way students interact with the course, teacher, and other students
7) **Ethical issues**: problems that come with the use data that is generated by students

Performance and dropout prediction are the most common topics. These two share many similarities and the main difference is in the goal of the prediction. The two prediction topics have a combined number of 24 articles which represent almost half of the selected pool. The least common topic, Ethical issues, is more important than its position might suggest. Ethical issues and data privacy concerns are always there when data is handled. Many of the articles discuss this topic but only two have it as a focus of the study.

Table 3 The articles clustered into topics and the indicators of each topic

Topic	Indicators
Performance prediction (16) (Gašević et al., 2016), (Tempelaar et al., 2015), (Romero et al., 2013), (Xing et al., 2015), (Dietz-Uhler & Hurn, 2013), (Asif et al., 2017), (Kabakchieva, 2013), (de Barba et al., 2016), (You, 2016), (Kotsiantis et al., 2010), (Zacharis, 2015), (Romero-Zaldivar et al., 2012), (Abdous et al., 2012), (Conjin et al., 2017), (Ifenthaler & Widanapathirana, 2014), (Kotsiantis, 2012)	• Predicting the grades • Identifying groups of students based on performance • Allocating resources effectively • Regression analysis
Dropout prediction (8) (Xing et al., 2016), (Márquez-Vera et al., 2013), (Costa et al., 2017), (Pursel et al., 2016), (Márquez-Vera et al., 2016), (Marbouti et al., 2016), (Lara et al., 2014), (Natek & Zwilling, 2014)	• Predicting if a student will complete the course or not • Understanding early signs • Giving support when needed • Binary classification
Course design (8) (Lockyer et al., 2013), (Macfadyen & Dawson, 2012), (Rienties & Toetenel, 2016), (Mor et al., 2015), (Gobert et al., 2013), (Scheffel et al., 2014), (Dyckhoff et al., 2012), (Ali et al., 2012)	• Finding effective teaching methods • Guiding students • Monitoring that the course works as intended • Combining pedagogical intent and data
Learning strategy (6) (Kizilcec et al., 2017), (Jovanović et al., 2017), (Berland et al., 2013), (Blikstein et al., 2014), (Tabuenca et al., 2015), (Cerezo et al., 2016)	• Clustering students by their actions • Following students' learning paths • Learning from the best performing students
Learning visualization (5) (Verbert et al., 2013), (Verbert et al., 2014), (Ruipérez-Valiente et al., 2015), (Muños-Merino et al., 2015), (Park & Jo, 2015)	• Giving teachers an overview of the course progress • Providing students visual feedback on their learning
Social learning analytics (5) (Shum & Ferguson, 2012), (Agudo-Peregrina et al., 2014), (He, 2013), (Gašević et al., 2013), (Fidalgo-Blanco et al., 2015)	• Understanding how students interact with each other and with the teachers • Creating sense of community
Ethical issues (2) (Slade & Prinsloo, 2013), (Pardo & Siemens, 2014)	• Privacy of the students • Data management

The articles in each topic, and the main indicators of the topics are shown in Table 3. The next chapters explain each topic and the research of the analyzed studies on a deeper level.

3.2.1 Performance prediction

Log data from learning management systems is a main data source for learning analytics and it can be utilized as a predictor for performance prediction. Using several data sources in performance prediction is advised to get both timely and predictive feedback. (Tempelaar et al., 2015) You (2016) found significant potential in using log data in the middle of an online course to predict student’s performance. However, it is not clear that early support guarantees improved results (Conjin et al., 2017). It is possible to build tools which allow not only data mining experts but also less experienced users to utilize the log data (Romero et al., 2013). The log data has been used by many researchers and the findings have been diverse which possibly is related to diversity in courses and on how the log data is processed into features (Conjin et al., 2017). In addition to log data, also demographic and academic data, admission/registration info and data gathered with surveys can be beneficial for performance prediction models (Kotsiantis, 2012). The use of external tools which do not leave traces in the log data should be noted when the performance prediction results are interpreted (Romero-Zaldivar et al. 2012).

In addition, for the model to be accurate it is also important that the model is comprehensible so that it is easier for the teachers to use it for decision making (Romero et al., 2013) Many prediction models are hard to understand for the teachers. This causes problems in the use of the models (e.g., personalizing education and intervention). Reducing data dimensionality and systematically contextualizing data in a semantic background can be used to create models which are easier to interpret. In the study Genetic Programming algorithm outperformed traditional models and it had better interpretability. (Xing et al., 2015)

Gašević et al. (2016) studied how instructional conditions influence the prediction of academic success. Course specific models were found to be more accurate than generalized models because the differences in technology usage affect the data. Ignoring the structure of the course can lead to over or under estimation of the effects the data has for students’ performance. (Gašević et al., 2016) Conjin et al. (2017) also found the portability of the prediction models between courses to be low. Educational data for learning analytics is context specific and the same variables can have different meanings across educational institutions and research areas (Ifenthaler & Widanapathirana, 2014). These findings indicate that the models must be built specifically for each course or study program.

Asif et al. (2017) found that on a four-year study program focusing on a few courses which are indicators of good or poor performance it is possible to provide timely support for low achieving students and give advice on new opportunities to high performing students. Kabakchieva (2013) researched students' performance at a Bulgarian university and found that university admission score and number of failures at the first-year university exams were the most influential factors in the predictions. Computer-assisted formative assessment was detected as the best predictor for underperforming students by Tempelaar et al. (2015). In blended learning courses forum usage, content creation, quiz efforts and number of files viewed were found as the most influential features (Zacharis, 2015). You (2016) identified regular studying, assignments submitted late, number of logins, and proof of reading the course material as significant predictors for performance in online courses. In their study Abdous et al. (2012) did not find students' forum usage or login times to have correlation with students' performance. It is important to remember that the scope of the research and data used have a major impact on which features work well. Overall, it can be summarized that active participation often leads to better performance and especially in online settings self-regulated learning skills are important.

Motivation is a strong predictor on how well a student performs in a MOOC. Motivation influences students' participation on the course which can be measured with log data. Motivational assessment at the early stage in the course can be leveraged for performance prediction. (de Barba et al., 2016) Even though motivation sounds like an obvious part of a student's performance measuring it is not easy and it is often ignored. Performance prediction can also be used to motivate students as providing proof on how the student's behavior affects their performance can work as a motivator (Dietz-Uhler & Hurn 2013).

3.2.2 Dropout prediction

High failure rates in introductory courses have alarmed many educators. To combat this problem a prediction model can be built to identify students who are at risk of failure. (Costa et al., 2017) Similarly MOOCs have recently raised concern in educator because of high dropout rates. The problem is often ignored as it can be described as a scale-efficacy tradeoff. MOOCs however generate huge amounts of data which can be utilized for the dropout prediction. (Xing et al., 2016) Dropout prediction can also help universities and other academic institutions to reduce the number of dropouts (Natek & Zwilling 2014). Predicting a student's possibility of failure or dropout is a similar task and in many cases failure and dropout are the same thing. Dropout prediction shares many qualities with performance prediction, but the main differences are that dropout prediction is bivariate classification tasks (students either drop out or do not) and in dropout prediction the focus is solely on the poorly performing students. The use of data sources and some of the variables with high prediction power are shared between dropout and performance prediction.

Identifying a small subset of at-risk students helps teachers to focus their support on the students who need it most (Xing et al., 2016). In dropout prediction it is most of the time best to minimize the number of at-risk students wrongly classified as students who will pass the course because giving unneeded support does not cause harm but ignoring students in need of help does (Marbouti et al., 2016). The dropout prediction should be carried out as early as possible because the more time the teacher and the students have for reacting to the alert the better (Márquez-Vera et al., 2013). Dropout prediction is further discussed in chapter 4.

3.2.3 Course design

Learning analytics can provide a wide variety of tools for teachers to improve the effectiveness of their courses step by step (Dyckhoff et al., 2012). LOCO-Analyst is a tool which provides teachers feedback on students' learning process and performance. On the development of the tool, it came clear that the user experience of the tool was important for the teachers. Enhancement of the tool's data visualization, user interface and supported feedback types helped teachers to interpret the results of learning analytics methods. (Ali et al., 2012) Dyckhoff et al. (2012) develop a learning analytics toolkit eLAT to process large datasets for teachers based on their individual interests and to take care of data privacy issues. With the toolkit teachers can evaluate their own technology-enhanced teaching methods to identify possible improvements. Creation of such tools allows teachers with limited technological knowledge to access learning analytics to improve their courses and their students' learning experience. Scheffel et al. (2014) introduced a five-dimensional framework for the evaluation of learning analytics tools. The five criteria and quality indicators are:

1) **Objectives** (Awareness, Reflection, Motivation, Behavioral Change)
2) **Learning Support** (Perceived Usefulness, Recommendation, Activity Classification, Detection of Students at Risk)
3) **Learning Measures and Output** (Comparability, Effectiveness, Efficiency, Helpfulness)
4) **Data Aspects** (Transparency, Data Standards, Data Ownership, Privacy)
5) **Organizational Aspects** (Availability, Implementation, Training of Educational, Stakeholders, Organizational Change)

For learning analytics to grow from small-scale practice to broad scale applicability, there is a need for a contextual framework which helps teachers to understand the results provided by the analytics. The study proposes learning design as a form of documentation of pedagogical intent that provides the context needed for making sense of the data analysis. Learning design includes resources which students can access, the tasks students are expected to complete, support mechanisms the teacher can use and checkpoints where the analytics can be applied. (Lockyer et al., 2013) Mor et al. (2015) suggested combining learning design, teacher inquiry and learning analytics to form a circle in which teacher defines meaningful questions to analyze then learning analytics provides possible improvement for learning design and the circle repeats. In their current state analysis of LMS use in a large research-intensive university Macdafyen & Dawson (2012) noticed that the lack of contextualization in the use of learning analytics led to the institution not gaining useful knowledge from the analyzes. Rientes & Toetenel (2016) combined learning design with dropout prediction in their study and found out that the primary predictor of academic retention was the amount of communication activities.

