Early Identification of Dropout Students in Massive Open Online Courses

AHMED SARHAN ALAMRI

Early Identification of Dropout Students in Massive Open Online Courses

Ahmed Alamri

A thesis presented for the degree of Doctor of Philosophy at Durham University

Supervised by: Prof. Alexandra I. Cristea

Department of Computer Sciences

Durham University

United Kingdom

December 2022

Abstract

Learning analytics (LA) provides the ability to understand the patterns of students' behaviour and improve their educational outcomes. Today, the capacity to retain more data has contributed significantly to the rapid growth of the field of LA. For instance, Massive Open Online Course (MOOC) platforms offer free courses for millions of students worldwide. Therefore, students who cannot afford the expense of higher education may benefit significantly from the available knowledge in MOOCs. This opens a door for educators and academic researchers with a fascinating variety of learning behaviour data that could be used to analyse students' activities and improve their outcomes.

While MOOCs platforms provide knowledge in a new and unique way, the very high number of dropouts is a significant drawback. Several variables are considered to contribute towards learner attrition or lack of interest, which may lead to disengagement or total dropout. In the past decade, many researchers have sought to explore the reasons behind learner's attrition in MOOCs. The jury is still out on which factors are the most appropriate predictors; nevertheless, the literature agrees that early prediction is vital to allow for a timely intervention.

This thesis aims to investigate the early prevention of dropout phenomenon in MOOCs by analysing the gaps in the current literature, identifying the under-researched areas, and developing continuous predictive models that can be used in real-time to identify students at risk of dropingout out of MOOCs. The current thesis explores a light-weight approach based on as little data as possible – since different MOOCs store different data on their users – and thus strive to create a truly generalisable method. Several features (e.g., registration date, students' jumping activities, and the times spent on every single task) have been proposed to predict at-risk students from an early stage. This goal was successfully achieved using different approaches such as statistical data analysis, machine learning and data visualisation.

The second aim of this thesis is to employ motivational theories, mapping online student behaviour onto them, to analyse the drives and triggers promoting student engagement. This thesis further contributes by building an Engage Taxonomy of MOOC engagement tracking parameters, mapped over four engagement theories: Self-Determination Theory (SDT),

Drive, Engagement Theory (ET), and Process of Engagement. The present thesis shows for the first-time metrics for measurable engagement in MOOCs, including specific measures for Autonomy, Relatedness and Competence. It also evaluates the parameters based on existing (and expanded) measures of success in MOOCs: Completion rate, Correct Answer ratio and Reply ratio.

Declaration

The work in this thesis is based on research carried out within the Artificial Intelligence and Human Systems Group (AIHS) at the Department of Computer Science at Durham University, UK.

List of Publications

No part of this thesis has been submitted elsewhere for any other degree or qualification, and it is all the author's own work unless referenced to the contrary in the text.

Accepted papers:

- **Chapter 5**: Cristea, A. I., Alamri, A., Kayama, M., Stewart, C., Alshehri, M., & Shi, L. (2018). Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses. In 27th International Conference on Information Systems Development (ISD2018), Lund, Sweden (Vol. 22). (**ISD conference, Core A)**
- **Chapter 6:** Alamri, Ahmed, et al. "Is MOOC Learning Different for Dropouts? A Visually-Driven, Multi-granularity Explanatory ML Approach." International Conference on Intelligent Tutoring Systems. Springer, 2020. (**ITS conference, Core A)**
- **Chapter 7:** Alamri, A., Sun, Z., Cristea, A. I., Stewart, C., & Pereira, F. D. (2021, June). MOOC next week dropout prediction: Weekly assessing time and learning patterns. In International Conference on Intelligent Tutoring Systems (pp. 119-130). Springer, Cham. (**ITS conference, Core B)**
- **Chapter 7:** Ahmed Alamri, Mohammad Alshehri, Alexandra Cristea, Filipe D. Pereira, Elaine Oliveira, Lei Shi and Craig Stewart. " Predicting MOOCs Dropout Using only two easily obtainable Features from the First Week's Activities." In International Conference on Intelligent Tutoring Systems, Springer, Kingston, Jamaica, 2019. (**ITS conference, Core A)**
- **Chapter 8:** Alexandra I. Cristea, Ahmed Alamri, Mohammed Alshehri, Filipe Dwan Pereira, Armando M. Toda, Elaine Harada T. The Engage Taxonomy: SDT-based measurable Engagement Indicators for MOOCs and their Evaluation. User modeling and user-adapted interaction (2023) (**UMUAI : 5 year Impact Factor: 5.383**)

Acknowledgements

First and foremost, I would like to thank my parents, who have always believed in me and encouraged me to pursue my dreams. Their unwavering love and support has been a constant source of motivation for me. I am also grateful to my siblings, who have always been there for me, offering their help and advice whenever needed.

I wish to thank my wife, Salma, who has stood by me through all my travails. She gave me support and help, discussed ideas and prevented several wrong turns. She also supported the family during much of my graduate studies. Also, I would like to thank our newborn son Battal, who always had a smile for me when I needed one and brought so many joyful moments throughout the last year of my PhD journey.

I am also deeply grateful to my supervisor, Professor Alexandra I. Cristea, who has been a constant source of guidance and support throughout my PhD studies. Her expertise and advice have been invaluable in shaping my research and guiding me through the challenges of completing my thesis.

I would like to extend my heartfelt gratitude to my dear friend Mohammed Alshehri for his constant support and encouragement throughout my PhD journey. His advice and guidance have been invaluable, and his belief in my abilities has kept me motivated and inspired. I am truly grateful for his friendship and for all that he has done to help me succeed in my studies. Also, I am grateful to my friend Dr. Zakhriya Alhassan for his constant support, encouragement, and valuable feedback during my PhD journey, which have greatly contributed to the success of my research.

I would also like to thank my friends, who have been a constant source of support and motivation during my PhD studies. Their encouragement and understanding have helped me stay focused and motivated during the challenging times.

Lastly, I would like to express my gratitude to all the members of the Computer Science

Department community, who have provided me with the necessary resources and support to complete my PhD studies.

In conclusion, I am truly grateful to all of the individuals mentioned above, who have contributed to my success and helped me achieve this milestone in my academic journey. Thank you all from the bottom of my heart.

Table of Contents

List of Figures

List of Tables

Thesis acronym

Acronyms	Meaning
AIED	Artificial intelligence in education
ANN	Artificial Neural Network
ASD	After start date
Backprop	Back propagation neural network
BERT	Bidirectional Encoder Representations from Transformers
BLS	Broad learning system
BLSTM	bidirectional long short term memory
BSD	Before start date
cMOOCs	Cnectivist Massive Open Online Courses
CNN	Convolutional Neural Networks
CRF	Conditional Random Fields
DNN	Deep feed-forward neural networks
EDM	Educational Data Mining
ELM	Extreme Learning Machines
FM	Factorization Machines
GBM	Generalised Boosted regression Models
GBN	General Bayesian Network
GCN	Graph convolutional networks
GLM	Generalized Linear Mode
GRU	Gated recurrent units
HAN	Hierarchical Attention Network
ICT	Information Communication Technology
ITS	Intelligent Tutoring Systems

LA	Learning Analytics
LDA	Linear Discriminant Analysis
LightGBM	Light Gradient Boosted Machine
LMS	Learning Management System
LR	Logistic Regression
LadFG	latent dynamic factor graph
MLP	Multiple Layer Perceptron with two hidden layers
MLR	MultiNomial Logistic Regression
MOOCs	Massive Open Online Courses
MSE	Mean Square error
NSSM	Nonlinear State Space Model
NTUSD	lexicon Sentiment Dictionary
OULAD	Open University Learning Analytic Dataset
RG	Regression
RNN	Recurrent Neural Networks
SDT	Self-Determined Theory
SVR	Support vector regression
SGD	Stochastic gradient descent gradient descent
SSAE	Stacked Sparse Autoencoder
SVM	Support Vector Machine
TAN	Tree Augmented Naive Bayes
Transformer	Transformer Encoder
TSF	Time series forest
XGBoost	Extreme Gradient Boosting
xMOOCs	Extended Massive Open Online Course

Chapter 1 : Introduction

Prologue

This chapter begins with (**i**) a brief introduction to the topic, (**ii**) followed by presenting the problem statement, (**iii**) describing the rationale for exploring this area, (**iv**) posing research questions, (**v**) discussing the aim and objectives, and (**vi**) detailing the research contributions and the thesis's outline.

1.1 Introduction

There is no doubt that the global digital revolution and the growing availability of broadband internet have paved the way for new forms of education (Phillips, 2005). These include, but are not limited to, online learning, digital educational content production and delivery, and mobile learning. The recent advent of massive online open courses (MOOCs, relatively short online courses), which target large student numbers and international audiences, has raised the interest of students, educators and researchers alike (De Freitas et al., 2015). MOOC as a term was first coined in 2008, followed by the naming of 2012 as the 'Year of the MOOC', when MOOC providers, such as Coursera, Udacity, edX and FutureLearn, were all launched (Bothwell and Havergal, 2016).

The goal of MOOCs is to provide open-access courses via the internet, regardless of the number of enrolled students (Ipaye, 2013). The vast potential of MOOCs has provided learning opportunities for millions of learners across the world (Kloft et al., 2014). This potential has engendered the creation of many MOOC providers (e.g. FutureLearn, Coursera,

edX, and Udacity),[1] all of which aim to deliver well-designed courses to a mass audience. MOOCs provide many valuable educational resources to learners who can connect and collaborate with each other through discussion forums (Yang et al., 2013).

In general, MOOCs have become a key mainstream approach to democratise knowledge (Atenas, 2015).

1.2 Problem Statements

Leading universities and colleges, which provide a wide variety of accredited degree and certificate programmes, sponsor the courses offered on MOOC platforms. This opens up a fantastic possibility for students from disadvantaged backgrounds who want to pursue high-quality education (Agrawal, 2018). While MOOC courses can scale their delivery to many tens of thousands of students (or more (Vivian et al., 2014)), only a small percentage of those students actually complete the course. Completion rates typically range from 3% to 15% (Jordan, 2015, Coffrin et al., 2014). This situation undermines the goal of making educational resources available to enable mass access and learning.

MOOCs' widespread adoption during their short history, has offered the opportunity for researchers and scientists to study them; with a specific focus given to their low rate of completion. Thus, there has been a great deal of interest and research in the reasons for the dropouts among these students and in developing strategies to keep students engaged with the course until completion (Balakrishnan and Coetzee, 2013, Hair et al., 2011, Koller et al., 2013). This has resulted in the creation of several predictive models that determine student success, with a substantial rise in the literature since 2014 (Gardner and Brooks, 2018). Predicting students' likelihood to complete (or not to complete) a MOOC course, especially from very early weeks, has been one of the hottest research topics in the area of learning

[1] https://www.mooclab.club/resources/mooclab-report-the-global-mooc-landscape-2017.214/

analytics. The main aim of these productive models is to provide systematic insight for instructors to target learners most in need of intervention (Xing, 2016).

1.3 Motivation

In the field of education, decision-support-related activities (e.g. predicting students at risk and student learning outcomes) primarily focus on easing access to early intervention preventative measures. This indicates that such predictive activities are highly important in improving the results for students, enhancing the performance of education services, reducing delays in supporting students with additional needs, and reducing education-related expenditures (Cui, 2019, Essa and Ayad, 2012, Gardner and Brooks, 2018).

In recent years, education analytics powered by data from online learning (and prediction tasks in particular) has emerged as a promising field of study, with significant improvements across various educational systems. In addition, the abundance of electronic learning record data is a gold mine of information that may be used to enhance prediction models and learning performance(Amane et al., 2020, Alharbi and Jacobsen, 2014).

Although MOOCs have emerged as a significant educational resource, they still have shortcomings and need significant advancements, especially in the area of the identification of at-risk students (see section 0).

1.4 MOOC dropout and associated factors

Identifying at-risk learners in a reasonable time frame might support instructors in delivering educational interventions and improving course structures (Hung et al., 2015). The researchers identified several causal factors that influenced MOOC dropout rates. These factors can be categorised into two groups: a) *student-related factors* and b) *MOOC-related factors*.

1.4.1 Student-related factors

- **Insufficient background knowledge**

Students may drop out of MOOCs if they do not have enough background knowledge or the skills needed. For example, students' inability to finish a course could be caused by the lack of knowledge in the mathematics required to complete the course (Belanger and Thornton, 2013). Furthermore, students must have strong reading and writing skills in addition to technical skills, as most interactions in MOOCs are text-based. There is a widespread belief that a lack of these abilities contributes to students dropping out (Murray, 2001).

- **Lack of motivation**

One of the most important reasons that prevent students from finishing a MOOC is insufficient motivation levels. Student motivation is impacted by a wide variety of factors such as the potential for future financial gain and the opportunity to grow personally and professionally. This makes investigating the factors that motivate students to participate in MOOCs quite intriguing (Dalipi et al., 2018).

According to a previous survey by Belanger and Thornton (2013), a general interest in a subject is considered a significant reason for registering in a MOOC, as indicated by 87% of students. However, only 15% of the participants signed up for the MOOC to pursue higher education.

- **Time constraints**

Time commitment is another aspect with a major effect on whether or not students drop out of a MOOC before finishing the course. According to a previous survey (Belanger and Thornton, 2013), students stop participating in MOOCs because they do not have enough time to attend online sessions, complete homework, and study for exams.

1.4.2 MOOC-related factors

- **Course design**

The design of a course is a primary reason for students to stop participating in the MOOC. Every course design has three main parts: the material covered in the course, the format in

which that material is presented, and the means through which students get information. The course content is one of the most crucial factors that affect students' decision to drop out of MOOCs (Hone and El Said, 2016, Dalipi et al., 2018).

- **Lack of discussions**

The other aspects that have been demonstrated to influence students dropping out of MOOCs are the sense of isolation and lack of course participation. According to a survey conducted by (Hone and El Said, 2016), the low levels of contact and inadequate feedback between the teacher and the students are also reasons for students dropping out of MOOCs. In addition, the surveyed students mentioned that the lack of collaboration and teamwork contributed to the atmosphere of isolation.

- **Hidden expenses**

Hidden expenses may contribute to the high student withdrawal rate in MOOCs. For example, in some MOOCs, students are required to pay money to receive course certificates or for the expensive textbooks suggested by instructors (Khalil and Ebner, 2014).

1.5 Research Questions

The research challenge and gaps found in the existing literature (detailed in more depth in section 0) were used as the basis for developing the research questions. T**he umbrella research question** that directed this research is: ***How can students' (interactional) data in a MOOC be used to identify and discover educational bottlenecks affecting student success?***

As the overarching research question is quite broad, the following sub-questions were formulated to help answer the umbrella question:

RQ1: Can a limited number of student data types be used for the prediction of success (as in completion)?

Here we interpret success as completion, as this is a preponderant way in learning systems (Moreno-Marcos, 2020a, Mubarak, 2020, Alami, 2021, Radovanović, 2021). We start by analysing a limited number of student data types, to understand if we can obtain the desired

results with limited computational effort. This research question is answered in Chapter 5, and its methodology is explained in Chapter 4 and Chapter 5.

RQ2: Can learning path visualisation of student interactional data be used to inform on student success (also seen as completion)?

Student success is not a simple variable, and often it is left to the instructor to interpret (Sunar et al., 2016). This is the approach we take in this research question, where we visualise student interactional data for an instructor, who would be making decisions based on their visual interpretation of the likelihood of student success.

RQ3: How does the time of student interactional data collection influence student success (completion) prediction and can early prediction be achieved?

One of the goals in student success prediction is to have timely (i.e. early) predictions. With this research question we tackle this problem, whilst returning to the simpler definition of student success as being completion (Moreno-Marcos, 2020a, Mubarak, 2020, Alami, 2021, Radovanović, 2021).

RQ4: Can engagement theories be applied to student interactional data, to help identify student success?

In the final research question of this thesis, we aim at generalising both the notion of student success which is automatically machine computable, as well as relating student interactional data to its underlying psychological causes, as described by engagement theories.

1.6 Research Aim and Objectives

The first aim of this research is to *develop a continuous predictive model that can be used in real-time to identify students at risk of dropping out of MOOCs*. The second aim is *to employ motivational theories, mapping online student behaviour onto them, to analyse the drives and triggers promoting student engagement*.

The following objectives were addressed in this research to answer the identified research questions:

- **O1** To conduct research on students' enrollment dates and evaluate the data for insights into their completion likelihood;
- **O2**: To investigate the possibility and efficacy of developing a system that sends automated, personalised messages to students depending on their enrollment dates;
- **O3**: To conduct a study utilising visualisation analytics to discover and compare the learning routes for those who completed and those who did not complete the course;
- **O4**: To conduct a study using statistical modelling approaches to examine and compare the educational paths of students who successfully finished their course with those who did not;
- **O5**: To perform a study on the utilisation of diverse visualisation approaches, such as fish-eye and bird's eye visualisations, to show students' learning routes at varying levels of granularity;
- **O6**: To conduct a study on the deployment of dynamic predictive models for dropout prediction, including both weekly and whole-course scenarios, and to assess their accuracy in predicting student dropout;
- **O7**: To conduct a study examining the impact of student jumping behaviours and catch-up learning patterns on the prediction accuracy of models used to detect future course dropouts;
- **O8**: To undertake research focusing on the development of predictive models for dropout using the student's number of accesses and time spent per access;
- **O9**: To undertake research focusing on mapping multimodal student behaviours and data according to temporal, action, and language modes in MOOCs over several motivational theories;
- **O10**: To conduct research to group students based on engagement factors and assess the link between these groups and performance indicators.

1.7 Research Contributions

{Chapter 4 }

- A collection and analysis of new educational data for 344,783 students across time (several runs over several years), subjects, universities, countries, and cultures. The data were obtained from two UK universities that delivered different courses in the FutureLearn MOOC platform and from the most massive Arabic MOOC (Rwaq).

{Chapter 5}

- An investigation of several approaches to predict dropout only from the very first interaction with the MOOCs system, the registration.

{Chapter 6}

- New insights into early learning behaviours exhibited by course completers and non-completers through bird- and fish-eye visualisations of partial or full learning graphs, with different levels of information disclosure.
- A proposal of a visual graph analysis as a pre-step to ML and prediction, here illustrated by discovering linear or catch-up behaviours, which then can be reliably predicted.
- A demonstration that theme-based visualisation (which can also be at bird- or fish-eye level) can detect other relations in the course, such as the effect of forums.

{Chapter 7}

- The implementation of eight ML algorithms.
- A comparison of the prediction of weekly and whole-course dropouts.
- A new feature incorporating students' learning patterns, specifically jumping behaviours, into the weekly predictive model and demonstrating its effectiveness.
- A lightweight approach based on tracking two early fine-grain learner activities (accesses to the content pages and time spent per access) to predict student non-completion.

- A map of multimodal student behaviours over several motivational theories.
- Engagement measures for MOOCs.
- Identification and semantically labelling student clusters using the engage taxonomy, specifically SDT mapping.
- A large-scale evaluation of the SDT theory for online learning and MOOCs, based on success measures.
- Application of ML techniques using SDT constructs as inputs.

1.8 Thesis Outline

The thesis is structured with ten chapters as follows:

Chapter 1: This chapter describes the research's area, motivation, and problems. In addition, it presents the research questions, the objectives required to conduct this research, and the contributions of the thesis.

Chapter 2: This chapter provides an overview of distance learning methods such as education through postal services, radio, TV, and E-learning. In addition, this chapter highlights the background of MOOCs, public datasets in MOOCs, Machine Learning (ML) techniques and Theories of Engagement.

Chapter 3: This chapter provides a review of the literature on ML approaches to identify students at-risk of dropping out from MOOCs. Moreover, this chapter presents related works, including the most influential works from the engagement literature.

Chapter 4: This chapter provides an overview of methodology used to answer each research question. Moreover, this chapter explains the dataset and tools used to achieve the aim of this thesis (e.g., feature extraction process, features selection, sentiment analysis, statistical analysis, visualisation tools, and predictive machine-learning techniques).

Chapter 5: This chapter presents the results of a study that was aimed at discovering factors (registration date) that can be identified before the students even start the course to predict which enrolled participants will not complete the MOOC.

Chapter 6: This chapter presents the results of visualising and comparing the different learning paths of completers and non-completers across four MOOCs and the learning theme from which learners tend to drop out. It shows how different granularity visualisations (fish eye and bird eye) allow both researchers and teachers to understand where issues occur and patterns emerge, supported by a statistical analysis.

Chapter 7: This chapter focuses on innovations in predicting student dropout rates by examining their next-week-based learning activities and behaviours. The study presented in this chapter aimed to build a generalised early predictive model for the weekly prediction of student completion using ML algorithms. Moreover, this chapter shows the ML prediction results (lightweight approach) based on two easily obtainable features (number of times content pages were accessed and time spent per access). This allows for easy and reliable implementation across various courses from different domains.

Chapter 8: This chapter proposes a novel, systematic way of analysing engagement, starting from multimodal tracking parameters, following established engagement and motivational theories. In addition, it proposes a concrete mapping between the tracking parameters and four of the most used theories of, or related to, engagement in digital systems, generating the engage taxonomy. Finally, it shows how such mapping can be practised by analysing the engaged and disengaged MOOC student behaviours in relation to the SDT theory.

Chapter 9: This chapter discusses the contributions and overall findings of the studies presented in this thesis (Chapter 5, Chapter 6, Chapter 7, and Chapter 8). This includes the limitations of a) proposed models to predict dropout students, b) MOOCs datasets, and c) the approach used to map large-scale student behaviour onto motivational theories.

Chapter 10: This chapter provides a general summary and conclusion the findings of the works presented in this thesis. Finally, it presents ideas for further research in this area.

1.9 Thesis Conceptual Structure

This thesis tackles the research topics described above. Its primary objective was to investigate the usage of the MOOC dataset to address the challenges of early identification of students at risk of dropping out from MOOCs. Figure 1.1 provides a visual representation of the overall conceptual structure of this thesis.

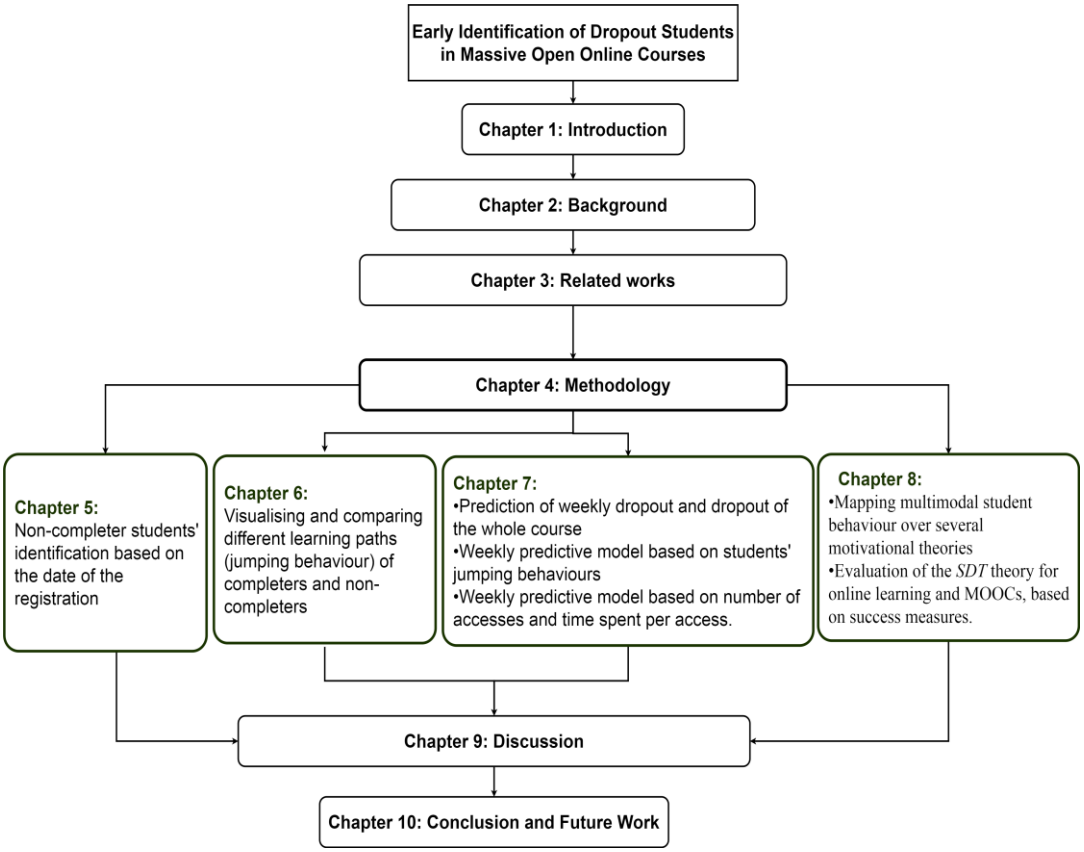


Figure 1.1 Overall conceptual structure

Finally, it is worth mentioning that the four main chapters (Chapter 5, Chapter 6, Chapter 7, and Chapter 8) that answer the thesis research questions are considered independent research regarding their goals, datasets, and methodologies.

Chapter 2 : Background

Prologue

This chapter outlines the background of distance learning technologies by discussing their development, advantages and challenges. In addition, we outline the background of MOOCs and public datasets in MOOCs. Finally, this chapter presents engagement theories.

2.1 Distance Learning

Whilst a variety of definitions have been used to describe ‘distance learning’, the use of technology is usually included in most definitions. For instance, Greenberg (1998) defined distance learning as ‘a planned teaching/learning experience that uses a wide spectrum of technologies to reach learners at a distance and is designed to encourage learner interaction and certification of learning’(Greenberg, 1998). However, postal correspondence is an example of an older distance learning paradigm. Therefore, Moore did not mention the use of technology in his definition, ‘Distance Learning is a learning environment in which students and teachers are separated by distance and sometimes by time’ (Moore and Kearsley, 1996).

When a new technology becomes useful for educational purposes, educators, practitioners, and decision-makers debate its efficacy in comparison with previous methods. This was the situation with film studies from the 1940s and 1950s, educational television (ETV) in the 1960s, computer studies in the 1970s, and teleconferencing studies in the 1990s (Smith and Dillon, 1999).

2.1.1 Postal services

Distance learning became possible only after postal services improved, allowing learning materials and student replies to be sent by post. As an early example, in 1728, Caleb Phillips

published an advertisement in the *Boston Gazette* newspaper, offering distance education services by sending weekly shorthand learning contents (Holmberg, 2005).

In 1844, Isaac Pitman (known as the Father of Distance Learning) started the first correspondence education in Europe, which took advantage of the new postal service in the United Kingdom, which provided faster and cheaper delivery than ever before across the country (Tait, 2003, Archibald and Worsley, 2019). Pitman's method was unique, as teachers could communicate, post resources, and receive students' answers by post. In the same time frame, several projects appeared in Germany that used Pitman's method to provide communication links between teachers and students by using the postal service (Tait, 2003).

Historically, the introduction of various external study programmes by the University of London was a crucial transition in the history of distance learning. The University of London is known as the first open university in the world. In 1858, the university offered a range of academic programs that did not require physical attendance. As a result, many students from different countries applied to study remotely (Bell and Tight, 1993).

Another example of distance learning is the tutoring team established by Thomas J. Foster in 1880 to assist in grading the tasks in booklets to help industrial workers understand mine safety (MacKenzie and Christensen, 1971). Foster expanded his project by opening the International Correspondence Schools, which offer various subjects. More than four million students registered in more than forty courses during the following half-century (Holmberg, 2005).

2.1.2 Radio

Over time, the educational community started exploring a new communication technology to reach out to more students. Therefore, distance learning took another turn with the advent of the radio. Educators transmitted information and courses via radio waves to broadcast to a vast audience (Buckland and Dye, 1991). The University of Wisconsin obtained the first license issued in the United States for a radio station dedicated to educational broadcasting. The university established an amateur wireless station to deliver lessons for students in 1919 (Engel, 1936). In the following three years, seventy-three educational establishments were granted broadcasting licences (Wood and Wylie, 1977).

The radio was and continues to be a powerful platform for informing and educating students in many countries owing to its low-cost and immediate ability to reach a considerable number of people (Kentnor, 2015).

2.1.3 Educational television (ETV)

Koenig and Hill defined ETV as "a medium which disseminates programs devoted to information, instruction, cultural or public affairs, and entertainment" (Koenig and Hill, 1967). The extensive use of audio-visual media in armed forces education has proved its educational efficacy; thus, the usage of video in the classroom has grown in popularity (Kentnor, 2015). Between 1932 and 1937, in the United States, the University of Iowa was the first to utilise television transmission for educational purposes. The University of Iowa has produced more than four hundred programs in a range of different subjects, such as engineering, art, drama, and botany (Koenig, 1969).

Meanwhile, educational television was gaining popularity in Europe and Asia. In the 1950s, BBC pioneered the introduction of enrichment ETV programmes for schools in the United Kingdom. During the same period, RTF–French Radio and Television Broadcasting produced many educational programmes for schools in France (Wallin, 1990). Between 1958 and 1962, education television was given considerable importance in several European countries such as Italy, the Soviet Union, Yugoslavia, and Poland. In China, in 1962, the Shanghai Television station started providing university-level education (Wallin, 1990).

2.1.4 Online Education (E-Learning)

The twenty-first century started with a change in the public perceptions of online learning (e-learning) (Harasim, 2000). The knowledge-based economy has seen a widespread and ever-increasing need for new methods of providing education in recent years, resulting in significant developments in learning technologies and organisations. These dramatic shifts in learning requirements have motivated educators to use information communication technologies to provide education over the internet, often known as online education or e-learning (Shea, 2002).

The term *online education* is described or defined as “a form of distance education that uses computers and the internet as the delivery mechanism, with at least 80% of the course content delivered online” (Allen and Seaman, 2011, Shelton and Saltsman, 2005).

Online education can provide new opportunities for academics to use artificial intelligence to investigate students’ progress. In addition, online teaching strategies may become more successful than traditional approaches when teachers become more aware of their students’ requirements in an online environment (Al-Shabandar, 2019, Ghaznavi et al., 2011).

2.1.5 Learning Management systems

The fast advancements in information and communication technology have greatly helped distance education by providing various frameworks and tools for delivering teaching material. ‘Learning management systems (LMSs)’ is the name given to these frameworks. LMS is a software application that helps administer one or even more courses for a group of students, such as the Blackboard system, which was launched in 1995 (Gallagher, 2008). The LMS provides a viable alternative to conventional methods of evaluating students, as it aims to alleviate the inherent constraints of traditional approaches (Al-Shabandar, 2019).

Through computer technology, teachers and students may participate in an online class at the same time. Nowadays, LMSs are extensively used as paradigm-integrated online education platforms in many universities and academic organisations worldwide (Weaver et al., 2008).

A good LMS provides several features that may contribute to the creation of a good learning environment in different ways. For example, providing a real-time report to monitor student performance and offering social learning tools such as chat, email, and forums, which help students share knowledge (Kulshrestha and Kant, 2013). In addition, teachers can benefit from using an LMS to create online exams. Whilst online assessments allow teachers to see students' responses and give instant and dynamic feedback to learners, these assessments also enable teachers to track students' progress (Botički et al., 2008). Although there are some free and open-source LMSs available, most LMSs options are designed for profit, with licencing fees(Gallagher, 2008).

In 2021, the LMS market was estimated to be valued at more than $15.72 billion. In the United States, 42% of the top 500 public businesses currently utilise educational technology

to train their workers. Hence, there is a growing need for LMSs to deploy and manage e-learning (Pappas, 2020).

2.1.6 Massive open online courses (MOOC)

Massive open online courses (MOOC) can be defined as "online courses that aim to have a wide appeal to people who are interested in learning about a specific subject on a course guided by subject experts as learning facilitators" (Atenas, 2015). *MOOC* as a term was first coined in 2008, followed by the naming of 2012 as the 'Year of the MOOC', when most popular MOOC platforms were launched and reached millions of learners across the globe (Dumouchel, 2015).

MOOCs can deliver learning content online to anyone interested to take a course, especially relatively short online courses with easy access, especially during the Covid-19 pandemic, MOOCs were the de facto platform for self-learning. MOOCs provide many valuable educational resources to learners, who can connect and collaborate with each other through discussion forums (Yang et al., 2013).

2.1.6.1 cMOOCs and xMOOCs

MOOCs can be classified into two main learning paradigms. The first paradigm is known as cMOOCs, which stands for connectivist MOOCs. The other is known as eXtended MOOCs (xMOOCs) (Sanchez-Gordon and Luján-Mora, 2014).

Although cMOOC was the first to be made available in 2011, it does not have the same degree of popularity as xMOOC (Gamage et al., 2016). The concept of cMOOC mimics the idea of the internet itself, as students learn by exchanging knowledge and engaging with each other through several channels such as discussions, collaborations, and discoveries (Sanchez-Gordon and Luján-Mora, 2014). The connectivism approach, in which the teacher does not give out the course learning content but allows students to ask questions and find answers among themselves. Therefore, learners do not gain knowledge via the transmission of information from the teacher directly but rather through the exchange of knowledge among participants (Siemens et al., 2015). cMOOCs allow students to have complete control over their learning experiences and the ability to be autonomous and build their own network of

peers. In addition, students can choose how much time they want to devote to the course (Wang et al., 2017).

However, academic institutions do not recognise cMOOCs as formal courses because of the lack of learning evaluations such as exams or coursework. Moreover, it is difficult to assess the student requirements in cMOOCs, as the course materials change regularly (Al-Shabandar, 2019).

Over the past decade, many cMOOCs have been introduced by scholars. For example, Dr Jim Groom offered a course in digital storytelling as a cMOOC at the University of Mary Washington in 2011. 'Social Media & Open Education' is another example of a cMOOC, which was offered by Dr Alec Couros (Sanchez-Gordon and Luján-Mora, 2014).

In contrast to cMOOCs, xMOOCs use the traditional lecture style, but courses are distributed via the internet in the form of downloadable recorded videos. Usually, a new video lecture series goes out every week for 10 to 13 weeks. In the past, many xMOOCs consisted of long 50-minute lectures, but owing to learning from previous experiences, some instructors now deliver videos that are just 15 minutes long. Nonetheless, the course duration is steadily shrinking over time; nowadays, some courses run for five weeks or less (Martín-Monje et al., 2018).

The multiple-choice test format is one of the most frequently utilised tools to evaluate student knowledge in xMOOCs. Thus, students can view their computer-marked results instantly after completing an online test (Lackner et al., 2014). Peer assessment is another technique used to evaluate student knowledge by placing students in small groups for peer evaluation (Elizondo-Garcia and Gallardo, 2020).

Although xMOOCs provide various communication tools (e.g. discussion forums, emails, and instant messaging), it is difficult for teachers to answer every single question because of the large number of student comments. Consequently, students depend on one another to answer questions and share knowledge (Dipinto and Principi, 2015).

Many xMOOCs have been introduced by academics during the last decade. For example, Sebastian Thrun and Peter Norvig performed an experiment with an online course titled, 'Introduction to Artificial Intelligence'. Approximately 160,000 learners from more than 190

countries participated in the course (Young, 2021). In 2012, MIT and Harvard University founded edX as a new learning platform that offers high-quality courses (xMOOCs) in collaboration with the world's top universities and institutions (Breslow et al., 2013). Up until November 2022, edX had introduced more than 3,600 courses for over 110 million learners[2]. MOOCs are still in their early stages of development, and many researchers are trying different approaches to improving learning and teaching in online courses. Thus, several examples of new MOOCs have appeared during the last few years. Table 2.1 provides a summary of MOOC types (c: cMOOCs, x: xMOOCs)

Table 2.1 Massive Open Online Couse

MOOC	Learning approach	Year	Type
CCK08	Connectivism and Connective knowledge offered by the University of Manitoba in 2008 (Uddin, 2021)	**2008**	**c**
EC&I 831	Social media and open education offered by University of Regina (Uddin, 2021)	**2008**	**c**
Project-based MOOC (pMOOC)	Typically requires students to submit a project and get feedback from other students to complete the course (Kjærgaard et al., 2013).	**2012**	**x**
Synchronous massive online course (SMOC)	Delivers live online classes for many students (Altinpulluk and Kesim, 2016).	**2013**	**x**

[2] **https://www.edx.org/about-us**

Adaptive MOOC (aMOOC)	A MOOC that customises learning styles to suit each student. Content is organised in learning patterns that may meet diverse learner preferences, and real-time personalised feedback is provided (Blanco et al., 2013).	**2013**	**x**
PLENN	Personal Learning Environments, Networks and Knowledge'' offered by the Athabasca University (Uddin, 2021)	**NA**	**c**
Distributed Online Collaborative Course (DOCC)	Allows all participants to contribute their expertise rather than having a single, centralised curriculum (Sanchez-Gordon and Luján-Mora, 2014).	**2013**	**c**
Vocational Open Online Course (VOOC)	An online course designed to assist younger students in making decisions about their future career paths (Sanchez-Gordon and Luján-Mora, 2014)	**2014**	**x**
Self-paced Online Course (SPOC)	These courses are available at any time throughout the year. Therefore, students should learn independently, as they have less chance to communicate with the teacher and other students (Southard et al., 2015).	**2015**	**x**
Personalised Open Online Course (POOC)	Provides a unique learning path for each student based on a continuous evaluation of student learning (Pilli and Admiraal, 2016).	**2016**	**x**
JMOOC and KMOOC	Named after specific countries such as Japan and Korea, respectively (Soraya et al., 2019).	**2017**	**x**

2.2 Learning Analytics and Educational Data Mining

Learning analytics (LA) is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and environments in which it occurs" (Long, 2011).

Educational data mining (EDM) is the process of applying computerised methods, such as machine learning and data mining, to an enormous volume of educational data. EDM and LA both use educational data to provide recommendations and advice to stakeholders (Liñán and Pérez, 2015).

Thus, LA is arguably mainly aimed at human consumption, whereas EDM is mainly aimed at computer processing. However, the boundaries are not very strict. In terms of applicable techniques for educational data, most are appropriate for both EDM and LA and encompass statistical methods, data mining, machine learning, network analysis and visualisation. The four techniques often used by both are as follows (Liñán and Pérez, 2015). Clustering methods are used to categorise groups of learners according to similar features. Prediction techniques are used to estimate a target variable on the basis of existing data of other variables. Relationship mining techniques are used to identify the relationships between variables such as learner behaviour and difficulties. Text mining is used to go through, extract, and analyse valuable contents from texts in, for example, web pages, documents, chats, and forums (Romero and Ventura, 2020).

In terms of popularity, the research trend is gradually moving towards LA rather than EDM, although both areas are still growing (Liñán and Pérez, 2015) (see Figure 2.1 and Figure 2.2).

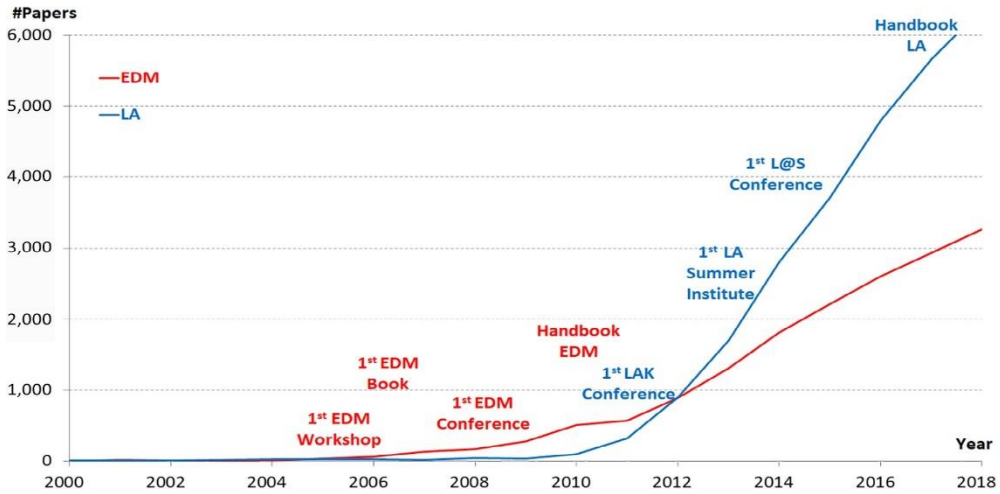


Figure 2.1 Number of papers and main events about EDM and LA according to (Romero and Ventura, 2020).

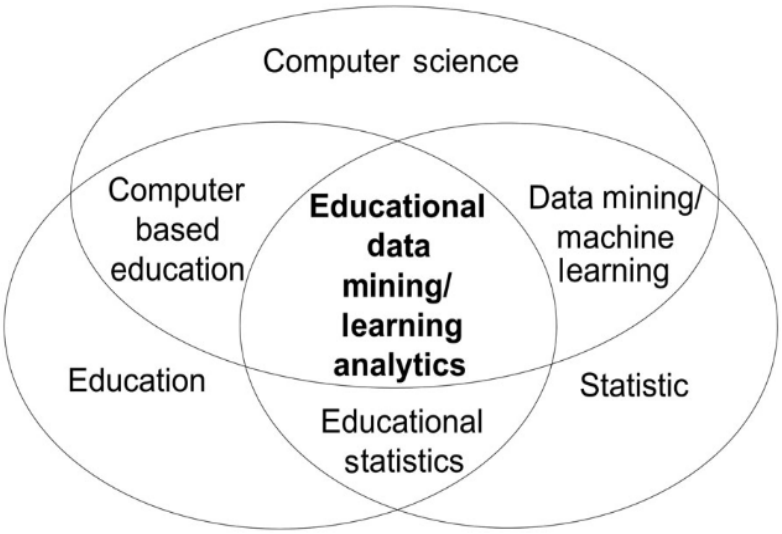


Figure 2.2 Main areas related to Learning Analytics and Educational Data Mining (Romero and Ventura, 2020)

2.3 MOOCs Datasets

MOOC platforms generate enormous datasets, which contribute significantly to the expansion of knowledge in the field of educational data mining (EDM). The data from MOOC platforms can be invaluable sources of information to assist researchers, educators, practitioners, and decision-makers in improving the quality of online learning (Alharbi and Jacobsen, 2014).

Generally speaking, educational data are subject to stringent privacy laws, which are designed to secure the confidentiality of student information. Therefore, researchers are often prevented from making their data publicly accessible under legal restrictions (Andres-Bray, 2021).