3.2.4 Learning strategy

Leaning analytics can be used to identify learning strategies in online and blended learning environments. Cerezo et al. (2016) identified four clusters in their study:

1) Cluster 1 – Non-Task or Theory Oriented Group (non-procrastinators)
2) Cluster 2 –Task Oriented Group (socially focused)
3) Cluster 3 – Task Oriented Group (individually focused)
4) Cluster 4 – Non-Task Oriented Group (procrastinators)

The biggest in the final marks of different clusters were between clusters 1 and 4. Procrastination clearly led to lower marks. (Cerezo et al., 2016)

Jovanović et al. (2017) examined log data in a flipped classroom to identify four learning strategies:

1) Cluster 1 (12.79 %): In the smallest cluster the actions of the students are focused on formative assessment and summative assessment actions are almost absent. Use of reading materials is not frequent.
2) Cluster 2 (41.85 %): In the biggest cluster the students had a trial-and-error learning approach and they focused on summative assessment. After exercises students tend to self-reflect.
3) Cluster 3 (28.63 %): The students focus on reading materials, course videos, and on some formative assessment tasks. The pattern indicates passive consumption of the given materials.

4) Cluster 4 (16.73 %): In these sessions the students mainly watch videos, then do the formative assessment tasks related to them and finally try the exercises.

In a study of self-regulated learning strategies in MOOCs goal setting and strategic planning were found as the best performing strategies. Help seeking was the weakest strategy in the study. The other strategies in the study from best to worst were self-evaluation, task strategies and elaboration. By identifying the students using weaker strategies it is possible to target support and advice for those students. (Kizilcec et al., 2017) Tabuenca et al. (2015) found that instructing students to track their time used in online courses positively affects the time management skill of the students and leads to the students using more effective learning strategies.

3.2.5 Learning visualization

Learning analytics dashboards can improve learning by giving teachers a better overview of the course, helping teachers to reflect on their teaching methods and by finding students who lack support. The dashboards can be utilized in face-to-face teaching, online learning as well as in blended learning settings. (Verbert et al., 2013) Having the dashboards scale from mobile use to larger desktop use cases enables great user experience. Visualizing traces in log data can help both teacher and students to have better awareness of the learning process and to reflect on the process. (Verbert et al., 2014)

ALAS-KA is a tool for Khan Academy platform which extends the learning analytics features already implemented on the platform. The tool includes more than 20 new indicators and new visualization for the entire class and for individual students. The tool helps teachers to make decisions supported by the information provided by the tool and allows students to have access to information which they can use for self-reflection. It also detects class tendencies and learner models. (Ruipérez-Valiente et al., 2015) The Learning Analytics Dashboard (LAD) visualizes students' online behavior patterns in a learning management system by mining the log data. While the newly developed tool did not significantly improve students’ learning results it was clear that its visualizations helped students to understand their learning process. For future development it is important that the visualizations are easy to interpret. (Park & Jo, 2015)

3.2.6 Social learning analytics

Shum & Ferguson (2012) list three challenged which social learning analytics offers for technology-enhanced learning research:

1) Educational landscape is changing constantly as new technologies are adopted. Online social learning is emerging as a significant part of research because online learning gets more and more traction.
2) Understanding the possibilities of different types of social learning analytics. Some learning methods are inherently social while some can be socialized.
3) Implementing analytics that satisfy concerns about the limitations and abuses of analytics.

The social network is built between students, teachers, and learning resources. As the data used is often the log data of learning management systems it is most of the time noisy. (Shum & Ferguson, 2012) The interactions in the network can be utilized as features in the prediction tasks (Agudo-Peregrina et al., 2014). By studying students’ online questions and chat messages valuable insights of the learning behavior can be found. The number of online questions students asked and students’ final grades were correlated. (He, 2013)

The social capital students accumulate during their studies is positively associated with their academic performance. Students with more social capital have significantly higher grades. The study implicates that degree programs should take into account the possibility for the students to build social capital during their studies. Data about cross-class networks can be used to support study planning in software systems. (Gašević et al., 2013)

3.2.7 Ethical issues and privacy concerns

When students interact with learning management systems, they generate highly sensitive data. Learning analytics uses this data to understand and improve the quality of learning experience. However, privacy and ethical issues should be considered when handling the data. To deal with these issues privacy principles are needed. (Pardo & Siemens, 2014). Slade and Prinsloo (2013) highlight the role of power, the impact of surveillance, and the need for transparency. They grouped the ethical issues in three broad categories:

1) The location and interpretation of data
2) Informed consent, privacy, and the deidentification of data
3) The management, classification, and storage of data

To build an ethical framework for learning analytics six principles are proposed. (Slade and Prinsloo, 2013)

1) Learning analytics should provide pointers for what is appropriate and morally necessary
2) Students should be thought of as collaborators instead of as sources of data
3) Students are evolving and that should be considered in the data collection and analysis
4) Student success is a complex and multidimensional phenomenon
5) It should be transparent what data is collected and who can access the data
6) Data is too valuable not to be used

3.3 Latent Dirichlet allocation (LDA) topic modelling

Topic models are algorithms which can automatically discover main themes from a large collection of documents. Latent Dirichlet allocation (LDA) is a common topic model. (Blei, 2012) For the LDA topic modelling the abstracts of 500 hundred most cited papers were exported from Scopus. Scikit-learn LDA Python library was used to perform the topic modelling. The algorithm takes a predetermined number of clusters as a parameter. To see how the number of clusters affect the results the modelling was performed multiple times with different numbers of clusters each time (Table 4).

Table 4 LDA topic modelling results

Clusters	*5*	*6*	*7*	*8*	*10*	*12*
1.	learning analytics	prediction	learning analytics	social learning analytics	performance prediction	social learning analytics
2.	prediction	visualization	social learning analytics	learning analytics	learning analytics	data mining
3.	data mining	learning analytics	educational data mining	prediction	visualization	prediction
4.	course design	social learning analytics	prediction	visualization	social learning analytics	visualization
5.			course design	mobile learning	data mining	Challenges/issues
6.						course platform
7.						digital learning/teaching

With every number of clusters, the smallest clusters were too small to form a clear topic, so they were discarded. The results show similar topics as the manually done literature review. Learning analytics and data mining are common topics found which is not surprising as both are visible in the search words. For this reason, they are not picked as topics in the manually done clustering. Prediction, visualization, and social learning analytics are common topics found with LDA and those are all also found manually. Prediction, which is the biggest cluster in the manual clustering, is also the only topic found with every number of clusters tested in LDA. Course design is another topic that comes up in both LDA (number of clusters 5 and 7) and subjective clustering. Learning strategy and ethical issues are the only topics found manually but not with LDA topics modelling. Learning analytics is a big topic in many of the LDA results which is a problem because the topics we would like to extract are clustered in this big topic we already knew to exist.

4. Methods for dropout prediction

This chapter focuses on explaining dropout prediction further and especially the methods used for dropout prediction. Earlier research results are examined to find out which methods are proved to work well and if there are some methods that need further research. The different sources for dropout prediction are discussed and the prediction power of features used in research are summarized. The methods used in the practical part of the thesis are explained in this chapter.

4.1 Benefits of dropout prediction

Web-based courses have higher dropout rates than traditional education courses. For universities, policymakers, higher education funding bodies and educator’s student retention rate is a measurement of the quality that an educational institute offers. This emphasis on retention and high dropout rates of e-learning courses makes the reduction of dropout rates an important task for online courses. Identifying at-risk students is a vital part of this task as it will help the instructors to provide better support for those who need it the most. (Lykourentzou et al., 2009) Machine learning algorithms and more specifically classification algorithms are used for detecting at-risk students. Dropout prediction is a binary classification problem as there are only two outcomes: the student either drops out or not. It is important to minimize the number of false negative errors and at the same time keep the number of false positives low (Marbouti et al., 2016).

MOOC platforms provide low level student behavioral trace data which opens opportunities for learning analytics and educational data mining methods to be utilized for identifying students at risk of dropping out. Effective prediction models must be able to detect at-risk students as early as possible. (Xing et al., 2016) The data available on the MOOC platforms is often referred to as log data. The benefit of log data is that it is always available because it automatically collects data whenever the student uses the platform. The drawback is that on bigger courses there will be a lot of data to go through and not all the data is useful for the dropout prediction task. Use of performance factors (i.e., grades) generated during the course or semester is proved to be beneficial for dropout prediction (Marbouti et al., 2016).

Dropout prediction is a challenging problem for multiple reasons. First, students have different levels of knowledge before the course. This means that a student who does not interact with the course often might be underperforming or be already familiar with the topic. This means that the data for dropout prediction is often noisy. Second, MOOC platforms log a lot of student activities but only a few of them might be important for the prediction. Third, the dropout rate is often very high (60-80%) which means that there is a lot more students who dropout than those who complete the course. This makes the data imbalanced. (Fei & Yeung, 2015) Because of imbalanced data a model can have a high accuracy but fail to identify the dropouts (Marbouti et al., 2016).

Based on the literature presented in chapter 3 and in this chapter the commonly used generalized features are collected in Table 5. The features are ones that are used in multiple studies so that the performance can be measured from multiple sources. The prediction power of these features is evaluated according to the results of the studies. The scale of evaluation is from 1 (low) to 5 (high). Features that require active participation from the student tend to have high prediction power. The use of passive learning materials such as video lectures have medium predictive power which can be explained by the fact that the student might not be focusing on the video even though they are playing it on their computer. Pre-course surveys measuring the motivation of the student are common in many institutions and those can be of use in prediction tasks. The problems of surveys are unstructured data and students might just answer quickly without properly thinking of the correct answer which creates noise in the data. Forum activity is a feature which has lots of variance in the prediction power. If the use of a forum is encouraged or even awarded the data generated from the forum can be really useful, but many courses have a forum just in case and it has little use and, in these courses, no useful data can be generated from the forums. General information (e.g., age, gender) in most cases was not useful for the prediction and it also increases privacy concerns and in some legislations the use of them is even prohibited.

Table 5 Prediction power of generalised features

Data	Prediction power (1-5)	Notes
Mid-term exams	5	
Exercises completed	5	
Earlier study performance	4	Depends on how closely related the studies are to the course
Motivation level	3	Pre-course survey is needed
Video view time	3	
Forum activity	3	Prediction power is higher when the use of discussion forum is encouraged
Text material downloads	3	
Registration date	2	
Gender	1	Privacy concerns
Age	1	Privacy concerns

There is no universal definition for dropout and different research groups have used different definitions in their research. When research is done on an institutional level, it can be considered as a dropout when an institution loses a student in whatever way (Márquez-Vera et al., 2016). For studies focusing on the dropout prediction on the course level there are more definitions for dropout and especially for the point of time at which the student is considered as a dropout. One common way is to classify all students who did not pass the course as dropouts (Marbouti et al., 2016). Fei & Yeung (2015) considered three different definitions for dropout:

- **Participation in the final week**: whether a student will stay to the end of the course
- **Last week of engagement**: whether the current week is the last week the student is active
- **Participation in the next week**: whether a student is active in the coming week

4.2 Scope of dropout prediction research

Dropout prediction can have different scopes of study as the research can focus on high school dropouts (Lykourentzou et al., 2009) or dropouts on individual courses on MOOC platforms (Xing et al., 2016). Even though the scope is different there are a lot of similarities between these tasks. The data available is often similar and the behavior of the student at-risk of dropping out has the same tendencies regardless of the scope. When the study focuses on high school dropouts there is often more historical data available (e.g., grades or education level) compared to MOOC platforms where the demographic and historical data is not compulsory for the student to input. Below papers with different scopes of study are summarized to give examples of different settings in the research area.

Xing et al. (2016) focused on a project management course with 3,617 registered students. The course lasted 8 weeks with 11 modules and it had online discussions and quizzes. Due to the high number of students the instructors had limited interaction experience with the students. The data obtained had click-stream data for the whole course, quiz scores and discussion forum data. General Bayesian Network (GBN) and decision tree (C4.5) algorithms were used for the prediction task. The predictions were calculated weekly and both Area Under the ROC Curve (AUC) and precision improved week by week as more data became available. GBN performed slightly better than decision tree as the average AUC for GBN was 89.0% and for decision tree it was 86.3%. Using a stacking method which utilizes both algorithms an average AUC of 90.7% was achieved. (Xing er al.)

Fei & Yeung (2015) studied two MOOCs one offered on the Coursera platform and the other on the edX platform. The Coursera course was a six-week course with 39,877 students with at least one activity. The edX course lasted for ten weeks and had 27,629 active students. On the Coursera course seven features were tracked from the log data while on the edX course there five features tracked. The focus of the study was temporal models. The models tested were Input-Output Hidden Markov Models (IOHMM), Vanilla Recurrent Neural Network (Vanilla RNN) and Long Short-Term Memory RNN (LSTM Network). These models were compared to baseline models which were Support Vector Machine and Logistic Regression.

The LSTM network outperformed other models consistently except in the first weeks where every model had a low AUC score. IOHMM models generally had the worst performance. The baseline models performed better on the Coursera course than on the edX course and hence were not as consistent as the temporal models. (Fei & Yeung, 2015)

Marbouti et al. (2016) used only in-semester performance factors in their study of predicting at-risk students in a course using standard- based grading. The course was held twice and there were 1,650 students on the first course and 1,413 students on the second. There were multiple graded components (e.g., quizzes, homework, exams, projects) on the course and most of these components were collected on a weekly basis. Course like this generates a large amount of performance data every week. Six common prediction methods were chosen: Logistic Regression, Support Vector Machine, Decision Tree, Multi-Layer Perceptron, Naive Bayes Classifier and K-Nearest Neighbor. The dropout rate on course was less than 10 % which means it is critical that the models can identify them. The best overall accuracy for classification was achieved with K-Nearest Neighbor but it failed to identify the at-risk students. Naive Bayes Classifier was the best model for identifying the dropouts with an accuracy of 86.2 %. (Marbouti et al., 2016)

Márquez-Vera et al. (2016) predicted dropouts at different steps of Mexican high school teaching. They gathered data from 419 high school students in Mexico. From these students 57 were considered as dropouts hence the dropout rate was 13.6 %. The dataset had previous scores/marks, general information, information about attendance and behavior of the students, survey of factors affecting school performance and final scores. The goal was to predict if the student continues to the next semester. The authors created a genetic programming algorithm and compared it to multiple classical classification algorithms: Bayesian classifier, Support Vector Machine, Instance-based lazy learning, Classification rules and Decision trees. The algorithm developed was able to predict student dropout in the first four to six weeks accurately enough to be used as a part of an early warning system. (Márquez-Vera et al., 2016)

All of these studies discuss dropout prediction and provide information on how to carry out the dropout prediction process. However, there are many notable differences in the studies. The number of students varies from hundreds to tens of thousands. This is caused by some studies focusing on MOOCs while others focus on face-to-face learning in schools. Different learning formats also affect the dropout rates which vary between ten and 90 percent. The data is imbalanced in both cases but on different sides. This affects the selection of the model and the tuning of it. The datasets used also differ because of differences in learning formats but also the learning institution affects the dataset as high schools usually have more information on the earlier studies of the students than for example MOOC websites. Which is why it is important to understand that dropout prediction can only be carried out with the data that is available. To achieve best dropout prediction results the organizing party of the course or study program must in advance think about what data can be collected and how the data collection is implemented.

4.3 Classification models

Dropout prediction is a binary classification task. Common practice is to label dropout class as 1 and students who pass as 0. The objective in classification tasks is to find a function that can identify the class of the observation (Hackeling, 2014). In the practical part three classification models are built and then compared in the dropout prediction case. These models are support vector machine, logistic regression, and random forest classifier. The logic of the models is presented in the following chapters. Performance metrics for the classification models are explained after models.

4.3.1 Support vector machine

The objective of the support vector machine (SVM) is to find a hyperplane in a N-dimensional that classifies the observations. As there are many suitable hyperplanes the objective is to find the best one. To find out which hyperplane fits the data best, the distances to each class should be maximized. If we define the distance from the hyperplane to the nearest observation vector as the margin of the hyperplane, then the SVM selects the maximum-margin separating hyperplane. Selecting the maximum-margin hyperplane maximizes the SVM’s ability to correctly classify new observations. However, the classes cannot always be completely separated by a hyperplane because real world datasets contain outliers. To handle these datasets a ‘soft margin’ can be added to the SVM algorithm. This allows some observations to the other side of the hyperplane without affecting the final result. The user-specified parameters that control how many observations are allowed to the wrong side and how far across the line they are allowed to go must be finely tuned as they affect the number of misclassifications. (Noble, 2006)

There are various SVM formulations for classification tasks and in this thesis C-Support Vector Classification is used. The SVM is created with the Python function SVC from the Scikit-learn machine learning library. The implementation of the function is based on libsvm (Chang & Lin, 2011).

4.3.2 Logistic regression

Logistic regression uses the principle of maximum likelihood estimation (Dangeti, 2017). In logistic regression, the function calculates the response variable which describes the probability that the observation belongs to the positive class. If the probability is equal to or exceeds a discrimination threshold, it is assigned to the positive class; otherwise, the negative class is assigned. The response variable is modeled as a function of a linear combination of the features using a logistic function. (Hackeling, 2014) Logistic regression transforms the dependent variable into a logit variable (which is the natural log of the odds of the dependent variable happening or not) with respect to independent variables. The equation can be written down as: (Dangeti, 2017)

$$\ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * x_1 + \cdots + \beta_n * x_n,$$

where p is the probability of the dependent variable being positive, β is a parameter of the model and x is an independent variable.