Table 2.2 An overview of the features provided by different MOOC platforms (Thakkar and Joshi, 2015)

Learning Methods	edX	Coursera	Udacity	MEC	FutureLearn	Canvas Network
Video with audio	✓	✓	✓	✓	✓	✓
Audio only	×	×	×	×	✓	×
Articles	✓	✓	×	×	✓	✓
Projects	×	×	✓	×	×	×
Discussions	✓	✓	✓	✓	✓	✓
Assignments	✓	✓	✓	✓	✓	✓
Quiz Tests	✓	✓	✓	✓	✓	✓
Transcripts	✓	×	✓	×	✓	×
Video with interactive transcripts	✓	×	×	×	✓	×
Certificate	✓	✓	✓	✓	✓	✓
Adaptive Learning	×	×	×	×	×	✓

Another challenge for researchers is the lack of a consistent data format across MOOCs. Here, we provide a quick overview of the raw data that can be pre-processed and used to predict student performance in a MOOC, including its typical forms, structure, and behavioural patterns. However, the raw data types (e.g. videos, quizzes, and adaptive learning) are not supported or generated by all MOOCs platforms (Table 2.2).

Each MOOC platform provides different features and uses different formats to represent its data. For example, Table 2.2 shows that all platforms have video together with audio, discussions, quizzes, assignments, and certificates. Articles are available on edX, FutureLearn, Coursera, and Udacity, while projects are available on Udacity. FutureLearn provides audio-only content, while transcripts are available on FutureLearn, edX, and Udacity. FutureLearn and edX deliver video with interactive transcripts. However, the only platform that provides adaptive learning is Canvas.

Therefore, analysing the data gathered from MOOC platforms can be challenging, as it requires many pre-processing steps (Gardner and Brooks, 2018). Consequently, in 2013, Veeramachaneni issued a call in his article, ‘Moocdb: Developing data standards for MOOC data science’, to develop data extraction techniques that would overcome the problem of data sharing. The author estimated that MOOC raw data processing takes up to 70% of the time

spent on analysis, which is much higher than the time spent on model development. To address this issue, the author developed a method known as MOOCdb. The aim was to develop data extraction tools that would operate with a regulated MOOC database schema (Veeramachaneni et al., 2013, Lohse et al., 2019). However, there is still a low number of publicly available datasets relevant to the field of MOOC research. Table 2.3 shows the publicly available datasets for MOOCs.

MOOCs data	Dataset description	#Students	# Courses
Standford MOOCPosts	Forum Discussion (students' posts)	29,604	6
KDD Cup 2015	Students learning log for 30 days	120,542	39
Khan Academy	Students exercises	47,495	-
OULAD	Learning Analytics	32,592	22
Coursera Forums	Forum Discussion (831,576 posts)	-	73

Table 2.3 Summary of publicly available datasets for MOOCs

2.4 Machine Learning (ML)

Artificial intelligence (AI) is a component of computer science that attempts simulate the mental processes of humans. Machine learning (ML) is a subset of artificial intelligence, which is a subject that encompasses a wide range of technologies. Fundamentally, machine learning is the act of providing a computer or model with access to data and allowing it to learn on its own. The aim of machine learning is to have a computer 'learn' via data or experience and then utilise that knowledge to solve a specific issue (Ciolacu et al., 2017). Since its inception, machine learning algorithms have been built specifically to tackle medical data sets (Adamopoulos et al., 2009). But over the past several years, many useful applications of machine learning appeared across various industries due to the digital

revolution that made data collection and storage more widely accessible (Ciolacu et al., 2017).

Since electronic computers began to be used in the 1950s and 1960s, algorithms were created that allowed for the modelling and analysis of enormous amounts of data. The starting point was when machine learning evolved into three main branches. The first branch is the neural networks by Rosenblatt in 1962; Rosenblatt is sometimes known as the "Father of Deep Learning". The second type is the statistical method by Nilsson in 1965. The third is symbolic learning by Hunt in 1966. Over time, all three branches came up with ever more sophisticated techniques; Table 2.4 provides examples for each type (Kononenko, 2001).

Table 2.4 Examples of Statistical, Neural Networks and Symbolic learning methods

Statistical Methods	**Neural Networks**	**Symbolic learning in**
k-nearest neighbours	Feedforward neural network with backpropagation learning	Decision trees and
Bayesian classifiers	Hopfield's associative memory	Decision rules
Discriminant analysis	Kohonen's self-organising network	Induction of logic programs

It is more challenging to investigate possible solutions to particular problems when people make decisions during the analysis phase or more precisely, investigate the relationship between numerous features in a large dataset. Therefore, machine learning may often be used effectively to increase system efficiency and solve these issues (Muhammad and Yan, 2015).

Statistics methods and machine learning have the same objective, but the two approaches vary. Statistical methods are considered to be mathematical models based on a hypothetical test where human judgment is needed to conclude the correlation between variables. On the other hand, machine learning allows computers to learn without direct human input (Al-Shabandar, 2019).

The ability to learn a task by adopting a specific learning algorithm is a fundamental characteristic of machine learning. Generally, the dataset contains various kinds of features,

such as binary, continuous, and categorical. Therefore, the machine learning approach should be chosen based on the type of dataset (Muhammad and Yan, 2015).

In machine learning, the data is typically split into two or more datasets. A subset of the dataset set is generally used to develop the machine learning model (training phase), while the remaining data is used to evaluate the model (testing phase) (Livingston, 2005).

Two main methods are widely used in machine learning: supervised learning and unsupervised learning. Supervised learning is often used when there is a target variable for each instance in the dataset — for example, classifying whether a student passed the exam or not. In contrast, unsupervised learning is the best option for unlabeled samples, such as clustering techniques (e.g. grouping students by using their learning behaviour into various groups) (Muhammad and Yan, 2015).

2.5 Engagement in MOOCs

Engagement is a complex concept, and there are several definitions for it (Alarcon and Edwards, 2011); indeed, some authors claim that there is no single definition suitable for all contexts (Witchel, 2013). One possible definition of the engagement of students in their learning is as the behavioural, cognitive, emotional and social connections that MOOC participants make with the course content, the instructor and/or other learners (Deng et al., 2020). Engagement, regardless of its definition, has been shown to be a significant attribute towards students' learning success (Kuh, 2003). Challengingly, engagement is notoriously difficult to be achieved, especially in learning environments (Willms, 2003, Shernoff et al., 2017). In MOOCs, sustaining engagement is even more difficult, as reflected by the higher rates of dropout (Jordan, 2015, Shi and Cristea, 2018b).

2.5.1 Theories of Engagement

A classic work on Engagement Theory (ET) stated that students are engaged, when they are intrinsically motivated to learn, and the activities they perform involve active cognitive processes, such as problem-solving. The ET framework promotes three main principles: Relate, which emphasises social interactions, mainly collaborating with other peers; Create, i.e. making learning a purposeful activity, related to the students' own pace; and Donate,

where the student makes useful contributions while learning, applying the knowledge onto something practical (Kearsley and Shneiderman, 1998).

Another highly influential conceptual framework for user engagement with technology, Process of Engagement, proposed four phases tied to the engaged state: Point of engagement, where the user gets acquainted with the application; the Period of engagement, where the user is using the tool; the Disengagement, where they present reasons on why the user would stop using the tool; and the Reengagement, which is an iterating process that returns to the first phase, Point of engagement. At the Point of engagement, the users start to use an application based on its aesthetics, novelty, extrinsic motivation to accomplish a task, interest, immersive experience with the product and the Autonomy provided to use the application. The users continue using the system during the Period of engagement, where the users' experience with the system is one of the engaging factors, as are realism, customisable interfaces, fun, time perception, connection with other people and feedback. They also conceptualised some attributes that are related to the disengagement factor, such as the inability to interact, or the lack of challenges and frustration within the system (O'Brien and Toms, 2008).

Another fundamental book of high influence and extensive implementation, by Deci & Ryan, proposed the Self-Determination Theory (SDT) (Deci and Ryan, 2013). This theory suggested that human motivation is sustained by three main constructs: Autonomy, which is related to the user control over their actions; Competence, related to the skills obtained and used to perform a certain task; and Relatedness, concerned with the users' interactions with others, as they perform the given task. The theory helps investigate why humans engage in certain activities and the purpose of those activities (Hofer and Busch, 2011). The use of SDT has become commonplace in the educational domain, as the theory supports the idea that the students' intrinsic motivation is a primary determinant in their engagement (Zhou, 2016).

A similar theory, Drive, described also in a recent, authoritative book, by Pink (Pink, 2011), brings to the fore similar constructs to SDT, like Autonomy and Mastery (the latter related to Competence); however, instead of Relatedness, the author proposes the construct of Purpose, which is tied to the personal desire of the user to do something meaningful for

themselves or the community. However, SDT has far more and deeper studies and instruments that can be used, e.g., for intrinsic motivation.

One of the latest proposals by Marczewski (Marczewski, 2015) merges the above two theories into a new intrinsic motivation theory, where all four concepts are supported (Autonomy, Mastery, Purpose and Relatedness). However, this theory has yet to gain much support (in terms of citations).

Epilogue

In this chapter, we have presented the background of distance learning and discussed its development, advantages and challenges. Furthermore, we provided an overview of MOOCs' platforms, data formats and the publicly available dataset. Finally, we discussed the theories of engagement. In the next chapter, we will review the current predictive models employed in the literature to predict at-risk students in MOOCs. In addition, we will present motivation studies in MOOCs.

Chapter 3 : Related works

Prologue

This chapter provides an overview and discussion of the current predictive models employed in the literature to predict students at-risk of dropping out from MOOCs (see section 3.1). In addition, this chapter presents engagement and motivation studies (see section 3.3). Finally, this chapter critically evaluates the published literature (see section 0).

3.1 Machine learning application in MOOCs

How to measure success in MOOCs is a debatable issue (Davis et al., 2013). Whilst completion is the most frequently used parameter for success (Mohamed and Salleh, 2021, Loizzo et al., 2017), it is not unanimously agreed upon as the best way of measuring the perceived, or even the actual success of MOOC students. Students may have different objectives when they embark on a MOOC-journey. Students may wish to learn a new topic in its entirety, but also, they may wish to use the obtained knowledge for different aims.

3.2 Related works

This thesis places significant importance on conducting a literature review pertaining to the use of ML methods for the purpose of predicting students who are at risk of dropout. In addition, the review has the potential to identify prospective directions for further investigation and the enhancement of more efficacious predictive models.

3.2.1 Previous literature surveys on MOOCs

Earlier literature surveys on MOOCs have covered a variety of pedagogical concerns, including the modelling of learning and evaluating in MOOCs (Joksimović et al., 2018). In a comprehensive evaluation of MOOC research in Mainland China (Cheng et al., 2022). Zhu et al. (2022) conducted a survey study focusing on trends and concerns in the empirical study

of MOOC learning analytics, providing a comprehensive mapping review of MOOC recommendation systems (Uddin, 2021). Saadatdoost (2015) carried out a survey study to explore MOOCs from the perspectives of education and information systems. There are several surveys focusing on machine learning methods for student dropout prediction in MOOCs. For example, Albreiki (2021) conducted a comprehensive assessment of the research on machine learning approaches for predicting student performance. The author exploited six research databases and collected 78 studies from 2009 to 2021. The study focused on different methodologies, such as early prediction, recommender systems, and dynamic approaches.

Dalipi et al. (2018) provided an overview of current studies on applying machine learning to predict, understand, and address the issue of student dropout in MOOCs. Only 25 studies conducted between 2013 and 2017 were reviewed in this survey.

In another survey paper by Moreno-Marcos (2019), the author used ISI Web of Knowledge and Scopus as data sources to extract 82 articles between 2014 and 2017.

3.2.2 Differences between the present and previous surveys.

One of the primary distinctions between the current survey and its predecessors is that the number of studies included in this review is larger than that of the previous reviews. A total of 950 studies were extracted from eight databases for education research, including Education Research Complete, Web of Science, Scopus, Emerald Insight, ERIC, Taylor & Francis, IEEE Digital Library, and ScienceDirect, resulting in 127 articles that were qualified for inclusion in the review. This enabled us to identify patterns and trends in the literature. A further difference between this survey and its antecedents is that we classified the evaluated studies into four categories based on their prediction targets (completion, performance, comments, and certificate). This approach offers a more comprehensive analysis of the literature and compares studies of machine learning methods for the prediction of at-risk students based on the prediction targets.

Moreover, to evaluate the temporality of prediction, we assessed the studies' prediction targets based on the timing of the prediction. This is an essential point, as previous literature

indicates that participants in MOOCs are more likely to opt out of a course during its initial weeks (Codish et al., 2019, Hung et al., 2015). Consequently, early intervention is essential to identify at-risk students early on, preferably within the first week of the course. Finally, another difference between this survey and previous ones is the evaluation of predictive models based on the possibility of putting them into practice in real-world settings (see Section 3.2.9). Assessing predictive models in real-world environments can shed light on the practicability and efficacy of implementing these models. This helps us evaluate the predictive models based on their applicability in real-world contexts, which is essential for strengthening their effectiveness and applicability.

3.2.3 Inclusion and Exclusion Criteria

In this section, we present a comprehensive review of the literature on the application of machine learning techniques to predict students at risk of dropping out from MOOCs.

A total of 950 studies were extracted from eight databases for education research. The filtering procedure removed 823 studies, leaving 127 that satisfied the inclusion criteria. The search keywords (‘MOOC’, ‘dropout’, ‘machine learning’, and ‘prediction’) were entered into each database.

This survey consists of academic journal articles and conference papers that satisfy particular requirements, such as being published in English and being peer-reviewed, to preserve rigorous research standards and credibility. Forming criteria for the inclusion and exclusion of studies in a review paper is a crucial step in ensuring that the final outcome is based on high-quality research.

In this review, the following procedures were employed to identify and select the most relevant studies: After conducting the initial search, along with eight additional database searches, there were a total of 950 papers. After this, the researcher eliminated any duplicate records. This phase is essential for avoiding duplication of records and ensuring that the results are based on original research. We used the EndNote reference management application to eliminate duplicates from the results. This step eliminated 253 duplicates, leaving a total of 697 studies.

Review papers were excluded, as generally speaking, they do not demonstrate a high level of originality, as their aim is to introduce the reader to the existing literature, including its gaps and limitations(Yarbrough, 1991). Therefore, we excluded any review papers that were identified during the search. This step resulted in the removal of 57 studies, leaving 640 studies for further screening. Finally, we scanned the abstracts of the research papers to determine those that employed machine learning algorithms to predict at-risk students in MOOCs by obtaining data from MOOC platforms. We concentrated on four primary factors to select the studies: (a) prediction of at-risk students in MOOCs based on their prediction targets (completion, performance, comments, and certificate), (b) data sources, (c) prediction models and evaluation metrics, and (d) temporal prediction techniques (see Sections 3.2.6, 3.2.7, 3.2.8, 3.2.10). All studies that did not include machine learning, experimentation, or technique validation were excluded from this review. The screening procedure eliminated 513 studies, leaving 127 that satisfied the inclusion criteria.. Figure 3.1 shows the flow of the studies through the review process.

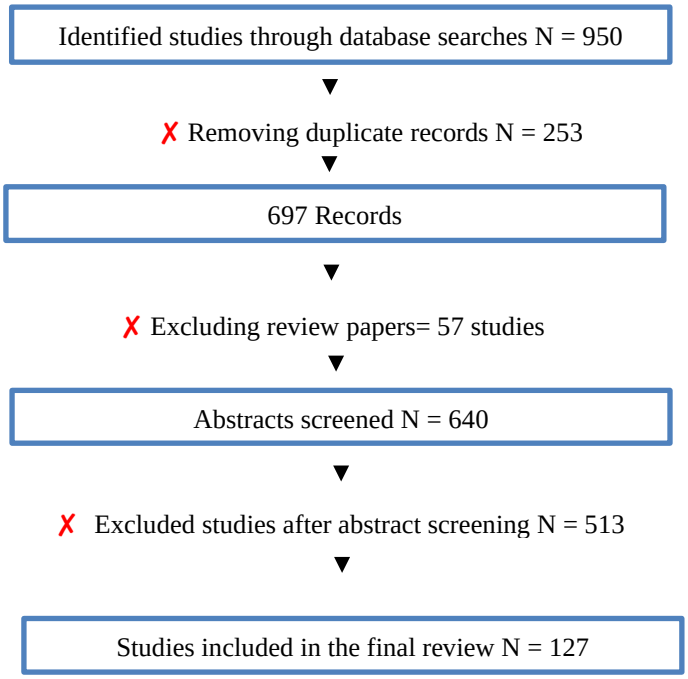


Figure 3.1 Flow chart of the review process

3.2.4 Students' at-risk predictions in MOOCs

ML models have shown the abilities to analyse and interpret complicated data in a broad range of applications. As MOOCs and other e-learning platforms provide a wealth of data unavailable in conventional school environments, many scholars have taken an interest in the prediction of at-risk students using ML (Dalipi et al., 2018, Chen et al., 2022).

In this survey, we reviewed works that contributed to predicting students at risk of dropping out from MOOCs using ML. The main focus was to compare the prediction targets and tools, methods, and datasets used. However, several factors complicated the head-to-head comparisons of predictive performance reported in the reviewed studies. For example, all researchers should use the same dataset, metrics, and protocols and predict the same output. In addition, many researchers cannot share their datasets because of legal limitations. Therefore, researchers on MOOCs have focused heavily on feature engineering from various data sources, and many breakthroughs in predictive modelling have relied on state-of-the-art feature extraction approaches.

3.2.5 Predictive models of students at-risk of dropping out from MOOCs: input features, ML, outcome, and dataset

Several works have attempted to answer the question of how to identify students potentially at risk of dropping out from MOOCs. For example, Monllaó used ML and extracted the learning behaviours (e.g. attempting a quiz and forum posting activities) of 46,895 students from eight courses to predict whether the students would participate in the last quarter of the course. This study focused on predicting dropout students in three phases: after the first quarter of the course, at the middle of the course, and at three-quarters of the course (Monllaó Olivé, 2020).

Another study conducted by Ye (2015) defined dropouts as students who accessed fewer than 10% of the course and did not have assessment activities. The author extracted data on the learning activities from the first week of the course (e.g. video, quiz, and peer-graded assignments data) and applied several ML models such as logistic regression (LR), support vector machines (SVM), and decision trees (DTs).

A small-scale study by (Xing, 2019b) presented a dataset from one course for 2084 students. The author focused only on the students' activities in the discussion forum and the emotional polarity of the students' posts (e.g. students' comments, positive comments, negative comments, and thread started). The target was to predict students who did not post comments in the last week of the course. This study used several ML modes such as naive Bayes (NB), LR, SVM, and DTs.

Sun et al. (2019) applied random forest (RF), gradient boosting (GBM), and XGBoost models to a longitudinal dataset for 12,847 students. The authors used different input features (e.g. number of access times, online time, session time, duration of watching the video, and forums visited) to predict the percentage of the course content completed for each student. This study identified students at risk of dropping out from an early stage by using a weekly prediction technique (prediction from weeks 1 to 12).

Many researchers (Şahin, 2021, Alsolami, 2020, Teruel, 2018, Mubarak, 2021c, Hong, 2017, Jin, 2021, Ardchir, 2020) have proposed different prediction models by using a public dataset known as XuetangX/KDD Cup 2015, which was provided by a Chinese MOOC learning platform initiated by Tsinghua University. In this dataset, students were considered dropouts if they did not have activities after 30 days from the official starting date of the course.

In another study, Xing et al. (2019) proposed a framework to predict whether a student will drop out in the following week on the basis of the weekly prediction technique (from weeks 1 to 7). The authors used a dataset extracted from eleven courses for 3,617 students with several input features such as the number of times course contents were accessed, number of assignments visited, number of assignments submitted, number of times the calendar was accessed, and number of times quizzes were accessed.

A study by Chen et al. (2016) used three classification algorithms (LR, RF, and k-nearest neighbours [KNNs]) to predict student dropouts in week 5. The dataset was extracted from two courses for more than 34,000 students. The predictive models were trained according to the students' activities in the first month of the course. In this study, the author used several features such as the number of watched videos, number of active days, number of rewatching records, number of skip records, and number of posts.

Drousiotis et al. (2021) extracted a dataset for more than 32,000 students from seven courses and applied three predictive models (RF, DTs, and BART) to predict four classes: withdrawn, fail, pass, and distinction. In addition, the author transferred multi-class datasets to binary classification problems using the one-versus-rest method. Several features were used in this study based on learning activity and demographic attributes (e.g. number of clicks until the course starts, first assignment mark, registration date, age, disability, and gender). In a recent study, Nitta et al. (2021) used the same dataset as that used in the study by Drousiotis et al. (2021) but implemented different ML models. The author applied graph convolutional networks to predict two classes: completion (distinction or pass) and dropout (fail or withdrawn).

Kim et al. (2018) used a dataset extracted from the Udacity platform for two courses and 10,154 students. In their study, the authors applied a deep learning model (long short-term memory) to predict whether the student will graduate from an early stage by using a weekly prediction technique (from weeks 1 to 8). The predictive models rely on the student's learning activities (e.g. number of videos watched, number of reading a text page, number of quiz attempts and grade) in previous weeks.

Babu et al. (2017) proposed a framework to predict whether the student will participate in the course until the last week and solve the final exercises. Selected predictors were used, such as the number of days of access, number of access times in the last two weeks, number of events in the final week, number of pages closed, and number of videos watched. In this study, the author gathered data for more than 200,000 students from 39 courses and then applied several ML classifiers (e.g. RF, DTs, NB, LR, SVM, DTs, GBM, and ensemble models).

Another study by Vitiello et al. (2017) applied SVM to predict whether the student will complete the final project and exam on the basis of a weekly prediction technique (from weeks 1 to 8). The authors used student tracking and discussion forum data extracted from 11 courses for 3,213 students.

Bote-Lorenzo et al. (2017) aimed to determine whether students' levels of engagement in the next chapter would be (higher/lower) than that in previous chapters. The authors extracted data for 26,947 students from one course and applied a set of ML algorithms such as the

stochastic gradient descent, RF, LR, and SVM. These predictive models were trained according to the students' learning behaviours (e.g. percentage of videos watched in each chapter, percentage of assignments submitted, and total grade of assignments).

A recent study by Alami et al. (2021) used the OULAD dataset, which included 32,593 students extracted from 7 courses. The authors used students' demographic and clickstream data as input features to predict four categories (withdrawn, fail, pass, and distinction). Another study by Xia et al. (2020) used part of the OULAD dataset (6,360 students from one course) to predict three categories (withdrawn, fail, and pass). However, the authors used only the clickstream features to train the predictive model in this study.

Cobos et al. (2017) applied LogitBoost, GBM, XGBoost, and KNN models to a longitudinal dataset for 12,465 students extracted from two platforms (FutureLearn and edX). The main task was to predict whether the student would complete 50% of the course on the basis of the weekly prediction technique (from weeks 1 to 7). This research used students' behavioural data such as number of access times, number of active days, total time spent on quizzes, and number of interactions in discussion forums.

In their study, Yu et al. (2021) used a dataset extracted from the OpenEdu platform for one course and 1,387 students. The authors proposed a framework to predict students' learning performance. First, they classified students' grades into a binary class (pass/fail) and then applied three classification algorithms (e.g. SVM, artificial neural network [ANN], and recurrent neural networks [RNN]). These predictive models were trained according to students' learning behaviours such as the numbers of videos watched, paused, and stopped, and clicking speed and correct answers.

A study by Moreno-Marcos et al. (2020) used a large dataset of 142,733 students. The dataset was extracted from three sources: 1) forums (e.g. the number of posts and votes, the positivity of the messages, threads a learner started, and the average number of characters in posts), 2) exercises (e.g. exercises attempted, exercises opened, correct exercises), and 3) learners' activities (e.g. number of video repetitions; number of times of access from a PC and during the weekend; number of videos watched, paused, and stopped; and clicking speed). They then applied five classification algorithms (DTs, RF, multi-layer perceptron [MLP], AdaBoost, and NB) to predict learners' performance in assignments (pass/fail).

Qu (2021) used data extracted from one course (C programming) for 1,528 students. The aim of the study was to investigate the usefulness of students' programming activities (e.g. invalid use of ‘struct data’, syntax error, and illegal statement in code). The author applied four prediction models (NB, LR, SVM, and MLP) to predict whether the student would pass the course.

Qiu et al. (2016) proposed the LadFG (latent dynamic factor graph) model, which takes into account the demographics, forum activity, and learning style of students. The author used a dataset from XuetangX, one of the largest MOOCs in China. The aim of their study was to predict students' performance on assignments and which students will obtain a certificate by the end of the course.

Another study by Wang (2016) used a dataset from the same source (XuetangX) for 2,633 students. The back-propagation neural network and linear regression algorithms were applied to predict students' grades on the basis of the weekly prediction technique (prediction from weeks 1 to 14). The author used students' demographics, videos, and exercise data (e.g. age, gender, watching ratio, and submission times).

A study by Kameas et al. (2021) used a dataset from DevOps MOOC for 936 students. The dataset consisted of students' clickstream and demographic features (e.g. number of video views, gender, educational level, number of posts, and number of assignments). The author applied several ML models such as AdaBoost, GBoost, Extra Trees (ExTree), LR, DT, linear discriminant analysis, and LightGBM to predict whether the student will pass the final exam and obtain the certificate. This study focused on predicting students in three phases. The first phase (week 0) was based on information collected before the beginning of the course, whilst the second and third phases (weeks 1 and 2) were based on information collected by the end of weeks 1 and 2.

Another study by Rawat(2021) sought to develop a predictive model for determining whether a student would complete a course. Several algorithms, such as deep learning, decision tree, logistic regression, and gradient boosting, were employed in this study. The prediction model included the variables start time, age, number of active days, gender, and forum posts. The dataset used in the study was obtained from two universities, Harvard and MIT, and included information from 367,375 students enrolled in 13 different courses.

Yu's study 2021 sought to construct a prediction model to predict students' course grades by using deep learning and machine learning algorithms, including ANN, KNN, and SVM. The research used a dataset from the OpenEdu platform that included 1,387 students who registered for one specific course. The study's prediction features included watching videos, pausing and stopping them, clicking speed, taking tests, and providing correct answers.

Edalati (2022) conducted a study to develop a prediction model for classifying student reviews into positive, negative, and neutral categories, using different machine learning and deep learning models, including RF, SVM, DT, CNN, and BERT. The study utilised a dataset obtained from the Coursera platform, consisting of reviews from 15 different courses.

Zhang (2021) carried out research to develop prediction models for identifying active students 30 days after the start of a course, using several machine learning and deep learning models, including RF, CNN, and LSTM. The study used a dataset from XuetangX, containing information regarding 120,542 students enrolled in 39 courses.

Qu (2021) conducted a study to build prediction models using various deep learning and machine learning algorithms, including NB, LR, SVM, and MLP, to predict whether students would pass or fail a C programming MOOC. The study extracted features from the programming activity of 1,528 students, such as invalid use of struct data, syntax errors, and illegal statements. The dataset was divided into a training set consisting of 80% of the data and a test set consisting of 20% of the data.

Hlioui (2021) carried out research by using the OULAD dataset, which consisted of data from 1,303 students enrolled in a single course, with the goal of building prediction models to predict withdrawal from the course. The study used demographic data, assessment scores, and four behavioural indicators: perseverance, autonomy, commitment, and motivation based on student interaction. The study employed various machine learning algorithms and deep learning algorithms, such as RF, DT, SVM, Bayesian classifier, and MLP, and assessed the performance of the algorithms using the F1 score metric. The results were validated using a five-fold cross-validation method.

Table 3.1 provides a comprehensive summary of the above-mentioned studies and additional studies that have been conducted in the area of predicting at-risk students in MOOCs from 2015 to 2022.

Table 3.1 Summarises the reviewed studies to predict at-risk students based on machine learning techniques from 2015 to 2022.

	Author	Platform / dataset	#students	#Courses	Input features	ML methods	Performance Metrics	Train/ Test sets	Prediction target	Temporal
1	(Şahin, 2021)	XuetangX	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	NN, FIS	PR, REC, F1, ACC	5 fold CV	Predict active students after 30 days.	✗
2	(Ai, 2020)	XuetangX	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	SVM, AdaBoost, NB	Rec ,F1 , Auc	NA	Predict active students after 30 days.	✗
3	(Chen, 2021)	XuetangX	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	**CNN LSTM,** LR, SVM .DT	AUC, Pr	NA	Predict students who do not participate in the following week.	From the first week up to week 5
4	(Mrhar, 2020)	Edx	11607	1	Clickstream and sentiment features (eg. Answered , comments, pages viewed, subsection viewed)	KNN, SVM ,DT, ANN	ACC, AUC	NA	Predict students who do not participate in the following week.	Prediction from week 1 to week 12
5	(Alsolami , 2020)	XuetangX	NA	10	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR,NB,DT,RF, SVM, GBoost	PR, REC AUC	10-fold CV	Predict active students after 30 days.	✗
6	(Olmos, 2018)	edX UAM	10919	3	Clickstream num events, connected days, video, forum events, forum time)	LR, RF, GBoost,NB, XGBoost, NR,SVM, KNN	AUC	75% TR_Set 25 % TS_set	Dropout prediction (Student tagged as a dropout when the number of activity days is less than six days).	Prediction from day 1 to day 60
7	(Ahmad, 2021)	OULAD	NA	NA	Demographic and Assessment Data Age, Gender, Region, Assessment Scores And Date of Submission	**ANN**	ACC, PR, RE, F1	10-fold CV	Predict the final result (Distinction, Pass or Fail).	✗
8	(Mourdi, 2021)	OpenEdx	3585	1	Performance data, navigation data, forum interaction data.	**MLP**, LR, NB, KNN, DT	ACC, PR, RE, F1	70% TR_Set 30 %TS_set	Predict students who do not participate in the following week.	Prediction from week 1 to week 9

9	(Wang, 2016a)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR, LR, LSTM and NSSM	AUC	70% TR_Set 30% TS_set	Predict students who do not participate in the following week.	Prediction from week 1 to week 5
10	(Batool, 2021)	OULAD-schools data	24086	NA	Demographic attributes (gender, age, family, parents' jobs, travel time to school, study time, health status etc.)	RF	F1, ACC	80% TR_Set 20% TS_set	Predict whether the student will Pass or Fail the course.	✗
11	(Rong, 2019)	China MOOC platform	NA	NA	Textual data 19,148 comments	NTUSD, **BERT HAN, CNN, Transformer**	ACC,F1	64% TR_Set 16%,validate 20%TS_set	Predict the sentiment polarity of students' comments.	✗
12	(Monllaó Olivé, 2020)	Moodle	46,895	8	Learning activity data, textual data	LR	average ACC ,F1	80% TR_Set 20% TS_set	Students who do not participate in the last quarter of the course.	25%,50% and 75% of the course
13	(Rawat, 2021)	Harvard and MIT	367375	13	Start time age, ,ndays_act, gender , forum posts etc)	**DL**, DT, LR and GBoost	ACC	75% TR_Set 25 % TS_set	Whether a student has completed the course.	✗
14	(Hlioui, 2021)	OULAD	1303	1	Demography, Assessments Scores and Behaviour(Four Behavioural Indicators, Such As Perseverance, Autonomy, Commitment, And The Motivation Based On Student Interaction)	RF, DT, SVM, Bayesian classifier, **MLP**	F1	5-fold CV	Withdrawal /completion.	✗
15	(Liu, 2020a)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	NB, DT, LR, **MLP**, GBoost, XGBoost	F1, ROC	5 fold CV	Predict active students after 30 days.	Prediction from week 1 to week 4
16	(Yu, 2021)	OpenEdu	1387	1	Learning activity data(video viewing, pause, stop, clicking speed, trying tests, correct answers etc)	ANN, KNN, and SVM	ACC	70% TR_Set 30% TS_set	Course grade.	✗
17	(Okereke, 2020)	Moodle	3617	6	Posts, video pages,quiz attempts, time spent on quizzes, views on lecture pages, views on assignment pages, view on wiki pages	RNN, GRU, LSTM	ACC	80% TR_Set 20% TS_set	Predict whether a student will take the next activity or not.	✗
18	(Narayanasamy, 2020)	OUs of China	22160	1	Self-characteristics and academic performance (gender, marriage status,age ,country,studied courses, average test results etc)	SVM, RF, CRF	ACC, PR, REC, and F1	60% TR_Set 40% TS_set	Dropout prediction.	✗

19	(Bote-Lorenzo, 2018)	edX	26947	1	Video watching activity regarding exercises assignments activity	LR	AUC	In situ prediction models	Predict if the engagement indicator decreases at the end of each chapter.	Chapter 1 to Chapter 11
20	(Tubman, 2018)	OULAD	32593	7	Clickstream Data (Lecture Video and Forum Data)	TSF	ACC	10-fold CV	Predict students' withdrawal from the course.	Start prediction with the first 5% of the data, then sequentially add 5% until the end of the course.
21	(Periwal, 2017)	edX (MITx and HarvardX)	1,69,621 - 66,151	2	Learning Activity Data and Demographic Attributes (Last Degree , Country , year of birth , Nchapters, Nforum, Ndays, Nvideos, Nevents, Ndays, Etc)	**KNN**. NB, DT, LR	ACC,confusion matrix	10-fold CV	Dropout prediction	✗
22	(Liu, 2018)	XuetangX	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR, SVM, RF	AUC PR, REC, F1	10-fold CV	Predict active students after 30 days.	✗
23	(Pei, 2021)	OULAD	5182	4	Demographic and Learning Activity(Age, Gender Educational Level, Region, Attempts)	RF,SVM,DT	ACC, PR, REC, and F1	80% TR_Set 20% TS_set	Predict students who do not participate in the following week.	Start prediction from week 5 up to week 35.
24	(Moreno-Marcos, 2018b)	edX	4358	1	Learning activity, graded assignments, forum data (main posts, reply posted, average positivity / negative messages)	RG,SVM, DT ,RF	AUC, F1	10-fold CV	Predict whether the student will Pass or Fail the course.	Predict seven graded assignments
25	(Moreno-Marcos, 2020b)	edX Java Programming	142733	2	Posts, Reply,Characters of Posts, Positivity, %Exercises Attempted, Completed Videos, Video Pauses, %Accesses During Weekend.	RG,SVM,DT ,RF	AUC	10-fold CV	Predict learners' performance in assignment grades	✗
26	(Panagiotakopoulos, 2021)	DevOps MOOC	961	1	Logins First Two Days and Demographic (Gender, Age, Nationality, Country, Mother Tongue, Education Level,Current Job)	DT, RF, MLP, AdaBoost, NB, LR	ACC	10-fold CV	Predict whether students will start a MOOC	✗

27	(Wenqing , 2020)	XueTang X	NA	28	Posts,%Video View, %Video Has Completed, Average Pauses of A Video, Average Repetitions of A Video.	DT, RF, SVM, LR, **MLP**, XGB	AUC	70% TR_Set 30% TS_set	Predict whether the student will Pass or Fail the course.	✗
28	(Dascalu, 2021)	Moodle	319	2	Learning activity data/forum posts	RNN	RMSE, R2	10-fold CV	Student's grades.	✗
29	(Qu, 2021)	C programming MOOC	1528	1	Features from programming activity (invalid use of 'struct data, syntax error, illegal statement in code)	NB, LR SVM **MLP**	ACC, REC	80% TR_Set 20% TS_set	Predict whether the student will Pass or Fail the course.	✗
30	(Ye, 2015)	Coursera	NA	2	Learning activity data(video, quizzes, peer-graded assignments)	LR, SVM, DT	F1	NA	Predict Students who accessed fewer than 10% of the course and do not have assessment activities.	✗
31	(Xing, 2019b)	Coursera	2084	1	Fourm data (avg post, avg positive & negative posts, avg thread started)	NB, LR, SVM, DT	ROC_AUC, Kappa	10-fold CV	Predict students who do not post comments in the last week.	✗
32	(Zheng, 2016)	Edx	200000	39	Enrollment Feature, User Feature, and Coursefeature	LR, SVM, RF, GBoost	AUC	5-fold CV	Predict active students after 30 days.	✗
33	(Lai, 2020)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	**ANN**	ACC	10-fold CV	Predict active students after 30 days.	✗
34	(d'Invern o, 2017)	Coursera programming course	993	1	Weekly activities of students(play videos, view set of shared files, view thread, comment when a video is playing etc)	SVM	ACC	50% TR_Set 50% TS_set	Predict whether the student will Pass or Fail the course.	Prediction from week 1 to week 7
35	(Khodeir, 2021)	Stanford	NA	11	29604 posts	**BERT**	AUC	Transfer learning	Predict students' urgent posts.	✗
36	(Wei et al., 2017)	Stanford	NA	11	29604 posts	**CNN** and **LSTM**	ACC	Transfer learning	Predict (Confusion, Urgency, Sentiment) of students' posts.	✗
37	(Yang, 2021)	icoursel63	NA	NAa	20000 posts	Glove, Word2vec, Skip-Gram	ACC	80% TR_Set 20% TS_set	Predict the emotional polarity of students' posts.	✗

38	(Tóth, 2018)	NA	1370	1	Mouse Behaviours (Move, Scroll, Click), Video Watching Attitudes and Text Inputs	**MLP**,**KNN** ,NB, RF, XGBoost, SVM, ctree	ACC, PR, REC, and F1	leave-one-out cross	Predict whether the student will Pass or Fail the final exam.	✗
39	(Wu, 2019)	Xuetang X	120542	39	Learning activity data	**CNN-LSTM-** SVM	AUC, PR, REC,F1	80% TR_Set 20% TS_set	Predict active students after 30 days.	✗
40	(Fu, 2021)	Xuetang X	12000	39	Learning activity data	**CNN, LSTM**	ACC, PR, REC, F1	80% TR_Set 20% TS_set	Predict active students after 30 days.	✗
41	(Teruel, 2018)	Xuetang X	120542	39	Learning activity data	**LSTM**	AUC,RSME, R2	70%TR_Set 10%,validation 20%TS_set	Predict active students after 30 days.	✗
42	(Liu, 2020b)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	DT ,NB, LDA, LR, SVM, RF, GBoost	ACC, PR, REC, F1	NA	Predict active students after 30 days.	Four stages of prediction, each stage includes ten days.
43	(Siddique , 2020)	OULA	32593	7	Demographic Info, Assessment Info and Interaction Data	DT,RF, **KKN**,LR	PR, REC, F1 and AUC	NA	Predict whether the student will Pass or Fail the course.	✗
44	(Mubarak , 2021c)	Xuetang / Stanford	199165	44	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	**CNN-LSTM , DNN,** SVM, LR	AUC, PR, REC, F1	65%TR_Set 15%,Val_Set 20%TS_set	Predict active students after 30 days.	✗
45	(Ren, 2021)	Xuetang/ MOOCCube2020	10421	1	Video clickstream data (number of views, play back, video start time, video end time etc)	**CNN**-SVM-LR-**DNN**	ACC, F1 REC	NA	Dropout prediction.	✗
46	(Sun, 2019)	Xuetang X	12847	1	Total online time, Eession Time, Duration of watching the video, Forums Visited, Access etc)	RF, GBoost, XGBoost	squared (R2)	80% TR_Set 20% TS_set	Percentage of course content completed in the whole course.	Prediction from week 1 to week 12
47	(Wang, 2017)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	SVM, LR, RF, DT, AdaBoost, Gboost, CNN	AUC, PR, REC, F1	80% TR_Set 20% TS_set	Predict active students after 30 days.	✗

48	(Hong, 2017)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	RF, SVM, MLR	AUC, PR, REC, F1, ACC	80% TR_Set 20% TS_set	Predict active students after 30 days.	✗
49	(Qiu et al., 2016)	xuetangx	88112	11	Gender, age, education level, post, replies received, replies received from well-performed students, chapters accessed, total time spent on videos, total time spent on assignments)	LR,SVM, FM LadFG	PR, REC, F1, Acc and AUC	60%TR_set 40%TS_set	1) Predicts students' performance on an assignment. 2) Predict whether the student will obtain a certificate.	✗
50	(Ortigosa, 2018)	edX	3353	1	Videos, time watching a video, videos backward movements, correct answer, comments etc.)	NB, DT, SVM, **LSTM**	ACC	NA	Predict whether the student will Pass or Fail the course	✗
51	(Xing, 2019a)	Canvas	3617	11	Acess the course access assignments, submit assignments, view the calendar, quizzes)	**KNN**,SVM, DT,**NN**	AUC,ACC	70% TR_Set/ 30% TS_set	Predict whether next week is the dropout week.	Prediction from week 1 to week 7
52	(Jin, 2021)	Xuetang X	53596	6	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	DT LR, RF , AdaBoost and SVM	AUC	67% TR_Set 33% TS_set	Predict active students after 30 days.	✗
53	(Fauziati, 2019)	Xuetang X	42490	5	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	Ensemble Learning	PR, REC, F1	70% TR_Set 30% TS_set	Predict active students after 30 days.	Prediction in 5 different time periods.
54	(Zhang, 2020b)	NA	26962	5	View course progress times, posts videos watched,replies etc	LR	ACC	10-fold CV	Predict whether the student will Pass or Fail the course.	✗
55	(Y. Chen; Q. Chen; Mingqian , 2016)	Edx/ Coursera	34064	2	Watched videos, active days, play records, rewatched records, skipped records, posts etc.	LR, RF,**KKN**	ACC	80% TR_Set 20% TS_set	Predict dropout students in week five.	✗
56	(Raj, 2021)	OULAD	NA	2	Learning activity data, demographic attributes	**CNN**	ACC	70% TR_Set 30% TS_set	Predict two classes, Pass and Withdrawn.	✗
57	(Drousiotis, 2021)	OULAD	32593	7	Learning activity, demographic attributes (first assignment mark, clicks	DT,RF,**BART**	PR, REC, F1, and ACC	70% TR_Set 30% TS_set	Predict four classes Withdrawn, Fail, Pass, and	✗

					till course starts, age, disability, gender etc				Distinction using “one-vs-rest”.	
58	(Ding, 2019b)	edX	5739	1	Playing video, load video, navigate-backward, pausing video, showing subtitles.	**LSTM**,LR, **CNN**, **MLP**	MSEs	NA	Predicting student grades in the next chapter.	Grade Prediction from chapter 2 to chapter 11
59	(Niu, 2018)	icourse163	17673	1	Clickstream, forum data, demographics (login, post, reply, video, age etc.)	XGBoost, RF, LR **MLP**, SVM, **KKN**, AdaBoost	AUC, ACC	70% TR_Set 30% TS_set	Predict students who do not access for more than 14 days.	✗
60	(Li, 2018)	Coursera	NA	3	Video views,forum views, video start-stop,video backward jump, indicator of the device, country, browser used, operating system,etc	SVM, **ANN**	ACC, REC, Kappa	NA	Predict students who do not participate in the following week.	Prediction from week 1 to week 19
61	(Robinson, 2016)	HarvardX	41946	1	Demographics and Pre-Course Survey(Age, Previous Moocs Enrolled, Previous Moocs Completed, Bachelor’s Degree, Parent Degree, Region)	NLP model	AUC	20-fold CV	Dropout prediction.	✗
62	(Er, 2020)	Canvas Network	12447	3	Discussion forum data, assignment data, quiz data, and peer-review data.	LR,RF, **MLP**	AUC	Transfer learning	Predict engagement levels in peer reviews.	✗
63	(Wang, 2016b)	xuetangx	2633	1	Demographics, video watching, exercise data (age, gender, Watching ratio, submit times, etc.)	**Backprop, Linear regression**	MSE	NA	Grade prediction.	Grade Prediction from week 1 to week 14
64	(Nitta, 2021)	OULAD	32593	7	Graph representation of clickstream data	**GCN**	PR, REC, F1, ACC, AUC	80% TR_set 20% TS_set	Completion = (Distinction or Pass), Dropout= (Fail or Withdrawn).	✗
65	(Kim, 2018)	Udacity	10154	2	Learning activity data (video, reading a text page, attempting a quiz, grade)	**LSTM**	AUC	5-fold CV	Predict whether the student will graduate.	Prediction from week 1 to week 8

66	(Prenkaj, 2021)	Xuetang X/	149542	71	Clickstream (navigational, video, homework, forum)	LR,NB,DT,SVM,KNN RF, DNN, CNN, LSTM, GRU	AUCPR	70% TR_set 10%,Val_set% 20 TS_set	Predict Passing the final examination.	Prediction in different time windows (day: (7,14,20,25,60,12, 180)
67	(Sheng, 2021)	Xuetang X	20238	1	Video Clickstream (Video View, Watch Time, %Watch, Access Time, Forward, Backward, Pause and Leave)	LR, DT, SVM, K-means	AUC	NA	Predict whether the student will not continue to watch future videos.	✗
68	(Babu, 2017)	Edx	200000	39	Events in the last week, days from access to the end of the course, accesses in the last two weeks, page closes, videos watched	DT ,NB, RF, LR, SVM, GBoost, Ensemble model	AUC	80% TR_set 20% TS_set	Predict whether the student will participate until the last week and solve the final exercises.	✗
69	(Ardchir, 2020)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR, AdaBoost, RF GBoost,	AUC	60% TR_set 40% TS_set	Predict active students after 30 days.	✗
70	(Getoor, 2020)	Coursera	NA	7	View posts, posts, vote in forums, take quizzes, view lectures, sentiment in posts, replies etc.)	LR, **MLP**, DT, Linear Regression	AUCPR	Transfer learning	1) Predict the students who will earn a certificate. 2) Predict the students who will follow the course until the end.	Prediction in three stages (33%, 67%, and 100% of course)
71	(Kameas, 2021)	DevOps MOOC	936	1	Demographic and Clickstream Dataset (Assignments, Video Views, Gender, Education Level, Posts)	AdaBoost, DT**,** ExTree, LR, LDA, LightGBM, GBoost	PR, REC, F1, MCC, AUC and Kappa	10-fold CV	Predict whether the student will pass the final exam and obtain the certificate.	Prediction in three stages Week 0, week 1, week 2
72	(Sehaba, 2020)	OULAD	32593	7	Attempts, access, avg score, submission data, region, gender, level of education, disability	DT, NB, SVC, KNN	PR	5-fold CV	Predict the final result (Fail, Pass or Excellent)	✗
73	(Wang, 2019)	NA	11227	5	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR, SVM, GBoost, RF,DT, **LSTM**	AUC	10-fold CV	Dropout prediction.	✗
74	(Liang, 2016)	Xuetang X	120542	39	Clickstream (such as watching video, posting a thread etc)	LR, SVM, RF, GBoost	ACC	60% TR_set 40% TS_set	Predict active students after 30 days.	✗

75	(Radi, 2017)	Harvard X-MITx	597,692	15	Behavioural and Demographic Features (Interact With The Chapter, Gender, Date Of Birth, and GPA).	DT ,RF, SVM, NB, **NN**, LR, LAD, SOM	Kappa, AUC, Acc,Spec , Sensy	10-fold CV	Predict whether the student will obtain a certificate.	✗
76	(Kashyap , 2018)	Harvard X-MITx	5200	NA	Behavioural and Demographic Features (Video Days, Access, Start Date, Gender, Date Of Birth Etc).	DT, SVM, NB, RF	PR, REC, F1, MCC, AUC	10-fold CV	Predict whether the student will obtain a certificate.	✗
77	(Zhu, 2021)	Xuetang	76843	12	Behavioural features (quizzes, homework, team participation, project milestones etc.	LR, DT, SVM, RF	PR, REC, F1, AUC	75% TR_set 25% TS_set	Predict whether the student will pass.	✗
78	(Chen, 2019a)	Xuetang X	NA	10	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	DT-ELM	AUC	NA	Predict students who do not participate in the following week.	Prediction from week 1 to week 5
79	(Fu, 2020)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR, NB, RF, SVM, DT, **CNN**	PR, REC, F1, ACC	80% TR_set 20% TS_set	Predict active students after 30 days.	✗
80	(Whitehill, 2017)	Harvard X	NA	40	Behavioural features(answers to the quiz, play/pause/rewind events videos etc.)	LR	AUC	Train on another course from the same field	Dropout prediction.	Prediction from week 1 to week 8
81	(Vitiello, 2017)	Telescope platform	3213	11	Learning Behaviour Records and Discussion Forum Data	SVM	PR, REC, and F1	75% TR_set 25% TS_set	Predict whether the student will complete the final project and exam.	Prediction from week 1 to week 8
82	(Kőrösi, 2020)	Stanford	12015	1	Learning behaviour records	XGBoost, **GRU**, RidgeRegression, XGBregression	ACC, RMSE	4-fold CV	Predict the final assessment quiz responses from 0–100%.	Prediction from week 1 to week 5
83	(Jin, 2020)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR, SVR, ELM, **Backprop**	Acc, AUC, F1	67% TR_set 33% TS_set	Predict Dropout week (one week after the learner last generating).	Prediction from week 2 to week 5
84	(Liu, 2020c)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	DT,NB, LDA, RF, SVM, CNN -**GRU**	PR, REC, F1 and ACC	NA	Predict active students after 30 days.	✗

85	(Zhang, 2020a)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	GBoost, XGBoost, LightGBM,CatBoost and AdaBoost	F1, AUC and ACC	75% TR_set 25% TS_set	Predict active students after 30 days.	✗
86	(Umer, 2017)	Coursera	167	1	Clickstream and Demographics (Age, Gender, Weekly Quiz, Attempts, Access)	NB, RF, LR, **KNN**	AUC, F1	10-fold CV	Predict whether the student will Pass or Fail the course.	Prediction from week 1 to week 8
87	(Goel, 2020)	Xuetang X	685387	NA	Clickstream and Demographics Information (Videos, Posts, Problems Solved, Gender, Education, Date Of Birth, Influence From Friends)	**NN**	F1 Score	80% TR_set 20% TS_set	Dropout prediction.	✗
88	(Pulikottil, 2020)	Xuetang X	403057	285	Clickstream (access, problem, pages close, delete a comment, load video, pause video, problem check, seek video, stop video, correct answers, wrong answers)	LR ,RF, SVM, **Gboost, GRU**	AUC and F1 score	80%TR_set 10%,Val_set 10%TS_set	Predict active students after 30 days.	✗
89	(He, 2020)	OULAD	32593	7	Students' Demographics Feature and Their Time-Series Logs	**RNN-GRU**	PR and REC	80% TR_set 20% TS_set	Predict whether the student will Pass or Fail the course.	Prediction from 5 weeks to week 39
90	(Jha, 2019)	OULAD	32593	7	Clickstream Demographic Info, and Interaction Data Assessment Scores.	RF,GBoost, DL, GLM	AUC	10-fold CV	1) Predict whether a student will drop out from the course, and 2) Predict whether a student will pass.	✗
91	(Borrella, 2019)	MITx	11715	5	Learning behaviour records (grade achieved, access, time spent during the past seven days, missed assignments, posts during the last seven days, problem attempts etc.)	LR, RF	PR and REC	older runs for training and recent runs for testing	Predict students who would skip the Midterm or Final Exam.	✗
92	(Waheed, 2020)	OULAD	32593	7	Demographics Feature and Logs Data	**ANN**, SVM, LR	PR , REC and ACC	10-fold CV	Predict 1)(PASS vs Fail), 2) (Distinction vs Fail), 3)(PASS vs Distinction), 4)(PASS vs Withdrawn)	✗

93	(Khan, 2021)	OULAD	32593	7	Demographics Feature and Logs Data	RF, SVM, K-NN, **ANN**, EextraTree, AdaBoost, GBoost	PR, REC, F1 and ACC	10-fold CV	Predict four categories (Withdrawn, Fail, Pass, Distinction)	Prediction in five stages (20%, 40%, 60%,80%, and 100% of course)
94	(Lemay, 2020)	EdX	10432	2	Video features (rewind, fast-forward, pause, the play of each video)	LR, SMO, NB, J48	ACC, AUC, KAPPA	10-fold CV	To predict assignment submission.	✗
95	(Radovanović, 2021)	OULAD	32593	7	Demographic features, learning behaviour records	LR	AUC and AUPRC	10-fold CV	1) Dropout prediction 2) Predict whether the student will Pass or Fail.	Prediction in eight stages during the course (days number 0,7, 15, 30, 45, 60,90, and 120)
96	(Yu, 2019)	OpenEdu	977	1	Learning behaviour (video, rewatching, skipping, fast watching, attempts, answers, test score etc.)	**KNN, ANN**, SVM	ACC	70% TR_set 30% TS_set	Predict whether the student will Pass or Fail	✗
97	(Wu, 2020)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	SVM, CNN-**RNN**, GBoost, LR, DT, RF, AdaBoost, NB	PR, REC, F1 and AUC	80% TR_set 20% TS_set	Predict active students after 30 days.	✗
98	(Qu, 2019)	C programming MOOC course	1528	1	Assignment-related behaviour	**MLP, LSTM**	ACC, REC	5-fold CV	Predict whether the student will Pass or Fail the Final exam.	✗
99	(Imran, 2019)	edX (MITx and Harvard X)	641138	NA	Clickstream(Viewed, Event, Days, Play Video,Chapters and Forum Posts)	**DNN**	PR, REC and AUC	60%TR_set 25%,Val_set 15%TS_set	Predicting certified students.	✗

100	(Chiu, 2018)	EDX	1313	9	Learning behaviour (video,avg video watched, median video watched, avg time spent, the median time spentm post like, days participating in the course	LR	PR, REC and ACC	10-fold CV	Predict whether the student will Pass or Fail.	Prediction from week 1 to week 5
101	(Bote-Lorenzo, 2017)	edX	26947	1	Learning behaviour (%videos watched in each chapter, % assignments submitted, total grade of assignments etc)	RF, SVM, LR and SGD	AUC	10-fold CV	Predict whether the level of student engagement in the next chapter will be lower than in previous chapters.	Prediction from chapter 1 to chapter 11
102	(Alami, 2021)	OULAD	32593	7	Demographic features, learning behaviour records.	SVC, LR, **KNN**, RF, AdaBoost	PR ,F1 and ACC	NA	Predict four categories (Withdrawn, Fail, Pass, and Distinction)	✗
103	(Xia, 2020)	OULAD	6360	1	Learning behaviour	RF	ACC	70%TR_set 30%TS_set	Predict three categories (Withdrawn, Fail and Pass)	✗
104	(Zhang, 2021)	Xuetang X	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	RF-**CNN-LSTM**	AUC	NA	Predict active students after 30 days.	✗
105	(Mubarak , 2020)	OULAD	32593	7	Learning Behaviour and Discussion Forum Data	LR, SVM, DT, RF	PR ,F1, ACU and ACC	5-fold CV	Predict students who do not participate in the following week.	Prediction in four stages (week 8, 16, 21 and final week)
106	(Doleck, 2020)	EDX	6241	1	Clickstream (videos viewed per week, stops, pauses, fast forwards, avg time spent watching, avg playback etc.)	SVM, NB, LR, **KKN**	ACC	NA	Predict students' performance in the assignment.	✗
107	(Mubarak , 2021a)	University of Stanford	1300	2	Clickstream events (play, search, pause, etc.)	**LSTM, ANN**, SVM, LR	PR, REC, F1, AUC and ACC	60%TR_set 10%,Val_set 30%TS_set	Predict students' performance (weekly quiz)	Prediction from week 1 to week 7
108	(Boyer and Veeramachaneni, 2016)	edX /Coursera	232174	15	Learning behaviour	RF, LR, SVM, **NN**	AUC	Transfer learning	Predict whether the student will be online next week.	Average AUC overall prediction

109	(Qiu, 2019)	Xuetang X	53596	6	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR, DT, GBoost AdaBoost, RF, NB, SVM, **CNN**	PR, REC, and AUC	75% TR_set 25% TS_set	Predict active students after 30 days.	✗
110	(Moreno-Marcos, 2020a)	Coursera	2035	1	Demographics and Learning Behaviour (Days, Time Spent, Videos Watched, Uncomplicated Video, attempts, Educational Level, Completed Assessments, Age, Gender, Job etc.)	DT, GLM, RF, SVM	AUC	10-fold CV	Predict dropout (if the student is not active for four weeks, considered as dropout).	Prediction from week 1 to week 5
111	(Qi, 2018)	icourse163	14393	5	Learning behaviour	**RNN** and **LSTM**,LR and SVM	AUC	60%TR_set 20%,Val_set 20%TS_set	1) Predict whether the student will be online next week. 2) Predict whether the student will Pass or Fail.	Prediction from week 1 to week 7
112	(Fei, 2016)	Coursera /edX	67506	2	Learning behaviour (videos downloaded, posts quizzes attempted, fourm accessed etc)	**RNN, LSTM**	AUC	5-fold CV	Predict whether a student has activities in the coming week.	Prediction from week 1 to week 9
113	(Xing, 2016)	Canvas	3617	11	Learning behaviour(days, access, posts, quiz views, fourm accessed)	GBN, DT	AUC	10-fold CV	Predict whether a student has activities in the coming week.	Prediction from week 1 to week 7
114	(Alshabandar, 2018)	OULAD	NA	2	Clickstreams data(number of activities during each session)	**KNN**, LR	AUC,F1, Sens, Spec, ACC	60%TR_set 40%TS_set	Predict whether a student will submit the assignment.	Prediction during six stages of the course.
115	(Edalati, 2022)	Coursera	NA	15	Students' reviews	RF,SVM ,DT, **CNN, BERT**	PR, REC, F1	70% TR_set 30% TS_set	Predict (Positive, Negative and Neutra) of students' Reviews.	✗
116	(Tang, 2018)	Xuetang X	79186	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	LR,RF,GBoost, **RNN-LSTM**	AUC	NA	Predict active students after 30 days.	✗
117	(Veeramachaneni, 2015)	edX	235197	1/ 3 Runs	Learning behaviour (attempt, submissions, correct answers, time spent to correct problems, total time spent on lecture, time spent on book	LR	AUC	Transfer learning	Predict whether a student will attempt at least one problem in the next week.	Prediction from week 1 to week13

					resources, average number of submissions per problem)					
118	(Ding, 2019a)	edX	98185	2/ 3 Runs	Learning behaviour (play video, show transcript, pause video, video speed, page close, stop the video, problem graded, etc.)	LR, **LSTM**, **CNN**	AUC	Transfer learning	Predict the dropout week of a student (Defined as the week after the student's last video interaction event).	Prediction from week 2 to week 9
119	(Itani, 2018)	Open-Classrooms	19493	2	(Learners' trajectory of engagement) (jumping activities)	RF, GBoost, DT, LR	F1	60%TR_set 40%TS_set	Predict dropout.	Prediction in three stages (25% and 50%, of course)
120	(Hassan, 2019)	OULA	32593	7	Learning and Assignment Behaviour (Quiz Activity, Discussion Forum, Video, Tutorial Sessions, PDF Resources, Wikipedia Content, Assignment etc)	**LSTM, ANN**, LR	PR, REC, and ACC	NA	Predict whether the student will drop out.	Prediction from week 5 to week 25
121	(Mubarak , 2021b)	Stanford	8368	2	Video-watching clickstream (play video, pause the video, rate-speed, stop the video, video forward, video back-ward)	**LSTM, GRU, RNN**	ACC, AUC	70% TR_set 30% TS_set	Predict whether the student will Pass or Fail.	Prediction from week 1 to week 7
122	(Lu, 2017)	Coursera	63441	2	Clickstream, assignment, forum activities (view quiz, view forum, view lecture, video view, video pause, post, downvote, upvote, etc.)	LR, SVM, **MLP**, **LSTM**	AUC	5-fold CV	Predict students who do not participate in the following week.	Prediction from week 3 to week 15
123	(Li, 2017)	XuetangX	120542	39	Clickstream (access, problem, pages close, navigate, video, discussion, wiki)	SSAE, softmax regression, SVM	AUC	5-fold CV	Predict students who do not participate in the following week.	Prediction from week 1 to week 5
124	(Hlioui, 2021)	OULAD	1303	1	Demographic (disability, region, gender, level of education, course previous attempts), activities related feature, students' performance)	DT, RF, SVM ,**MLP** ,TAN	F1	5-folds cross	Predict whether a student will drop out of the Course.	✗

125	(Gitinabard, 2018)	Edx/Coursera	66203	2	Learning and social features (video download, video view, total attempts, total posts)	RF, SVM, LR	F1 and AUC	Transfer learning	Predict whether a student will stop engaging at some point.	Prediction from week 1 to week 5
126	(Drousiotis et al., 2021)	Xuetang X	2000	1	Learning activities (14 different types of unique actions)	DT, RF, **BART**, **LSTM**	PR, REC, F1 ACC and AUC	70% TR_set 30% TS_set	Predict whether a student will stop engaging at some point.	✗
127	(Cobos et al., 2017)	FutureLearn / edX	12465	2	Learning activities (access, days, total time spent on quizzes, interactions in discussions forums)	**KNN**, GBM, LogitBoost, XGBoost	AUC	NA	Predict whether a student will complete 50% of the course.	Prediction from week 1 to week 7

3.2.6 Prediction targets in MOOCs

Student dropout is a complicated topic involving students' human behaviours and emotional and cognitive involvement (Hew, 2016). Although a considerable amount of literature has been published on the prediction of MOOC dropout, no formal definition of dropping out has been established. The intricacy of the phenomenon of students being at risk of dropping out from MOOCs is exacerbated by a lack of academic consensus on the dropout, success, completion, and certification categories. Therefore, at-risk students have been analysed from different perspectives, and researchers have used several targets to predict students dropping out from MOOCs (Sunar et al., 2016).

During a preliminary examination of 127 selected studies, several prediction targets in MOOCs were identified to predict at-risk students. Therefore, we grouped the reviewed studies into four categories according to their prediction target (completion, performance, comments, and certificate). Table 3.2 provides a summary of the terminologies and concepts reported in the reviewed studies to predict at-risk students using ML techniques.

Table 3.2 Summary of the prediction targets in the reviewed studies to predict at-risk students

Prediction target	Category
Predict active students after one month.	Completion
Predict students who do not participate in the following week.	
Predict whether a student will complete the whole course.	
Predict withdrawal from the course.	
Predict the percentage of course content completed in the entire course.	
Predict engagement levels in peer reviews.	
Predict whether the student will not watch future videos.	
Predict whether the student will follow the course until the end.	
Predict the level of student engagement in the next chapter.	
Predict whether a student will complete 50% of the course.	
Predict students' course grades.	Performance
Predict students' Pass/Fail status.	
Predict students' performance in the assignment.	
Predict whether the student will succeed in the final exam of the MOOC.	
Predict students' grades in the next chapter.	
Predict the final result.	
Predict whether the student will complete the final project.	
Predict whether the student will skip the midterm or final exam.	
Predict whether the student will submit an assignment.	

Predict whether a student will attempt problems in the following week.	
Predict the sentiment of students' posts (positive, negative or neutral).	Comments
Predict students who do not post comments.	
Predict confusion/urgency of students' posts	
Predict whether a student will obtain a certificate.	Certificate

As can be seen in Figure 3.2, among the 127 studies, 93 aimed to predict some form of student completion or dropout. The most common dropout definition used in 26 works was 'the students who do not have activities after 30 days from the course started'. This is because most of these works used a publicly accessible dataset known as 'KDD CUP 2015 Competition', which is provided by XuetangX, the largest MOOC platform in China, and dropout was previously defined in the dataset. The second popular definition of a dropout is 'the student who does not participate in the following week', which was used in 13 studies. Figure 3.2 shows that 39 studies targeted student performance, such as predicting students' course grades, pass/fail statuses and final project completion. However, only six studies targeted students' comments to indicate confusion/urgency or sentiment in students' posts (Positive, Neutral, or Negative). Moreover, six studies focused on predicting students who will obtain a certificate by the end of the course.

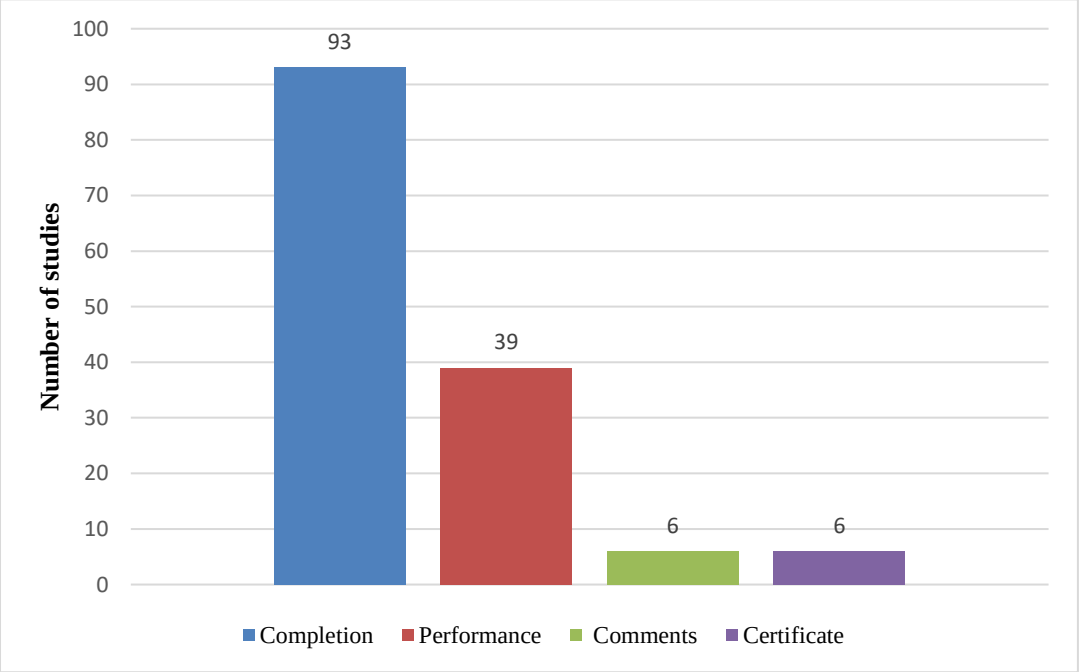


Figure 3.2 Prediction Targets in MOOCs

3.2.7 Data sources for at-risk prediction In MOOCs

Most surveyed works proposed at-risk predictive models using more than one data source such as clickstreams, assignments, forums, and demographic data. Figure 3.3 illustrates that clickstreams are the most common data source, used by more than 88% of the surveyed works. This is not surprising because clickstreams provide a lot of rich, detailed data that the research community is just starting to discover as how to represent all its complexity. However, clickstreams are unprocessed text files that require substantial time to interpret manually and computationally (Gardner and Brooks, 2018).

In addition, it can be seen from Figure 3.3 that assignment and forum data are the second and third most used data (55% and 51%, respectively), and demographic information is the least frequently used source (used by only 24% of the surveyed works).

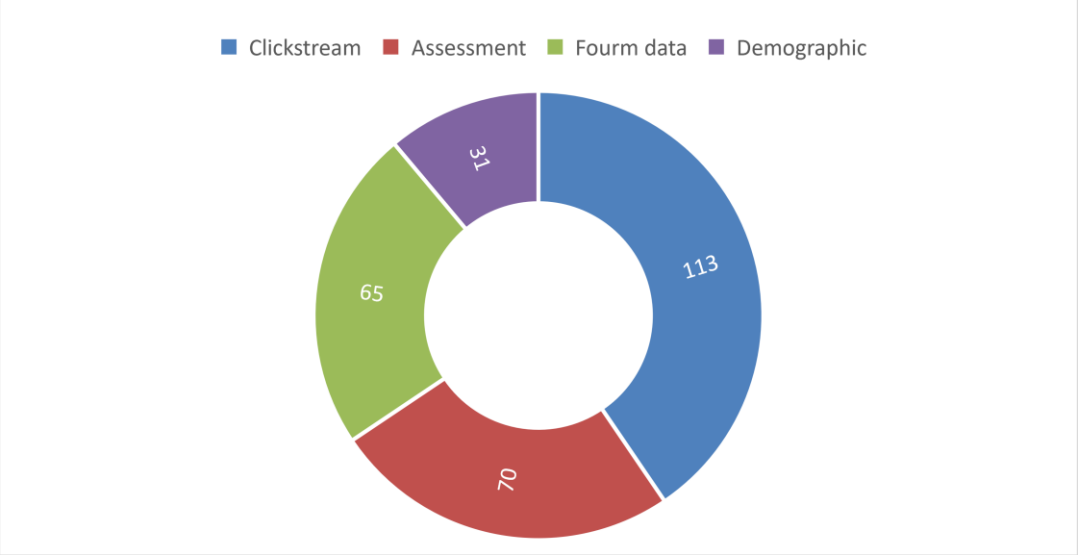


Figure 3.3 Data sources used in reviewed studies for predictive Modelling

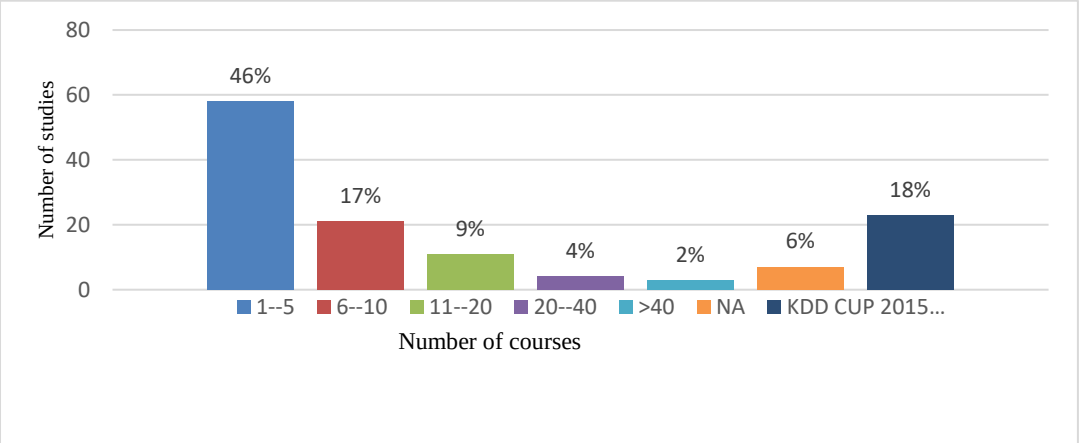


Figure 3.4 Number of courses evaluated across works surveyed

Figure 3.4 shows that 58 studies evaluated between one and five courses, whilst more than half of these studies extracted datasets from only one course (32 studies). The most likely causes of this include the lack of available MOOC datasets (Gardner and Brooks, 2018). On the other hand, approximately 18% of the surveyed studies used a publicly accessible dataset known as KDD Cup 2015, which was provided by XuetangX, the largest MOOC platform in China. This dataset contains students' logs for more than 120,000 students extracted from 39 courses. In addition, Figure 3.5 below illustrates the sizes of the datasets that have been explored in previous studies on MOOCs. Generally, the size of the dataset is based on the number of students, as well as the duration of the course. It can be seen that more than 50% (69 studies) of the surveyed works used a dataset that contained less than 50,000 students, and only 6% (8 studies) used a dataset for more than 200,000 students.

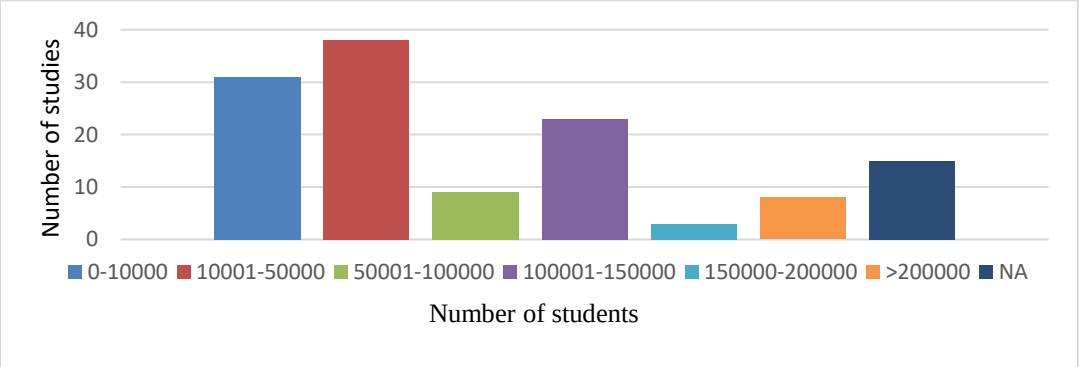


Figure 3.5 Number of students across works surveyed

3.2.8 Prediction models and evaluation metrics

The reviewed studies showed a long tail of modelling techniques. This also reflects that the research area is still in its infancy, with experts disagreeing on how best to tackle the challenge of making accurate predictions. Figure 3.6 compares the summary statistics for several predictive models in the reviewed studies. Approximately half of the proposed models appeared in only one study (34/64 models, represented by other). This means that most previous studies have emphasised novelty to present a new predictive model. On the other hand, some modelling algorithms such as LR, RF, and SVM have been used in more than 50 studies.

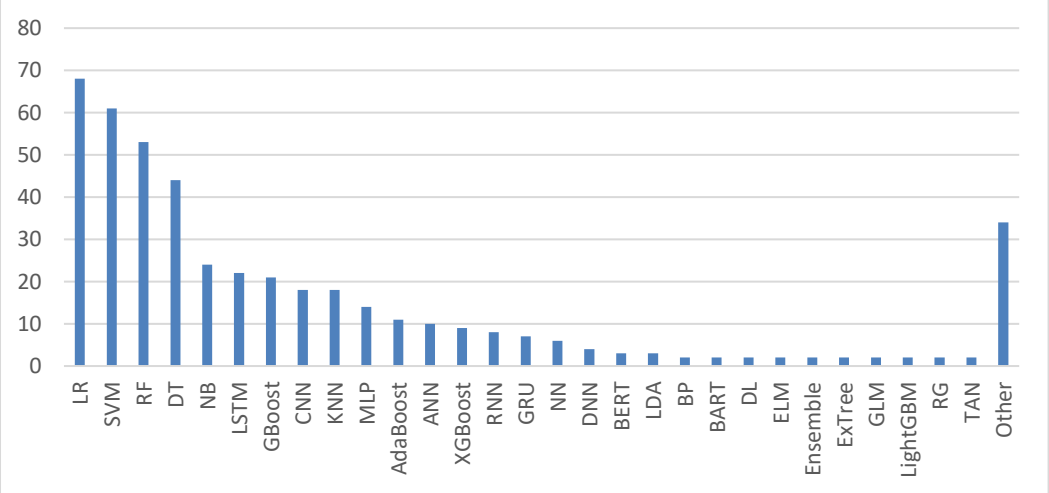


Figure 3.6 Modelling algorithms presented in MOOC studies

Figure 3.7 shows a substantial disagreement among the researchers on the best measures for evaluating models in MOOCs. Predictive effectiveness can be measured in various ways depending on the target of the study; thus, without a common baseline, it is impossible to make meaningful comparisons between studies.

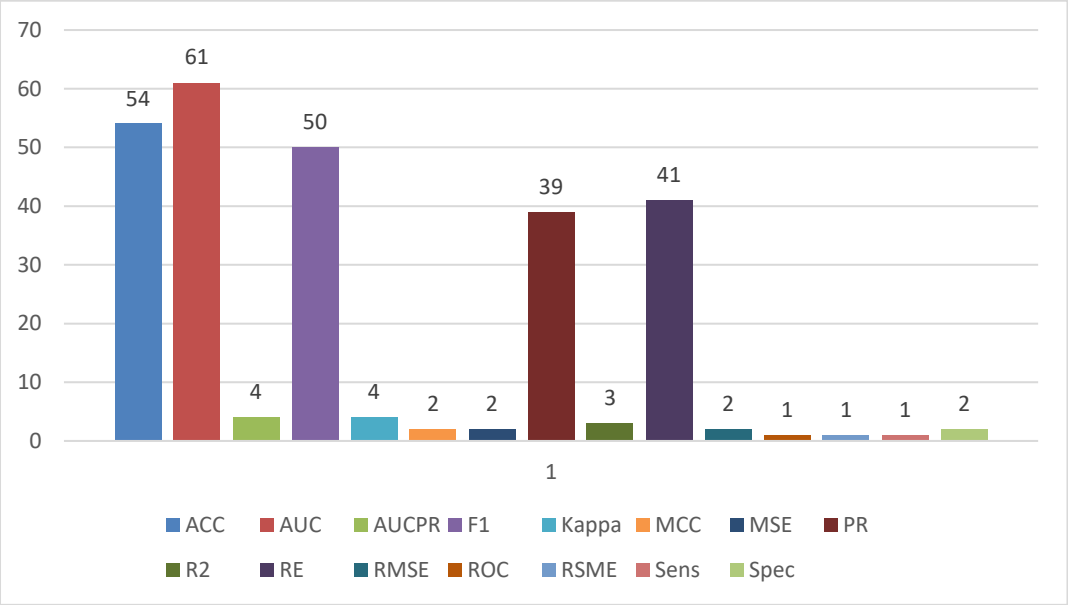


Figure 3.7 Evaluation Metrics in MOOCs Prediction

3.2.9 Experiments with a realistic environment

As predictive modelling studies have become more commonplace, they provide a fertile setting for developing MOOC modelling techniques and experimentation. Most surveyed works intended to design predictive models that could be applied to improve the learning environment of MOOCs and provide appropriate early intervention for at-risk students. Nevertheless, the predictive models in most of the reviewed studies were evaluated/tested using a subset of data extracted from the same course data used for training the model rather than using future data. Therefore, these works cannot be put into practice in real-world settings because applying predictive models in MOOCs requires a setting that is as close to reality as possible, especially regarding the data that can be collected at the time of prediction.

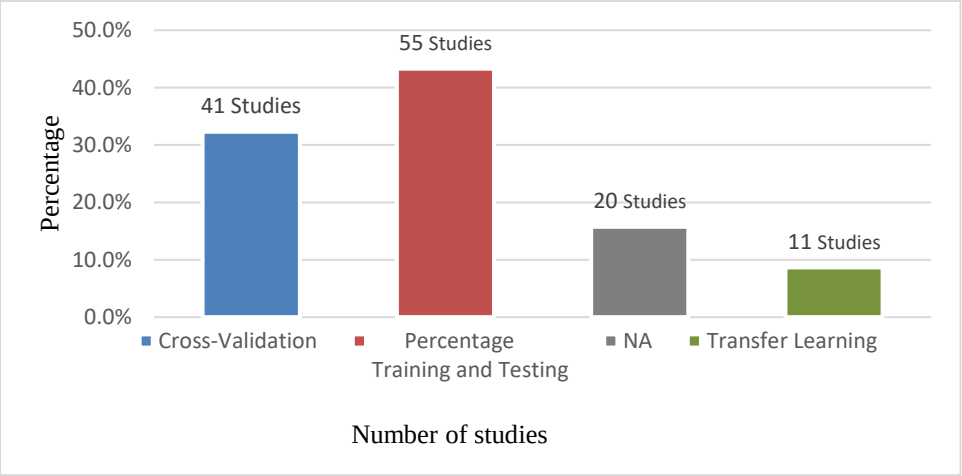


Figure 3.8 Training and Testing techniques across all the works surveyed

Figure 3.8 shows that more than 75% of the surveyed works split previously labeled data into train/test sets by using either the k-fold cross-validation or percentage-splitting technique. Furthermore, 20 studies did not mention the type of data they used to train and test the predictive models.

On the other hand, a small body of research (only 11 studies [8.7%]) has used transfer learning or in situ learning to provide some approaches for predicting at-risk students in MOOCs. For example, a study by (Getoor, 2020) proposed a ML model to predict two targets: whether the student will a) follow the course until the end and b) earn a certificate. Getoor used the transfer learning technique by training the model on one course and testing the model on another course data. Another study by (Bote-Lorenzo, 2018) used an in situ learning approach to predict students' levels of engagement in each chapter. For example, to predict students' levels of engagement in Chapter 3 (unknown target at the time of prediction), the model was trained by using data from the second chapter (as the target variable is known) and then using the trained model to predict the level of engagement in Chapter 3. This method is based only on data accessible at the prediction point, making the prediction models appropriate for an ongoing MOOC.

3.2.10 Temporal modelling techniques

Students' activities in MOOC courses take place during the course duration, which is several weeks for most courses. Therefore, student data can be collected sequentially, with limited data obtainable during the early stages of a course. In addition, the behaviours of learners change over time (Codish et al., 2019). As a result, models that consider time provide a more accurate illustration of students' behaviours during different periods. In our surveyed works, temporal prediction techniques were used in 41% of the analysed works to predict at-risk students at different periods of the course. For example, (Monllaó Olivé, 2020) attempted to predict students who did not participate in the last quarter of the course at different periods (25%, 50%, and 75% of the course). Another study by (Getoor, 2020) used temporal prediction in three stages (33%, 67%, and 100% of the course) to indicate whether the student will pass the final exam and obtain a certificate. In addition, a study by (Kameas, 2021) applied several ML classifiers at different times (weeks 0, 1, and 2) to predict whether the student would pass the final exam and obtain the certificate. The author established the prediction even before the course started by using only data accessible at the prediction time, such as students' demographic data. On the other hand, only 26% of the surveyed works applied prediction techniques from the first week of the course, where participants are most likely to drop out in the first few weeks. Therefore, early intervention is essential to identify those students at an early stage (week 1).

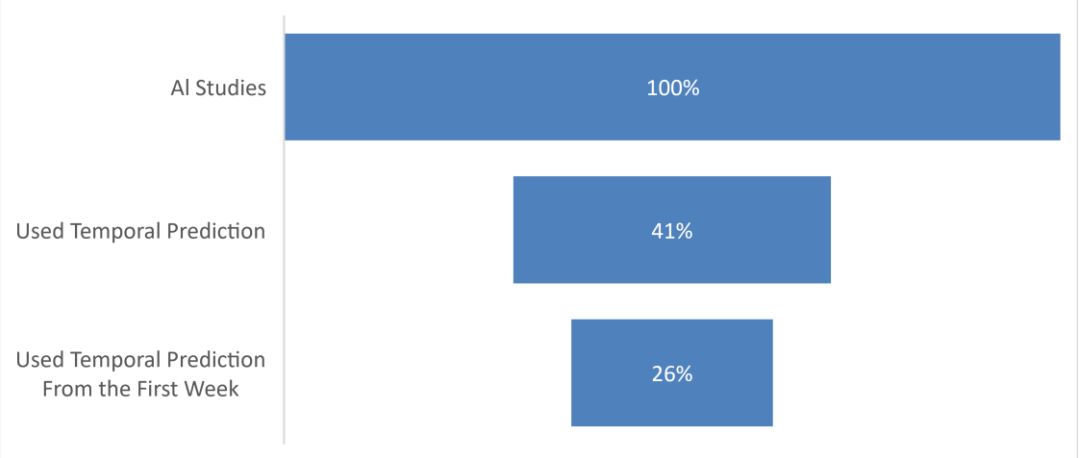


Figure 3.9 Temporal prediction across surveyed studies

3.3 Engagement and Motivation

3.3.1 Engagement and Motivation in AIED and ITS studies

Engagement and motivation have been studied in the areas of AIED and ITS, where the connection to theories is more or less explicit. One of the earliest studies, by Arroyo (Arroyo et al., 2007), showed that students were more likely to get re-engaged with the system by providing monitoring interventions to students in between problems. In addition, negative feedback messages may have motivated some students to be more attentive and avoid receiving such feedback in the future.

Another early study was conducted by (Cocea, 2007) to detect students' engagement levels in an e-learning system. Students were labelled (engaged/ disengaged) based on their performance. For instance, a student who took longer or less than the necessary time to perform a given task was considered a disengaged student. The study showed that average time spent reading was the best indicator of engagement.

Jackson divided students into two groups based on their expectations before engaging with the system. The first group included students who were "sure" that using the e-learning system would help them improve their knowledge, and the second group had students who were unsure whether a system could assist them. The results of the post-survey indicated that the students in the first group thought that the system was significantly more enjoyable and motivating. In contrast, the students in the second group did not like engagement and found it less motivating (Jackson et al., 2009).

In the last few years, the movement has been more towards educational data mining. As such, there have been studies that have proposed using several machine learning techniques at the same time to build their prediction models. One very recent study (Khan, 2021) used seven different machine learning techniques, including RF, SVM, KNN, ANN, ExtraTrees, AdaBoost and gradient boosting to predict four categories (withdrawn, fail, pass and distinction). Additionally, another very recent study (Khodeir, 2021) deployed state-of-the-art language modelling transformers (BERT) for the natural language processing task, using the student comments in a MOOC as input, to predict students who require urgent

intervention. A study (Mubarak, 2020) targeted struggling learners who needed early intervention – to keep the engagement – by designing a prioritising at-risk student temporal model to predict whether the student will drop out next week.