4.3.3 Random forest classifier

To understand how a random forest classifier (RFC) works it is first important to know what decision trees are. Decision tree is one of the simplest classification algorithms. A decision tree is a set of rules each of which divides the dataset in two sets. The tree (Figure 4) starts from the root node and the nodes following are called children. There is a rule attached to each node and the observations are divided in two sets at each node. The leaf nodes are the last set of nodes which are in the end classified to belong in one of the classes. Starting from the root node an observation goes through the nodes until it ends up in a leaf node where it is classified. (Jolly, 2018) The growing phase stops when a stopping criterion is met. Rokach & Maimon (2008) list the following common stopping rules:

1) All observations of the training set belong to single class
2) The maximum tree depth is reached
3) The number of observations in a lead node is less than set number required for a parent node
4) If a node is split and the number of observations in one split is less than the minimum required for a new split
5) The best splitting criterion is lower than a certain threshold

Two common splitting criteria used in decision trees are information gain and Gini index. They are both impurity-based criteria. Information gain uses entropy as the measure of impurity. Gini Index measures the divergences between the probability distributions of the target classes. (Rokach & Maimon, 2008)

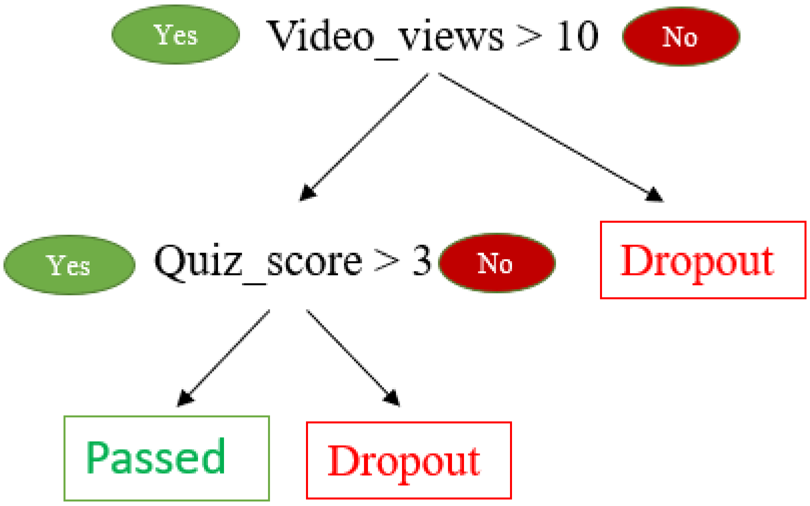


Figure 4 Example of simple decision tree

RFC is an ensemble of decision trees. Decision trees tend to have high variance, which makes them unstable. By creating an ensemble of decision trees, we can minimize the variance of the model and create a new model that is nearest to an ideal model. RFC samples both the features and the observations of the training data so that the developed decision trees are independent. The prediction is then decided by majority vote or by averaging the probabilities. (Dangeti, 2017) To build the trees a random vector is generated independently for each tree from the same distribution. These vectors are used with the training dataset to train the individual trees. Random forests do not overfit because of the Law of Large Numbers. (Breiman, 2001)

4.3.4 Performance metrics for binary classification models

Performance metrics are needed for the comparison of different classification models. As dropout prediction is a case of binary classification only metrics used for binary classification models are presented here. Classification results can be evaluated by calculating the number of observations in four groups:

1) **True positives (TP)**: the model correctly predicted the positive class

2) **True negative (TN)**: the model correctly predicted the negative class

3) **False positive (FP)**: the model incorrectly predicted the positive class

4) **False negative (FN)**: the model incorrectly predicted the negative class

These four numbers can be summarized in a table called Confusion Matrix (Table 6). Confusion Matrix is useful for a quick visualization of the classification model's performance.

Table 6 Confusion matrix for binary classification

	Classified as Positive	Classified as Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

When the counts for these variables are known, it is possible to calculate different metrics to evaluate the classification model's performance further. The most commonly used metrics for binary classification are presented in Table 7 (Sokolova, 2009). When comparing the classification models with these metrics the evaluation focus and class imbalance must be considered. For example, accuracy can give a high performance when the more common class is predicted often (Luque at al., 2019).

Table 7 Metrics for binary classification

Metric	**Formula**	**Evaluation focus**
Accuracy	$\frac{TP + TN}{TP + FN + FP + TN}$	Overall effectiveness of the model
Precision	$\frac{TP}{TP + FP}$	How many of the instances classified as positive are actually positive
Recall or Sensitivity or True Positive Rate	$\frac{TP}{TP + FN}$	What proportion of the positive class got correctly classified
F-score	$\frac{(\beta^2 + 1)TP}{(\beta^2 + 1)TP + \ \beta^2 FN + FP}$	Relations between data's positive observations and those given by the model
Specificity	$\frac{TN}{FP + TN}$	How effectively the model identifies the negative class
False Positive Rate	$\frac{FP}{FP + TN}$	How many of the instances classified as negative are actually negative

A receiver operating characteristics (ROC) graph is a technique for visualizing classification models based on their performance. ROC graph is a two-dimensional graph in which true positive rate (TPR) is plotted on the Y axis and False Positive Rate (FPR) is plotted on the X axis. Discrete classification models (e.g., decision trees and rule sets) produce a single point in the ROC space. Some classifiers (e.g., Naive Bayes classifiers and neural networks) which give an instance a probability or a score that is a number which represents a degree to which an instance is a member of a class. These numbers and a selected threshold can then be used for producing a discrete classifier. If the instances score is over the threshold, it is classified as positive. The selected threshold affects the TPR and FPR of the classifier which means that by changing the threshold multiple points can be calculated for the ROC graph. (Fawcett, 2006)

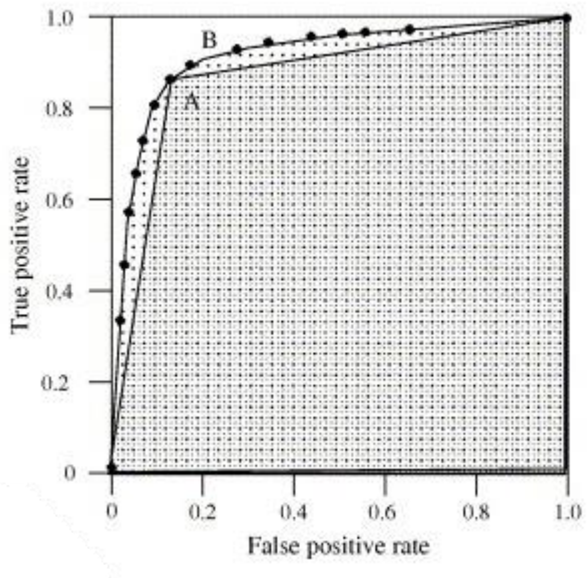


Figure 5 Area under the curve of a discrete classifier (A) and a probabilistic classifier (B) (Fawcett, 2006)

As ROC curve is two-dimensional representation of classifier performance it is often beneficial to reduce ROC performance to a single scalar by calculating the area under an ROC curve (AUC) (Figure 5). AUC is a portion of the area of the unit square so its value is always between 0 and 1. However, random guessing results in an AUC of 0.5 which is why a classification model should not have an AUC under 0.5. (Fawcett, 2006)

4.4 Feature importance

Feature importance (also called feature detection, feature attribution, and model interpretability) outputs a score or metric which allows the features to be ranked from largest to smallest contribution to the machine's prediction. This can be achieved by permuting features to understand which features affect the prediction power the most. (Musolf et al., 2021) When the feature importance is known it is easier to understand the model's logic and it gives the predictions an interpretable reasoning. The feature importance scores can also be used for feature selection. With the scores it can be possible to understand the causes which lead to the predicted variable to be positive, for example, in dropout prediction the scores might be useful in determining what causes students to drop out (or complete the course).

Mutual information (MI) is a measure of independence between random variables X and Y. MI is zero only in cases where the two variables are completely independent. It is based on entropy estimates from k-nearest neighbor distances. The MI is defined as

$$I(X,Y) = \int\int dxdy\, \mu(x,y) log\, \frac{\mu(x,y)}{\mu_x(x)\mu_y(y)},$$

where $\mu_x(x)$ and $\mu_y(y)$ are the marginal densities of X and Y, and where $\mu(x,y)$ is the joint probability density function. The base of the logarithm decides the units in which the information is measured. (Kraskov et al., 2004)

5. Results

In this chapter the results of the dropout prediction model are presented. The chapter goes through all parts (Figure 6) of the process of turning the data into predictions. These include data preprocessing, building, and tuning the machine learning models, and finally the models are compared, and the features evaluated.

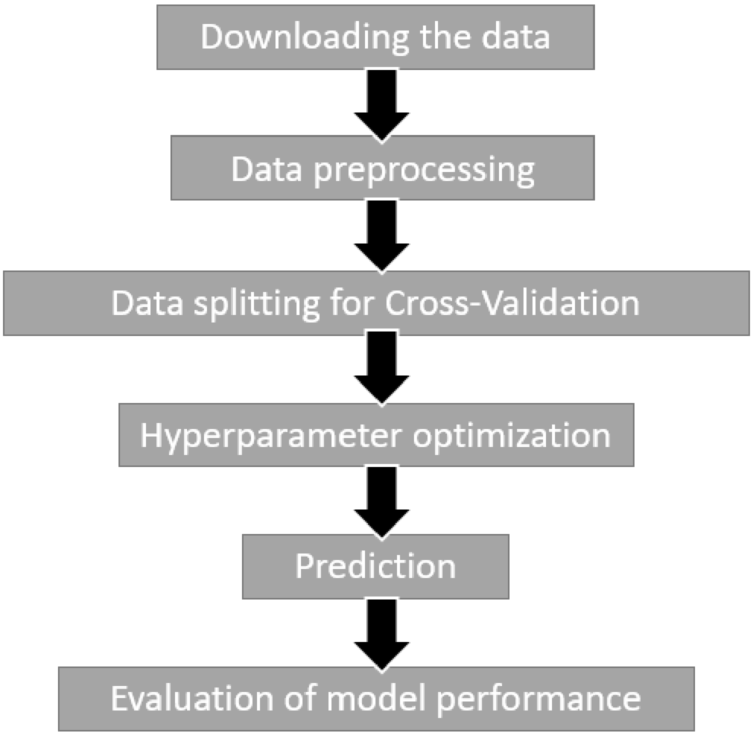

Downloading the data
Data preprocessing
Data splitting for Cross-Validation
Hyperparameter optimization
Prediction
Evaluation of model performance

Figure 6 The process of building a machine learning model

The model building process is fully carried out with Python using Jupyter Notebook. The following chapters present the steps of the process and its results in the order of which they were completed.

5.1 Dataset

The dataset used in this thesis is an open dataset of tracking log files from XuetangX platform. The dataset was first introduced by Feng et al. (2019) at AAAI-19 conference and is made available at moocdata.cn. It contains data from over 300,000 users and has over one thousand courses. All users' learning activities on the platform are included and for some users also profile information (e.g., age and gender) is available.

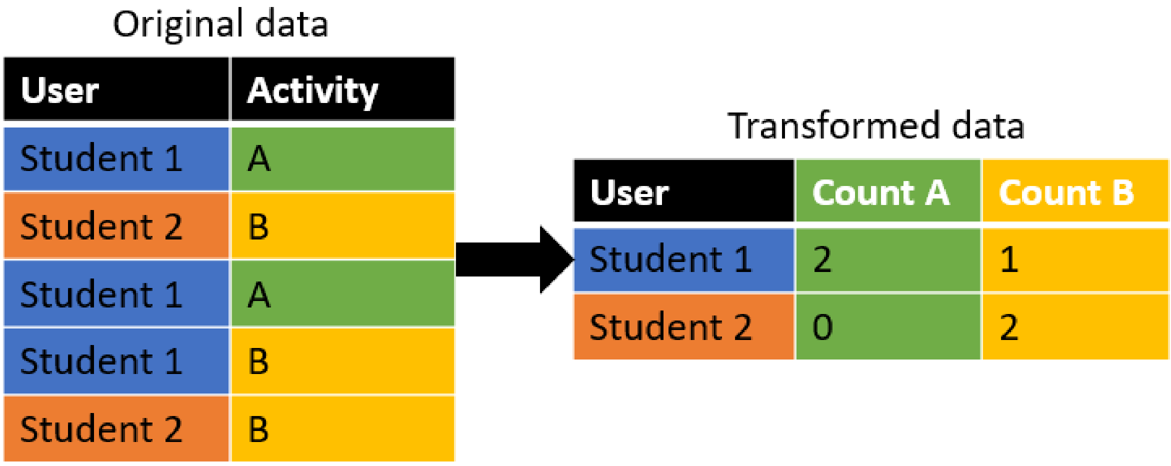
The object of this thesis is to study the use of log data from a single course. For this purpose, the biggest course on the dataset is selected and only the tracking log data is utilized, meaning that the profile information is discarded. In the selected course there were 3700 users of which 1205 dropped out of the course meaning that the dropout rate was 32.6 %. Every time a user interacts with the learning platform the user's id, activity and timestamp are recorded in the log data. In the log data of the course there are a total of 3,077,367 interactions. By looking at the total counts of the activities (Table 8) it is clear that the discussion forum use in this course is minimal which means that the benefits of forum usage cannot be assessed based on this data. The course is mostly based on videos and problems which have different activities for correct and incorrect answers.

Table 8 Total counts of the activities

Activity	Total count
stop_video	1,188,696
click_courseware	541,383
pause_video	407,855
load_video	252,032
play_video	246,041
close_courseware	221,585
seek_video	59,539
problem_get	36,711
click_info	29,928
problem_check	25,795
problem_check_correct	18,961
click_about	15,400
click_progress	10,905
problem_check_incorrect	8,489
problem_save	7,622
click_forum	5,989
create_comment	339
create_thread	81
delete_comment	16

5.2 Data preprocessing

To test how early the dropouts can be identified the data was cut after every week to create new datasets. Datasets were created for the first five weeks and the datasets included all the weeks before the cut point, for example, on week three the data from weeks one and two is also used. For the machine learning algorithms used in this thesis the data must be in a format where each row contains all features for one user. However, in the dataset there is a row for every activity performed on the learning platform. To turn the data into a format that can be used for machine learning it is transformed as illustrated in Figure 7. This is done by calculating the number of times a user performed each activity. Now there are 19 features for each user. In addition to these a new feature is calculated to study if it matters when a student engages with the learning platform for the first time. This is done by calculating how many days passed after the course started before the student's first activity is logged. Finally, the data is split into training and test datasets with a 70-30 split.



Original data

User	Activity
Student 1	A
Student 2	B
Student 1	A
Student 1	B
Student 2	B

Transformed data

User	Count A	Count B
Student 1	2	1
Student 2	0	2

Figure 7 Data transformation

5.3 Building the models

Three models were built for the dropout prediction: SVM, RFC and Logistic regression. All these models were implemented with the scikit-learn library for Python. For all the models the same dataset was used, and the same performance metrics are measured so the models can be compared. Before training the SVM and Logistic regression models the data was standardized because these algorithms generally perform better with standardized data. The scaling is done by using the mean and standard deviation of the training dataset for both training and test datasets. With this method the training dataset has a mean of 0 and a standard deviation of 1.

Hyperparameter optimization is the process of selecting an optimal set of hyperparameters for the learning algorithm. Hyperparameters are the parameters which are set manually before the training is started. Tuning the hyperparameters is one of the most important steps in building an efficient machine learning model. (Agrawal, 2020) The goal of the hyperparameter optimization in this project was to tune the most influential hyperparameters of SVM and RFC models with reasonable optimization times. Default values are used for all other hyperparameters. Logistic regression is not as sensitive to hyperparameter optimization as the other two models which is why default values were used for all hyperparameters when building the model. The models were optimized to achieve as high recall as possible. The final hyperparameters are presented in Table 9 and the optimization process is explained below.

Table 9 Functions and hyperparameters used in model building

Model	Scikit-learn function	Hyperparameters

SVM	SVC	C = 1.0
		kernel = ‘rbf’
		gamma = 0.05

Logistic Regression	LogisticRegression	Default

RFC	RandomForestClassifier	criterion = ‘gini’
		max_depth = 15
		max_leaf_nodes = 50

To build the SVM model function called SVC was used. Three hyperparameters of SVC function that were tuned: kernel, regularization parameter (C), and kernel coefficient (gamma). Linear and Radial basis function (RBF) kernels were first tested to find the base model. With the linear kernel the model was unable to separate the classes on the data of the first two weeks. Without changing other hyperparameters RBF outperformed linear kernel on all five datasets so it was selected. To find optimal C and gamma values grid search was utilized. With this method the optimal C was found to be 1 and gamma equal to 0.05.

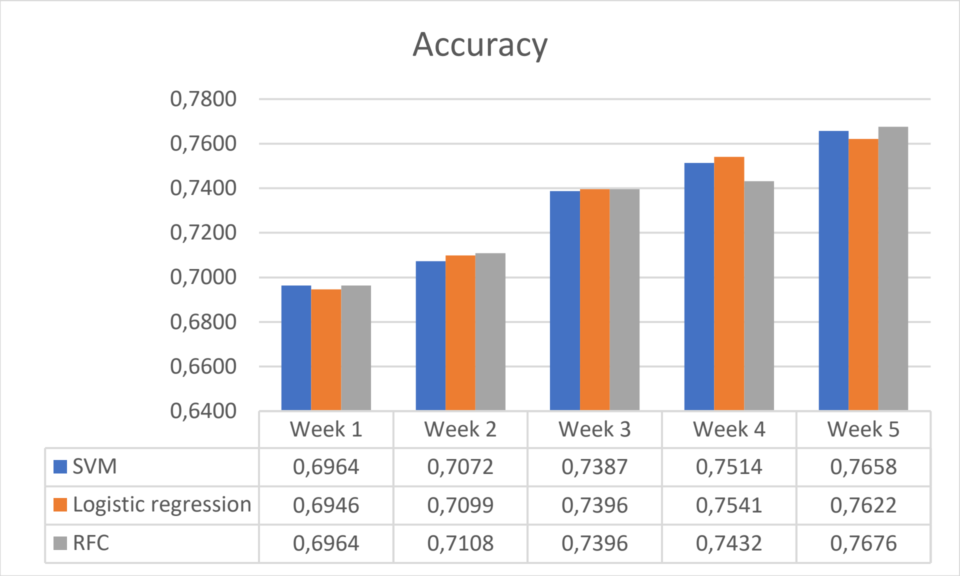
RFC model was built with the RandomForestClassifier function. Hyperparameters selected for tuning were criterion, max_depth and max_leaf_nodes. Criterion determines the function used for measuring the quality of the split. The supported options are Gini impurity and information gain. Max_depth sets the maximum number of nodes in a tree and max_leaf_nodes sets the maximum number of leaf nodes in a tree. The optimal values were selected by testing both criterions, max_depths from three to 15 and max_leaf nodes from five to 500. The best result was achieved with Gini impurity, max_depth set to 15 and max_leaf_nodes set to 50.