3.3.2 Engagement and Motivation in MOOCs

Here, we analyse the state-of-the-art in engagement and motivation-related studies in MOOCs (see Table 3.3). In a recent work conducted by (Sunar et al., 2016), the authors investigated how social interactions impact on course completion in MOOCs. According to the authors, dropout rates could be reduced by increasing the users' engagement with social interactions in the systems. The authors presented descriptive statistics and a literature review on the prediction of user behaviours in MOOCs. However, the authors did not base their investigation on any existing motivational theories. Their survey showed that many works (8 out of 15 of their reviewed studies) concerned with predicting students' behaviour, focused on course attrition, by analysing clickstream data and student activity within the system. They argued that social interaction also needed to be analysed. However, none of the surveyed works focused on extracting measurable engagement indicators.

A recent work by Nam et al. (2017) predicted disengaging behaviour using data-driven methods on a small scale (25 students). The authors developed a model to predict when the students were not engaged. This resrech conducted a comparative analysis between two distinct families of statistical models, namely logistic Regression (LR) and mixed-effects logistic Regression (MLR). LR commonly utilised regression model in the field of data analysis for the purpose of classifying data with binary labels. The MLR uses random effect variables to account for differences in repeated measurements. According to the authors, different attributes could be used to predict engagement and context-sensitive information improved the prediction accuracy. They performed a feature analysis, to select the best prediction features, but they only used supervised learning (with manually labelled data). They did not systematically employ engagement theories, but instead labelled the data manually as engaged/disengaged.

Another recent work of Chen (2019b) proposed the prediction of learning outcomes through learning behaviour, including engagement, in short online courses.The author evaluated six

classifiers, namely K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis, Random Forest (RF), Forward Artificial Neural Network (ANN), and Gradient Boosting (XGB). In this work, the authors adopted time spent and completion rate to model engagement. They stated that by using social learning network features, they could increase the prediction accuracy over time. Although relevant, the attributes used to model engagement were not based on any engagement or motivational theory.

In the study conducted by Pardo et al. (2016), the authors focused on creating an approach that combined data about self-regulated learning skills and online activities in a blended learning course. The researchers did not apply machine learning models; they opted for hierarchical clustering analysis to determine the factors that more effectively differentiate the variability in student performance. Furthermore, the selection of regression analysis was made to investigate the possible linear correlation between the numerical variables and academic achievement. In this study, engagement was measured through a self-reported questionnaire at the end of the course, and not in real-time, using variables from the system.

Another study conducted by Barak et al. (2016), explored the motivation to learn in MOOCs. They analysed existing theories and created a model of motivational components to inspect the influence of language and social engagement in a MOOC. The authors developed a model to measure students' motivation, based on pre- and post- course questionnaires; only 325 out of 13,405 students participated in the questionnaire (which is not that surprising, considering the response rates in MOOCs are generally lower for the questionnaire (Mihalec-Adkins et al., 2016)). The authors defined success as students completing essay questions, 10 weekly quizzes and a final project. However, they did not focus on predicting success by using machine learning models or tying the motivation to success measures. According to their findings, the larger the number of posts, the higher the students' motivation.

Another study by Lan and Hew (2020) used self-reported questionnaires and interviews to measure learner engagement and motivation for both completers and non-completers. They also received a low response rate from participants (82 out of 693 students agreed to be interviewed). The authors employed a statistical methodology known as Multiple Regression to evaluate the predictive capacity of the three dimensions of engagement, namely behaviour,

emotion, and cognition, on students' perceived learning. The study showed that completers are more willing to participate in the social network and receive a certificate.

Moreover, research by Stone (2021) used self-reported questionnaires with only 68 learners to examine "why students are motivated to enrol". The scope of this research focuses only on the motivation that led students to enrol in MOOC courses based on the questionnaire's outcome. The author did not analyse students' interaction data to explore the link between student activities and their motivation. Additionally, the author did not use machine learning techniques; only statistical methods, such as Correlation coefficients, were used to measure the association of input and output variables.

In the same line of research, de Barba et al. (2016) focused on using student motivation as a predictor for learning performance. However, the authors didn't use machine learning to predict learning performance. Instead, descriptive statistics and Spearman correlations were used for data analysis.

They created a theory-rooted predictive model, to estimate students' intrinsic motivation and participation, and found that the number of attempts in answering quizzes was a good predictor of their final grades. However, this study included only students who were active during the last three weeks of the course and responded to the questionnaire on the last week. Meanwhile, the majority of participants in MOOCs will drop out much earlier. As a result, analysing such learners at an early stage is critical, in order to give early assistance and maintain engagement.

Wang et al. (2015) analysed comments within a MOOC and aimed at creating a predictive model that relates the learning gains with social interactions. The author employed a logistic regression model and a 10-fold cross-validation approach to assess the model. The authors defined a taxonomy for the comments and concluded that students that present active (who actively practiced the learning material via quizzes) and constructive discussions (who produced content, e.g. explanations and examples, based on the course material) had significant learning gains. According to the authors, constructive behaviours produced more learning gains than active behaviour.

The study conducted by Brinton et al. (2014) focused on providing statistical evidence and a Generative Model for extracting important topics in each forum (by considering two models

naive Bayes (NB) and support vector machine (SVM)) based on user interaction within a MOOC). The authors performed a large-scale study (73 courses) and found that teachers' participation within forums was positively associated with the level of interaction (e.g. increased discussion threads), but did not affect the decline rate of participation in the course. By conducting this analysis, the authors focused on improving social learning in MOOCs.

Another similar work, conducted by Wen et al. (2014b), focused on sentiment analysis in a MOOC, aiming to understand the relationship between students' comments and course success. The authors developed a model to identify motivated students based on their comments in a course forum using logistic regression for binary classification. According to the authors, it would be difficult to categorise each forum post made by a student as a motivational statement. Numerous online postings lack discernible indications of user motivation. The statsical analysis shows that students who had positive behaviour towards the course (according to their motivation model) had lower rates of dropping out from the course.

3.4 Critical Evaluation

Although the literature identifies several variables that contribute to student dropout on MOOC platforms, few researchers have investigated how these issues influence at-risk students. Moreover, most solutions involve a large number of parameters and are what we call ‘heavy weight’. While such approaches may provide higher accuracy, they are less applicable in real life, as they require real-time processing of large quantities of data and may provide results too late in the course cycle to be of real effect. *To the best of our knowledge, none of the previous studies attempted to forecast attrition based only on the participant’s first point of contact with the educational system, which is the registration date.* In addition, *there have been no controlled studies visually comparing the differences in the learning paths of completers and non-completers for the entire course session and investigating the performance of weekly prediction models with and without students’ jumping behaviours.*

Additionally, Research is still undergoing on whether the low rate of completers indicates a partial failure of MOOCs, or whether the diversity of MOOCs learners may lead to this phenomenon world (Kloft et al., 2014). In the meantime, this problem has attracted more

attention from both MOOC providers and researchers, whose goal is to investigate methods for increasing completion rates. This starts by determining the *indicators of student dropout*. Previous research has proposed several indicators. Ideally, the earlier the indicator can be employed, the sooner the intervention can be planned (Ye and Biswas, 2014). Often, combining several indicators can raise the precision and recall of the prediction (Coates et al., 2011); however, such data may not always be available. For example, a linguistic analysis of discussion forums showed that they contain valuable indicators for predicting non-completing students (Wen et al., 2014c). Nevertheless, these features are not applicable to the majority of the student population, as only five to ten percent of the students post comments in MOOC discussion forums (Wen et al., 2014a).

Therefore in this thesis, *we present a first of its kind research into a novel, light-weight approach based on tracking two (accesses to the content pages and time spent per access) early, fine grained learner activities to predict student non-completion*. Specifically, the machine learning algorithms take into account the first week of student data and thus are able to 'notice' changes in student behaviour over time. It is noteworthy that we apply this analysis on a MOOC platform firmly rooted in pedagogical principles, which has seen comparatively less investigation. Based on an in-depth feature analysis, we found that time spent and number of accesses are important attributes, not only because they are simple to obtain for the majority of courses but also because the data demonstrates that the amount of time spent at every step is a significant factor in predicting student completion (see Section 7.3.3).

Finally, It is important to examine the relationship between the levels of students' motivation and cognitive learning outcomes. One of the most significant limitations of previous studies is the absence of a methodology for evaluating motivation levels in an online environment. In most cases, researchers use questionnaires or interviews as tools for data collection to assess students' motivation in MOOCs. *To the best of our knowledge, to date, no studies have systematically employed motivational theories, mapping online student behaviour onto them, to analyse the drives and triggers promoting student engagement.* The work in this thesis is pioneering in that there are no works that create an ***engagement* taxonomy** with measurable engagement parameters and allowing for an engagement theory (here, SDT) to be evaluated on a large scale. A comparison of the state-of-the-art in engagement and

motivation-related studies in MOOCs and our work depicted in this thesis can be observed in Table 3.3.

Table 3.3 Summary of comparison of our Engage Taxonomy with related work and state of the art.

Study	Theory-driven	Data-driven	Motivation and engagement analysis	Prediction (Learning outcomes and Success)	Sentiment analysis
(Sunar et al., 2016)		X	X	X	
(Nam et al., 2017)		X	X	X	
(Chiang, 2019)		X	X	X	
(Pardo et al., 2016)	X		X		
(Barak et al., 2016)		X	X		
(de Barba et al., 2016)	X	X	X	X	
(Wang et al., 2015)		X	X	X	
(Brinton et al., 2014)		X	X		
(Moreno-Marcos, 2018a, Moreno-Marcos et al., 2018, Kloos, 2018)		X	X		X
(Adamopoulos, 2013)		X		X	X
(Wen et al., 2014b)		X	X	X	X
(Arroyo et al., 2007)		X	X		
(Cocea, 2007)			X	X	
(Jackson et al., 2009)		X	X		
(Khan, 2021)		X		X	
(Khodeir, 2021)		X			X
(Mubarak, 2020)		X		X	
Our Engage Taxonomy	**X**	**X**	**X**	**X**	**X**

The proposed method, the *Engage Taxonomy* (see definition in sections 4.5.2 and 8.2), brings together all these strands: theory-driven and data-driven approaches, as well as the important motivation and engagement analysis, prediction in terms of learning outcomes and success, as well as emotional engagement (via sentiment analysis). This is done in a focused attempt to express motivational engagement theories via data-driven approaches, and then evaluate them.

Epilogue

This chapter provides an overview and discussion of the current literature to predict students at-risk of dropping out from MOOCs. Furthermore, this chapter presented motivation-related studies in MOOCs. The following chapter will give a summary of the approach taken to achieve the aim of this thesis.

Chapter 4 : Methodology

Prologue

This chapter provides an overview of the methodology used to answer each research question. Moreover, this chapter explains the dataset and tools used to achieve the aim of this thesis (e.g., the feature extraction process, feature selection, sentiment analysis, statistical analysis, visualisation tools, and predictive machine-learning techniques).

4.1 Introduction

The driving force behind this research was the need to discover and extract latent data patterns and solve complicated issues using MOOC datasets. This thesis was driven by a desire to enhance student outcomes in MOOCs by addressing several categorisation challenges linked to high student dropout rates.

To achieve the aim of this thesis and answer the research questions posed in Chapter 1, several methods have been proposed, including machine ML algorithms, visualisation techniques, and statistical methods to detect at-risk students in MOOCs. Chapters 5–8 are the four main chapters of this thesis; they fit together and help us reach our primary goal. Figure 4.1 shows the emphasis of each of these core chapters. Firstly in chapter 5, we will analyse the first interaction data with the MOOC system – the registration date to predict students' completion using statistical methods (see sections 4.6, 5.3 and 5.4). After that, in chapter 6, the power of visualisation will be used to analyse students' learning patterns in the course and compare the differences between completers and non-completers. Next, in chapter 7, we focus on innovations in predicting student dropout rates by building a generalised early predictive model for the weekly prediction of student dropout using machine learning algorithms (see sections 4.5.2, 4.8, 7.3.1, and 7.3.2). The final contribution (chapter 8), we propose a concrete mapping between the tracking parameters and four of the most used theories related to engagement in digital systems, generating the engage

taxonomy. It will show how such mapping can be practised by analysing the engaged and disengaged MOOC student behaviours in relation to the SDT theory (see section 8.3).

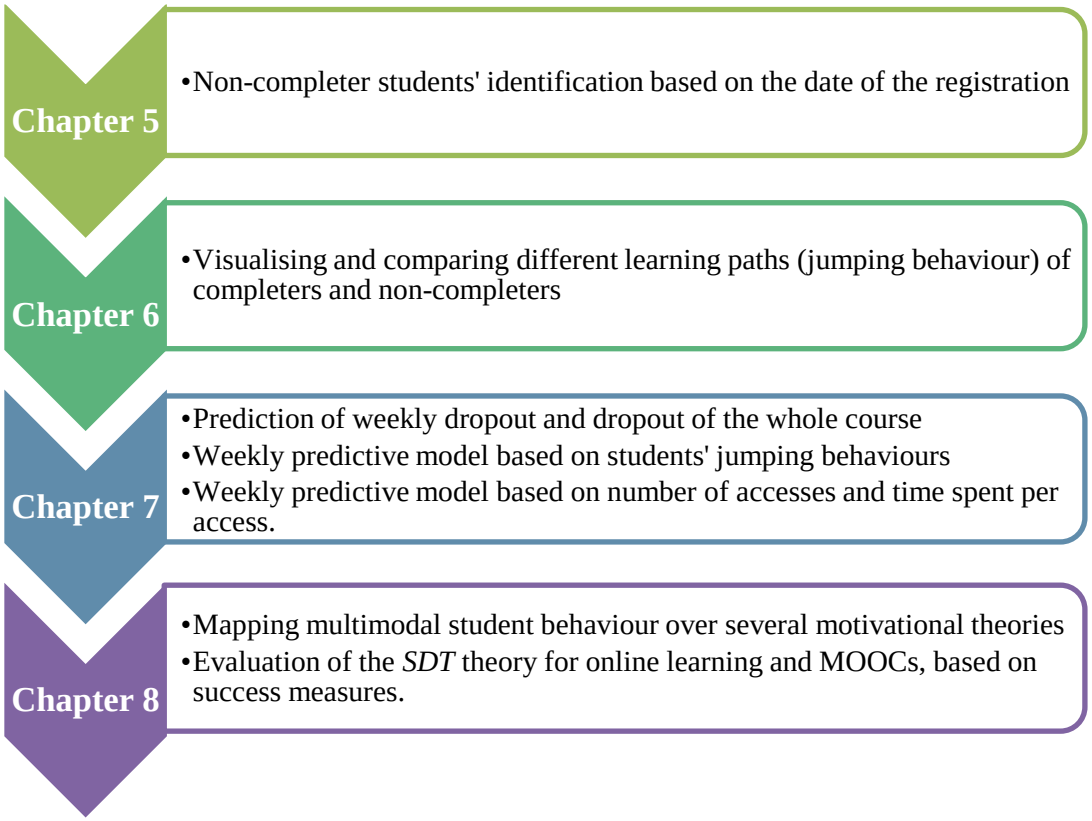


Figure 4.1 Thesis' core chapters

4.2 Addressing Research Questions

Each core chapter in this thesis will focus on one of the main research questions. For example, **Chapter 5** explores the correlation between student success and the registration date. In this chapter, we answer the first research question (**RQ1**): 'Can a limited number of student data types be used for the prediction of success (as in completion)?'

This initial question is divided into two sub-questions:

- **RQ1.1:** Can the date of registration (in terms of distance from the course start) of students predict their completion (or non-completion)?

- **RQ1.2:** How can dropout rates be alleviated based on the registration date?

Therefore, we extracted one feature (***the registration date***) as a calculated value which presents the number of days between the registration date and the starting date for a given student. **RQ1.1** will be addressed using statistical methods (see sections 4.6, 5.3, and 5.4), and **RQ1.2** will be answered by implementing rules to automatically deliver personalised messages to students according to their registration dates (see section 5.4).

Chapter 6 further illustrates different granularity visualisations for learning patterns to compare the differences between completers and non-completers. This chapter aims to answer the second research question (**RQ2)**: 'Can learning path visualisation of student interactional data be used to inform on student success (also seen as completion)?'. Accordingly, **RQ2** is split into three sub-questions:

- **RQ2.1**: How can learning paths be visualised to differentiate between completers' and non-completers behaviours (to inform teachers for early interventions)?
- **RQ2.2**: Are there (significant) differences in the learning paths of completers and non-completers and can they be deduced from visualisation early on in the course?
- **RQ2.3**: What kind of level of granularity is necessary for the visualisation of significant differences between completers and non-completers?

RQ2.1 will be answered using the visualisation technique (see section 4.7). We will visualise and compare the different learning paths of completers and non-completers. For **RQ2.2**, a statistical test will be used to determine whether there is a significant difference between completers and non-completers in terms of learning paths (see sections 4.6 and 6.2.3). Finally, **RQ2.3** will be answered by visualising different granularities (*fish eye*, *bird eye*) of of students' activities (see section 6.3).

The aim of **Chapter 7** is to answer the third research question (**RQ3), :** How does the time of student interactional data collection influence the student success (completion) prediction, and can early prediction be achieved? We will explore the predictability of students' completion based on students' weekly learning activities. Several ML algorithms will be used to build generalised early predictive models. Therefore, three sub-questions have been formulated:

- **RQ3.1**: Are there (high) differences in the prediction of weekly dropouts and whole-course dropouts?
- **RQ3.2**: Will the weekly predictive model be more accurate after considering students' jumping behaviours and catch-up learning patterns during the course?
- **RQ3.3**: Can MOOC dropout be predicted within the first week of a course, based on the learner's number of accesses and time spent per access?

For **RQ3.1**, we will compare two prediction methods: whole-course dropout prediction (**CP**) and weekly dropout prediction (**WP**). For the **CP**, the students are labelled as dropouts if they do not access 80% of the steps in the whole course. For the **WP**, the focus will be only on students who will drop out in the near future (next week) (see section 7.3.1). To address **RQ3.2**, we will incorporate the students' learning patterns (jumping behaviours) into the weekly dropout predictive model and compare the model's performance with and without students' jumping behaviours (see section 7.3.2).

To answer **RQ3.3**, we will investigate students' access features and the time spent on each access (see section 4.5.2) and, using a lightweight approach for prediction, build a predictive model based exclusively on those two features.

The final research question of this thesis **RQ4 (Chapter 8)** is as follows: 'Can engagement theories be applied to student interactional data, to help identify student success?'. This question is quite broad, so the following sub-questions were formulated :

- **RQ4.1**: How are engagement theories applicable in MOOCs?
- **RQ4.2**: Can engagement theories help identify student success in MOOCs?

RQ4.1 can be answered with the help of three experts by mapping the motivation concepts and the potential indicators within MOOCs. Moreover, Fleiss' kappa will be applied to measure the rate of agreement between experts (see section 8.3.2 and Table 8.1). To address the second question (**RQ4.2)**, students will be clustered based on their engagement and analysed via the connection to their success. In addition, ML will be used to evaluate the predictability of the extracted SDT constructs as early predictors of student activity (see sections 8.4.1, 8.4.2 and 8.4.3).

Finally, it is worth mentioning that the methodology of each study in this thesis are presented in more detail in each chapter.

4.3 Datasets

Unlike most of the surveyed works (see 3.2.7), this thesis analysed a massively large dataset across time (several runs over several years), subjects, universities, countries, and cultures. The data were obtained from two UK universities that delivered different courses in the FutureLearn MOOCs. In addition, another type of student activity data were acquired from the most massive Arabic MOOC (Rwaq).

We obtained interactional educational data (not publicly available) for 344,783 students. The data contained students' activities such as social interactions, topic access, quiz attempts, correct answers, and wrong answers. The studied courses had lengths ranging from 3 to 10 weeks and were delivered between 2013 and 2019.

4.3.1 MOOC Dataset Challenges

Although several MOOC datasets have been used in the literature, we could not identify a dataset that is open to the public and has a data structure comparable to ours. This is mostly due to the fact that each MOOC platform stores student behaviour differently, making it challenging to locate datasets that are similar to one another (Veeramachaneni et al., 2013, Lohse et al., 2019).

Numerous researchers have used the KDD Cup 2015 competition dataset to predict student dropout in MOOC. In this thesis, we cannot use the KDD Cup 2015 dataset, as several factors prevent us from using our dataset for comparison with the findings of the KDD Cup 2015 dataset.

Initially, it is essential to note that our dataset’s structure exhibits dissimilarities compared to the KDD Cup 2015 dataset, potentially resulting in our prediction model’s incompatibility. For the purpose of applying our model to different datasets, it is crucial to acquire data that is structured in a comparable manner to our original dataset so that we can ensure reliability and the consistency of outcomes.

In addition, our research aims to examine the sentiment of students in the context of predicting student attrition. The KDD Cup 2015 dataset lacks the inclusion of students' comments. The insufficient availability of the requisite data may render the utilisation of the KDD Cup 2015 dataset unsuitable for our research investigation.

Finally, the utilisation of the KDD Cup 2015 dataset without appropriate permissions may give rise to ethical and legal issues, as the dataset was removed from the official website after the end of the competition. Unauthorised utilisation of the dataset may result in contravention of ethical regulations and data usage policies.

4.3.2 FutureLearn

FutureLearn is one of the youngest massive online learning platforms (since 2012) and the European counterpart of the United States's Coursera, EdX, and so on. FutureLearn started as a partnership between several UK universities, the BBC, and the British library, expanding later to include courses from international institutions, non-government organisations, and businesses, which now supports 2,400 courses created by 175 partners and reached 10 million students by (Chastney, 2019).

As it is a newer platform, fewer studies have been conducted on it. We fill this gap by selecting courses delivered through the platform. FutureLearn courses are delivered by two universities in the United Kingdom (University of Warwick and Durham University).

The first dataset was obtained from 25 runs of 7 FutureLearn-delivered courses via the University of Warwick between 2013 and 2017. The number of enrolled students reached 276,491.

The second dataset was extracted from two FutureLearn courses delivered by the Business School at Durham University. The dataset was obtained from 7 runs between 2016 and 2018, with 16,488 students. The dataset format is similar to the Warwick dataset, as the courses were delivered through the same learning platform.

4.3.3 Rwaq dataset

Rwaq[3] is the most massive Arabic MOOC provider, with more than 882 active courses. Well-known professors and teachers deliver Rwaq courses from all over the Arab world (Khan et al., 2022). Considering the breadth of the course offering, very little research has been done to analyse the success of Rwaq. Indeed, a quick literature search on Google Scholar[4] (since 2012) on 'Rwaq analytics' rendered only 395 results, as opposed to 55,000 on the 'Coursera analytics' and 20,200 on the 'Udacity analytics'. Thus, a further contribution of this thesis is to study a popular and growing platform, albeit less explored.

The third dataset used in this thesis was acquired from the Rwaq platform, which, to the author's best knowledge, has not been used in previous research in relation to completion predictability based on ML. The dataset contains students' activities such as students' comments, correct answers, wrong answers, and accessed topics that were extracted from three courses. The number of enrolled students in the three courses was 51,804.

4.4 Dataset Formats

The dataset acquired comprises raw data regarding students' learning behaviours in MOOCs extracted from their log files. The data is explained in detail in Sections (4.4.1, 0, 4.4.3, and 4.4.4), which present information on students' behaviours and the format in which the educational contents were delivered (e.g. videos, articles, quizzes, assignments, and comments). Despite the unavailability of educational materials or course content, it is still possible to conduct an analysis based on the presentation format of the educational materials for each course. One possible approach is to conduct an analysis of various metrics related to student engagement, such as the time spent viewing videos or reading articles, the number of attempts made to answer quizzes or assignments, the number of correct and wrong

[3] www.rwaq.org

answers, and the number of comments provided by each student. In addition, sentiment analysis can be applied to review the students' comments.

The following sections provide an overview of the raw logged data used in this thesis that can be pre-processed and analysed, including their typical forms, structures, and behaviour patterns.

4.4.1 Clickstream data

This kind of data commonly includes every integration with the web server of the platform offering the course. Clickstream data are a record of the course navigational footprints left by students. Most of the time, clickstream data provide information about how students interacted with the learning platform, including clicks and page views of course materials such as videos and quizzes. The clickstream data are the most popular source of information for predicting student performance in MOOCs (Gardner and Brooks, 2018) (see Section 3.2.7). However, the clickstream data cannot be used directly as input data for most prediction models; consequently, the pre-processing stage becomes critical to extracting the features for prediction. The screenshots in Figure 4.2 and Figure 4.3 provide examples of students' clickstream data stored in two different MOOC platforms (FutureLearn and Rwaq).

learner_id	step	week_number	step_number	first_visited_at	last_completed_at
f5241163-3894-4a36-83	1.1	1	1	2017-07-18 09:57:53 UTC	2017-09-12 13:04:16 UTC
991380a2-aeec-462b-b	1.1	1	1	2017-08-07 10:50:16 UTC	
779d3fef-a24e-41e5-93	1.1	1	1	2017-08-24 10:51:21 UTC	
d475e4e9-b332-4b0c-9	1.1	1	1	2017-09-03 17:27:55 UTC	
2c3443ec-845b-47c9-90	1.1	1	1	2017-09-08 12:58:28 UTC	2017-09-18 07:35:53 UTC
268ea3d0-6a8a-42d2-a	1.1	1	1	2017-09-11 13:33:15 UTC	
aefdf564-1c8a-48f0-82	1.1	1	1	2017-09-18 00:12:04 UTC	
a2ae01de-41c2-4ca2-b	1.1	1	1	2017-09-18 00:14:56 UTC	2017-09-18 01:23:35 UTC
c7ba2166-08b8-4955-b	1.1	1	1	2017-09-18 00:19:41 UTC	2017-09-18 00:21:15 UTC

Figure 4.2 Example of student's clickstream data (FutureLearn platform)

student_id	enrollment_date_time	lecture_completion_time	lecture_view_time	lecture_id
130	2016-02-14 09:59:30 +0300	2016-03-23 11:48:25 +0300	2016-02-23 12:00:29 +0300	4064
246	2015-11-03 09:31:23 +0300	2016-02-09 07:51:42 +0300	2016-02-09 07:29:18 +0300	4064
322	2015-10-12 15:17:13 +0300		2016-03-26 20:47:57 +0300	4064
367	2015-10-20 11:59:27 +0300		2016-02-11 02:26:06 +0300	4064
408	2016-02-03 09:18:12 +0300		2016-02-03 09:18:39 +0300	4064
445	2016-06-14 05:58:01 +0300		2016-06-14 05:59:20 +0300	4064
543	2016-03-27 21:27:12 +0300		2016-03-27 21:28:19 +0300	4064
562	2016-01-25 09:49:28 +0300	2016-02-02 08:03:51 +0300	2016-02-02 07:49:29 +0300	4064
655	2015-09-11 19:53:22 +0300		2016-03-02 21:48:15 +0300	4064
712	2016-02-02 13:09:13 +0300		2016-02-02 13:11:33 +0300	4064

Figure 4.3 Example of student's clickstream data (Rwaq platform)

The screenshots show that the structure of FutureLearn courses is based on a weekly learning unit (week_number). Every learning week includes so-called steps (step_number), which cover images, videos, articles, and quizzes. Having joined a given course, students can access (first_visited_at) these steps and optionally mark them as completed (last_completed_at). These steps also allow comments, replies, and likes on these comments from different users enrolled in the course.

Rwaq platform provides similar information about the students' activities, such as the student ID (learner_id), steps (lecture_id), visited time, and completion time. On the other hand, the structure of Rwaq courses is not based on a weekly learning unit. Therefore, to match the Rwaq dataset with the FutureLearn dataset, we grouped every 14 steps (the average number of weekly steps in FutureLearn courses) in 1 week.

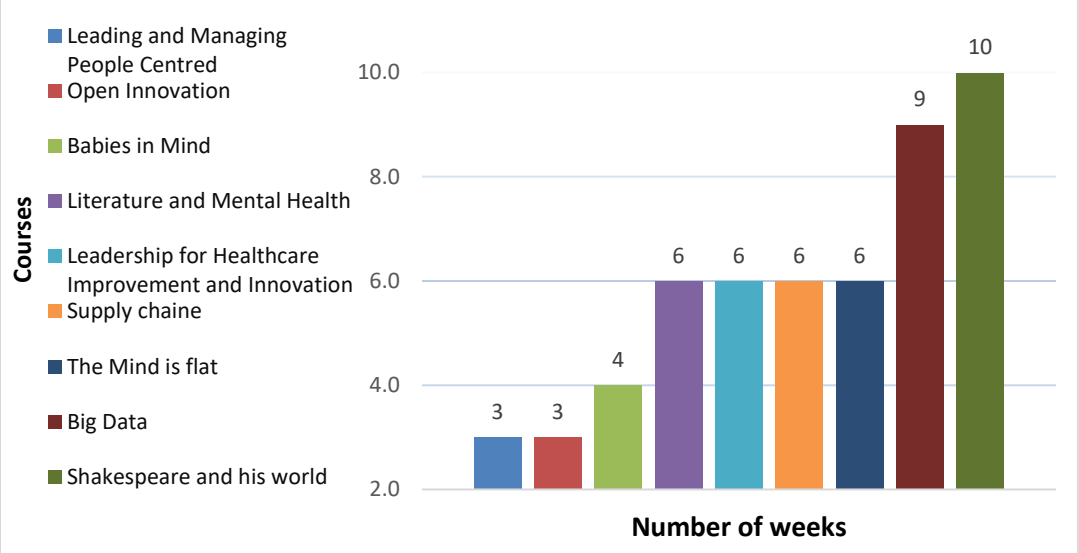


Figure 4.4 shows the number of weeks in each course and shows the lengths of each course. By analysing the duration of the courses with respect to the courses' provider, it can be observed that the courses delivered by Durham University show a relativity shorter period compared to the courses offered by Rawaq and Warwick University. Durham University provides two courses, Open Innovation in Business and Leading and Managing People-Centred Change, which contain 38 and 30 steps, respectively.

Warwick University provides a variety of courses, including Shakespeare and His World, which comprises 134 steps, and Babies in Mind, which consists of 75 steps, making it the

shortest course offered. Rawaq offers courses of varying length: Java Programming comprises the longest course with 95 steps, while the Excel and Self Confidence courses consist of 78 and 75 steps, respectively.

Table 4.1 presents the number of steps in each course.

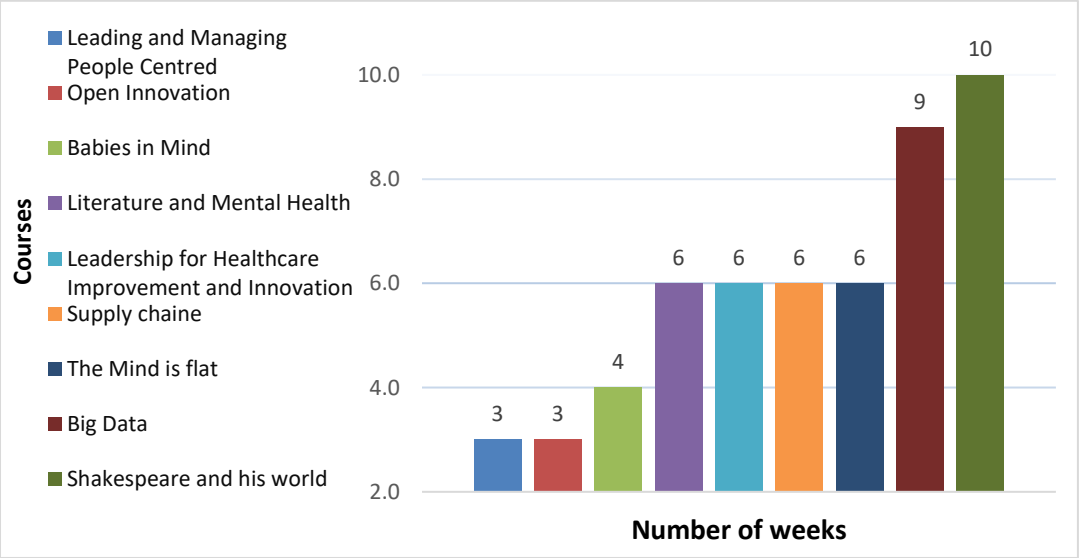


Figure 4.4 Number of weeks in each course.

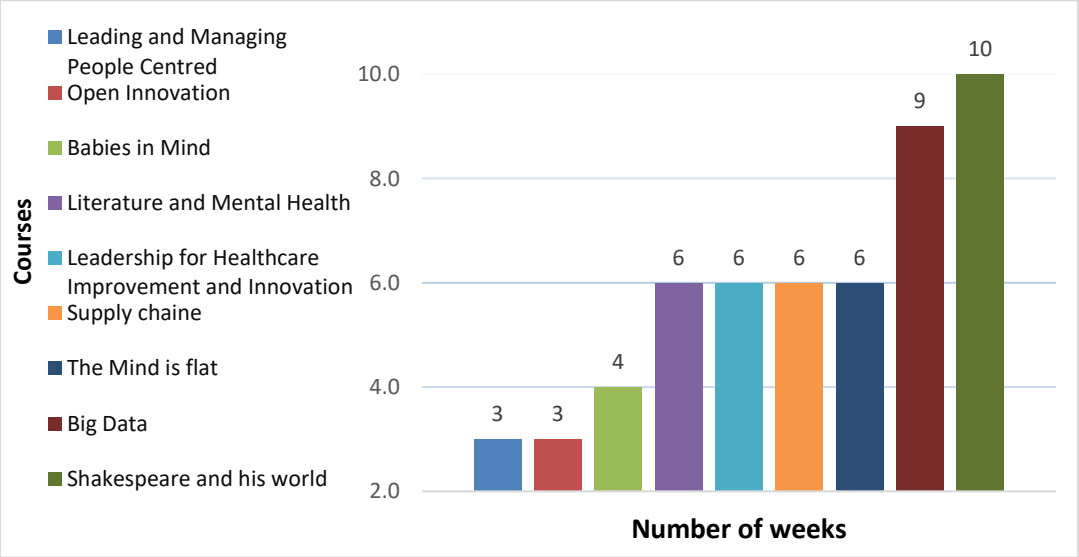


Figure 4.4 shows information about the duration of various MOOCs provided by different sources, such as Warwick University, Durham University, and the Rawaq platform. The figure illustrates the number of weeks for each of the listed courses. The duration of the courses offered differs in length. The Leading and Managing People-Centred Change course and Open Innovation course have a shorter time frame of three weeks each, while Shakespeare and His World has the longest duration (ten weeks). The remaining courses exhibit varying durations: the Babies in Mind course spans four weeks, while the rest of the courses extend for six to nine weeks. The number of steps listed in shows the lengths of each course. By analysing the duration of the courses with respect to the courses' provider, it can be observed that the courses delivered by Durham University show a relativity shorter period compared to the courses offered by Rawaq and Warwick University. Durham University provides two courses, Open Innovation in Business and Leading and Managing People-Centred Change, which contain 38 and 30 steps, respectively.

Warwick University provides a variety of courses, including Shakespeare and His World, which comprises 134 steps, and Babies in Mind, which consists of 75 steps, making it the shortest course offered. Rawaq offers courses of varying length: Java Programming comprises the longest course with 95 steps, while the Excel and Self Confidence courses consist of 78 and 75 steps, respectively.

Table 4.1 shows the lengths of each course. By analysing the duration of the courses with respect to the courses' provider, it can be observed that the courses delivered by Durham University show a relativity shorter period compared to the courses offered by Rawaq and Warwick University. Durham University provides two courses, Open Innovation in Business and Leading and Managing People-Centred Change, which contain 38 and 30 steps, respectively.

Warwick University provides a variety of courses, including Shakespeare and His World, which comprises 134 steps, and Babies in Mind, which consists of 75 steps, making it the shortest course offered. Rawaq offers courses of varying length: Java Programming comprises the longest course with 95 steps, while the Excel and Self Confidence courses consist of 78 and 75 steps, respectively.

Table 4.1 Number of Steps in each course

Source of the Data	Course	Number of steps
Durham University	Leading and Managing People Centred	30 steps
	Open Innovation in Business	38 Steps
Rawag	Excel	78 Steps
	Java programming	95 Steps
	Self confidence	75 Steps
Warwick University	Babies in Mind	75 steps
	Shakespeare and his world	134 steps
	Supply Chains	118 steps
	The Mind is flat	93 steps
	Literature and Mental Health	88 steps
	Leadership for Healthcare Improvement and Innovation	79 steps
	Big Data	105 steps

4.4.2 Discussion forum data

In MOOC platforms, students may engage in social interactions with one another and participate in discussions that are either specifically relevant to the content of the course or more general in scope. In addition to the main course material, MOOC forums could be a great place for students to ask and respond to questions from their peers.

Discussion forum data include students' postings, responses to those posts, ‘likes’ and ‘dislikes’ of other comments, and timestamps for these activities. These data are the third most frequently used by researchers to predict at-risk students (see Section 3.2.7). Figure 4.5 and Figure 4.6 show an example of the discussion forum data extracted from the FutureLearn and Rawq platforms. Some pieces of information have been intentionally blurred for privacy.

id	Author_id	parent_id	step	week_number	step_number	Text	timestamp	moderated	likes
[illegible]	[illegible]	[illegible]	1.1	1	1	Thanks Suzy Tobias ... looking forward to the course	2016-03-07 01:19:16 UTC		0
[illegible]	[illegible]	[illegible]	1.5	1	5	I think it can be a very useful tool for entrepreneurs and inve	2016-03-07 01:32:16 UTC		6
[illegible]	[illegible]	[illegible]	1.1	1	1	Hi All, thanks to the presenters, it's a great start and my	2016-03-07 01:40:22 UTC		0
[illegible]	[illegible]	[illegible]	1.4	1	4	Interesting research . I think people in the society that	2016-03-07 01:43:51 UTC		3
[illegible]	[illegible]	[illegible]	1.5	1	5	Perhaps it would also be interesting, using the future orienta	2016-03-07 02:00:49 UTC		7
[illegible]	[illegible]	[illegible]	1.8	1	8	Travel Data, Transaction Data, Social Data, Lifestyle, Preferen	2016-03-07 02:16:29 UTC		7
[illegible]	[illegible]	[illegible]	1.4	1	4	I wonder if this could be correlated to Gallup's research on gl	2016-03-07 02:52:39 UTC		3
[illegible]	[illegible]	[illegible]	1.4	1	4	Really interesting. So did the searches cover calendars and or	2016-03-07 02:55:31 UTC		1
[illegible]	[illegible]	[illegible]	1.5	1	5	law enforcement and military for terms involving illegal activ	2016-03-07 03:00:42 UTC		2
[illegible]	[illegible]	[illegible]	1.4	1	4	I wonder if there is a correlation between investment and "pr	2016-03-07 03:33:40 UTC		2
[illegible]	[illegible]	[illegible]	1.4	1	4	Interesting the link between forward looking and Gross Dom	2016-03-07 03:35:11 UTC		3

Figure 4.5 Example of discussion forum data (FutureLearn platform).

course_slug	user_id	comment	comment_text	lecture_id	comment_date_time	parent_re	reply_con	likes_numbe
human_nutrition	[illegible]	152471	Thanks Suzy Tobias ... looking forward to the course	4064	2016-02-01 19:39:33 +0300	parent_comment		6
human_nutrition	[illegible]	152474	Hi All, thanks to the presenters, it's a great start and my	4064	2016-02-01 19:45:15 +0300	parent_comment		3
human_nutrition	[illegible]	152478	Interesting research . I think people in the society that	4064	2016-02-01 19:51:59 +0300	parent_comment		2
human_nutrition	[illegible]	152479	Perhaps it would also be interesting, using the future orienta	4064	2016-02-01 19:53:59 +0300	parent_comment		4
human_nutrition	[illegible]	152486	I wonder if this could be correlated to Gallup's research on gl	4064	2016-02-01 20:04:14 +0300	parent_comment		2
human_nutrition	[illegible]	152507	Really interesting. So did the searches cover calendars and or	4064	2016-02-01 20:29:04 +0300	parent_comment		3
human_nutrition	[illegible]	152515	law enforcement and military for terms involving illegal activ	4064	2016-02-01 20:33:34 +0300	parent_comment		2
human_nutrition	[illegible]	152533	I wonder if there is a correlation between investment and "pr	4064	2016-02-01 20:52:42 +0300	parent_comment		1

Figure 4.6 Example of discussion forum data (Rwaq platform)

4.4.3 Assignment data

Like traditional classroom settings, MOOCs generally include coursework and save information about student submissions in the database. Such information may be automatically gathered via the use of tools such as multiple-choice quizzes or manually collected by uploading written assignments such as essays. Taking into consideration the massive number of students enrolled in MOOCs, peer review has become an increasingly common practice on many platforms, rather than relying only on the course instructors to evaluate students' works (Er et al., 2020). Figure 4.7 shows an example of peer review data extracted from the FutureLearn platform. Each student received feedback comments from three reviewers. Figure 4.8 and Figure 4.9 show an example of multiple-choice quiz data from both the Rwaq and FutureLearn platforms.

id	step	week_number	step_number	reviewer_id	assignment_id	feedback_one	feedback_two	feedback_three	created_at
289021	10.3	10	3	[illegible]	179262	[illegible]	[illegible]	[illegible]	2017-03-30 15:52:51 UTC
289038	10.3	10	3	[illegible]	180353	[illegible]	[illegible]	[illegible]	2017-03-30 17:16:55 UTC
290743	10.3	10	3	[illegible]	180353	[illegible]	[illegible]	[illegible]	2017-04-06 12:18:42 UTC
292732	10.3	10	3	[illegible]	181671	[illegible]	[illegible]	[illegible]	2017-04-14 05:49:23 UTC
292843	10.3	10	3	[illegible]	183001	[illegible]	[illegible]	[illegible]	2017-04-14 17:05:15 UTC
293017	10.3	10	3	[illegible]	183080	[illegible]	[illegible]	[illegible]	2017-04-15 19:45:54 UTC

Figure 4.7 Example of peer review data (FutureLearn platform)

learner_id	quiz_question	week_number	step_number	question_number	response	submitted_at	correct
[redacted]	2.11.1	2	11	1	1	2015-11-25 13:31	TRUE
[redacted]	2.11.1	2	11	1	2	2016-01-13 11:27	FALSE
[redacted]	2.11.1	2	11	1	3	2016-01-13 11:27	FALSE
[redacted]	2.11.1	2	11	1	4	2016-01-13 11:27	FALSE
[redacted]	2.11.2	2	11	2	1	2016-01-13 11:30	FALSE

Figure 4.8 Example of multiple-choice quiz data (FutureLearn platform)

course_slug	student_id	attempt_id	question_id	assignment_id	question_solution_result
human_nutrition	[redacted]	404796	4605	859	correct
human_nutrition	[redacted]	404796	4606	859	wrong
human_nutrition	[redacted]	404796	4607	859	correct
human_nutrition	[redacted]	404799	4605	859	correct
human_nutrition	[redacted]	404799	4606	859	wrong

Figure 4.9 Example of multiple-choice quiz data (Rwaq platform)

4.4.4 Demographics data

Learners are often asked to enter their demographic information during registration for a MOOC. This information may include gender, age range, and occupation. Nonetheless, because these details are mostly voluntary, MOOC demographic data are hardly obtained in a considerable percentage compared with automatically collected data such as clickstream data (Morris et al., 2015). However, the unavailability of this kind of data has lately motivated several research studies to profile learners and anticipate their demographics using other data types such as students' clickstreams and forum data (ALJOHANI and MUSLIH, 2022). Figure 4.10 and Figure 4.11 show examples of demographic data obtained from the Rwaq and FutureLearn platforms.

course_sl	student_i	student_a	country	education	job
java_prog	[redacted]		SA		التسويق
java_prog	[redacted]		SA	ثانوى	
java_prog	[redacted]	27	IN	جامعى	Software testing Enginreer
java_prog	[redacted]		DZ		
java_prog	[redacted]		MA	جامعى	طالب
java_prog	[redacted]		SA	جامعى	
java_prog	[redacted]		PS		
java_prog	[redacted]	24	JO	جامعى	طالب جامعي
java_prog	[redacted]	37	SA	جامعى	مدير مشاريع
java_prog	[redacted]	25	YE	جامعى	طالب

Figure 4.10 Demographic data on the Rwaq platform.

learner_id	gender	country	age_range	highest_education_level	employment_status
[illegible]	female	RU	18-25	university_masters	working_full_time
[illegible]	female	GB	56-65	secondary	retired
[illegible]	female	SA	18-25	secondary	full_time_student
[illegible]	female	GB	36-45	tertiary	looking_for_work
[illegible]	female	GB	56-65	secondary	working_full_time
[illegible]	female	GB	26-35	secondary	unemployed
[illegible]	female	GB	56-65	university_degree	not_working
[illegible]	male	GB	>65	university_masters	retired
[illegible]	male	GB	>65	apprenticeship	retired

Figure 4.11 Demographic data on the FutureLearn platform.

4.4.5 Questionnaire data

When researchers aim to collect data, questionnaires are beneficial ‘to obtain information about the thoughts, feelings, attitudes, beliefs, values, perceptions, personality, and behavioural intentions of research participants (Johnson and Christensen, 2019). For example, administering questionnaires is one method for gauging how students feel about the course being offered (Liyanagunawardena et al., 2015). Several researchers studying MOOCs have used questionnaires to collect data from students; for example, (Barak et al., 2016) conducted a pre- and post-course questionnaire survey to investigate the influence of motivation on MOOCs. Another study by (Arnavut et al., 2019) used a questionnaire to evaluate the perspectives of students who participated in MOOCs and viewed the videos.

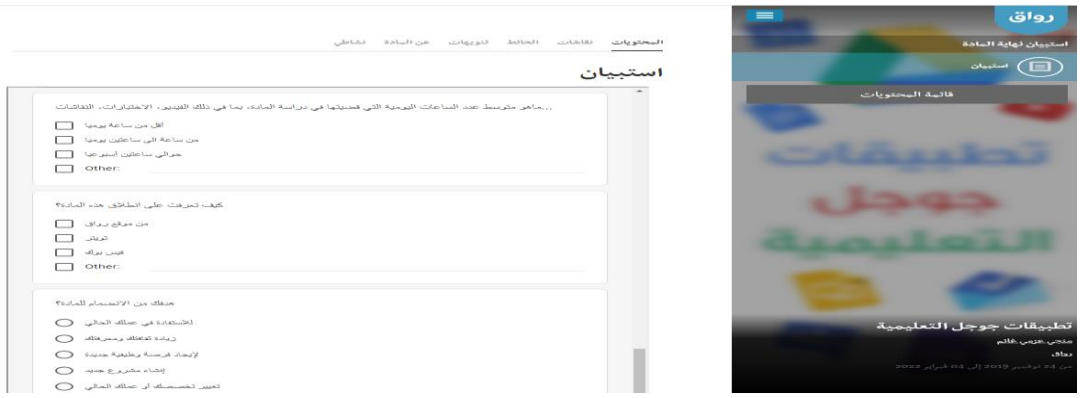


Figure 4.12 A screenshot of a short post-course questionnaire on the Rwaq platform.

4.5 Extracting Raw and Computing Aggregated MOOC Indicators

4.5.1 Sentiment analysis

Sentiment analysis has become valuable to solving a wide range of problems, extracting opinions, and making decisions across different disciplines and fields, including sociology, marketing and advertising, psychology, economics, and political science (Bakharia, 2016). Only relatively few studies have employed sentiment analysis in MOOCs. In this thesis, we used the outcomes of sentiment analysis to generate potential indicators of student behaviours, such as the number of positive/negative/neutral comments or replies. To achieve this, a natural language processing (NLP) tool called TextBlob[4] has been employed to classify students' comments into three categories: positive, neutral, and negative. TextBlob is an NLP-oriented Python library trained on a movie review corpus. TextBlob offers a simple API to measure the polarity and subjectivity of a textual dataset for certain tasks such as sentiment analysis and more complex text processing tasks (Saha et al., 2017, Gujjar and HR, 2021). The tool has been widely used on similar datasets extracted from several social media platforms and proved to be an effective tool for sentiment analysis (Vyas and Han, 2019, Dutta, 2021, Lohar et al., 2021, Gryllos et al., 2017). In addition, the TextBlob tool can be used for Arabic text classification (Hadwan et al., 2022). This would help in understanding student expectations and overall satisfaction with the course contents and outcomes in FutureLearn and Rwaq platforms.

4.5.2 Features extraction

We have considered it best to analyse each course independently, merging only the data from different runs of each course, as some courses offer quizzes every week on subjects of

[4] **https://textblob.readthedocs.io/en/dev/**

different nature and/or difficulty levels, whereas others skip some weeks. The latter was made possible, as all courses had runs (within that course) of similar structure.

From the first interaction with the platform, each student's activities were logged, along with a timestamp, into our dataset, with the unique student ID. Our longitudinal dataset, which consisted of the repetitions of the 12 courses delivered over consecutive years, has enabled a deep exploration of student behaviours. In this in-depth analysis, we used multi-modal data from many perspectives. First, we use time-related, numerical, and textual data. Second, we used data of different granularity. From a time-related perspective, we analysed data at the level of a timestamp: a day, a week, several weeks, or the whole course. In total, we extracted 21 features to be used in different experiments and gathered the following:

1. ***Registration day:*** as students can register before or after the course's official starting date, this feature presents how many days between the registration date and the starting date for a given student.
2. ***Number of accesses steps per week****:* how many steps are accessed by a given student per week.
3. ***Number of days:*** how many days the student accessed the course per week.
4. ***Number of correct answers per week*****:** questions within quizzes answered correctly by a given student.
5. ***Number of wrong answers per week*****:** as students can have multiple attempts to a question, before they get it right, this counter controls how many of those wrong attempts were made by a given student.
6. ***Number of attempts per week***: number of wrong answers per week plus the number of correct answers per week. This is a measure of the total activity of a student in terms of quizzes per week, showing their engagement along this axis.
7. ***Time spent:*** this feature measures the time spent by given student to complate each step accessed.

8. ***Number of comments per week***: as students can comment on any 'step', they can produce varying numbers of comments each week. This is the most straightforward way to measure their social contribution and engagement.
9. ***Number of likes received per week***: this is a clearly positive social construct; it is a measure of the influence, popularity, as well as of engagement of other students with a given student.
10. ***Number of positive comments per week***: this is an aggregate measure, derived based on sentiment analysis, to measure the (positive) nature of engagement of a student with peers.
11. ***Number of negative comments per week***: this is a similar measure as the one above, measuring (negative) nature of engagement of a student with peers.
12. ***Number of neutral comments per week***: this is a similar measure as the one above, measuring (neutral) nature of engagement of a student with peers.
13. ***Number of replies posted per week***: to incorporate part of the social interactive engagement element, we track how many replies a student places to others, thus where they go beyond uttering their opinions in public, but instead, consider the opinions of others.
14. ***Number of replies received per week***: this is a social construct, but also a measure of the popularity, and hence influence of a given student on their peers - receiving comments shows how engaged other students are with this particular one.
15. ***Number of positive replies posted per week***: this is based on sentiment analysis, a measure of the nature of the social engagement of the current student with their peers.
16. ***Number of negative replies posted per week***: this is a similar measure as the one above, measuring negative engagement.
17. ***Number of neutral replies posted per week***: this is a similar measure as the one above, measuring neutral engagement.

18. ***Number of positive replies received per week***: measure of the impact of a student on others, as well as of the nature of this impact.
19. ***Number of negative replies received per week:*** this is a similar measure as the one above, for negative engagement.
20. ***Number of neutral replies received per week:*** this is a similar measure as the one above, for neutral engagement.
21. ***Number of jumping activities:*** to incorporate students' learning patterns element, we track how many jumping activities a student made per week (non-sequential movement between the course contents).

4.6 Statistical Analysis and visualisation

The Shapiro test is used to determine whether or not data are normally distributed. For normally distributed data, a p-value of ≥ 0.05. Otherwise, the data are not normally distributed (Zhou, 2022).

In this thesis, the Shapiro test was used to determine the normal distribution of the variables in each group (completers and non-completers). Depending on this, we then used a *T*-test for the normally distributed data or the Wilcoxon rank-sum/Mann-Whitney test for the non-normally distributed data (Massimiani et al., 2019).

Moreover, the Pearson correlation coefficient test (Benesty et al., 2009) was used to assess the relationship between the SDT constructs and the success measures.

4.7 Visualisation tools

Graphviz is an open-source graph drawing tool in Python. Graphviz can be used to construct graph objects composed of various nodes and edges (Ferreira, 2017). In this thesis, the Graphviz package was used to create graphical representations of students' activities in MOOCs. For example, by using Graphiz, the thickness of the edge can be modified according to the total number of transitions. This will be useful in visualising students' activities, specifically jumping behaviours, according to the corresponding transition count.

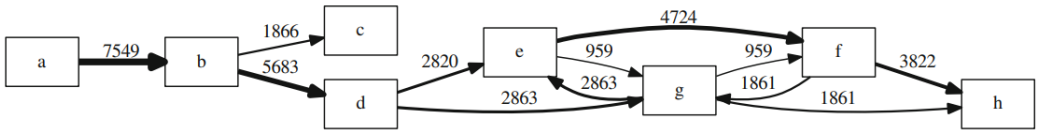


Figure 4.13 Example Graphviz transition graph based on the transition count (Ferreira, 2017)

4.8 Predictive Machine Learning Approaches

According to the no free lunch (NFL) theory, there is no one model that is optimal for resolving all types of problems (Adam et al., 2019). Therefore, the determination of machine learning algorithms depends on several factors, including the data type, the number of input attributes, and the complexity of the prediction task (Hafez et al., 2017). Borisov et al. (2022) conducted a survey of machine learning techniques for tabular data across different real-world datasets. The results indicate that classical machine learning models exhibit superior performance compared to deep neural network-based methods for datasets of small and medium sizes (less than 1 million samples).

The dataset used in this thesis is a tabular dataset with numeric features, and the records range between 6,071 and 83,543; therefore, this thesis focuses on classical machine learning models, given their compatibility with tabular datasets. The decision to put more emphasis on conventional machine learning models is consistent with recent research indicating that advanced deep learning models may not necessarily exhibit superior performance compared to traditional machine learning algorithms in MOOC prediction.. For example, Aljohani and Cristea (2021) and Sebbaq and El Faddouli (2022) have reported that implementing complex deep-learning models may not necessarily result in improved prediction performance.

On the other hand, Fu (2021) used a dataset with more features extracted from students' activities in the first 30 days of the course to predict active students after one month. The author shows that the deep-learning model (which combines CNN and LSTM) outperformed MOOC's baseline model for dropout prediction. Despite the fact that using more features enhances the performance of the prediction models, these models cannot be put into practice in real-world settings because participants are most likely to drop out in the first few weeks. Therefore, early prediction should take place in Week 1 of the courses.

Another study from 2018 highlighted that a predictive model's efficacy is not directly related to its level of complexity. The author showed that shallow algorithms, such as decision tree models, exhibited better performance in predicting dropout rates in MOOCs compared to long short-term memory (LSTM) (Ortigosa, 2018).

Furthermore, the efficacy of the shallow machine learning model was observed in the KDD Cup 2015 competition to predict dropout students in MOOC. In this competition, a shallow machine learning model (XGBoost) was utilised by all the winning teams in the top 10 (Chen and Guestrin, 2016).

Additionally, shallow machine learning models require lower computational resources and exhibit faster training and prediction capabilities compared to deep learning algorithms, making them easier to apply in real-life scenarios, which was a consideration for our research (Velu et al., 2023, Yang et al., 2021, Özdaş et al., 2022).

For comparison with deep learning models, we will apply multilayer perceptron (MLP). This is one of the most popular deep-learning models, so it is well-suited for tabular data (Balles et al., 2021, Si, 2022).

The machine learning models that are most frequently employed for tabular datasets have been chosen based on the review conducted in Section 5.2. These models include random forest (RF) (Breiman, 2001), gradient boosting machine (gradient boosting) (Friedman, 2001), adaptive boosting (AdaBoost) (Freund and Schapire, 1997), XGBoost (Chen and Guestrin, 2016), ExtraTrees (Geurts et al., 2006), logistic regression (LR) (Rawlings et al., 1998), K-nearest neighbour (Anchalia and Roy, 2014), and multi-layer perceptron (MLP) (Gardner and Dorling, 1998). These models have been widely applied and have demonstrated good results across different (see section 3.2.5). The following sections present a summary of the predictive machine learning techniques used to achieve the goals of this thesis.

4.8.1 Decision Tree

A decision tree is one of the most frequently used algorithms for prediction purposes (Mazraeh et al., 2019). It relies on a set of question-and-answer processes to categorise data (Aitkenhead, 2008). The algorithm is built in the form of a tree structure with three major elements: the root node, the internal node, and the leaf node (Yu et al., 2010).

The root node is the starting point of the tree, where the data is divided into different subsets depending on specific feature values (known as cutoff or threshold values); each subset contains distinct instances (Molnar, 2020). The feature in the root node is chosen based on Attribute Selection Techniques (e.g. Gini index technique) (Dorfman, 1979).

The internal nodes are in the middle of the tree between the root node and the leaf nodes. Through splitting, all instances are distributed among various subsets until they reach the last point. The leaf node (also called the terminal) is the last point on the tree; at this point, there is no more branching on the decision tree (Chauhan, 2019). Figure 4.14 shows an example of a decision tree structure.

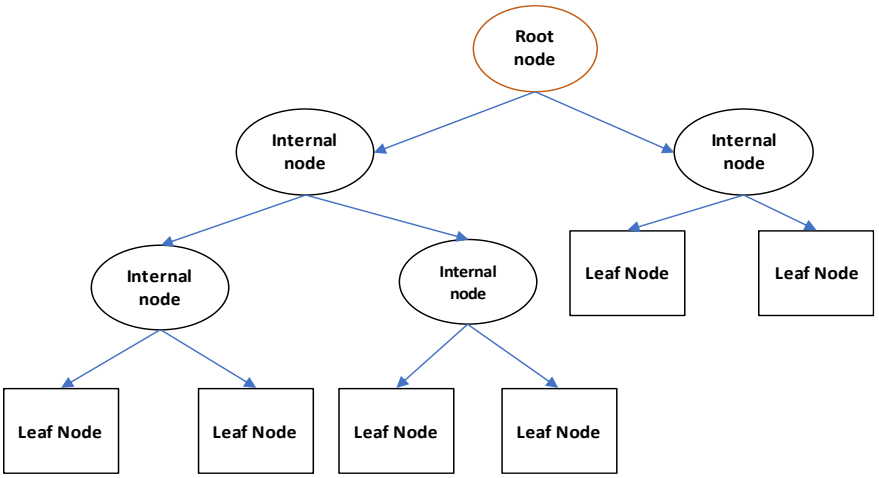


Figure 4.14 Example of a Decision Tree structure

4.8.2 Random Forest and Extra Tree algorithm

The random forest (RF) algorithm is known as a part of the supervised ML group of algorithms that use the bagging technique. It is used to handle regression and classification

problems by employing an ensemble methodology that involves multiple classifiers (Azar et al., 2014). The idea of creating RF is derived from the decision tree algorithm by combining a large number of decision trees (Mishina et al., 2015). However, since no optimal number of trees applies in all models, the number of trees should be chosen on a trial-and-error basis until satisfactory performance is obtained. Typically, the starting point for the number of trees used by most developers is 100 – the default number (Oshiro et al., 2012).

In contrast to the decision tree, the features are chosen randomly during the splitting of the nodes, and the training dataset is extracted randomly from the original dataset; this process is known as bootstrap aggregation (Chauhan, 2019). The random features and the random dataset are repeated with all trees until a forest is built. Consequently, each decision tree in the forest is unique in terms of the tree's structure and the dataset used to train the tree. The final decision for classification problems is determined based on the majority vote of each decision tree (Liaw and Wiener, 2002). Figure 4.15 shows an example of an RF structure and a voting system, where N refers to the number of trees.

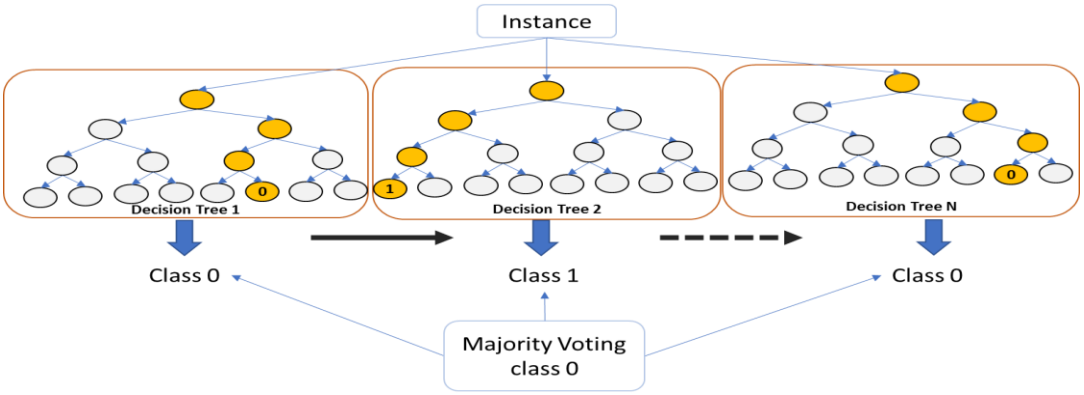


Figure 4.15 Random Forest structure

The 'extremely randomised tree' is another name for the 'extra tree' algorithm. Extra tree also creates a prediction model for classification and regression problems. It is similar to the RF algorithm because it constructs several trees (Ekanayake et al., 2022). Nevertheless, extra trees and RF may be distinguished from one another in two ways: the extra tree algorithm does not use bootstrap replicas, and the nodes are separated using randomised feature selections (Shang et al., 2022).

4.8.3 Boosting algorithms

The boosting approach refers to a set of algorithms that merge many weak models into a single robust model (McDonald et al., 2003). In contrast to many other ML models that concentrate on just one model performance, boosting algorithms seek to obtain a more robust prediction capability by training a series of models. Therefore, each model helps its predecessor by correcting errors. For example, during the data training phase, when the first decision tree is built, the wrongly categorised records in this tree are given precedence over other records before they are sent to the second tree. Therefore, each tree focuses on high-weight instances from the previous tree to classify them correctly(Schapire, 1999, Nikolaou et al., 2016).

Figure 4.16 illustrates the structure of boosting algorithms. This figure shows that the instances from the first model are sent to the next model as inputs with new weights. The procedure continues until the specified condition is met. This is the basic concept for all kinds of boosting algorithms, such as AdaBoost, gradient boosting, and extreme gradient boosting (Schapire, 1999, Sigrist, 2018).

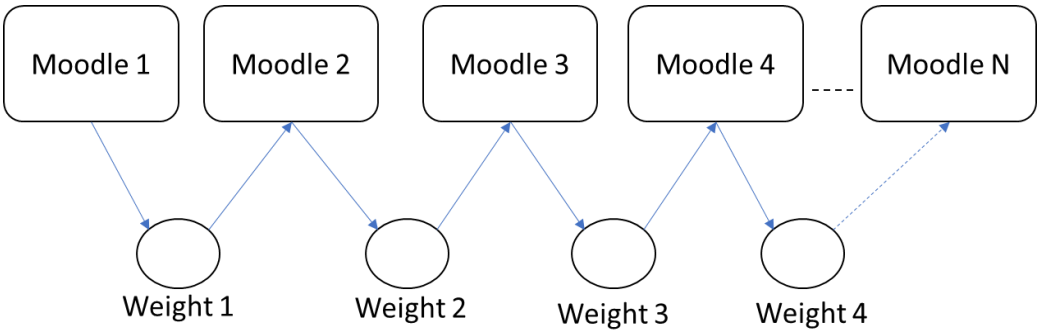


Figure 4.16 Boosting Algorithms

4.8.3.1 AdaBoost

AdaBoost is part of the boosting algorithms family. AdaBoost is known as an adaptive algorithm because the weights are redistributed to each instance (Mazini et al., 2019, Kumar et al., 2011).

By comparing the RF algorithm with the AdaBoost algorithm, RF creates many different decision trees that vary in size. However, AdaBoost makes a forest of trees of only one size, known as 'stumps' or 'weak learners', that consist of only one node linked with two or more leaves (Wang et al., 2016, Hu et al., 2008). Figure 4.17 shows the structure of the AdaBoost algorithm, where W denotes the weighted instances that are sent to the next stump. For the sake of illustration, each stump's size in the figure refers to the influence in making the final decision.

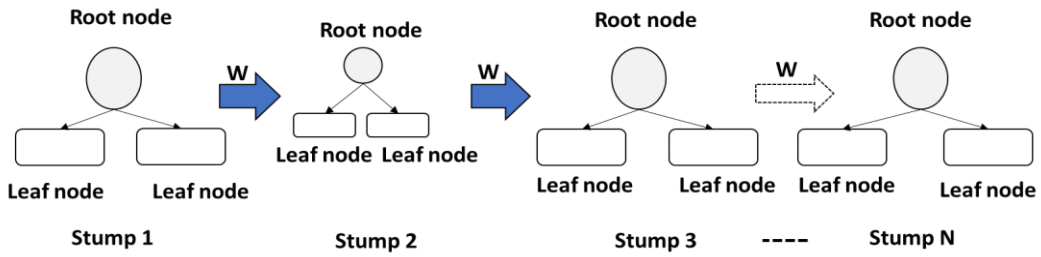


Figure 4.17 AdaBoost Algorithm

The first step of Adaboost is to measure the impurity of each feature in the dataset; the attribute with the least impurity is chosen as the first stump in the chain and in the initial stump, and all instances have equal weight (Cahyana et al., 2019).

The wrongly classified instances by the earlier stump obtain a higher weight than the correct ones. Thus, the following stump emphasises instances with higher weight to avoid making the same errors (Shen and Bai, 2004). The instance weight on the first stump can be calculated using Equation 4.1 where N refers to the total number of instances (Rojas, 2009).

$$w = \frac{1}{N} \in [1,0] \tag{4.1}$$

The power to make the final decision in the AdaBoost algorithm is based on the misclassification rate (sum of all errors) for each stump. Therefore, the stumps with a high weight have more influence than others on the final decision of the model. The stump's weight (SW) is calculated using Equation 4.2, where TR refers to the sum of all errors in a particular stump (Rojas, 2009).

$$SW = \frac{1}{2} \log\left(\frac{1 - TR}{TR}\right) \quad (4.2)$$

4.8.3.2 Gradient Boosting and XGBoost

In 2000, Friedman established a connection between AdaBoost and vital statistical concepts, enabling the boosting framework to include loss functions (known as gradient boosting machines, or GBM)(Li et al., 2022).

Gradient boosting is similar to AdaBoost in that it combines a large number of weak learners sequentially; the learners concentrate on the errors made by the preceding learners (Aldor and Helle, 2021). However, in gradient boosting, each learner attempts to predict the residual of the previous learner in order to minimise the difference between the actual and expected values (Cui et al., 2018).

Extreme Gradient boosting (XGBoost) is an enhanced, efficient form of gradient boosting; it has been built to be a faster and more flexible method (Chen and Guestrin, 2016). The XGBoost algorithm possesses advantages over conventional boosting algorithms. Because gradient boosting is sequential, it is very difficult to synchronise the procedures. Therefore, XGBoost is designed to enable parallel processing using strong distributed processing engines such as Spark and GPU(Boehmke and Greenwell, 2019). In addition, to prevent overfitting, XGBoost offers a number of regularisation parameters, including gamma, alpha, and lambda, to avoid excessive model complexity. Moreover, unlike Gradient boosting, XGBoost uses the dropout approach, a commonly employed method in deep learning to minimise overfitting and ensure that the model does not evaluate more trees if the tress do not provide an improvement (Boehmke and Greenwell, 2019).

4.8.4 Logistic Regression

The first recorded use of the logistic regression (also known as logit regression, LR) model was in 1845, when it was used in mathematical analyses of population expansion (Cokluk, 2010). LR is a statistical technique widely used in ML to estimate the likelihood that an instance belongs to a specific class (e.g., the probability that an email is a cyberattack). For

instances in which the estimated probability ($\hat{p}$) is greater than 0.50, the model predicts that the instance belongs to the positive class (labelled '1'). Otherwise, the class labelled '0' indicates that the instance is a member of the negative class (see equations 4.3 and 4.4) (Géron, 2019).

$$\hat{p} = \mathrm{h}\theta(X) = \sigma(\theta^T . x) \tag{4.3}$$

Where $\mathrm{h}\theta$ is the hypothesis function and $\sigma(\cdot)$ is a sigmoid function

$$\hat{\mathrm{y}} = \begin{cases} 0 & \hat{p} < 0.5 \\ 1 & \hat{p} \geq 0.5 \end{cases} \tag{4.4}$$

Logistic Regression model prediction where outputs a number between 0 and 1

4.8.5 K-Nearest Neighbour algorithm

The K-Nearest Neighbour (KNN) algorithm was developed by Hart in 1968 (Hart, 1968). It is a fundamental instance-based learning approach that is extensively used in various domains because of its great efficiency and robustness. KNN is a non-parametric approach for dealing with classification and regression issues based on the idea of the least distance between comparable items, which posits that objects with similar properties remain close together (Deekshatulu and Chandra, 2013).

For simplicity, consider the following example, which involves a training dataset (TRD) and a testing dataset (TSD). The sample in the data set is represented in a vector format (x_1,x_2, ...,x_n,L). Each instance in the *TRD* is associated with a label (L). On the other hand, the class label (L) is not known for *TSD*.

In order to predict the label for *TSD*, the KNN algorithm calculates the distance between each instance in the *TSD* and *TRD*. There are various methods for calculating the distance between instances; the Euclidean distance is one of the most common measure methods (see equation 4.5) (Triguero et al., 2016).

$$D(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2} \quad (4.5)$$

Next, the KNN algorithm groups the *k* samples from the *TRD* that are nearest to the instance in the *TSD* and ranks them in ascending order based on their distance. Finally, the algorithm calculates a majority vote using the class label of the closest neighbours to assign the label (*L*) to an unlabelled instance (Triguero et al., 2016).

Figure 4.18 shows the KNN algorithm when k = 3 and when k = 5. The value of *k* may have an effect on the overall performance of the algorithm.

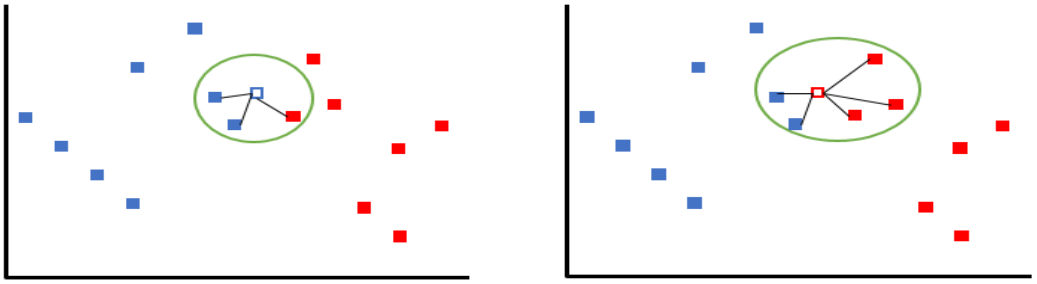


Figure 4.18 Nearest Neighbour Classification (*k*=3and *k* =5)

4.8.6 Multi-Layer Perceptron (MLP)

One of the most popular deep learning algorithms is the Multi-Layer Perceptron (MLP). It's primarily used to deal with supervised learning issues. It predicts unknown data by determining the relationship between input and output variables (Cunningham et al., 2008). Most MLP models are constructed based on a layered network architecture (an input layer, one or more hidden layers, and an output layer). The MLP structure is illustrated briefly in Figure 4.19 for a binary classification problem.

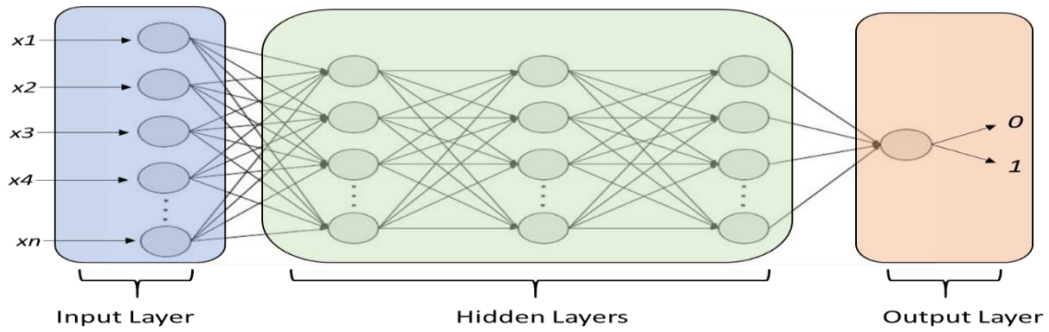


Figure 4.19 Example MLP with 1 Input Layer, 3 Hidden Layers, and 1 Output Layer

The primary goal of the fully connected layers in the middle of the model (hidden layers) is to determine the relationships between the input features (input layer) and the classification layer (output layer). Neurons in each layer communicate with neurons in the preceding and following layers through means of weight and function. The MLP employs a backpropagation technique to continuously adjust the weights and biases inside the hidden layers in order to reduce output error (ALHASSAN and NASSER, 2021).

Epilogue

In this chapter, we have introduced the methodology used to answer the research questions. In addition, we have presented the dataset and tools used to achieve the aim of this thesis. The employed technologies include feature selection, sentiment analysis, statistical analysis, visualisation, and predictive machine learning algorithms.

The following chapter will be the first core chapter of this thesis. We will explore the predictability of students' completion based on students' registration dates.

Chapter 5 : Earliest Predictor of Dropout in MOOCs: A Longitudinal Study of FutureLearn Courses

Prologue

This chapter presents the results of the analysis of data set of FutureLearn MOOC users over several runs, focusing specifically on non-completion, to determine if there are factors that can be identified before the students even start the course, that can guide teachers to target and support these students, so that they do not disengage from their learning.

5.1 Introduction

To analyse students' non-completion, various variables can be considered, such as student profile data (e.g., age, gender, country), behavioural patterns related to the consumption and generation of data when interacting with the course (e.g., reading, watching, writing, taking quizzes). This research instead, however, uses only one relatively simple variable, which, to the best of our knowledge, has not been studied in prior research in relation to completion predictability: ***the registration date***.

The advantage is that this data is available, in most of cases, even before the course starts; thus, if completion can be predicted from it, very early intervention is possible. Hence, this research targets the following research questions:

- **RQ1.1**: Can the date of registration (in terms of distance from the course start) of students predict their completion (or non-completion)?
- **RQ1.2**: How can dropout rates be alleviated based on the registration date?

5.2 MOOCs Analytics and Mining

The work presented in this chapter is closest related to the area of retention-versus-dropout in MOOCs. The issue of massive open online course systems (MOOCs) having high dropout rate, as said, by many researchers (Dillon et al., 2016, Jordan, 2015, Liang et al., 2016). Various solutions have been proposed, such analysing a students' activity in online forums (Wen et al., 2014a), or analysing the students' click-stream (Kloft et al., 2014, Sinha et al., 2014), classification methods of longitudinal engagement trajectories (Coffrin et al., 2014) and monitoring video views (Guo et al., 2014). However, most of these approaches are only able to predict retention or dropout once a student has started learning and, importantly, interacting with a MOOC. For example, in (Kloft et al., 2014), correlation was observed between activities in the latter part of the course and dropping out. Also (Wen et al., 2014a) investigated the relationship between learner sentiments expressed on forums and the chance of dropping out.

To the best of our knowledge, none of the published evaluation studies attempts to predict dropout only from the very first interaction with the system – ***the registration***.

Moreover, the FutureLearn platform has not been studied as frequently as other MOOC platforms (e.g., Coursera and edX) (Vigentini et al., 2017). Recent work on FutureLearn data exploration includes social aspects (Chua et al., 2017a), dashboard development (Vigentini et al., 2017), pedagogies on MOOCs (Mohamed and Hammond, 2018), and reviews of empirical MOOC literature(Sinha et al., 2014, Zhu et al., 2018).

Another common point for prior researches is that they have analysed only a few courses in a MOOC (e.g. (Guo et al., 2014) claims to be the largest study, with only 4 courses, with only one run each; one MOOC with one run in (Barba et al., 2016); three courses used in (Atapattu et al., 2016)). They have often only analysed courses on the same, or similar, subjects (Chua et al., 2017b).

This research performed comparative longitudinal studies of several runs of a large number of different courses on varied subjects.

5.3 Setup: Terms and Methodology

Firstly, a few definitions are required, as follows. Here we are studying MOOC courses which have an official starting date (considered as date 0) and which are expected to run over a specific number of weeks, after which they end. Non-completing students are students who have not completed the course. Enrolled students are students that completed enrolment. Note however that these can be also students who have never logged into the course, but just have enrolled for it. Completing students are students who have completed the course.

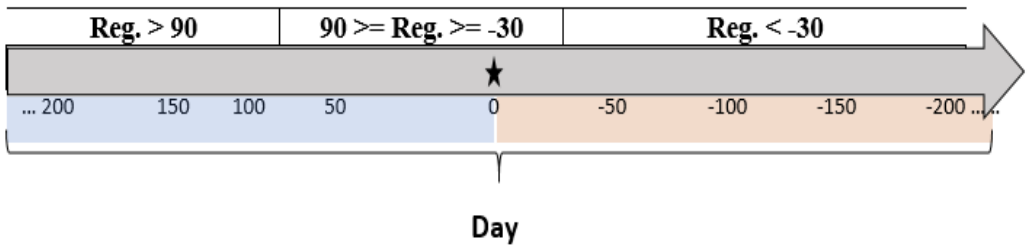
To address the research questions set in this chapter **RQ1.1** and **RQ1.2**, we analyse six courses on different subjects delivered by the University of Warwick (from literature to computer science to social sciences: Literature and Mental Health, Shakespeare and His World, Big Data, Supply Chains, Babies in mind, The Mind is Flat), each with several runs, for a total of 23 runs, for a total of 240,568 students, employing a variety of statistical methods. These courses were freely available for anyone and allowed for enrolment at any time.

5.4 Results

Results show that out of 240,568 registered students, only 7,437 (~3%) complete (see Table 5.1); thus, this highlights an extreme MOOC non-completion issue, at the lower end of the boundary of 3-15% (Jordan, 2015, Coffrin et al., 2014). We further analyse the normality of the registration data, results showing that registration is not normally distributed ($p < 2.2e-16$). Thus, the T-test cannot be applied. Thus, we have applied the non-parametric Wilcoxon test instead – firstly, to all data across all courses and runs (column 'Total' in the table below). We notice that students register, on average, 1 month (30.47 days) in advance of their FutureLearn courses start. We can also see that non-completers tend to register, possibly non-intuitively, around 3.5 days earlier, on average, than completers (see more discussion on this in section 5.5 and that this difference is statistically significant. We also can estimate that non-completers are the ones influencing the overall average (due to their larger number) and the large variance (with a maximum of 256 days in advance, up to 809 days after the course starts).

Table 5.1 Initial analysis of impact of registration date (Reg.) versus course starting date (here, 0), onto completion.

		Total	Reg. > 90	90 >= Reg. >= -30	Reg. < -30
ALL	**Data size**	240568	16522	214676	9370
	Avg.	30.47	142.14	25.37	-49.39
	Var.	2142.46	1160.13	1008.42	893.49
	Max.	256	256	90	-31
	Min.	-809	91	-30	-809
Completers	**Data size**	7437	279	7016	142
	Avg.	27.06	126.85	24.54	-44.69
	Var.	1459.73	1206.21	990.29	131.03
	Max.	210	210	90	-31
	Min.	-83	91	-30	-83
Non-completers	**Data size**	233131	16243	207660	9228
	Avg.	30.58	142.4	25.39	-49.46
	Var.	2163.86	115.33	1009.01	904.9
	Max.	256	256	90	-31
	Min.	-809	91	-30	-809
Who registers earlier?		Non-Completers	Non-Completers	Non-Completers	Completers
Wilcoxon's p		*p = 0.0010*	*p =1.94e^{-13}*	*p = 0.093*	*p = 0.033*



Before the course starts - After the course starts - ★Official starting day

Figure 5.1 Initial periods of the registration date

This large spread becomes more obvious in the box diagram (Figure 5.2). The figure shows that non-completers are responsible for most outliers, as well as the largest spread. It becomes visible from the figure that registering too early - or too late - will possibly result in non-completion; this is studied further, below, in order to quantify what 'too late' or 'too early' means in this context.

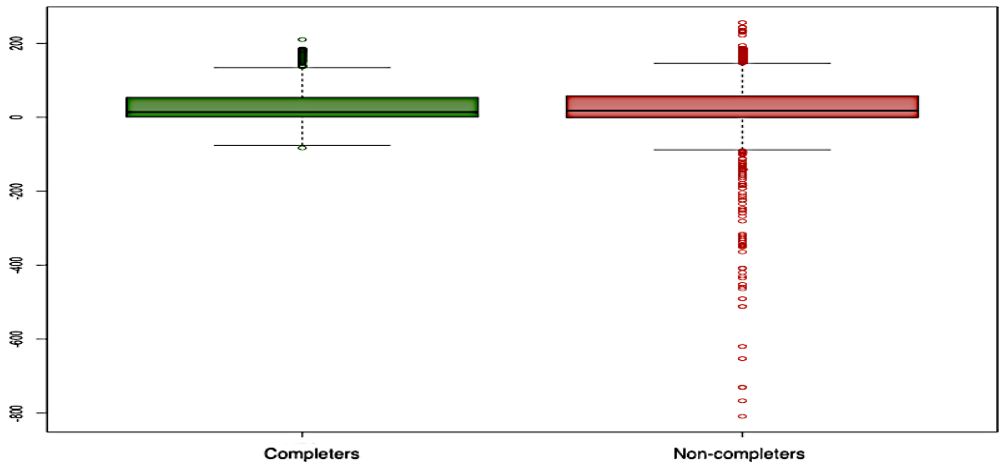


Figure 5.2 Box diagram for registration date for completers and non-completers across all courses and runs, in absolute values.

A further visual analysis of the spread of registration dates is shown in Figure 5.3 where the number of completers (in small green dots) and non-completers (in large red dots) are placed side-by-side, for each registration date. Beside the larger spread of the non-completers in terms of date, it can be clearly seen that their numbers are much larger as well (visually confirming that only less than 3% of the students actually complete). As these two spreads are at such very different scales, this data is further analysed separately in Figure 5.4 (for completers) and Figure 5.5 (for non-completers). The images show that, surprisingly, the shapes of the two graphs are relatively similar: beside the peak around the actual course starting date, there is a peek somewhere around 90-100 days before the course starting date.

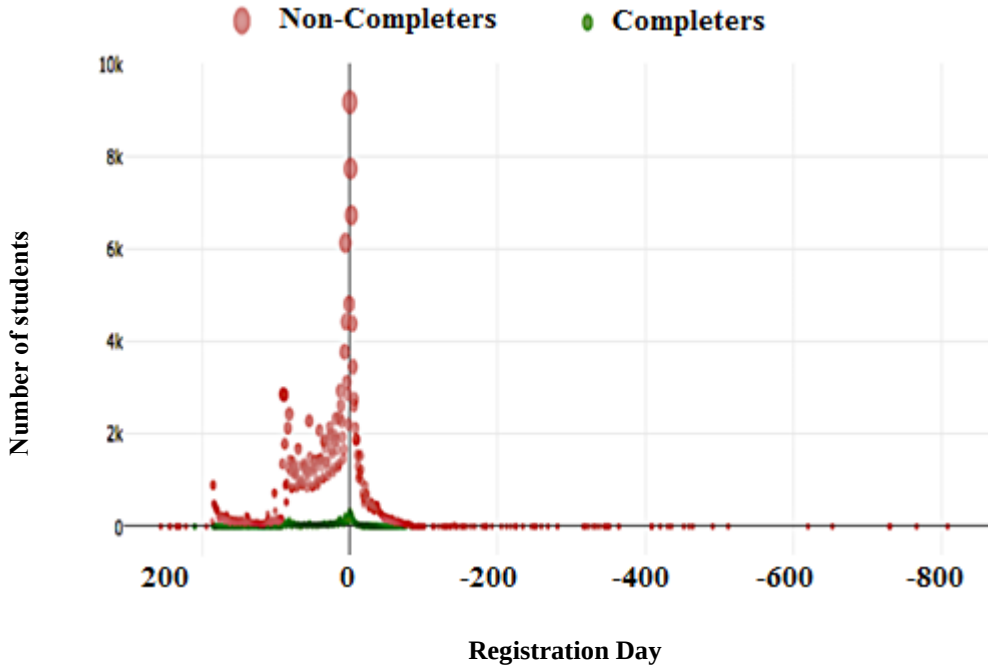


Figure 5.3 Completers (green) versus non-completers (red) across all courses and runs, in absolute values.