5.4 Dropout prediction models performance

To measure how well different methods perform compared to each other three measurements are evaluated:

- **Accuracy:** accuracy measures the percentage of correctly identified observations
- **Recall:** the fraction of actual positives which are correctly classified (in this case what percentage of the dropouts were identified)
- **Area Under the ROC Curve (AUC):** AUC ranges from 0 to 1. If all predictions are wrong AUC is 0 and in case all predictions are correct, then AUC is 1. If AUC is 0.5 it means that the model can’t separate the two classes at all.

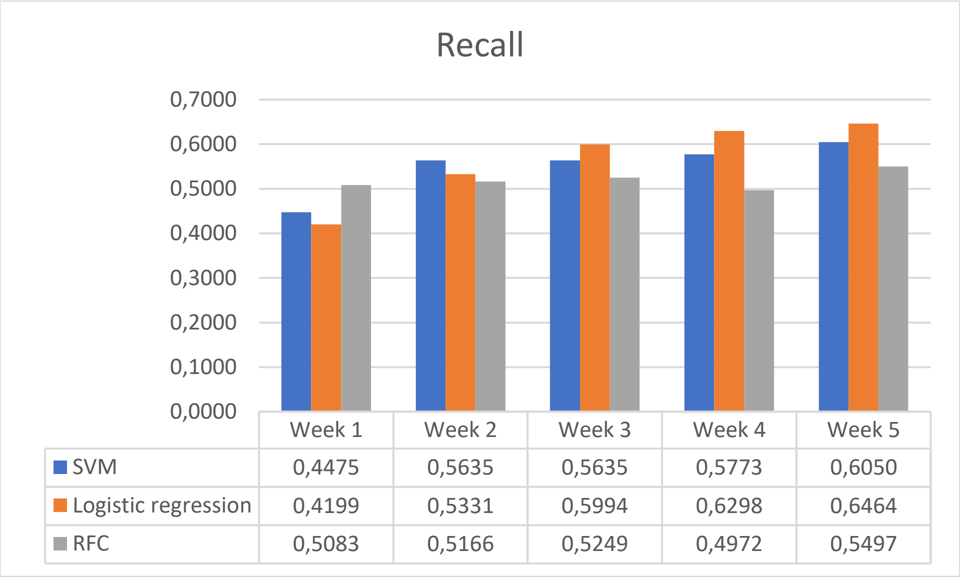
The accuracy (Figure 8) of each model increased every week as expected. The differences between the accuracy of the three models in the first three weeks are small. The largest difference in accuracies is on week four where RFC has the lowest accuracy (74.3 %) while SVM is in the middle (75.1 %) and logistic regression has the highest (75.4 %). Then on week five SVM and RFC have similar accuracy and logistic regression is slightly behind. There is significant improvement with all models between week two and three.



	Week 1	Week 2	Week 3	Week 4	Week 5
🟦 SVM	0,6964	0,7072	0,7387	0,7514	0,7658
🟧 Logistic regression	0,6946	0,7099	0,7396	0,7541	0,7622
⬜ RFC	0,6964	0,7108	0,7396	0,7432	0,7676

Figure 8 Accuracy

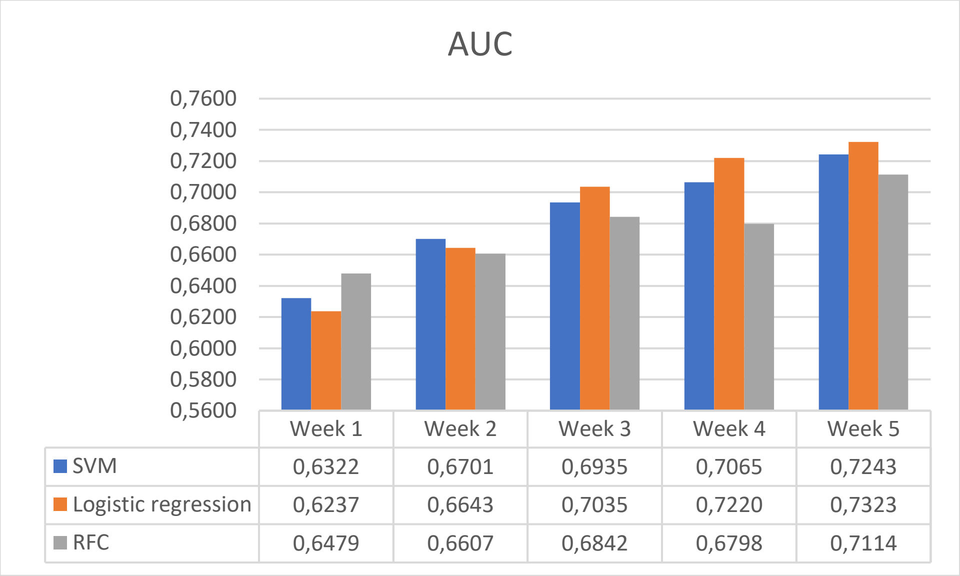
There were no big differences in the accuracies, but the recalls have more dispersion. When comparing the recalls (Figure 9) random forest classifier has the worst performance. Even though it has the best recall on the first week it falls behind after that. SVM leads logistic regression on weeks one and two, but its performance does not improve much on weeks three to five. Logistic regression starts with the lowest recall, but it improves steadily week after week and on week three onwards it has the highest recall. As we want to identify the students at-risk of dropping out as early as possible the first weeks are more important than the later weeks. Week one seems to be too early to identify the at-risk students. SVM has a big improvement on week two from 44.8 % to 56.3 % which is already beneficial. On week three at the halfway point of the course logistic regression clearly becomes the best option.



	Week 1	Week 2	Week 3	Week 4	Week 5
■ SVM	0,4475	0,5635	0,5635	0,5773	0,6050
■ Logistic regression	0,4199	0,5331	0,5994	0,6298	0,6464
■ RFC	0,5083	0,5166	0,5249	0,4972	0,5497

Figure 9 Recall

AUC results (Figure 10) look similar to the recalls of the models. RFC starts strong but it does not improve significantly and does not achieve high scores. On week two SVM has the highest AUC and from week three to five logistic regression has the highest AUC.



	Week 1	Week 2	Week 3	Week 4	Week 5
■ SVM	0,6322	0,6701	0,6935	0,7065	0,7243
■ Logistic regression	0,6237	0,6643	0,7035	0,7220	0,7323
■ RFC	0,6479	0,6607	0,6842	0,6798	0,7114

Figure 10 AUC

Overall, the results show that in the first week it is too early to make accurate predictions. SVM has promising recall on week two, but the low overall accuracy means that many students who will pass the course are classified as dropouts. As the accuracies of all the models on week three improve and logistic regressions recall improves, it becomes the model to use. Week three is also in the middle of the course so if the instructor wants to provide at-risk students with extra support there is still time for it. There is still some improvement on the performance of SVM and logistic regression on weeks four and five but as the course comes closer to the end identifying the dropouts is not as beneficial as before. With these models and this dataset, the best single point for the dropout prediction is after three weeks with the logistic regression model.

5.5 Feature comparison

Mutual information (MI) is a measurement which tells how much one random feature depends on another feature. It is a non-negative value which is equal to zero if the two random features are independent.

Table 10 Mutual information between the features and dropout

Action	Score
close_courseware	0.148194
click_courseware	0.147094
problem_get	0.123247
problem_check_correct	0.121944
load_video	0.115592
pause_video	0.114892
problem_check	0.108848
play_video	0.100793
stop_video	0.100433
seek_video	0.055819
days_after_start	0.045266
click_info	0.041786
click_progress	0.040675
problem_check_incorrect	0.031095
problem_save	0.027257
click_forum	0.006751
click_about	0.005462
delete_comment	0.000000
create_thread	0.000000
create_comment	0.000000

Higher mutual information value means higher dependency. Calculating the mutual information between the features and the prediction target makes it possible to rank the features. In Table 10 the mutual information between each feature and the dropout vector are sorted from highest to lowest. The first nine features on the table have notably higher MI values than the rest. These features are all generated when a student access the learning resources. Interactions with the forum of the course are at the bottom of the list. This could mean that the use of the forum is not instructed properly as earlier studies have shown activity in the forum to be a good predictor for dropouts.

6. Conclusions

This thesis was inspired by the growing use of digital tools in learning and teaching. Digitalization of teaching provides new data about the behaviour of the students and the efficiency of learning methods. COVID-19 pandemic accelerated the implementation of digital tools in many institutions and created demand for technologies to transform the raw data to information and knowledge. High dropout rates have been a problem especially in online learning formats and machine learning models have been built to identify the students at-risk of dropping out at an early stage. The thesis looked for answers to the three research questions by conducting a literature review and by testing several prediction methods with a MOOC log dataset.

1) What are the main topics of learning analytics?

Seven topics were identified as the main research areas of learning analytics. In order from the most popular to the least popular the topics are performance prediction, dropout prediction, course design, learning strategy, learning visualization, social network analysis, and ethical issues. Almost half (24 out of 50) of the articles selected for the literature review focused on the prediction tasks. The interest in machine learning has increased in many fields of research lately and learning analytics is one of them. The prediction models must be suited for different teaching formats and education levels. Course design focuses on utilizing the data to improve the existing courses and to create better new courses which can take the needs of the students into account. The data can be studied to understand how the students can learn more efficiently and guidance can be then given according to the findings. Visualization is an important part of the learning analytics as if the teachers and students cannot understand the results, they can’t change their actions accordingly. Implementing dashboards, which visualize the learning progress, to learning management systems will help the teacher to allocate the resources efficiently and help the students to better plan their studies. Social network analysis creates information on how the students interact with each other, the instructor, and the learning platform. This information can be used to create a sense of community in online learning and to generate features for the prediction models. Ethical issues and data privacy is always a concern when data is handled. Learning management systems should be transparent of their data usage and the student should be considered as collaborators in learning analytics instead just as a source of data.