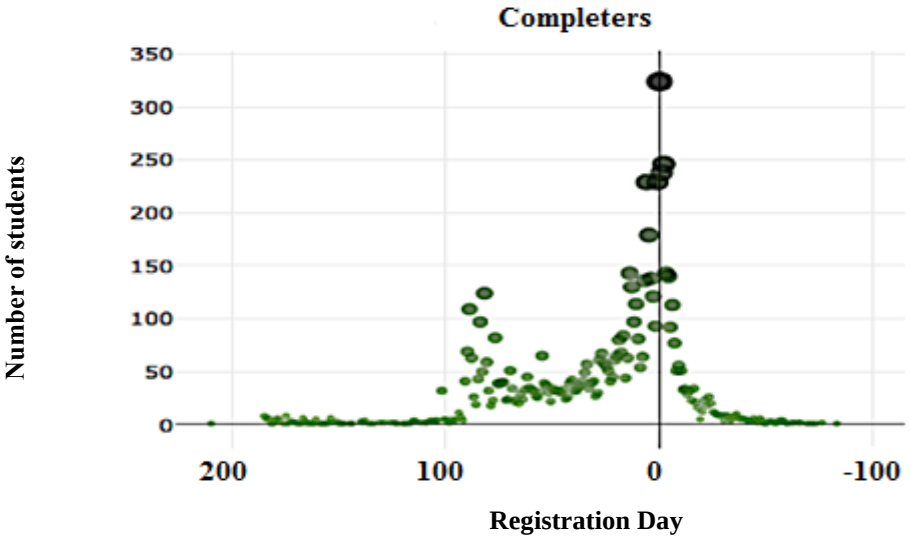


Figure 5.4 Completers and their registration dates.

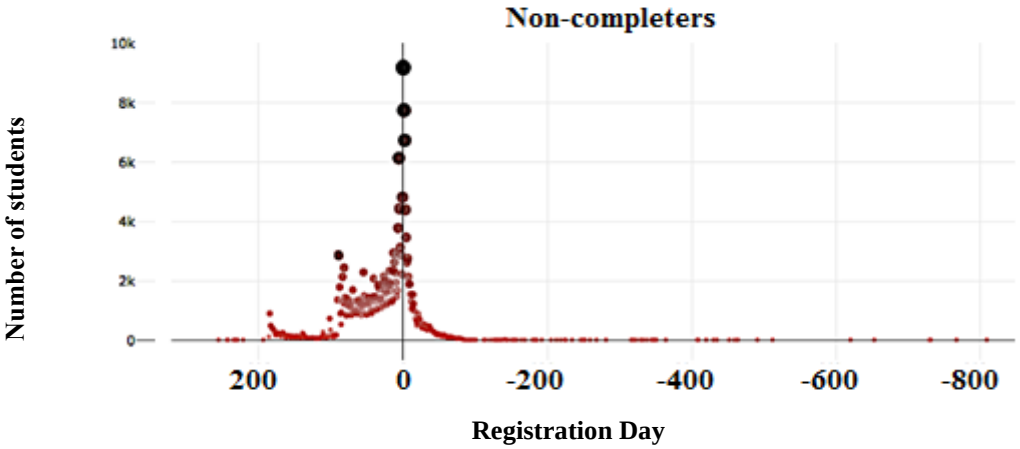
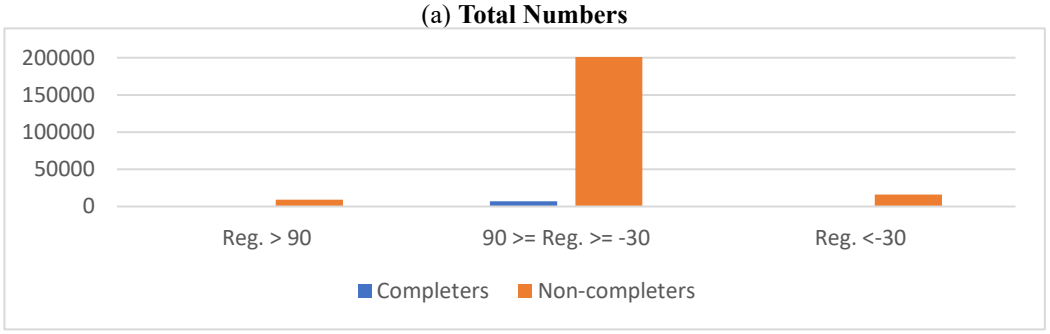


Figure 5.5 Non-completers and their registration dates.

Thus, we analyse this data further, taking 90 days later as one transition point, and using its symmetrical counterpoint of 90 days earlier as another. The latter results from Figure 5.4, where completers tend to disappear around that date. Thus, we specifically look at very early registrations (initially about 3 months – 90 days – in advance), late ones (1 month – 30 days – after course start) as well as the period in-between (see Figure 5.1). Table 5.1 further shows these initial results for the overall cohort for all registrations. It can be seen that the averages shift considerably, with early registrations averaging at 142.4 days before course start for non-completers, who register a significant 15 days earlier than completers; late registrations averaging at 49.46 days after course start, with non-completers registering about 5 days later, on average, than completers (significant at $p<0.05$). The overwhelming majority of completers, however, are in the middle region (7016/7437 or 94.34%). They register, on average, 24.54 days in advance, with completers registering about 1 day later than non-completers (but this is not significant). Figure 5.6 helps visualise this data, for the three periods. The total numbers are less informative (Figure 5.6, a), as the number of non-completers dominates the overall numbers. Thus, we use the percentage view (Figure 5.6, b), which shows that there is a larger percentage of completers than non-completers who register closer to the course starting time, and a smaller percentage of completers who register very early, or very late. However, the figure also shows that the majority of both completers and non-completers register in the identified central period.

(a) **Total Numbers**



(b) **Percentage**

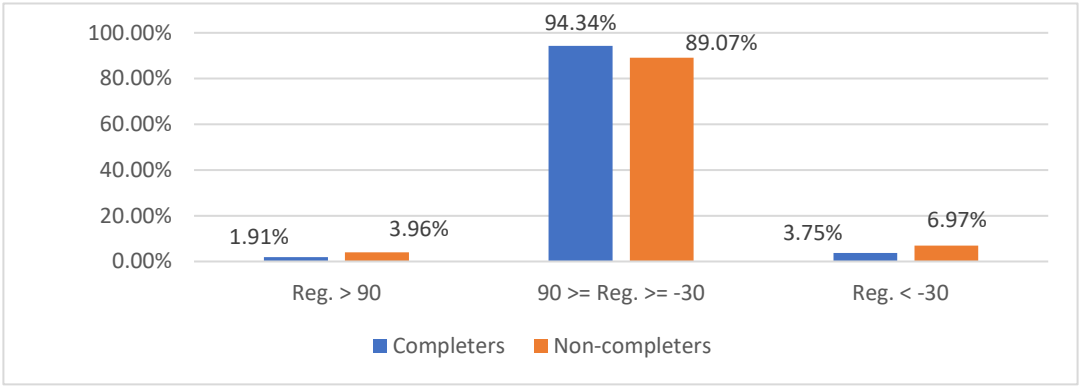


Figure 5.6 Completers (in blue) and non-completers (in orange) visualised as total numbers (a) and as a percentage (b) for the initial three periods identified in Table 5.1

Thus, it is clear that the central period needs further analysis, as additionally, the statistical results (as shown in Table 5.1). Also, the labels 'early' and 'late' have been applied based on visual information only. Thus, we considered defining periods more rigorously, based on the features of the data, starting with Avg.=30.47, the overall average number of days in advance of the course start that students register on, as well as the overall standard deviation, σ (see Figure 5.7). Interestingly, the 'early' (P1) and 'late' (P5) periods remain relatively similar - albeit better supported by the data - at 99.9 days in advance and 38.96 days after, respectively, confirming our initial intuition. The results for these periods also remain relatively similar, as can be seen in Table 5.2.

Table 5.2 Periods identified based on σ, the standard deviation; 'Reg.' stands for registration date; 'Avg.' stands for average.

Avg	30.47					
σ	46.29	P1 Reg.> Avg. + 3/2σ =~99.9	P2 [99.9, 53.6] Avg.+3/2σ >= Reg.> Avg. + 1/2σ	P3[53.6, 7.33] Avg. + 1/2σ >= Reg.> Avg. - 1/2σ	P4 [7.33, -38.96] Avg.-1/2σ >= Reg.> Avg. - 3/2σ	P5 Reg. < Avg.- 3/2σ =~-38.96
ALL	Data size	13941	52744	74489	93264	6130
	Avg.	151.25	74.27	27.97	-4.59	-57.17
	Var.	842.91	143.78	170.81	99.97	1187.73
	Max	256	99	53	7	-39
	Min.	100	54	8	-38	-809
Completers	Data size	198	1669	2392	3091	87
	Avg.	140.75	75.5	25.44	-2.92	-51.24
	Var.	1030.78	133.96	166.96	72.96	99.58
	Max.	210	99	53	7	-39
	Min.	100	54	8	-38	-83
Non-completers	Data size	13743	51075	72097	90173	6043
	Avg.	151.4	74.23	28.06	-4.65	-57.26
	Var.	838.67	144.06	170.72	100.8	1202.9
	Max.	256	99	53	7	-39
	Min.	100	54	8	-38	-809
Who registers earlier?		Non-Completers	Completers	Non-Completers	Completers	Completers
Wilcoxon's p		p =1.29e-06	p=4.75e-05	p= 2.2e-16	p= 2.2e-16	p= 0.041

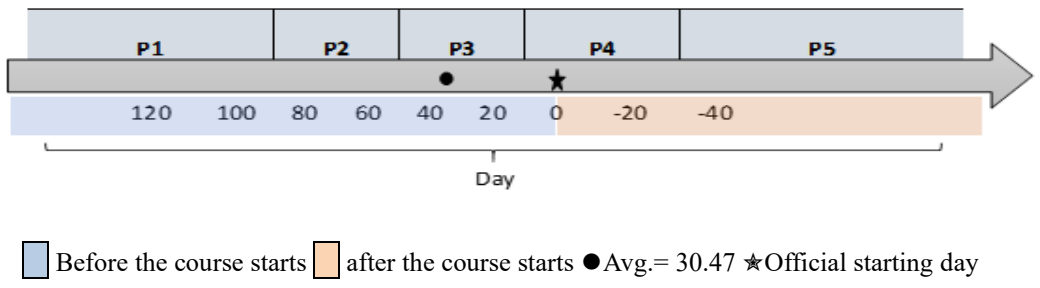


Figure 5.7 Early (P1) and Late (P5) periods

However, now we can analyse in more details the middle period by splitting it into three parts: the centre is half a deviation (1/2 σ) from the overall average Avg. both ways, and the sides contain the remaining periods, up to 3/2 σ, each way (as shown in Figure 5.7).

Here, we see some very interesting and potentially surprising fine-grained results: P2 completers and non-completers have only 1, however, significant, day, on average, between them. Interestingly, P3 and P4 show strong significant differences between the completers and non-completers, of opposite signs (2.62 and -1.73, respectively). Thus, completers register earlier in P2, later in P3 and earlier in P4. Figure 5.8 shows that, for all three middle periods, the percentage of completers is consistently larger than the percentage of non-completers. However, it shows that, for both completers and non-completers, both their absolute numbers (left) and their percentages (right) grow steadily between periods P2, P3 and P4, with their peak in P4.

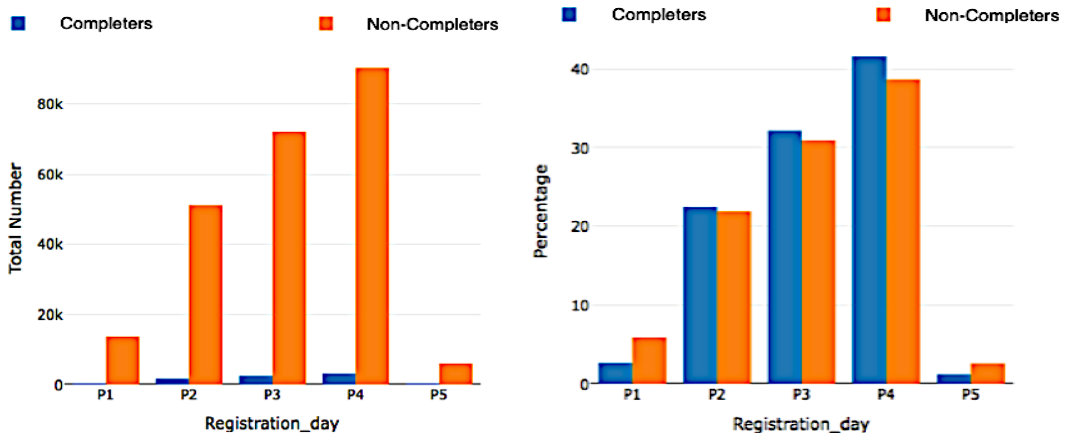


Figure 5.8 . Completers (in blue) and non-completers (in orange) visualised as total numbers (left) and as a percentage (right) for the five periods identified in Table 5.2.

5.5 Discussion and Extracted Rules

The results obtained are worthy of discussion, because some were not as straightforward as we initially expected. As in previous research by (Balakrishnan and Coetzee, 2013, Cristea et al., 2018, Hair et al., 2011, Jordan, 2015, Kloft et al., 2014, Koller et al., 2013, Rosé et al., 2014), in our Warwick courses, there are a substantial number of students who don't complete. In answering **RQ1.1**, indeed, registration time is a significant predicting factor. However, there is, for instance, not a simple answer to the question if the students should register early or late. Whilst we initially expected that "the early bird catches the worm" and thus students who register early would have a higher chance of completion, the general answer is in fact that, on average, registering later seems to be more advisable. Specifically,

and interestingly, we were able to find explicit periods of time, related to the course starting date, for which this question can be answered in a statistically significant way, i.e.: **P1** (99.9 days before the start of the course); **P2** (99.9 to 53.6 days before); **P3** (53.6 to 7.33 days before); **P4** (7.33 before to 38.96 days after).

Intuitively, if students register too early (here, above 3 months in advance, covering P1), this is not very beneficial, as they may possibly forget that they have done so in the first place, so it is 'better' during this period to be one of the ones who register later, rather than earlier. Thus, being somewhat closer to the actual start of the course, when registering, is more desirable. On the other hand, if students register during a slightly closer time to the start (between 3 and above 2 months before the course has started), then, students completing or not are close, whilst it is slightly better to register earlier. However, even closer to the starting point (about 2 months before course start, up to about one week before), students enrolling slightly later are again more likely to complete. Interestingly, just around course start (1 week before course start up to about 1 month after course start) it suddenly becomes better to enrol slightly earlier. This surprising result may be explained by noting that, for most of the time, it translates to registration being more likely to lead to a successful outcome if it is closer to the course starting date. The only exception is period P2, which would need further analysis in future research. Finally, for students enrolling too late (over 1 month after the course has started), it is again better for students who enrol closer to the starting date (thus earlier).

Based on these results, we can further address **RQ1.2** as follows. The teacher could analyse the data very early on, and give specific customised feedback to students. As the registration time is known before or just after the start date, students could be advised to only register when they are quite sure about attending the course, and as close as possible to the start of the course. They should possibly be also given the choice to deregister, if they have lost the interest, to give a teacher a better and more realistic picture of the actual cohort to follow the course. For students who are registering too late, they should possibly be notified of further times the course is run, and let known in advance that they would have to put in an additional effort, if they really want to complete the course - possibly offering them simplified material, or other type of support, to catch up.

Alternatively (or at the same time with the tutor intervention), an intelligent tutoring MOOC extension could implement some rules to automatically deliver such messages to the students (e.g., Table 5.3). The table shows also that students can not only be given messages, but also be supported with additional resources, or more tailored resources, when they are starting, for instance, late. For early registration, storing the course information as soon as possible in the agenda of the students and ensuring that no other overlap is occurring by omission is important (of course, other overlaps outside the influence of the students may still exist). Further development of such adaptation rules remains for future research, although defining periods centred on the start date, and moving standard deviations from it, seems promising. Finally, whilst the research questions posed are answered, further answers can be sought in future work.

Table 5.3 Rules in pseudo-code based on registration date.

IF registration_date > 3 months before start date (BSD) **THEN** recommend ("Please consider registering closer to the start of the course. Would you like to be reminded of this a couple of months before the course starts? Would you like to have the date automatically registered to your calendar?")
IF registration_date in (2 months BSD to 1 week BSD) **THEN** recommend ("Please consider confirming your registration closer to the course start. Would you like a reminder a week before the course starts? Would you like to have the date automatically registered to your calendar?")
IF registration_date in (After start date (ASD)to 1 month ASD) **THEN** recommend ("As you have registered to the course after its start, please note that you would have to put in much more focussed work in order to complete. If you prefer to opt to enrol the next time this course starts, please let us know. Please also consider visiting these links for additional support.")
IF registration_date > 1 month ASD **THEN** recommend ("You have registered very late to the module. You are strongly recommended to consider taking this course at a later date.")

Epilogue

In this chapter, we analysed a large data set of FutureLearn MOOC users over several courses, each with several runs. Our results show that completion can be predicted based on the date of the registration. We were able to identify specific periods when it was likelier for the students registered (relatively) early to complete, as well as periods for which the opposite was true. In the following chapter, we will investigate the clickstream data to visualize students' learning patterns and compare the differences between completers and non-completers.

Chapter 6 : Is MOOC Learning Different for Dropouts? A Visually-Driven, Multi-granularity Explanatory

Prologue

This chapter presents different granularity visualisations for learning patterns to compare the differences between completers and non-completers. It shows how different granularity visualisations (fish eye and bird eye) that allow both researchers and teachers to understand where issues occur and patterns emerge, supported by a statistical analysis.

6.1 Introduction

Since MOOC platforms offer free courses for millions of students, the retention rate of learners is notoriously low. The majority of the research work on this issue focuses on predicting the dropout rate, but very few use *explainable learning patterns* as part of this analysis. However, *visual representation of learning patterns* could provide deeper insights into learners' behaviour across different courses.

Thus, this chapter proposes and compares *different granularity visualisations for learning patterns* (based on clickstream data) for both *completers* and *non-completers*. In the large-scale MOOCs we analysed, across various domains, our *fine-grained, fish-eye visualisation* approach showed that non-completers are more likely to jump forward in their learning sessions, often on a *'catch-up' path*, whilst completers exhibit *linear* behaviour. For *coarser, bird-eye granularity visualisation*, we observed learners' transition between types of learning activity, obtaining *typed transition graphs*. The results, backed up by statistical significance analysis, provide insights for course instructors to maintain the engagement of learners by adapting the course design to not just 'dry' predicted values, but explainable, visually viable paths extracted.

This chapter addresses the following research questions:

RQ2.1: How can learning paths be visualised to differentiate between completers' and non completers behaviours (to inform teachers for early interventions) ?

RQ2.2: Are there (significant) differences in the learning paths of completers and non-completers and can they be deduced from visualisation early on in the course?

RQ2.3: What kind of level of granularity is necessary for the visualisation of significant differences between completers and non-completers?

Existing studies on MOOCs analytics mostly focus on finding reliable completion indicators from learner behaviour patterns, using data of forum activities(Santos et al., 2014), clickstreams(Kloft et al., 2014), assignment activities (Coffrin et al., 2014) to name a few. Other predictive studies (Ye and Biswas, 2014, Greene et al., 2015),(Li et al., 2017) attempt to forecast the performance, including final grade and pass/fail in exams. Overall, existing research usually does not disregard the potential of using visualisation as an initial step before prediction, nor do they consider the granularity of visualisation as a factor.

6.1.1 Visualisation

The explanatory power of visualisation has been stated to be crucial to provide more insights for module instructors and designers, next to completion prediction (Davis, 2016). A learning path is defined as the learning trajectory through a course by learners; according to (Guo and Reinecke, 2014), participants generally study web courses in a non-linear manner. We are specifically interested in comparing the transition behaviour between completers and non-completers, as suggested by (Jiang et al., 2014). Wen and Rosé (2014) investigated the typical learning activity sequences across two MOOC datasets and mined the difference in learning themes among groups with different grades, by visualising the distribution. However, they mainly focused on which topics were more popular, instead of visualising the entire learning paths of their four groups of learners (none, fail, pass and distinction). Later (Davis, 2016) visualised log traces of learners across four edX MOOC datasets, using discrete-time Markov Models and observed that learners were more likely to jump forward than backwards from video content. Additionally, they found that non-passing learners were more likely to exhibit binge video watching, i.e., transit from one video to another without

answering questions, deviating from the designed linear learning path. However, they only visualised the video interaction activity of passing and non-passing learners, instead of the whole learning sessions, like in our work.

Recently, (Shih, 2019) used visualisation software, Gephi, to visualise clickstream-based learning paths. They observed that learners are more likely to skip the quizzes at the end of each chapter, by watching the beginning videos of the next chapter, but learn linearly within chapters. However, neither did such previous studies explore the phenomenon in-depth, nor provided a convincing explanation.

This chapter validates that completers behave differently from dropout learners by visualising the entire learning paths of participants. We also statistically analyse students' movements by combining courses' themes and content.

6.1.2 Statistical Analysis

According to (Zheng and Yin, 2015), statistical analysis can be divided into descriptive statistics, to summarise the demographics information of learners and inference statistics, to explore behaviours exhibited by participants. For instance, (Zhao et al., 2017) firstly examined if there is any impact on the behaviour of learners, after they reached the passing state. They examine weekly quiz score distribution for all learners by K-means clustering, which showed that early passers obtained relatively lower scores immediately after passing, compared with their previous performance. Later, (Watted and Barak, 2018) investigated the motivation of two groups of completers: university students and general participants, using Mann–Whitney U test, and concluded that as participants' ages increased, earning a certificate was less significant. Recently, (Peng and Xu, 2020) explored different behaviours of course completers and non-completers from a discourse perspective on course content review, via a chi-square test. They found that completers were more likely to post original and less negative opinions, whilst non-completers were more willing to reply to and criticise others' posts. Inspired by their work, we implement a Wilcoxon test for completers and non–completers to learn their transition behaviours among different learning themes, separately.

6.2 Methodology

To explore if completers and non-completers behave differently, we apply visualisation analytics firstly, to identify the different learning paths executed by these groups and then implements statistical modelling to analyse their learning behaviour (transition from themes such as *video, quiz, discussion, review and article*) across different courses. Additionally, by comparing dropout learners' transition from different learning activities across different courses, this research offers insights into the impact of the designed course learning path on the learning behaviour of participants.

6.2.1 Dataset

The dataset of learner activities has been extracted from 8 runs of 4 Future Learn-delivered courses from The University of Warwick between 2013 and 2017, with over than 106,036 learners. We investigate learner activities across different domains: Psychology (The Mind is Flat and Babies in Mind), Literature (Shakespeare and His World) and Computer Science (Big Data). The learner activities include watching videos, taking a quiz, discussion, submitting assignments, viewing assignment feedback, reviewing another learner's assignment and giving feedback.

Table 6.1 The dataset of learner activities

Course Title	Run	Enrolled	Accessed	Dropout	Completers
Babies in Mind	Run 1	12651	5841	4634	1207
	Run 2	9740	4924	4030	894
Big data	Run 1	11281	4715	4202	513
	Run 2	5761	3840	3583	257
Shakespeare and His World	Run 1	15914	9050	7170	1880
	Run 2	12692	6902	5804	1098
The Mind is Flat	Run 1	22929	8198	6858	1340
	Run 2	15068	6760	5743	1017

Of all participants, 50,230 learners have accessed at least one step of the course, but over half, 53%, have never accessed the course after registration. The 42,024 learners who accessed less than 80% are defined as *dropout learners*. The 8206 learners who accessed

80% or more of the materials in one run are defined as *completers*. The 80% threshold (as opposed to, e.g., 100% completion), the total number of those who completely accessed 100% of the steps was relatively low. Completers represent 7% of the total and 16% of the learners who accessed at least once.

6.2.2 Visualisation of High & Low Granularity Levels

After analysing the learning path of learners, we have divided the dataset of learners' activities into two components: *linear* and *catch-up* (see Figure 6.1). The former shows the linear path between two sequential steps and the latter shows the catch-up activities, i.e., jumping behaviour. This research implements a flow network analysis, to present the learning pattern for linear movem(from source: $x_{i,}$ to destination: x_{i+1}) and catch-up (from source: $x_{i,}$ to destination $\neq x_{i+1}$). Depending on the granularity, the learning pattern is further identifiable as *bird's eye view*, i.e., high granularity (a node representing multiple steps), and *fish-eye view*, i.e., low granularity view (step-level representations). In addition, course activities have been grouped by different colours based on their themes see Figure 6.2.

Figure 6.1 Examples: a) Linear activities b) Catch-up activities

The size of the circle represents the number of learners who accessed the course content, and the thickness of the arrow shows the percentage of learners' movements.

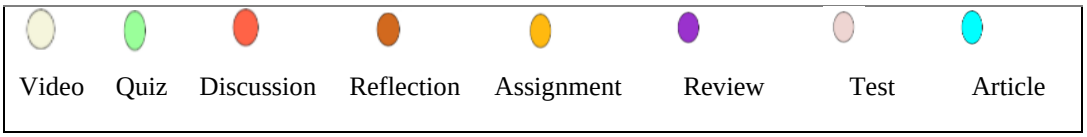


Figure 6.2 Colour codes indicating the type of course content

6.2.3 Statistical Analysis

We further analyse the normality of the data, we report results as percentages instead of the total number of learners to eliminate the effects of having different numbers of learners for

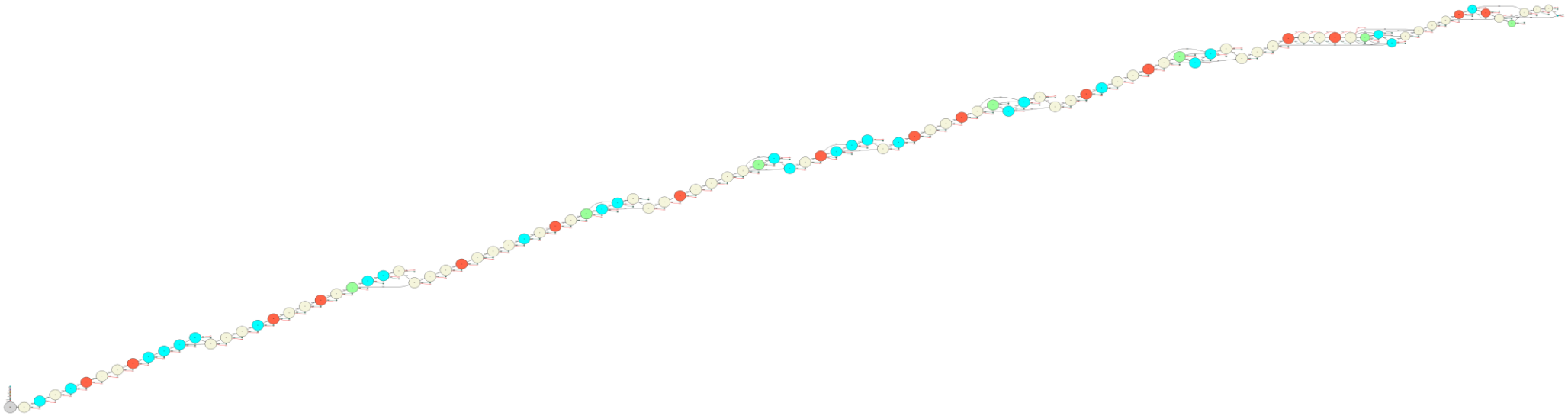
each group. The results show that learners' learning paths were not normally distributed ($p < 2.2e-16$). Thus, we have applied the non-parametric Wilcoxon test instead.

6.3 Results and Discussion

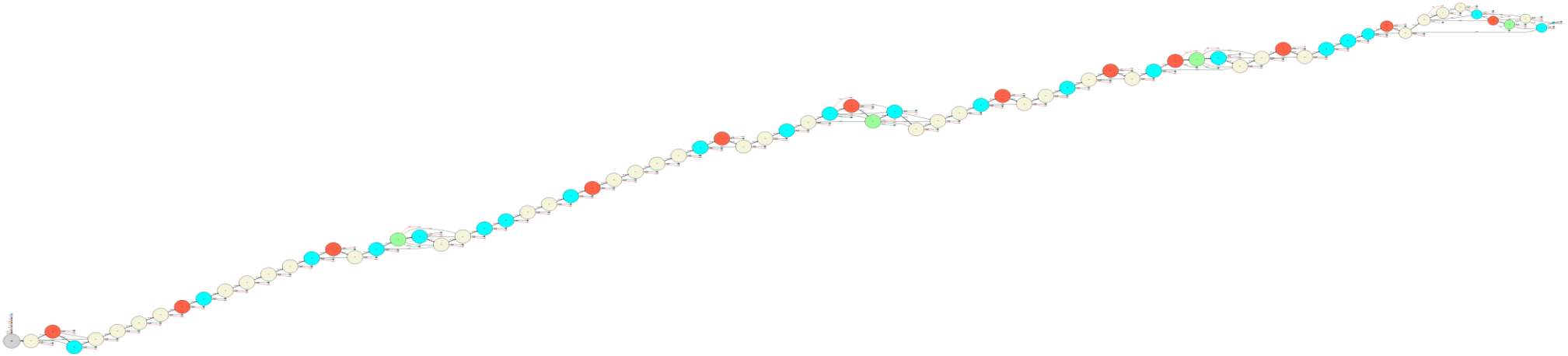
The flow network analysis shows that completers are more likely to complete the courses linearly (see Figure 6.3 and Figure 6.5), whilst non-completers are more likely to skip quizzes and assessments (the catch-up learning pattern); which mainly shows the overall learning pattern instead of providing a clear view in details due to the course length). For instance, non-completers have various learning paths; some of them may directly jump to lessons in week two after accessing the first lesson Figure 6.6. Instead, completers are much more "obedient", as they mainly follow the designed learning path, compared with dropout learners; interestingly, this holds true across different domains - as shown in the bird eye view in Figure 6.4. The Statistical analysis results in Table 6.2 further confirm that these learning paths are significantly different.

Table 6.2 P-values of linear and catch-up learning activities

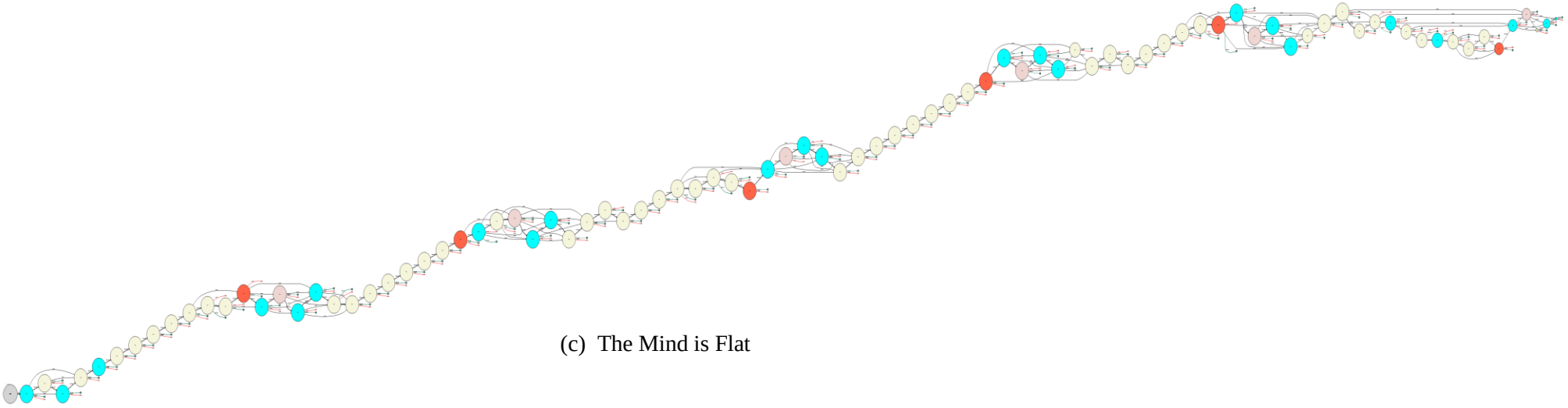
Courses/ Run	**P-value**	
	Catch-Up activities	**Linear activities**
Babies in Mind Run 1	1.13E-13 ($p<0.001$)	2.46E-85 ($p<0.001$)
Babies in Mind Run 2	7.74E-14 ($p<0.001$)	2.32E-62 ($p<0.001$)
Big Data Run 1	2.66E-018 ($p<0.001$)	5.97E-110 ($p<0.001$)
Big Data Run 2	1.35E-68 ($p<0.001$)	2.66E-18 ($p<0.001$)
Shakespeare Run 1	1.130E-13 ($p<0.001$)	2.09E-23 ($p<0.001$)
Shakespeare Run 2	7.73E-14 ($p<0.001$)	1.87E-09 ($p<0.001$)
Mind is Flat Run 1	2.66E-018 ($p<0.001$)	5.21E-74 ($p<0.001$)
Mind is Flat Run 2	6.62E-63 ($p<0.001$)	2.51E-16 ($p<0.001$)



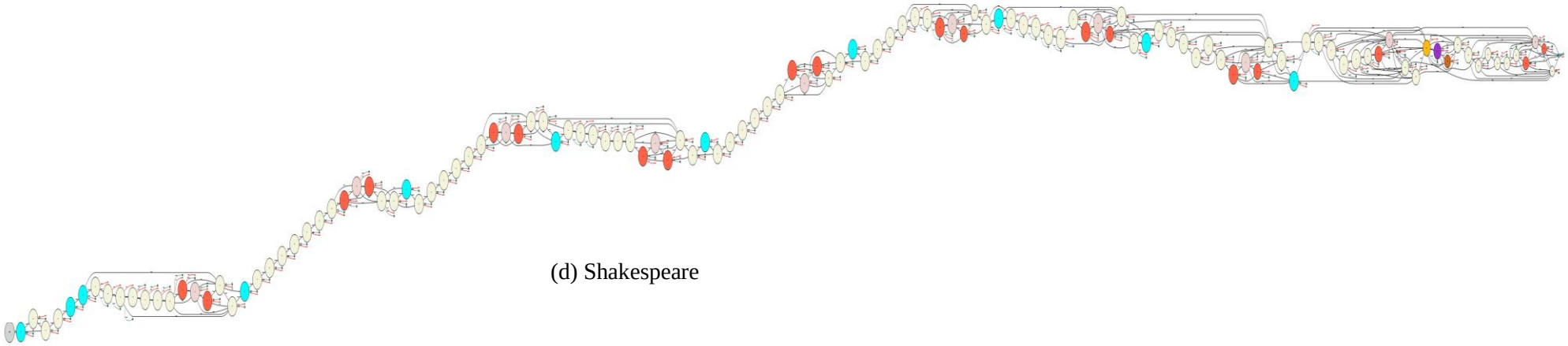
(a) Babies in Mind



(b) Big Data



(c) The Mind is Flat



(d) Shakespeare

Figure 6.3 Completers learners learning path (Bird eye view (a-d)).

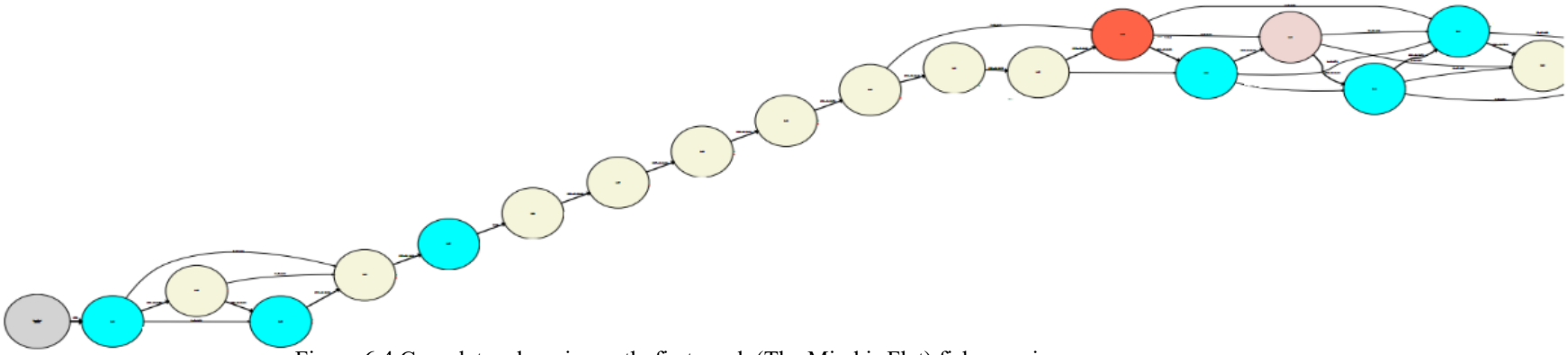
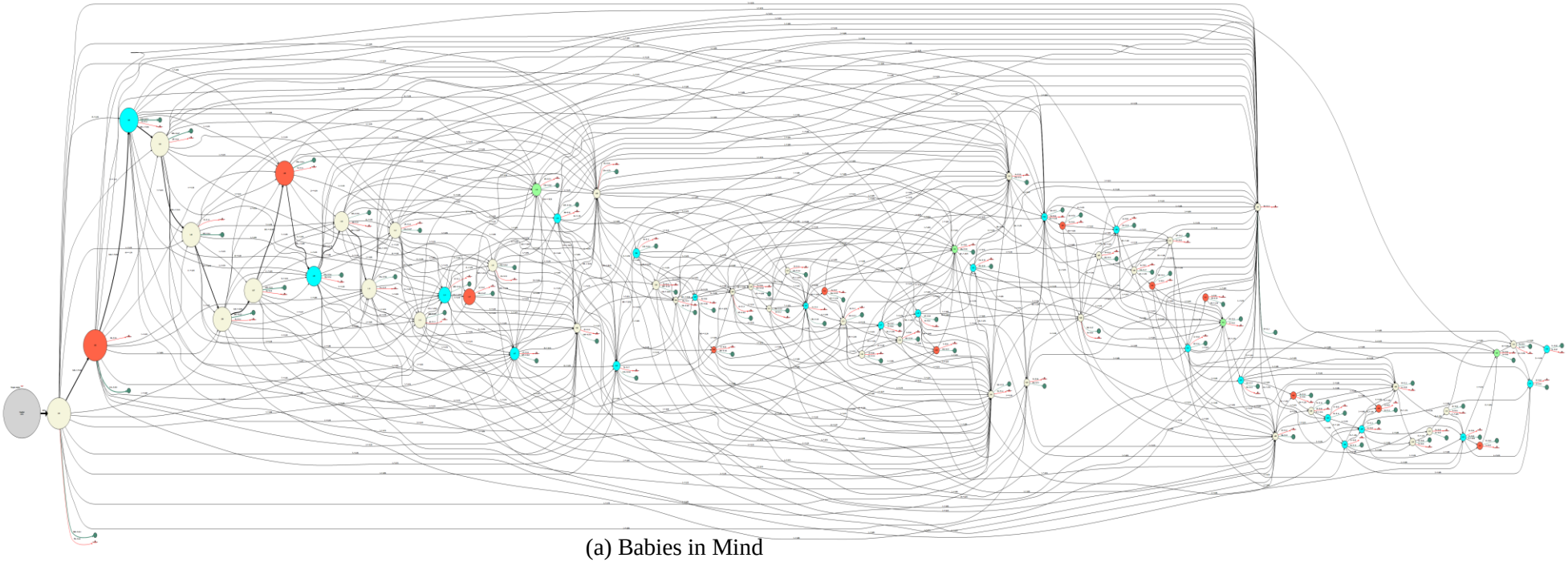
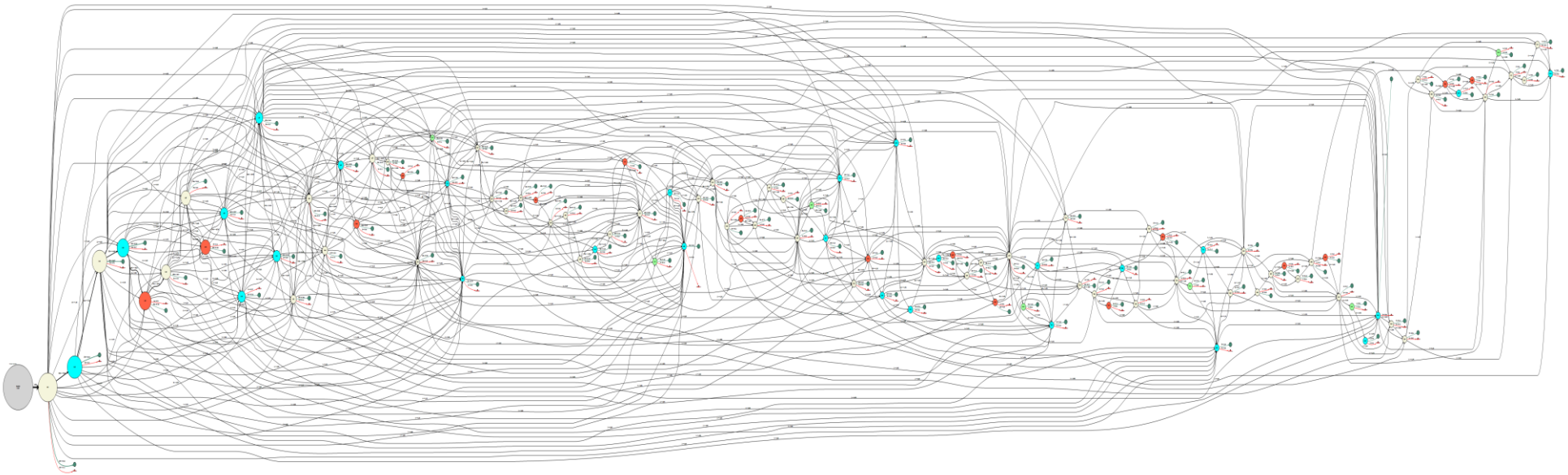


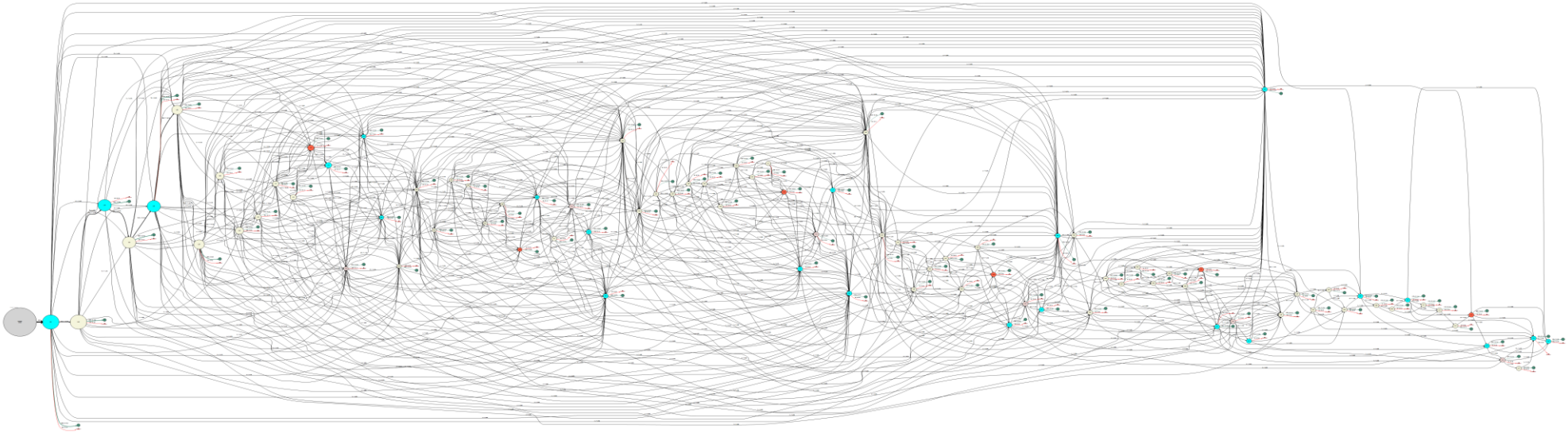
Figure 6.4 Completers learning path, first week (The Mind is Flat) fish eye view



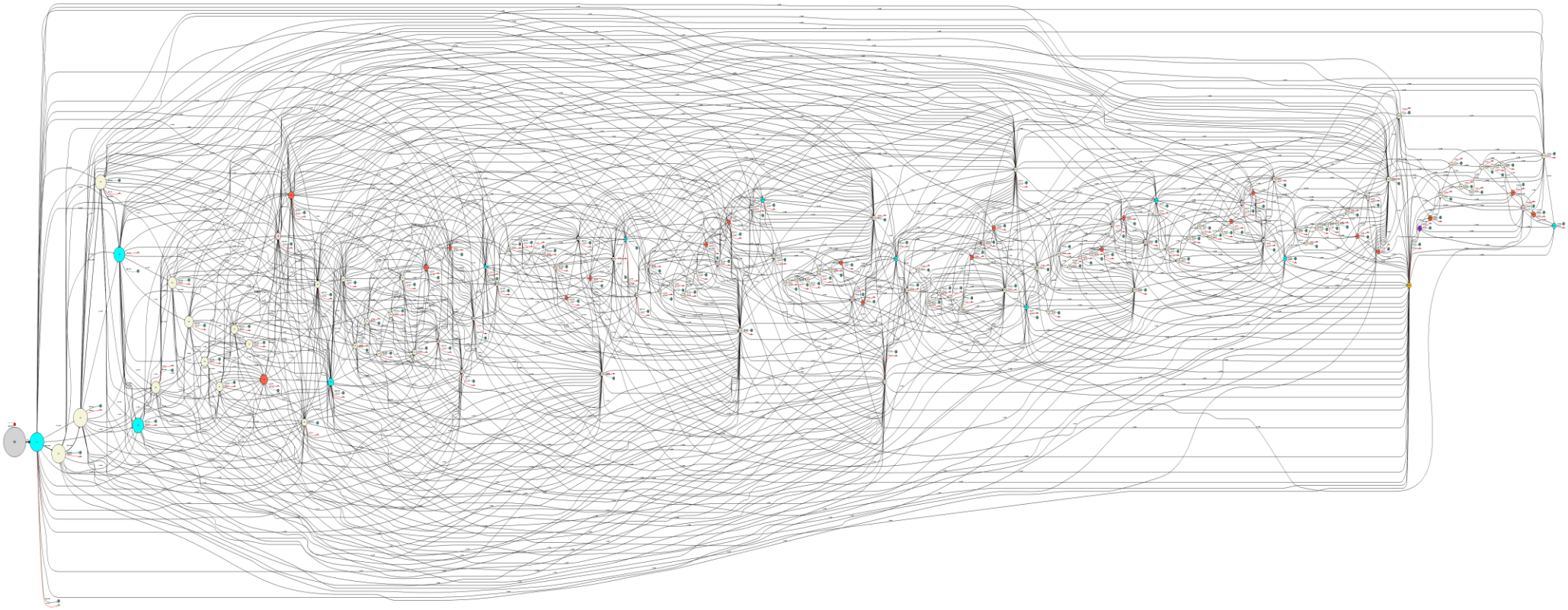
(a) Babies in Mind



(b) Big Data



(c) The Mind is Flat



(d) Shakespeare

Figure 6.5 Dropout learners learning path (Bird eye view (a-d)).