2) What data should be collected for dropout prediction?

In dropout prediction the data used has a significant impact on the performance of the prediction model. The amount of data changes in institutions and in different courses. To successfully use dropout prediction in an attempt to lower dropout rates the data collection should be considered in the course's design phase. If the data collection is not planned there might be some important data missing even though it could have been easily collected, e.g., with a pre-course survey. In the practical part of the thesis only the log data of a single course was used to predict the dropouts. Even though the results showed some promise, the accuracy of 74 % at the course's halfway point is not as high as in some other studies which used larger sets of data. In the practical part the data from the use of text materials was the most useful followed by exercises and videos. It should be noted that the discussion forum of the course was not actively used. From literature it was found that mid-term exams, quizzes and exercises are the strongest indicators for dropouts. The student's activity on the discussion forum was useful when the discussion forum was an active part of the course. All these variables directly show either the student's progress or activity. On the low end of useful data were often personal information such as gender and age. The use of personal information also often raises privacy concerns so considering these facts it is not advisable.

3) How to use the learning management system log data to predict dropouts?

In the thesis the performance of three models were tested: support vector machine, logistic regression, and random forest classifier. The data was from Chinese MOOC platform XuetangX. The study focused on a single course with 3700 students of which 32.6 % dropped out of the course. The log data was transformed to features by counting the number of times a student performed an activity. The prediction was done after the first five weeks and the largest improvements in the models' performance were between the first three weeks which are also the most interesting weeks as the goal in dropout prediction is to identify the at-risk as soon as possible. Out of the three models support vector machine, had the best performance in the first two weeks and logistic regression took first from week three onwards. In the first week all the performance metrics were quite low and in the second week accuracies and AUC of the models were low. After three weeks logistic regression was the model with the best performance with an accuracy of 74.0 % and recall of 60.0 %. In weeks four and five there were some improvements but, in this case, it is more beneficial to take early action than to wait for these small improvements. These results are not as high as in other studies. It is possible that there were not strong enough patterns in the log data used to train the models. The studies which achieved over 90 % accuracy with their models used other data than just the tracing log collected by the learning management system.

6.1 Limitations and future research

In the building process of three machine learning models the dataset had some limitations. It did not contain information about the course structure and goals. The data was also limited to only learning management system log data and better results would have been likely achieved if other data sources were available. The dataset is from a Chinese learning site and there might be cultural differences in learning which affect the prediction power of the features. Another limitation in the study was the models used which are classical classification algorithms and there are studies which utilize more advanced models like temporal models and genetic programming models. In future research utilizing multiple data sources to generate the features for the models could be tested. Use of more advanced machine learning models could also be an objective for future research. Creating new features and utilizing feature selection could also be part of future research.

References

Abdous, M., He, W. & Yen, C.-. 2012. Using data mining for predicting relationships between online question theme and final grade. Educational Technology and Society. Vol. 15, no. 3, pp. 77-88.

Agudo-Peregrina, Á.F., Iglesias-Pradas, S., Conde-González, M.Á. & Hernández-García, A. 2014. Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. Computers in Human Behavior. Vol. 31, no. 1, pp. 542-550.

Ali, L., Hatala, M., Gašević, D. & Jovanović, J. 2012. .A qualitative evaluation of evolution of a learning analytics tool. Computers and Education. Vol. 58, no. 1, pp. 470-489.

Asif, R., Merceron, A., Ali, S.A. & Haider, N.G. 2017. Analyzing undergraduate students' performance using educational data mining. Computers and Education. Vol. 113, pp. 177-194.

Bach, C. 2010. Learning Analytics: Targeting Instruction, Curricula and Student Support. IMSCI 2010 - 4th International Multi-Conference on Society, Cybernetics and Informatics, Proceedings. 1.

Berk, J. 2004. The state of learning analytics. T+D: better performance through workplace learning. Vol. 58, no. 6, pp. 34–37.

Berland, M., Martin, T., Benton, T., Petrick Smith, C. & Davis, D. 2013. Using Learning Analytics to Understand the Learning Pathways of Novice Programmers. Journal of the Learning Sciences. Vol. 22, no. 4, pp. 564-599.

Blei, D.M. 2012. Probabilistic topic models. Communications of the ACM. Vol. 55, no. 4, pp. 77-84.

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S. & Koller, D. 2014. Programming Pluralism: Using Learning Analytics to Detect Patterns in the Learning of Computer Programming. Journal of the Learning Sciences. Vol. 23, no. 4, pp. 561-599.

Boelens, R., Voet, M. & De Wever, B. 2018. The design of blended learning in response to student diversity in higher education: Instructors' views and use of differentiated instruction in blended learning. Computers and education. Vol. 120, pp. 197–212.

Breiman, L. 2001. Random Forests. Machine learning. Vol. 45, no. 1, pp. 5–32.

Chang, C.-C. & Lin, C.-J. 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2, no. 3, pp. 1-27.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M.P. & Núñez, J.C. 2016. Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. Computers and Education. Vol. 96, pp. 42-54.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U. & Thüs H. 2012. A reference model for learning analytics. International Journal of Technology Enhanced Learning. Vol. 4, no. 5-6, pp. 318–331.

Conijn, R., Snijders, C., Kleingeld, A. & Matzat, U. 2017. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. IEEE Transactions on Learning Technologies. Vol. 10, no. 1, pp. 17-29.

Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F. & Rego, J. 2017. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior. Vol. 73, pp. 247-256.

Dangeti, P. 2017. Statistics for Machine Learning. Birmingham: Packt Publishing, Limited.

de Barba, P.G., Kennedy, G.E. & Ainley, M.D. 2016. The role of students' motivation and participation in predicting performance in a MOOC. Journal of Computer Assisted Learning. Vol. 32, no. 3, pp. 218-231.

de Freitas, S. I., Morgan, J. & Gibson, D. 2015. Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. British journal of educational technology. Vol. 46, no. 3, pp. 455–471.

Dietz-Uhler, B. & Hurn, J.E. 2013. Using learning analytics to predict (and improve) student success: A faculty perspective. Journal of Interactive Online Learning. Vol. 12, no. 1, pp. 17-26.

Dillenbourg, P. 2016. The Evolution of Research on Digital Education. International journal of artificial intelligence in education. Vol. 26, no. 2, pp. 544–560.

Dyckhoff, A.L., Zielke, D., Bültmann, M., Chatti, M.A. & Schroeder, U. 2012. Design and implementation of a learning analytics toolkit for teachers. Educational Technology and Society. Vol. 15, no. 3, pp. 58-76.

Dziuban, C., Graham, C. R., Moskal, P. D., Norberd, A. & Sicilia, N. 2018. Blended learning: the new normal and emerging technologies. International Journal of Educational Technology in Higher Education. Vol. 15, no. 1, pp. 1–16.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern recognition letters. Vol. 27, no. 8, pp. 861–874.

Fei, M. & Yeung, D., Y. 2016. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. Data Mining Workshop (ICDMW), 2015 IEEE International Conference on. pp. 256–263.

Ferguson, R. 2012. Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning. Vol. 4, no. 5-6, pp. 304–317.

Fidalgo-Blanco, Á., Sein-Echaluce, M.L., García-Peñalvo, F.J. & Conde, M.Á. 2015. Using Learning Analytics to improve teamwork assessment. Computers in Human Behavior. Vol. 47, pp. 149-156.

Gašević, D., Dawson, S., Rogers, T. & Gasevic, D. 2016. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. Internet and Higher Education. Vol. 28, pp. 68-84.

Gašević, D., Zouaq, A. & Janzen, R. 2013. Choose Your Classmates, Your GPA Is at Stake!": The Association of Cross-Class Social Ties and Academic Performance. American Behavioral Scientist. Vol. 57, no. 10, pp. 1460-1479.

Gobert, J.D., Sao Pedro, M., Raziuddin, J. & Baker, R.S. 2013. From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining. Journal of the Learning Sciences. Vol. 22, no. 4, pp. 521-563.

Graham, C. R. 2013. Emerging practice and research in blended learning. Handbook of distance education. 3rd ed. New York, Routledge.

Hackeling, G. 2014. Mastering machine learning with scikit-learn: apply effective learning algorithms to real-world problems using scikit-learn. 1st edition. Birmingham, Packt Publishing.

He, W. 2013. Examining students' online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior. Vol. 29, no. 1, pp. 90-102.

Hew, K. F. & Cheung, W. S. 2014. Students’ and instructors’ use of massive open online courses (MOOCs): Motivations and challenges. Educational research review. Vol. 12, pp. 45–58.

Ifenthaler, D. & Widanapathirana, C. 2014. Development and validation of a learning analytics framework: Two case studies using support vector machines. Technology, Knowledge and Learning. Vol. 19, no. 1-2, pp. 221-240.

Jolly, K. 2018. Machine Learning with Scikit-Learn Quick Start Guide: Classification, Regression, and Clustering Techniques in Python. Birmingham: Packt Publishing, Limited.

Jovanović, J., Gašević, D., Dawson, S., Pardo, A. & Mirriahi, N. 2017. Learning analytics to unveil learning strategies in a flipped classroom. Internet and Higher Education. Vol. 33, pp. 74-85.

Kabakchieva, D. 2013. Predicting student performance by using data mining methods for classification. Cybernetics and Information Technologies. Vol. 13, no. 1, pp. 61-72.

Kaplan, A. M. & Haenlein, M. 2016. Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster. Business horizons. Vol. 59, no. 4, pp. 441–450.

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 and Education. Vol. 104, pp. 18-33.

Kotsiantis, S., Patriarcheas, K. & Xenos, M. 2010. A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. Knowledge-Based Systems. Vol. 23, no. 6, pp. 529-535.

Kotsiantis, S.B. 2012. Use of machine learning techniques for educational proposes: A decision support system for forecasting students'' grades. Artificial Intelligence Review. Vol. 37, no. 4, pp. 331-344.

Kraskov, A., Stögbauer, H. & Grassberger, P. 2004. Estimating mutual information. Physical review E. Vol. 69.

Kumar Basak, S., Wotto, M. & Bélanger, P. 2018. E-learning, M-learning and D-learning: Conceptual definition and comparative analysis. E-learning and digital media. Vol. 15, no. 4, pp. 191–216.

LAK '11 (Learning Analyics and Knowledge). 2011. Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada, February 27-March 1, 2011.

Lara, J.A., Lizcano, D., Martínez, M.A., Pazos, J. & Riera, T. 2014. A system for knowledge discovery in e-learning environments within the European Higher Education Area - Application to student data from Open University of Madrid, UDIMA. Computers and Education. Vol. 72, pp. 23-36.

Lockyer, L., Heathcote, E. & Dawson, S. 2013. Informing Pedagogical Action: Aligning Learning Analytics With Learning Design. American Behavioral Scientist. Vol. 57, no. 10, pp. 1439-1459.

Luque, A., Carrasco, A., Martín, A. & de las Heras, A. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern recognition. Vol. 91, pp. 216–231.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. & Loumos, V. 2009. Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers & education. Vol. 53, no. 3, pp. 950–965.

Macfadyen, L.P. & Dawson, S. 2012. Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. Educational Technology and Society. Vol. 15, no. 3, pp. 149-163.

Maity, S., Sahu, T. N. & Sen, N. 2021. Panoramic view of digital education in COVID-19: A new explored avenue. Review of education. Vol. 9, no. 2, pp. 405–423.

Marbouti, F., Diefes-Dux, H. A. & Madhavan, K. 2016. Models for early prediction of at-risk students in a course using standards-based grading. Computers & education. Vol. 103, pp. 1–15.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Mousa Fardoun H. & Ventura, S. 2016. Early dropout prediction using data mining: A case study with high school students. Expert systems. Vol. 33, no. 1, pp. 107–124.

Márquez-Vera, C., Cano, A., Romero, C. & Ventura, S. 2013. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Applied Intelligence. Vol. 38, no. 3, pp. 315-330.

Mor, Y., Ferguson, R. & Wasson, B. 2015. Editorial: Learning design, teacher inquiry into student learning and learning analytics: A call for action. British Journal of Educational Technology. Vol. 46, no. 2, pp. 221-229.

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

Musolf, A. M., Holzinger, E. R., Malley, J. D., & Bailey-Wilson, J. E. 2021. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. Human genetics.

Natek, S. & Zwilling, M. 2014. Student data mining solution-knowledge management system related to higher education institutions. Expert Systems with Applications. Vol. 41, no. 14, pp. 6400-6407.

Nielsen, W., Miller. K. A. & Hoban G. 2015. Science Teachers’ Response to the Digital Education Revolution. Journal of science education and technology. Vol 24, no. 4, pp. 417–431.

Noble, W. S. 2006. What is a support vector machine? Nature biotechnology. Vol. 24, no. 12, pp. 1565–1567.

Panigrahi, R., Srivastava, P. R. & Sharma, D. 2018. Online learning: Adoption, continuance, and learning outcome—A review of literature. International journal of information management. Vol. 43, pp. 1–14.

Pardo, A. & Siemens, G. 2014. Ethical and privacy principles for learning analytics. British Journal of Educational Technology. Vol. 45, no. 3, pp. 438-450.

Park, Y. & Jo, I.-. 2015. Development of the learning analytics dashboard to support students' learning performance. Journal of Universal Computer Science. Vol. 21, no. 1, pp. 110-133.

Pursel, B.K., Zhang, L., Jablokow, K.W., Choi, G.W. & Velegol, D. 2016. Understanding MOOC students: Motivations and behaviours indicative of MOOC completion. Journal of Computer Assisted Learning. Vol. 32, no. 3, pp. 202-217.

Rienties, B. & Toetenel, L. 2016. The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. Computers in Human Behavior. Vol. 60, pp. 333-341.

Rokach, L. & Maimon, O. Z. 2008. Data mining with decision trees theory and applications. Singapore, World Scientific.

Romero, C., Espejo, P.G., Zafra, A., Romero, J.R. & Ventura, S. 2013. Web usage mining for predicting final marks of students that use Moodle courses. Computer Applications in Engineering Education. Vol. 21, no. 1, pp. 135-146.

Romero, C. & Ventura, S. 2010. Educational data mining: A review of the state of the art. IEEE transactions on systems, man and cybernetics. Vol. 40, no. 6, pp. 601–618.

Romero-Zaldivar, V.-., Pardo, A., Burgos, D. & Delgado Kloos, C. 2012. Monitoring student progress using virtual appliances: A case study. Computers and Education. Vol. 58, no. 4, pp. 1058-1067.

Ruipérez-Valiente, J.A., Muñoz-Merino, P.J., Leony, D. & Delgado Kloos, C. 2015. ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. Computers in Human Behavior. Vol. 47, pp. 139-148.

Scheffel, M., Drachsler, H., Stoyanov, S. & Specht, M. 2014. Quality indicators for learning analytics. Educational Technology and Society. Vol. 17, no. 4, pp. 117-132.

Shen, D., Cho, M., Tsai, C. & Marra R. 2013. Unpacking online learning experiences: Online learning self-efficacy and learning satisfaction. The Internet and higher education. Vol. 19, pp. 10–17.

Shum, S.B. & Ferguson, R. 2012. Social learning analytics. Educational Technology and Society. Vol. 15, no. 3, pp. 3-26.

Siemens, G. 2013. Learning Analytics: The Emergence of a Discipline. The American behavioral scientist. Vol. 57, no. 10, pp. 1380–1400.

Siemens, G. & Baker, R. 2012. Learning analytics and educational data mining: towards communication and collaboration. Proceedings of the 2nd International Conference on learning analytics and knowledge. 2012 ACM. pp. 252–254.

Slade, S. & Prinsloo, P. 2013. Learning Analytics: Ethical Issues and Dilemmas. American Behavioral Scientist. Vol. 57, no. 10, pp. 1510-1529.

Sokolova, M. 2009. A systematic analysis of performance measures for classification tasks. Information processing & management. Vo. 45, no. 4, pp. 427–437.

Tabuenca, B., Kalz, M., Drachsler, H. & Specht, M. 2015. Time will tell: The role of mobile learning analytics in self-regulated learning. Computers and Education. Vol. 89, pp. 53-74.

Taglietti, D., Landri, P. & Girmaldi, E. 2021. The big acceleration in digital education in Italy: The COVID-19 pandemic and the blended-school form. European educational research journal EERJ. Vol. 20, no. 4, pp. 423–441.

Tempelaar, D.T., Rienties, B. & Giesbers, B. 2015. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. Computers in Human Behavior. Vol. 47, pp. 157-167.

UNICEF. 2020. Covid-19: Are children able to continue learning during school closures? A global analysis of the potential reach of remote learning policies using data from 100 countries. UNICEF, New York.

Verbert, K., Duval, E., Klerkx, J., Govaerts, S. & Santos, J.L. 2013. Learning Analytics Dashboard Applications. American Behavioral Scientist. Vol. 57, no. 10, pp. 1500-1509.

Verbert, K., Govaerts, S., Duval, E., Santos, J.L., Van Assche, F., Parra, G. & Klerkx, J. 2014. Learning dashboards: An overview and future research opportunities. Personal and Ubiquitous Computing. Vol. 18, no. 6, pp. 1499-1514.

Voigt, I., Stadelmann, C., Meuth, S. G., Funk, R. H. W., Ramisch, F., Niemeier, J. & Ziemssen, T. 2021. Innovation in Digital Education: Lessons Learned from the Multiple Sclerosis Management Master's Program. Brain sciences. Vol. 11, no. 8, pp. 1110.

Xing, W., Chen, X., Stein, J. & Marcinowski, M. 2016. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. Computers in human behavior. Vol. 58, pp. 119–129.

Xing, W., Guo, R., Petakovic, E. & Goggins, S. 2015. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. Computers in Human Behavior. Vol. 47, pp. 168-181.

Yashalova, N. N. & Vasiltsov, V. S. 2020. Digital Education: New Challenges and Opportunities. Scientific and technical information processing. Vol 47, no. 4, pp. 260–265.

You, J.W. 2016. Identifying significant indicators using LMS data to predict course achievement in online learning. Internet and Higher Education. Vol. 29, pp. 23-30.

Zacharis, N.Z. 2015. A multivariate approach to predicting student outcomes in web-enabled blended learning courses. Internet and Higher Education. Vol. 27, pp. 44-53.