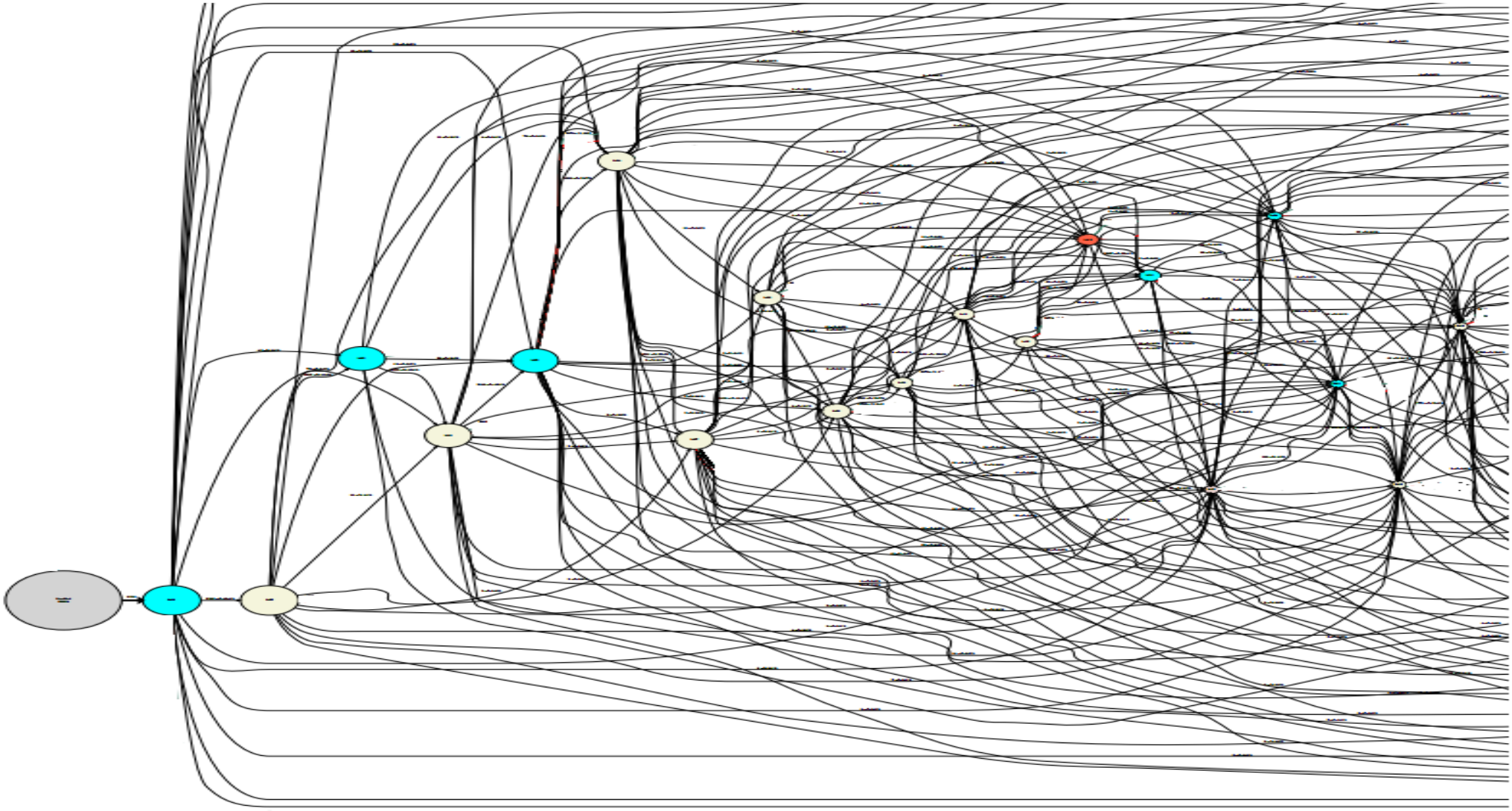
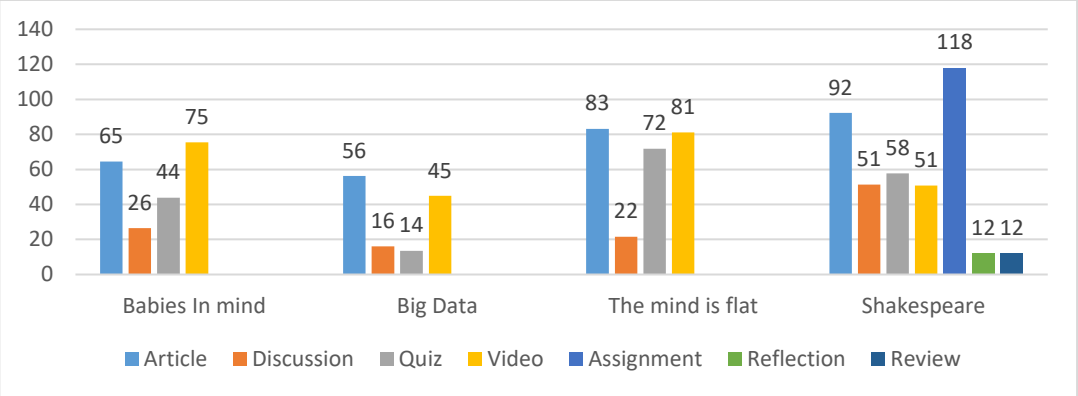
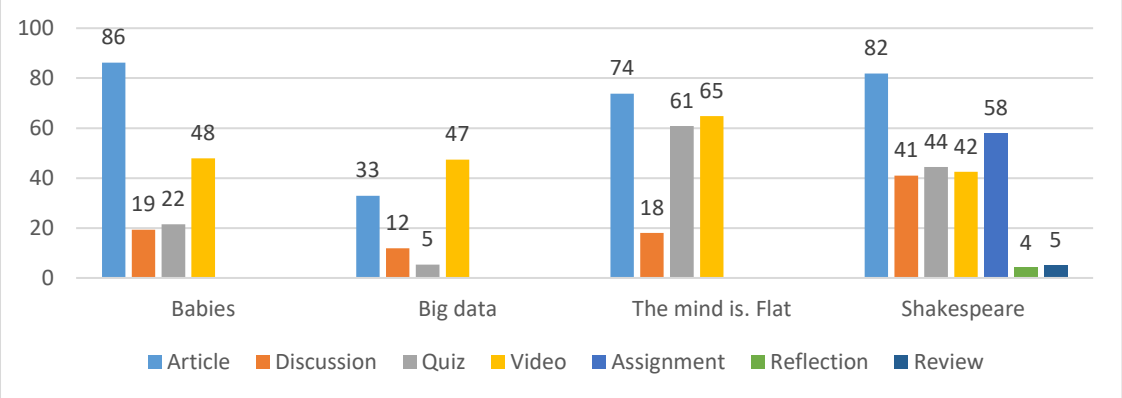


Figure 6.6 Dropout learning path, first week (The Mind is Flat course) fish eye view

Then, we compared dropouts after their last accessed activity: video, discussion, quiz and article for the four courses. Figure 6.7 illustrates that learners are more likely to drop out after articles and videos. Interestingly, participants drop out the least after discussion. The attraction of discussion has been confirmed by our analysis across courses. The reason may be that participants feel encouraged to share their knowledge and can gain support (Warren et al., 2014). In the Literature course, Shakespeare, participants were more likely to drop out after the assignment. The reason may be the difficulty of the creative writing at the final week that learners are required to write their film, book, ballet or musical; this is useful feedback for the course designers to change assignments potentially. The figures also suggest that the dropout patterns, according to themes are similar across runs.



(a) First run



(b) Second run

Figure 6.7Average number of dropout per topic (course contents type)

Epilogue

In this chapter, we used data visualisation to examine and compare the learning routes for completers and dropouts across four MOOCs (8 Runs). We focused on which learning theme students likely tend to drop out. The results of this research show that students who completed the course are more likely to learn in a linear way, whereas students who drop out are more likely to engage in what we term the "catch-up" learning pattern.

In the next chapter, we will use students' activities to predict the non-completers by using machine learning algorithms. In addition, we will examine the effect of using learners' jumping behaviour ("catch-up" learning pattern) as a feature to predict non-completers students.

Chapter 7 : Next week MOOC dropout prediction: Weekly assessment time and learning patterns

Prologue

This chapter focuses on innovations in predicting student dropout rates by examining their *weekly learning activities*. This study is based on three MOOC platforms, including 303,466 students from 10 courses with 41 runs spanning 2013–2019. This study aims to build a generalised early predictive model for the weekly prediction of student dropout using machine learning algorithms.

7.1 Introduction

Moocs researchers intend to find the most predictive feature(s) of students' dropout activity and thus enable early intervention. One usual way is to identify learning behaviour indicators to raise the accuracy of MOOCs' completion prediction (Prenkaj et al., 2020). However, data is not always available for such indicator analysis. For instance, non-completion can be predicted by a linguistic analysis of discussion forum data (Wen et al., 2014b). Nevertheless, as students' comments only amount to 5-10% of posts in discussion forums, this feature is not applicable universally and more features are needed (Rose and Siemens, 2014).

Additionally, numerous variables can be considered for non-completion analysis, such as student profile data (e.g., country, age, gender) (Kameas, 2021, Hlioui, 2021, Moreno-Marcos, 2020a, Robinson, 2016) and course-attended related data (e.g., reading, watching, writing, taking quizzes) (Ding, 2019a, Gitinabard, 2018, Qiu, 2019, Mubarak, 2020, Doleck, 2020).

This research investigates different methods of targeting dropout students by comparing the prediction of student dropout in the whole course (**CP**) and the prediction of student dropout in the following week (**WP**).

In addition, based on the visualisation findings in Chapter 6, students' learning path is an insightful dropout prediction feature, as successful learners will follow the instructed path and exhibit 'linear learning behaviours'. Conversely, learners may jump forward and backward in their learning sessions (Gardner and Brooks, 2018), which is defined as exhibiting jumping behaviour, and they are likely to quit in the process. To the best of our knowledge, this research is the first to consider participants' learning paths and the associated behaviours in weekly dropout prediction.

Moreover, we conducted a first-of-its-kind study of a novel, lightweight approach based on tracking two features (accesses to the content pages and time spent per access), fine-grained learner activities to predict student non-completion. Hence, our research questions are formulated as follows:

RQ3.1 Are there (high) differences in the prediction of weekly dropout and whole course dropout?

RQ3.2 Will the weekly predictive model be more accurate after considering student jumping behaviours and catch-up learning patterns during the course?

RQ3.3 Can MOOC dropout be predicted within the first week of a course, based on the learner's number of accesses and time spent per access?

This chapter presents three experiments as shown in Figure 7.1. *First,* we will compare two methods for predicting dropout students from an early stage: we will compare the weekly prediction (**WP**) approach with the more traditional approach to predicting student dropout in the whole course (**CP**). For example, with **WP**, we will only indicate students' completion of the second week by using their previous learning behaviours in the first week. The model will also predict students' completion of the fifth week by using their previous four weeks' learning pattern. This experiment aims to find the most accurate method of predicting dropout students, so we compare the prediction performance for two methods (**CP** vs **WP**).

In the *second experiment*, building on the previous work in Chapter 6, we will investigate jumping behaviours of completers and non-completers in MOOCs. We will incorporate the students' learning patterns (jumping behaviours) into the weekly dropout predictive model and gauge the impact in the model's performance with and without students' jumping behaviours.

In the *third experiment*, the prediction model will be based on only two independent variables (the number of accesses and the time spent on a page). Importantly, unlike our approach, most prior research has used many independent features (see Table 3.1). For example, (Kloft et al., 2014) employed 19 features, including technical features and those that captured the activity level of learners. Promisingly, our model, despite using only two features from the first week of each course, can also achieve a 'good enough' performance, as shall be further shown.

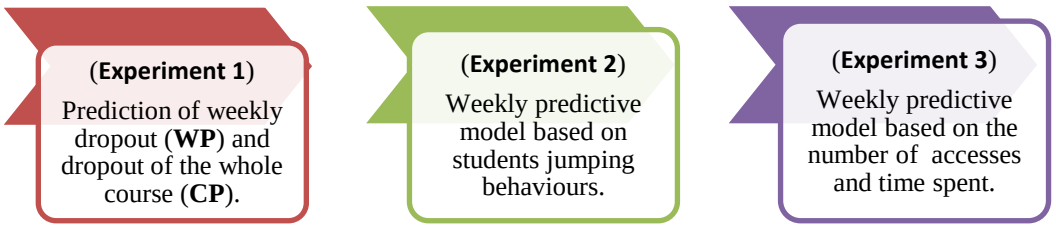


Figure 7.1 Dropout prediction experiments

7.2 Methodology

This study analysed a massively large dataset of 41 runs (each course has run several times over the years) of 10 multidisciplinary courses, which fall under four main categories – computer science, literature, business and psychology. The FutureLearn courses were delivered by two universities in the United Kingdom (University of Warwick and Durham University), and Rwaq courses were delivered by well-known professors and teachers from all over the Arab world.

The courses studied were three to ten weeks long and were delivered between 2013 and 2019. The structure of these courses was based on a weekly learning unit. Every learning week included 'steps' covering images, videos, articles and quizzes. Having joined a given course,

learners can access these steps and optionally mark them as completed. These steps also allow comments, replies and likes on these comments from different users enrolled in the course. Moreover, quizzes can be frequently attempted until the correct answer is obtained.

Table 7.1 Courses' Summary

Course	Enrolled	Accessed	% Dropout	Run
Open Innovation in Business (OI)	6071	2792	54.0%	4
Leading and Managing People-Centred Change	10417	6566	37.0%	3
Babies in Mind (BIM)	48771	26175	46.3%	6
Big Data (BD)	33427	16272	51.3%	3
Shakespeare and His World (SHK)	63625	29432	53.7%	4
Supply Chains (SUP)	5808	2912	49.9%	2
The Mind is Flat (THM)	83543	39894	52.2%	7
Java Programing (JAV)	22419	10862	51.6%	4
Self Confidence (SLC)	14757	8242	44.1%	4
Excel (EXC)	14628	8297	43.3%	4
Total	**303,466**	**151,444**		**41**

Overall, we have acquired educational data that is not open to the public for 303,466 students, (shown above in Table 7.1). Enrolled refers to registered students, and accessed refers to students who have accessed the course at least once. It can be seen from the data that about half of enrolled students in MOOC do not access the course contents after the course has started.

Based on the data, it can be observed that the Open Innovation course exhibited the highest rate of attrition at 54%, followed closely by Shakespeare and His World at 53.70% and The Mind Is Flat at 52%. The Leading and Managing People-Centred Change and Excel courses exhibited the lowest rates of student attrition, with figures of 37.0% and 43.30%, respectively. However, each course has several runs, as they are popular and held for more than one term. The Mind Is Flat is the largest of the courses in terms of enrolled students, student accesses, and number of runs.

7.2.1 Data pre prossing

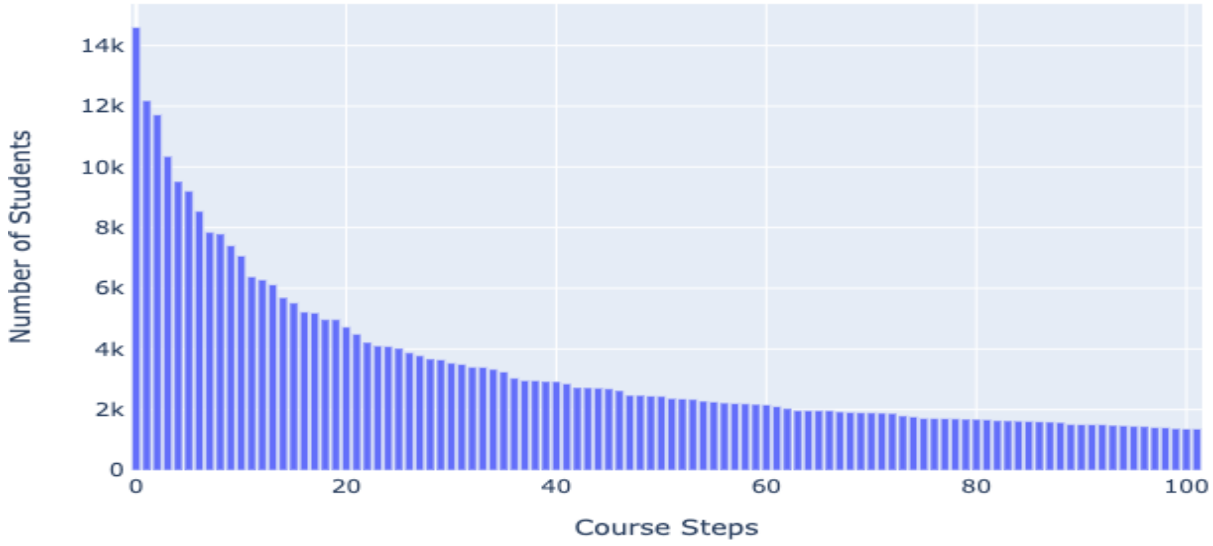
To begin with, the raw dataset was refined, removing all students who enrolled but never accessed any material. We dealt with those learners separately, based on even earlier

parameters (such as the registration date, see Chapter 5). Subsequently, there were 151,444 remaining learners to be studied. The reason of selecting 80% completion as a sufficient level of completion (as opposed to, e.g., 100% completion) is because some course steps include optional reading lists. Moreover, the total number of those who completely accessed 100% of the steps was relatively low.

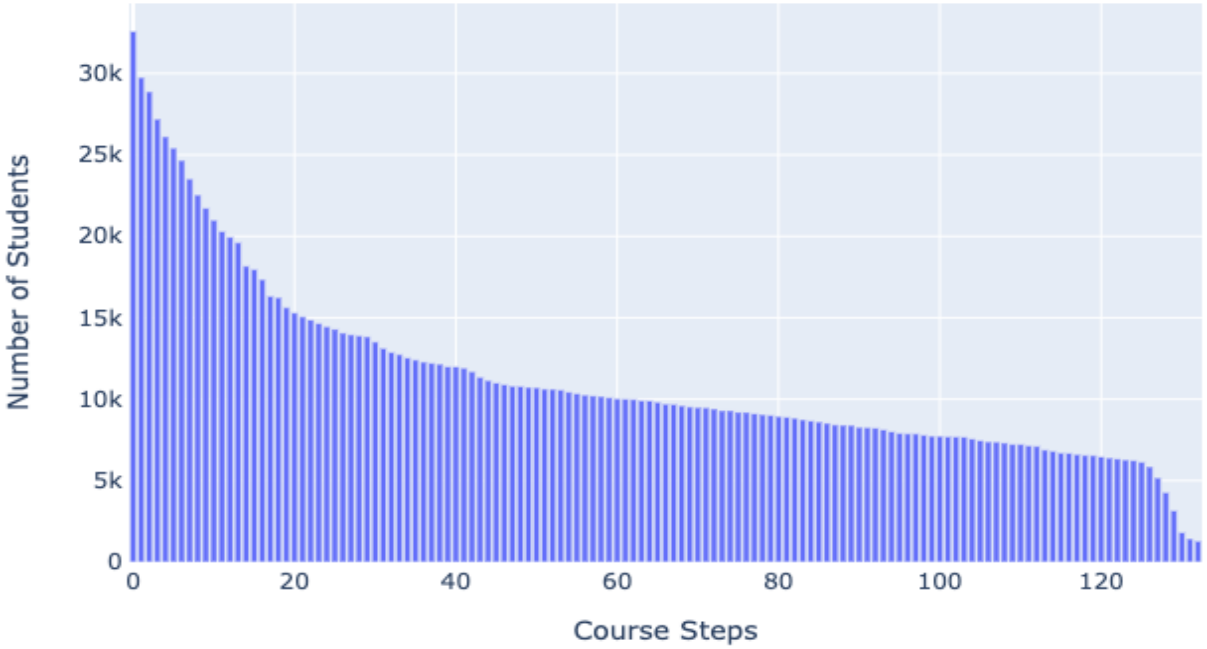
Regarding early prediction, we have opted to start the prediction from the first week, as this methodology is one of the most difficult and least accurate approaches when compared with the current state of the art in the literature. Alternatively, a relative length (e.g., 1/n days of the total length of each course) could have been used. However, in practice, this tends to use later prediction data than our approach (e.g., 1/4th of a course is one week for Babies in Mind, but 2.5 weeks for Shakespeare and his Work).

7.2.2 Prediction targets

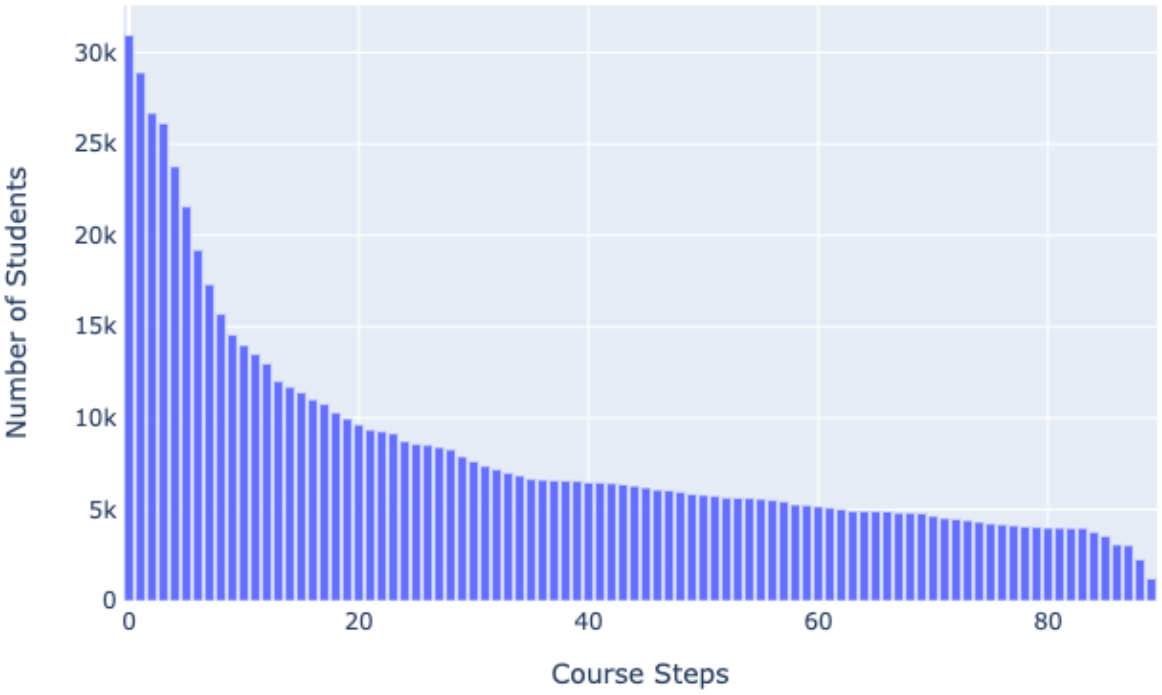
Although, about 3-15% of participants complete their courses in MOOC (Coffrin et al., 2014), dropout is a gradual process. Therefore, we are interested in analysing and predicting those weekly dropouts in the first experiment. Figure 7.2 presents the number of weekly dropouts and persisting students over time. Clearly, participants are most likely to drop out in the first few weeks in all courses. Therefore, identifying those early dropouts is important for prediction.



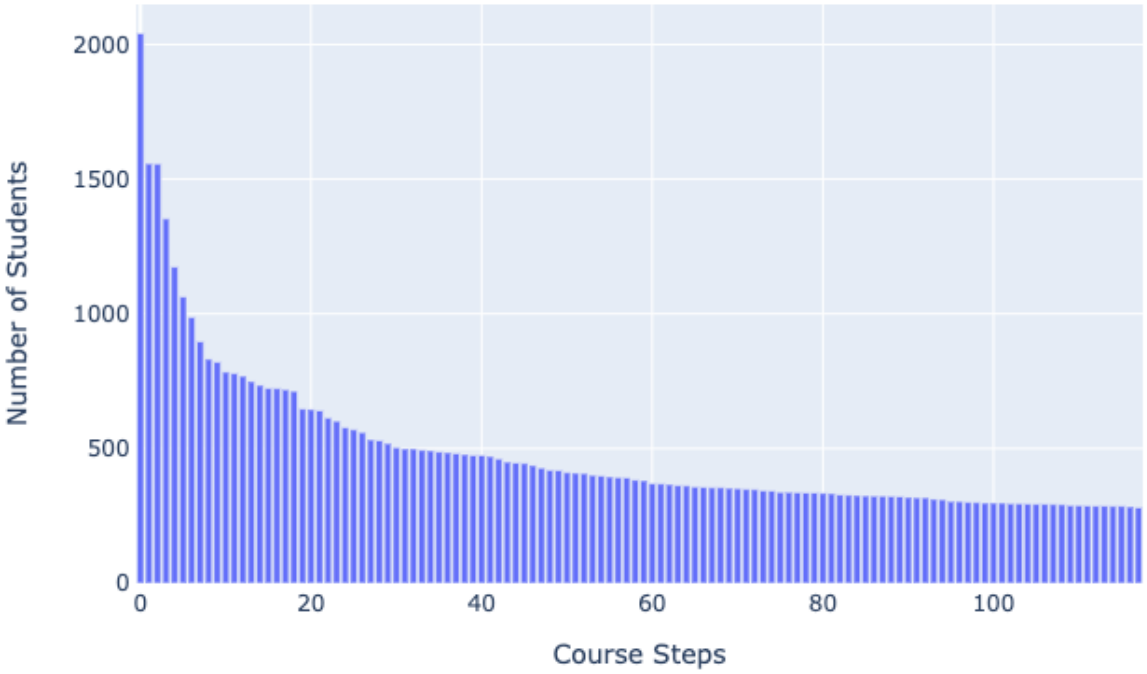
(a) BD


Number of Students
30k
25k
20k
15k
10k
5k
0
0
20
40
60
80
100
120
Course Steps

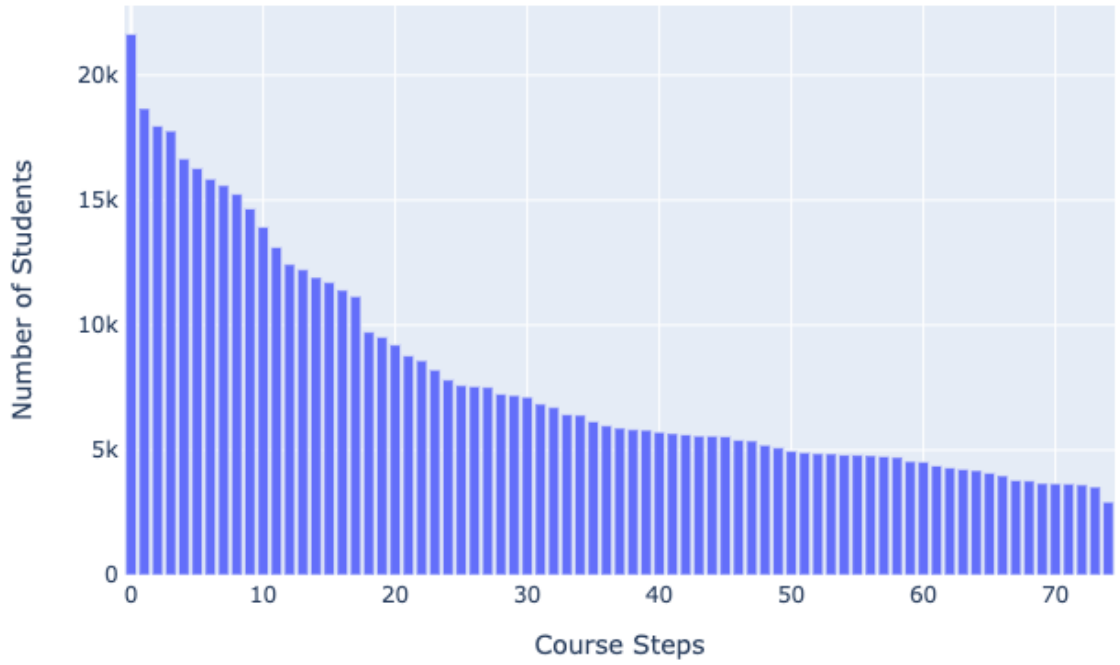
(b) SHK


Number of Students
30k
25k
20k
15k
10k
5k
0
0
20
40
60
80
Course Steps

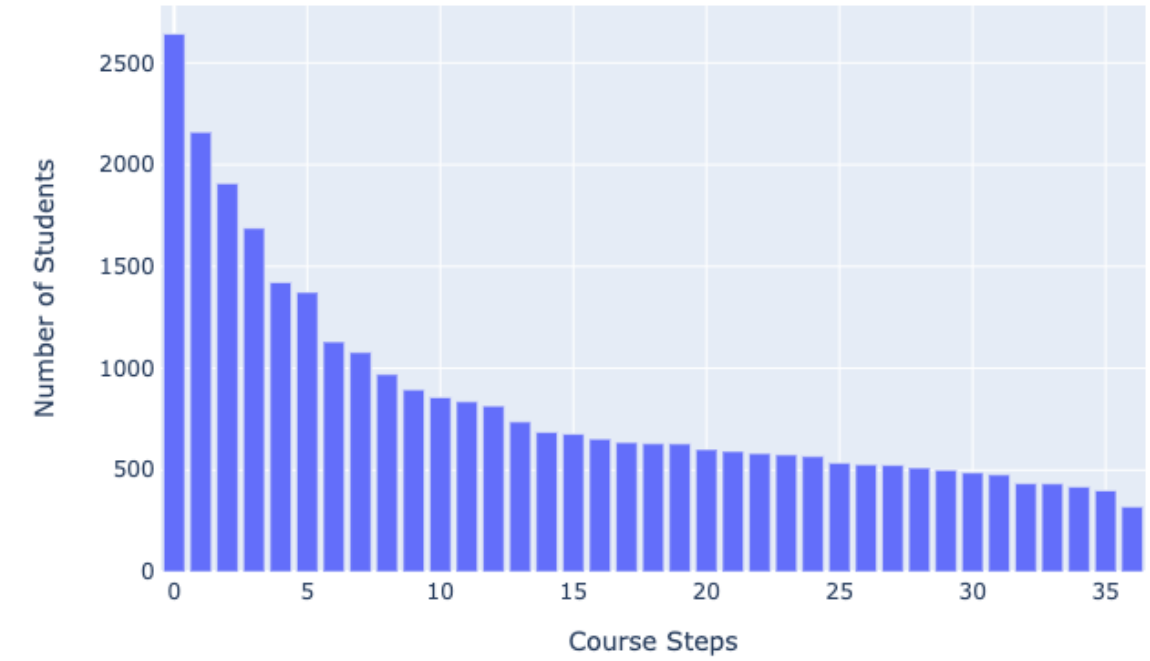
(c) TMF


2000
1500
1000
500
0
Number of Students
0
20
40
60
80
100
Course Steps

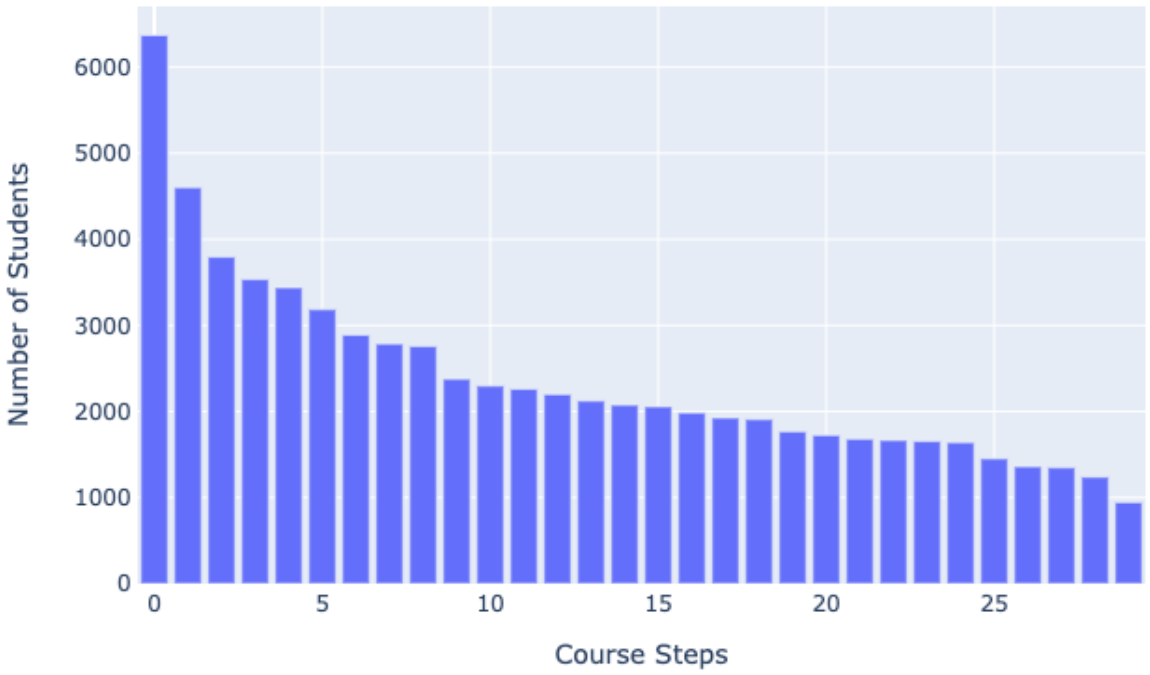
(d) SUP


20k
15k
10k
5k
0
Number of Students
0
10
20
30
40
50
60
70
Course Steps

(e) BIM



(f) OI



(g) LMPCC

Figure 7.2 Remaining students over time in different courses (a-g)

In the current research, we prepared a dataset based on the **WP** technique to determine at-risk students at an early stage. It is believed that predicting at-risk students from their previous weeks' activities may improve the model's prediction performance. Therefore, to address the first research question (**RQ3.1**), we implemented weekly dropouts and whole-course dropout prediction models based on students' activities. For the whole course dropout prediction (**CP**), the students were labelled as dropouts if they did not access 80% of the topics in the whole course. On the other hand, for the weekly dropout prediction (**WP**), we focussed only on students who would drop out in the near future (next week). Therefore, the students were labelled as dropouts if they did not access 80% of the topics in the next week. We compared our **WP** method (see equation 7.1) with the more traditional method of predicting student dropout from the whole course (the students who did not access 80% of the whole course) (see equation 7.2).

$$\boldsymbol{WP}(\mathbf{s},\mathbf{w}) = \begin{cases} \mathbf{1}, \boldsymbol{if\ TAS}(\boldsymbol{s},\boldsymbol{w}) < \boldsymbol{TS}(\boldsymbol{w}) * \mathbf{0.8} \\ \mathbf{0}, \boldsymbol{rest} \end{cases} \quad (7.1)$$

$$\boldsymbol{TAS}(\boldsymbol{s},\boldsymbol{w}) = \sum_{\boldsymbol{j}=\mathbf{0}..\boldsymbol{TS}(\mathbf{w})} \boldsymbol{AS}(\boldsymbol{s},\boldsymbol{w},\boldsymbol{j})$$

$$\boldsymbol{AS}(\boldsymbol{s},\boldsymbol{w},\boldsymbol{j}) = \begin{cases} \mathbf{1}, \boldsymbol{if\ student\ s\ accessed\ step\ j\ in\ week}\ \mathbf{w} \\ \mathbf{0}, \boldsymbol{rest} \end{cases}$$

Where s: student, TAS(s,w): total steps accessed by student s in week w; TS(w): total course steps available in week w; AS(s,w,j): step j accessed by a student.

Where TAS(s,w)<= TS(w), as the maximum number of steps a student could access in week w are all available ones.

Example of *WP*:

Given a student s, and student's activities from the current week (week w=1), predict if the same student s is a dropout in the following week (i.e., week w+1=2).

$$\boldsymbol{CP}(\boldsymbol{s}) = \begin{cases} \mathbf{1}, \boldsymbol{if\ TASC}(\boldsymbol{s}) < \boldsymbol{TSC} * \mathbf{0.8} \\ \mathbf{0}, \boldsymbol{rest} \end{cases} \quad (7.2)$$

$$TASC(\boldsymbol{s}) = \sum_{\boldsymbol{j}=\mathbf{0}..\boldsymbol{TSC}} \boldsymbol{ASc}(\boldsymbol{s},\mathrm{w},\boldsymbol{j})$$

$$\boldsymbol{ASc}(\boldsymbol{s},\boldsymbol{w},\boldsymbol{j}) = \begin{cases} \mathbf{1}, \boldsymbol{if\ student\ s\ accessed\ step\ j} \\ \mathbf{0}, \boldsymbol{rest} \end{cases}$$

Where s: student, TASC(s): total steps accessed by student s in the whole course; TSC: total course steps available; ASc(s,w,j): step j accessed by a student.

Where TASC(s) <= TSC, as the maximum number of steps a student could access are all available ones in the whole course.

Example of CP:
Given a student s, and student's activities from the current week (week w=1) predict if the same student s is a dropout from the whole course ((i.e., accessed less than 80% of steps available in the whole course).

7.2.3 Jumping behaviours

To address the second research question (**RQ3.2**), we incorporated students' learning patterns, specifically *jumping behaviours*, into the weekly predictive model by adding a new column that presented the number of jumping activities for each student each week. We will compare the performance of weekly dropout prediction (**WP**) and weekly dropout prediction with jumping activities (**WPWJ**) to demonstrate the effectiveness of jumping behaviours.

7.2.4 Sentiment Analysis

In this research, the power of Natural Language Processing (NLP) has been used to analyse student comments and use them as features to predict their dropout activities. Textblob tool has been employed to classify students' comments into three categories: *positive*, *neutral* and negative (see section 4.5.1).

7.2.5 Time spent feature

To address the third research question (**RQ3.3**), we employed Time Spent feature (TimeS) that reflects a calculated value (rather than a log parameter within the collected data set). This feature defines the difference between the access time to a certain step by a particular student and the point at which that student completes that step (see equation 7.3).

$$\mathrm{TimeS}(\mathrm{st},\mathrm{s}) = \mathrm{CMT}(\mathrm{st},\mathrm{s}) - \mathrm{ACT}(\mathrm{st},\mathrm{s}) \qquad 7.3$$

Where s is student, TimeS(st,s): is the total time spent by a given student s in step st, CMT(st,s): is the first time stamp for a given student s accessing step st, ACT(st,s): is the completed time stamp for a given student s completed step st (by clicking the button labelled 'Mark as Completed').

7.2.6 Feature Selection

The methods of feature extraction have been explained in section 4.5. Table 7.2 shows the set of features used in each experiment to train the dropout prediction model.

Table 7.2 Features used for Dropout prediction

Input Feature	Experiment Number
1. Number of access steps per week	1,2,3
2. Number of correct answers per week	1,2
3. Number of wrong answers per week	1,2
4. Number of attempts per week	1,2
5. Number of comments per week	1,2
6. Number of likes received per week	1,2
7. Number of positive comments per week	1,2
8. Number of negative comments per week	1,2
9. Number of neutral comments per week	1,2
10. Number of replies posted per week	1,2
11. Number of replies received per week	1,2
12. Number of positive replies posted per week	1,2
13. Number of negative replies posted per week	1,2
14. Number of neutral replies posted per week	1,2
15. Number of positive replies received per week	1,2
16. Number of negative replies received per week	1,2
17. Number of neutral replies received per week	1,2
18. Number of Jumping activities per week	2
19. Time spent	3

7.2.7 Proposed Machine Learning Model

We proposed several models to predict students' future activities, such as *next week's dropout*. The first phase is to clean the datasets, by removing the blank values and missing data. To build our model, we employed several competing ML methods, as follows: Random Forest (RF) (Breiman, 2001), Gradient Boosting Machine (GBM), (Friedman, 2001), Adaptive Boosting (AdB) (Hastie et al., 2009), Logistic regression (LR) (Cokluk, 2010), K-Nearest Neighbor (KNN) (Hart, 1968), Extra Tree (EX)(Geurts et al., 2006), Multi-layer perceptron (MLP) (Gardner and Dorling, 1998) and XGBoost (Chen and Guestrin, 2016). Please note that all these classification algorithms have been explained in section 4.8.

Still, the literature has reported that class imbalance can affect ML algorithms' performance. Due to the massive different completers' ratio to non-completers in our dataset, we set the class weight (Sozykin et al., 2018, Rasouli et al., 2022) to the inverse of the frequency of different classes.

As we have used a massive data set for different courses, we have prepared the training and testing sets based on the last Run of the course. For example, in The Mind Is Flat course, we extracted data from several runs (1-6), with students' activities between 2013 to 2016, to train our models, and to test the model, we used a new data set from a different Run (Run 7) that contains students' activities in 2017 - see Figure 7.3- which is similar to some extent to Transfer learning models (Getoor, 2020, Bote-Lorenzo, 2018).

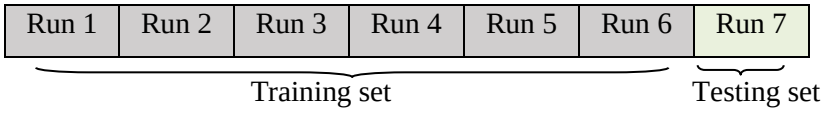


Figure 7.3 The Mind Is Flat course training and testing sets

The current research used the ***F1*** score to measure the performance of the models as it depends on precision and recall; this metric is widely used to evaluate the model's performance when dealing with an imbalanced dataset, by preventing the majority of negative samples from biasing the result (Dutta et al., 2018, Kodiyan et al., 2017).

7.3 Results and Discussion

This section shows the performance results generated by our eight chosen ML algorithms. As mentioned before, we examine students' learning patterns and accessing time for the coming week's dropout prediction. Figure 7.4 shows that participants are more likely to complete the weekly learning activities at the beginning and drop out as time has passed. Around 7500 students have completed the first week of the Big Data course. In contrast, only 2223 completed week 5. Therefore, weekly prediction is a reasonable approach to determining at-risk students at an early stage.

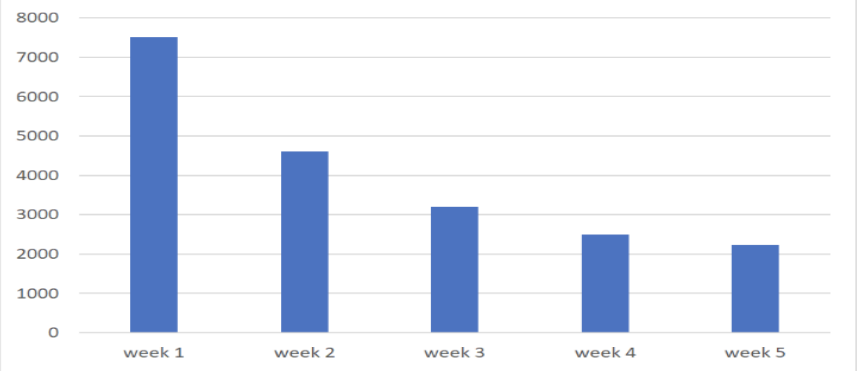


Figure 7.4 Number of completers students in each week (Big Data course)

7.3.1 Results of Weekly Prediction

We selected eight of the most successful methods for classification problems, applying them in the domain of learning analytics in general and on completion prediction in particular. Another candidate was SVM (Mahesh, 2020), which was less successful. Table 7.3 shows the models' performances for both **CP** (if the learner did not access 80% of the topics in the whole course) and **WP** (if the learner did not access 80% of the topics in the next week). In this experiment, we used students' activities in the first week to predict students' dropping out. The classification performance was evaluated using the F1 score.

In general, the most robust model is RF, as it outperforms in four courses in **WP**: 'SLC, 'BIM, 'BD, and 'SUP. The results show that all eight models performed better with weekly predictions (**WP**) using the same input features and achieved higher accuracy.

It is worth mentioning that the grey highlighted values indicate a higher F1 score between **CP** and **WP** (see Table 7.3). In addition, the table shows the prediction results differences (∆) between the two methods.

Table 7.3 Results (F1 score) for prediction models in week 1 for both "weekly dropout prediction" and "dropout from the whole course"

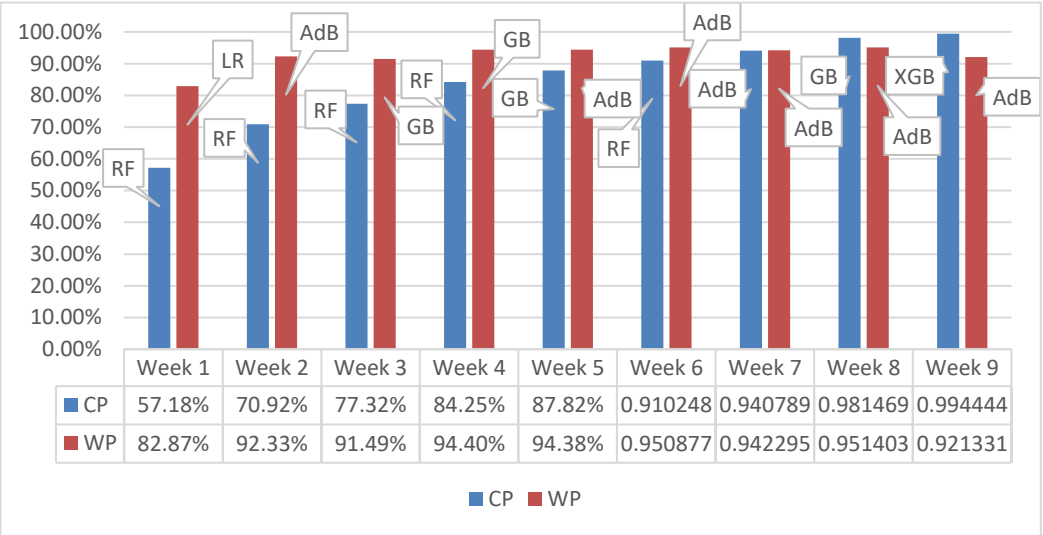
	Courses	Prediction Method	Testing F1-Score Accuracy (F1)													
			AdB	Δ	EX	Δ	GB	Δ	KNN	Δ	LR	Δ	MLP	Δ	RF	Δ	XGB	Δ
Rwaq Platform	SLC	**CP**	75.99%	11.45	72.00%	8.82	74.80%	12.51	63.81%	20.26	75.56%	7.31	0.00%	84.88	75.49%	11.95	75.44%	12.01
		WP	87.44%		80.82%		87.31%		84.07%		82.86%		84.88%		87.44%		87.44%	
	EXC	**CP**	53.01%	34.76	75.61%	4.07	54.00%	33.83	54.17%	32.91	76.18%	10.91	51.25%	28.84	78.62%	8.75	57.18%	30.65
		WP	87.77%		79.68%		87.83%		87.08%		87.09%		80.09%		87.37%		87.83%	
	JAV	**CP**	0.00%	68.79	34.78%	35.04	1.00%	67.96	4.52%	64.33	37.11%	36.09	1.01%	70.44	35.59%	36.07	0.00%	68.68
		WP	68.79%		69.82%		68.96%		68.85%		71.75%		71.45%		71.66%		68.68%	
FutureLearn Durham University	OI	**CP**	59.49%	16.00	75.47%	7.14	66.67%	14.78	55.26%	5.51	65.14%	18.34	65.50%	1.86	74.15%	8.46	65.88%	12.62
		WP	75.49%		82.61%		81.45%		60.77%		83.48%		63.64%		82.61%		78.50%	
	IMPCC	**CP**	72.91%	7.61	71.13%	8.67%	70.56%	9.77	61.51%	11.43	72.81%	7.27	61.90%	15.61	71.24%	7.14	69.14%	11.23
		WP	80.52%		79.79%		80.33%		72.94%		80.08%		77.51%		78.38%		80.38%	
FutureLearn Warwick University	BIM	**CP**	9.96%	52.57	55.17%	12.45	10.32%	43.62	28.99%	15.18	55.44%	12.27	30.68%	8.97	55.53%	12.39	11.92%	38.77
		WP	62.53%		67.62%		53.94%		44.17%		67.71%		39.65%		67.92%		50.69%	
	BD	**CP**	0.77%	78.09	37.23%	41.2	2.26%	76.53	6.99%	69.13	37.34%	40.30	0.00%	77.91	37.34%	41.58	5.88%	72.79
		WP	78.86%		78.51%		78.80%		76.12%		77.64%		77.91%		78.93%		78.67%	
	SHK	**CP**	0.00%	82.05	56.62%	25.00	33.27%	49.53	36.55%	42.05	56.99%	25.89	14.85%	67.36	57.18%	25.18	28.97%	53.20
		WP	82.05%		81.63%		82.80%		78.60%		82.87%		82.21%		82.36%		82.17%	
	TMF	**CP**	0.00%	81.46	47.25%	33.45	13.22%	68.02	23.93%	51.28%	47.32%	33.97	0.00%	81.45	47.17%	33.38	15.14%	65.63
		WP	81.46%		80.70%		81.24%		75.21%		81.29%		81.45%		80.55%		80.77%	
	SUP	**CP**	25%	52.19	58.75%	18.84	37.65%	40.99	40.00%	29.27	59.04%	18.79	50.42%	19.74	58.02%	20.64	38.71%	39.89
		WP	77.19%		77.59%		78.63%		69.27%		77.82%		70.16%		78.66%		78.60%	

In answering **RQ3.1**, the results clearly show how the **WP** method contributed to increasing the F1 score of the classifiers' performance from an early stage (week 1). The **WP** technique generally performed better than the **CP** method in all courses**.** One reason may be that the students who drop out in later weeks behave similarly to completer students in the first week.

However, this is not a problem with the **WP** method, as the model here only predicts dropout students in the upcoming week (week 2). We concluded that the early prediction model should focus only on students who drop out soon. Therefore beyond this section, we will use only the **WP** method to predict dropout students.

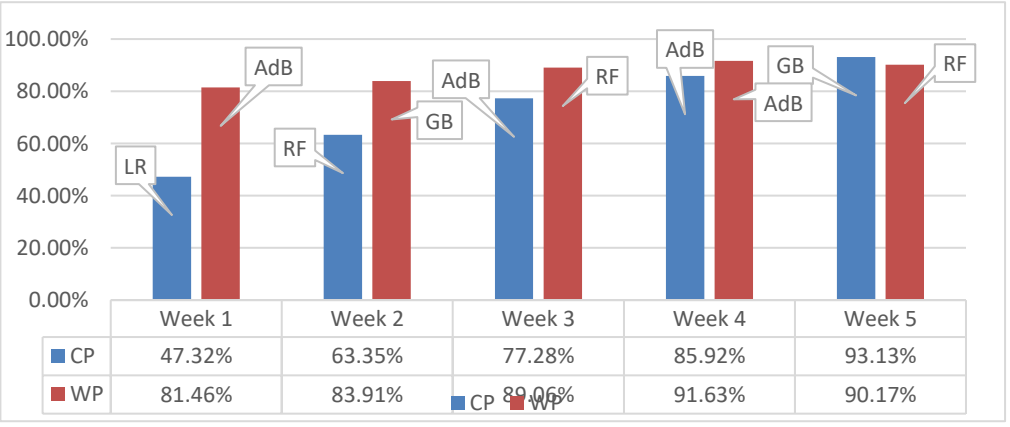
Next, we present the prediction results for more weeks. Figure 7.5 (a–e) shows the most robust prediction models of **WP** and **CP**. It can be seen that the F1 score of **CP** increases in the later weeks of the course.

a) SHK



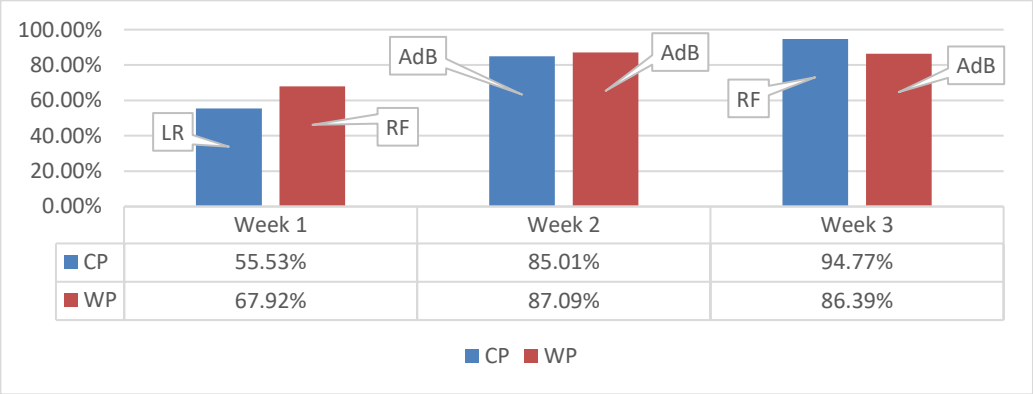
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
CP	57.18%	70.92%	77.32%	84.25%	87.82%	0.910248	0.940789	0.981469	0.994444
WP	82.87%	92.33%	91.49%	94.40%	94.38%	0.950877	0.942295	0.951403	0.921331

b) TMF

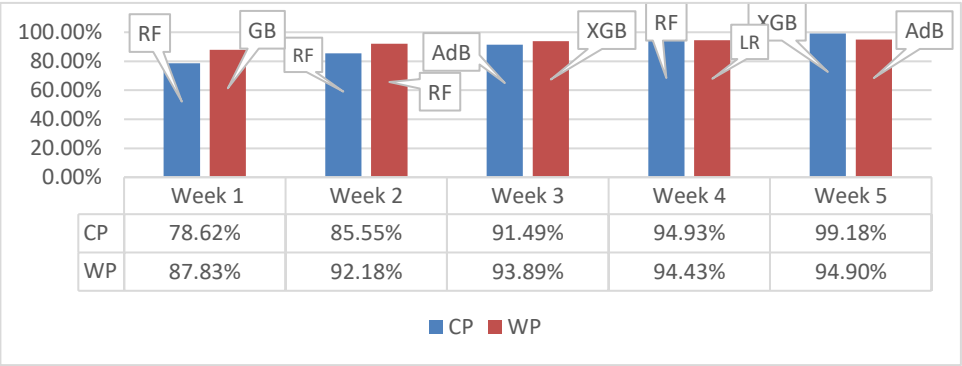


	Week 1	Week 2	Week 3	Week 4	Week 5
CP	47.32%	63.35%	77.28%	85.92%	93.13%
WP	81.46%	83.91%	89.06%	91.63%	90.17%

c) BIM



d) EXC



e) JAV

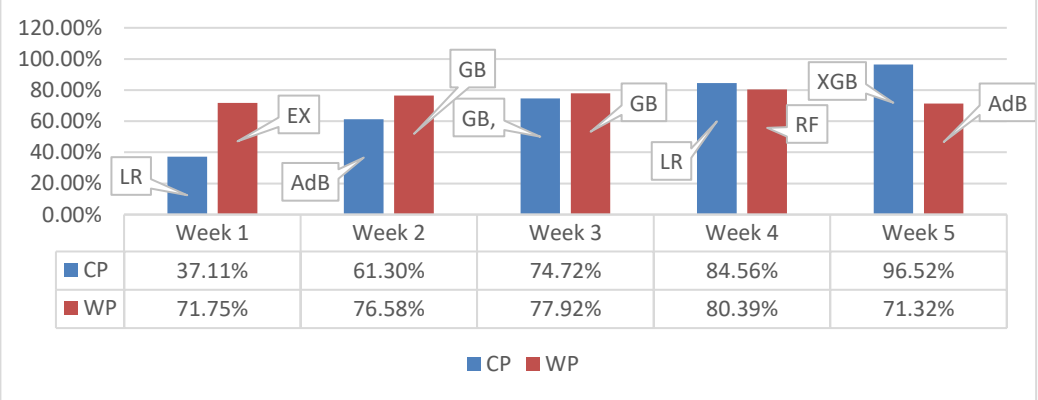


Figure 7.5 weekly prediction vs entire course prediction per week with the best-performing model

7.3.2 Weekly Prediction with Jumping Activities

Building on the previous experiment, this section compares the performance of weekly dropout prediction (**WP**) and weekly dropout prediction with jumping activities (**WPWJ**). In answering **RQ3.2**, we verified the improved prediction performance after considering learners' jumping behaviour in 10 courses. These results are supported by a strong inter-model consensus, with the prediction performance of 71 out of 80 models increasing between 0.07% and 15.98%. For example, after incorporating the jumping learning pattern as a new feature into the dataset, the F1 score increased by 12.25%, from 78.99% to 90.91%, in the RF model in the Supply Chains (SUP) course. In the Shakespeare and His World (SHK) course, accuracy improved by nearly 1.55% to 83.72% for the XGB classifier. This weekly dropout prediction improvement was even more generalised in four courses, where all eight models implemented were more insightful, and the highest F1 score was 90.91% after considering the jumping learning behaviours. Table 7.4 shows the prediction results based on the first-week activities. In light of this analysis, module instructors could implement early interventions, judged on a weekly basis, to improve students' engagement at risk for the upcoming week's dropout.

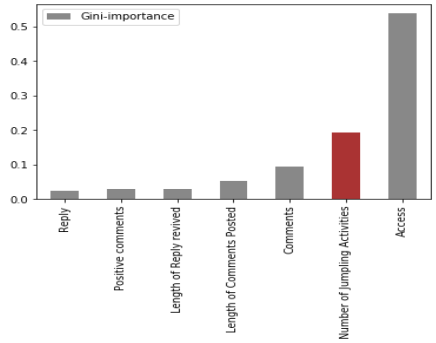
Table 7.4 Results (F1) of prediction models in week 1 for both weekly dropout prediction (*WP*) and weekly dropout prediction with jumping activities (*WPWJ*)

	Courses	Prediction Method	Testing F1-Score Accuracy (F1)															
			AdB	Δ	EX	Δ	GB	Δ	KNN	Δ	LR	Δ	MLP	Δ	RF	Δ	XGB	Δ
Rwaq Platform	SLC	**WP**	87.44%	1.74%	80.82%	2.84%	87.31%	1.87%	84.07%	1.50%	82.86%	6.32%	84.88%	-0.70%	87.44%	1.74%	87.44%	1.74%
		WPWJ	89.18%		83.65%		89.18%		85.58%		89.18%		84.17%		89.18%		89.18%	
	EXC	**WP**	87.77%	0.86%	79.68%	7.86%	87.83%	1.14%	87.08%	1.18%	87.09%	0.12%	80.09%	8.36%	87.37%	1.32%	87.83%	0.78%
		WPWJ	88.63%		87.54%		88.97%		88.26%		87.21%		88.46%		88.69%		88.61%	
	JAV	**WP**	68.79%	0.98%	69.82%	2.25%	68.96%	0.92%	68.85%	0.92%	71.75%	0.41%	71.45%	-1.83%	71.66%	1.32%	68.68%	1.08%
		WPWJ	69.77%		72.07%		69.88%		69.78%		72.17%		69.62%		72.98%		69.77%	
FutureLearn Durham	OI	**CP**	75.49%	-0.73%	82.61%	0.00%	81.45%	0.58%	60.77%	0.34%	83.48%	0.22%	63.64%	10.51%	82.61%	0.51%	78.50%	4.06%
		WPWJ	74.76%		82.61%		82.03%		61.11%		83.70%		74.15%		83.12%		82.57%	

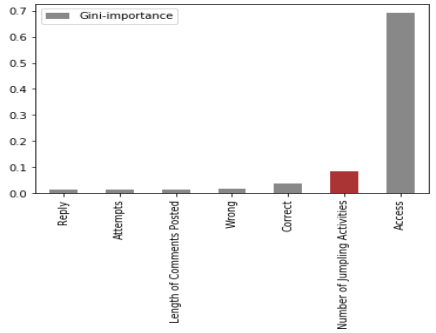
FutureLearn Warwick University	IMPCC	**WP**	80.52%	0.12%	79.79%	0.31%	80.33%	1.00%	72.94%	3.37%	80.08%	0.25%	77.51%	-5.14%	78.38%	2.09%	80.38%	0.93%
		WPWJ	80.65%		80.10%		81.33%		76.30%		80.33%		72.37%		80.47%		81.31%	
	BIM	**WP**	62.53%	7.82%	67.62%	0.23%	53.94%	15.85%	44.17%	7.81%	67.71%	0.57%	39.65%	7.31%	67.92%	2.73%	50.69%	15.98%
		WPWJ	70.35%		67.85%		69.79%		51.98%		68.28%		46.97%		70.65%		66.67%	
	BD	**WP**	78.86%	0.00%	78.51%	1.43%	78.80%	1.62%	76.12%	1.16%	77.64%	0.15%	77.91%	1.56%	78.93%	1.49%	78.67%	1.58%
		WPWJ	78.86%		79.93%		80.42%		77.28%		77.79%		79.47%		80.42%		80.24%	
	SHK	**WP**	82.05%	0.00%	81.63%	0.16%	82.80%	0.89%	78.60%	1.67%	82.87%	0.25%	82.21%	-0.84%	82.36%	-0.14%	82.17%	1.55%
		WPWJ	82.05%		81.79%		83.69%		80.26%		83.12%		81.38%		82.22%		83.72%	
	TMF	**WP**	81.46%	0.07%	80.70%	0.17%	81.24%	0.96%	75.21%	1.80%	81.29%	0.10%	81.45%	0.32%	80.55%	0.29%	80.77%	1.45%
		WPWJ	81.54%		80.87%		82.21%		77.01%		81.39%		81.77%		80.85%		82.22%	
	SUP	WP	77.19%	10.96%	77.59%	7.63%	78.63%	9.74%	69.27%	13.14%	77.82%	8.38%	70.16%	10.68	78.66%	12.25%	78.60%	10.29%
		WPWJ	88.15%		85.22%		88.37%		82.41%		86.21%		80.83%		90.91%		88.89%	

Following this, we analysed the most important features to predict students who did not access 80% of the topics in the second week. Figure 7.6 shows the feature importance (Gini importance) (Dorfman, 1979) for the most robust model in each course. We present the seven most important features.

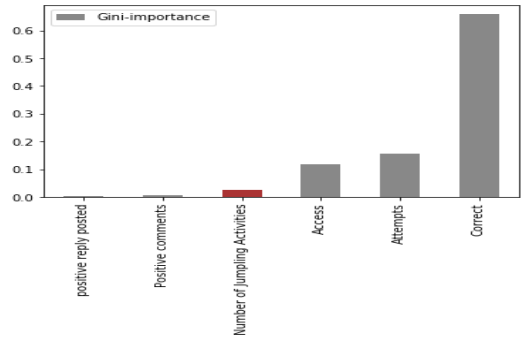
From the figure, it can be seen that the *number of jumping activities* feature is ranked as number one in terms of its importance in predicting student dropout in three courses (SLC, EXC and SUP) and the second most important feature in four courses (LMPCC, TMF, BD and SHK).



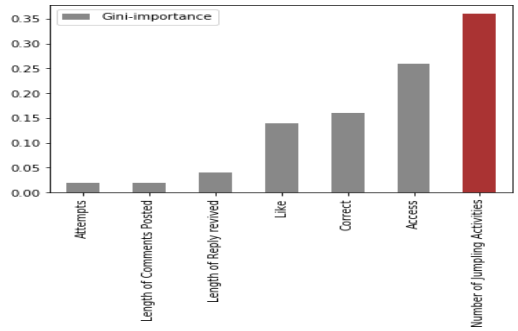
a) BD



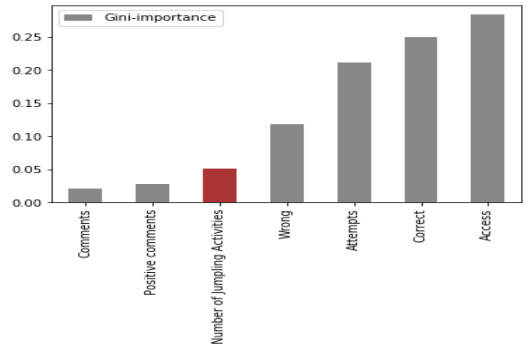
b) SHK



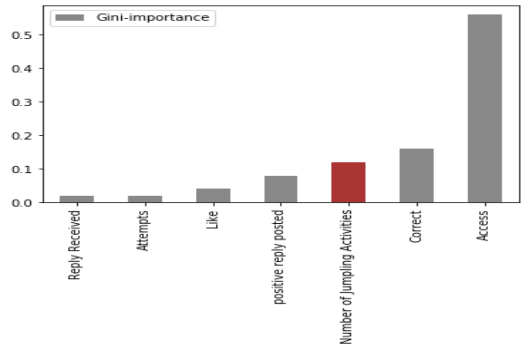
c) TMF



d) SUP



e) BIM



f) OI

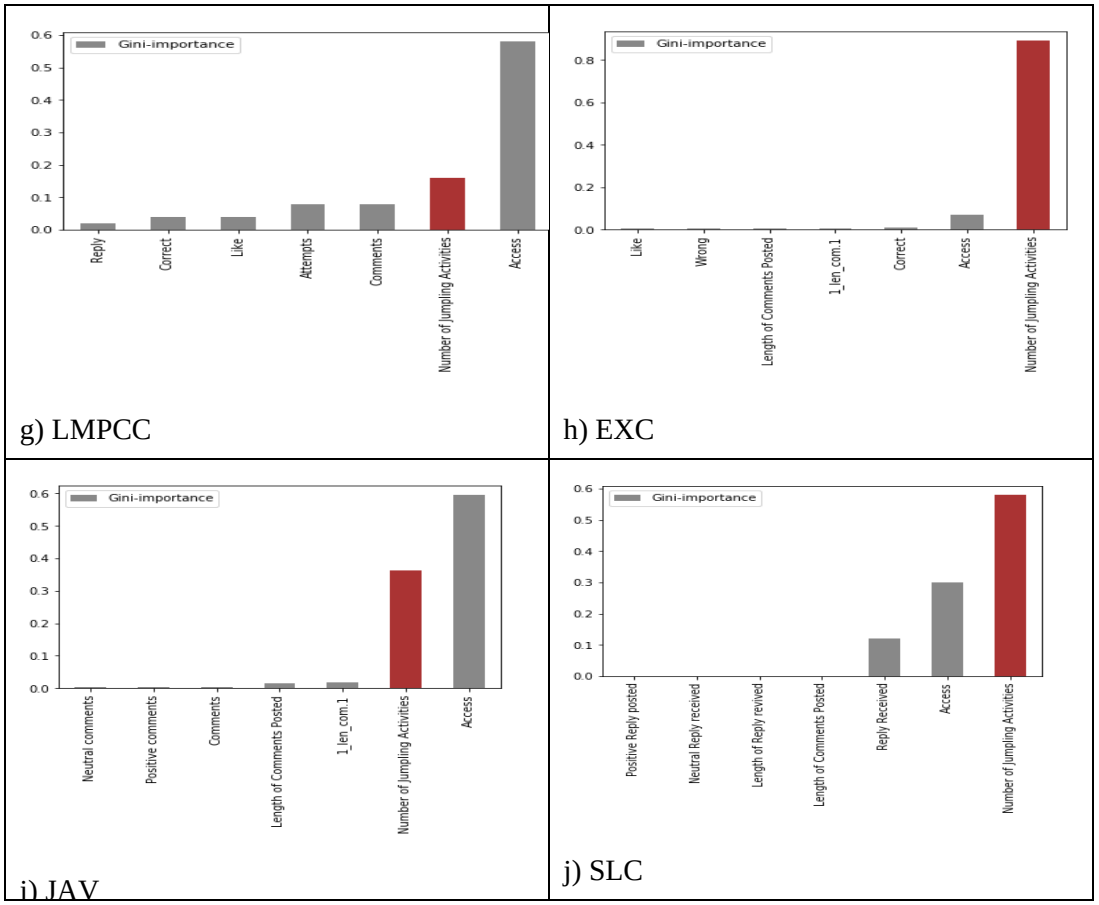


Figure 7.6 Importance of predictive features (a-j)

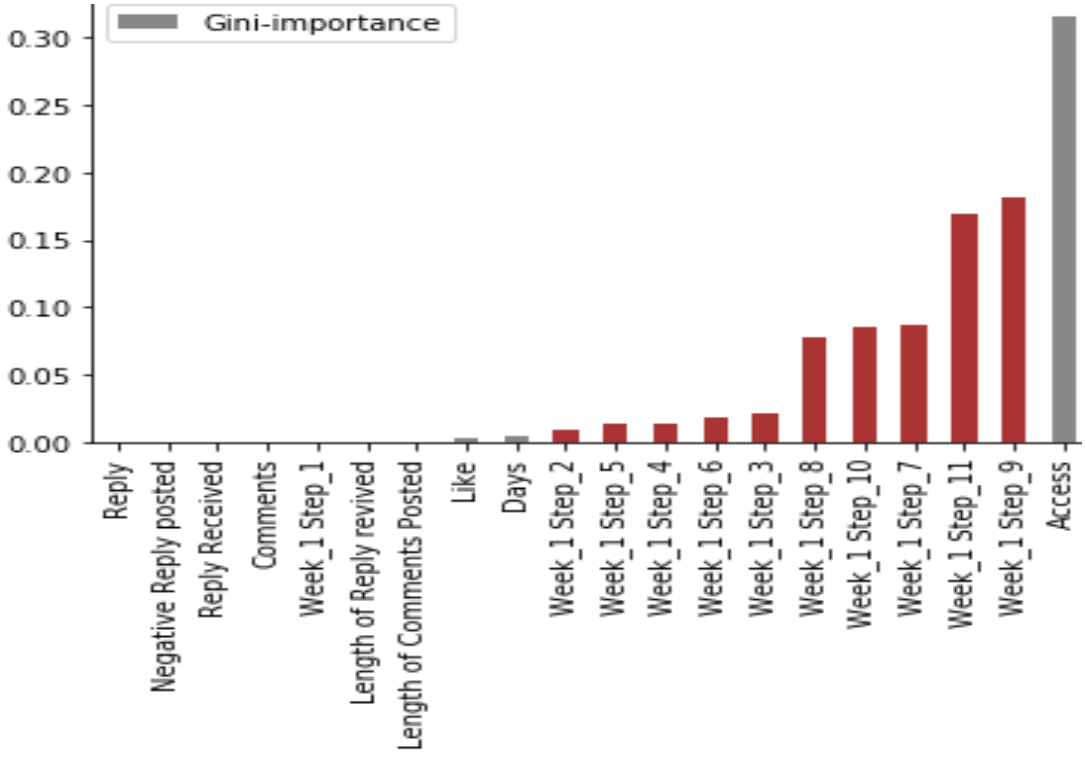
7.3.3 Two easily obtainable features

One of the goals of this research was to create a simple model. From the data in Figure 7.6, it is apparent that the ***number of accesses*** feature is considered significant, as it ranked at the top among the six courses. Therefore, to answer **RQ3.3**, we will investigate students' ***access*** features and the ***time spent*** on each access. In addition, we focussed on specific features that could be used for various MOOCs – this was done to enhance the generalisation and applicability of the findings for the providers.

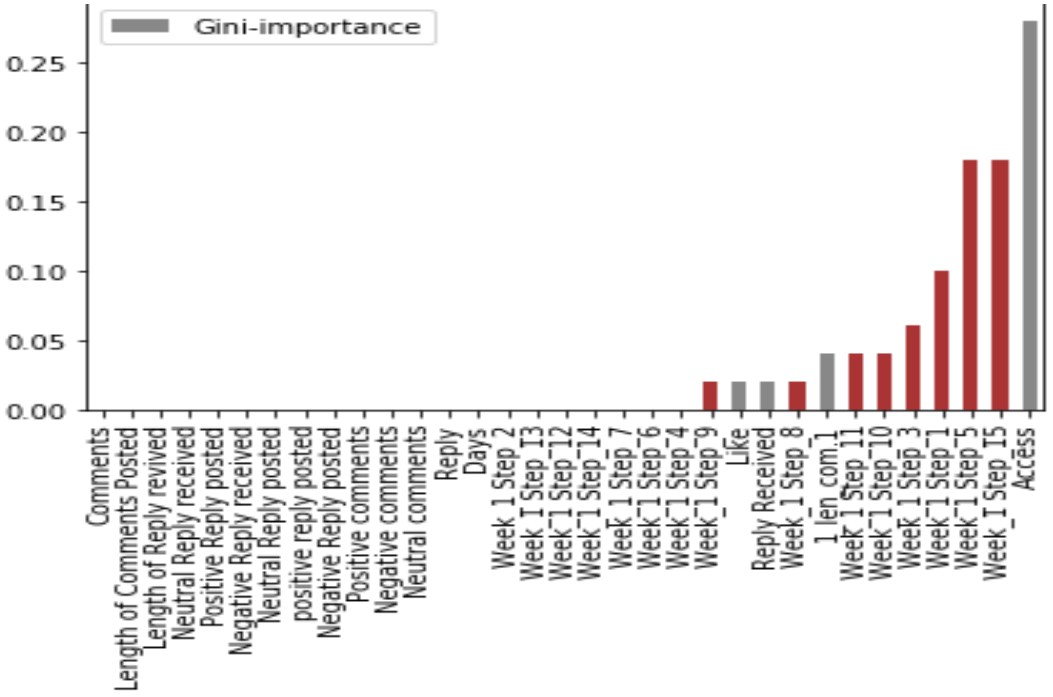
Therefore, we first applied all the features to predict dropout students in the second week (see Figure 7.7 Gini importance for five courses. Due to space limitations, Appendix C

presents the other courses). In this experiment, we included time spent as a feature that represents the total time spent completing each step (see Section 7.2.5).

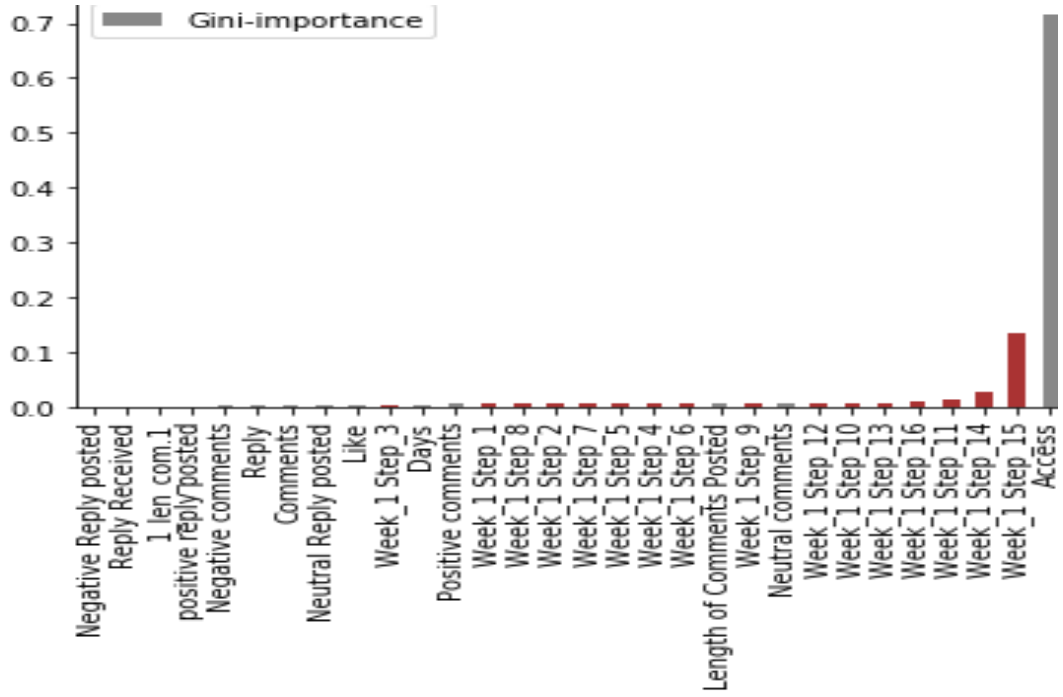
We concluded that *time spent* and *number of accesses* are vital features not only because they are easy to obtain for most courses, but also because the results show that time spent in each step plays a critical role in predicting student completion. Moreover, the ***number of accesses*** was, in general, an essential feature of all the courses. Furthermore, it should be taken into consideration that some courses do not have quizzes every week, and only 5% to 10% of students posted comments in MOOC discussion forums (Wen et al., 2014a); in this case, the **w*rong answers***, ***correct answers*** and all the features related to students' comments did not play an important role in predicting student completion in those courses (see Big Data (**a**) and Self Confidence (**b**) courses in Figure 7.7). Finally, all features in Figure 7.7 that begin with "week_1_steps_" represent the total time spent by a given student in that step.



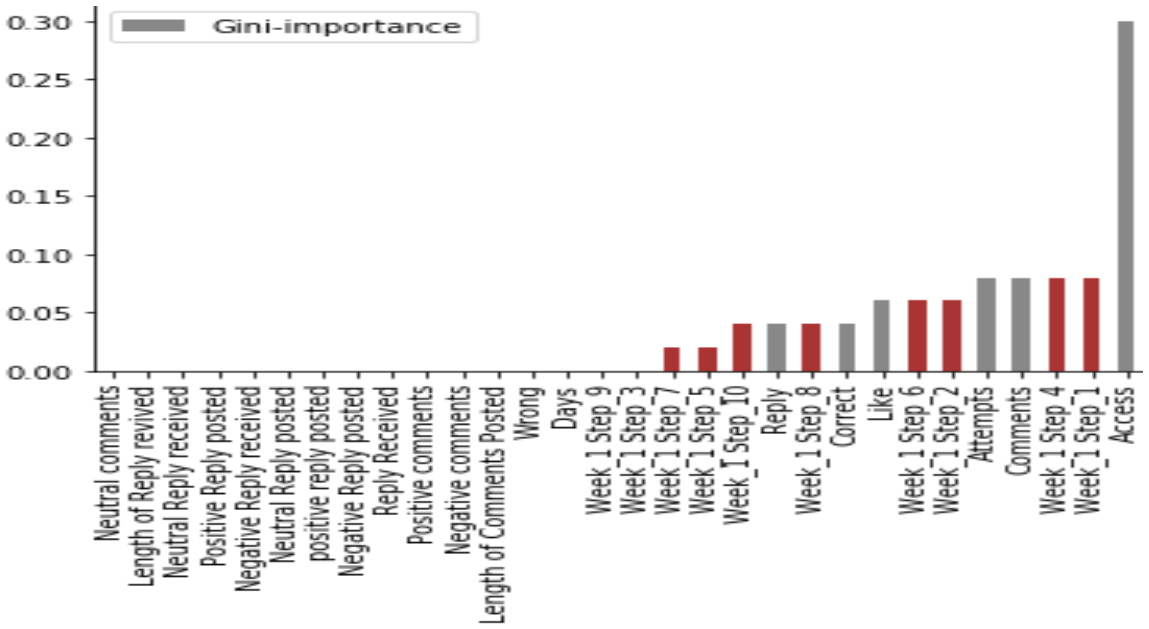
a) Big data


Gini-importance
0.25
0.20
0.15
0.10
0.05
0.00
Comments
Length of Comments Posted
Length of Reply revived
Neutral Reply received
Positive Reply posted
Negative Reply received
Neutral Reply posted
positive reply posted
Negative Reply posted
Positive comments
Negative comments
Neutral comments
Reply
Days
Week 1 Step 2
Week 1 Step 13
Week 1 Step 12
Week 1 Step 14
Week 1 Step 7
Week 1 Step 6
Week 1 Step 4
Week 1 Step 9
Like
Reply Received
Week 1 Step 8
1 len com.1
Week 1 Step 11
Week 1 Step 10
Week 1 Step 3
Week 1 Step 1
Week 1 Step 5
Week 1 Step 15
Access

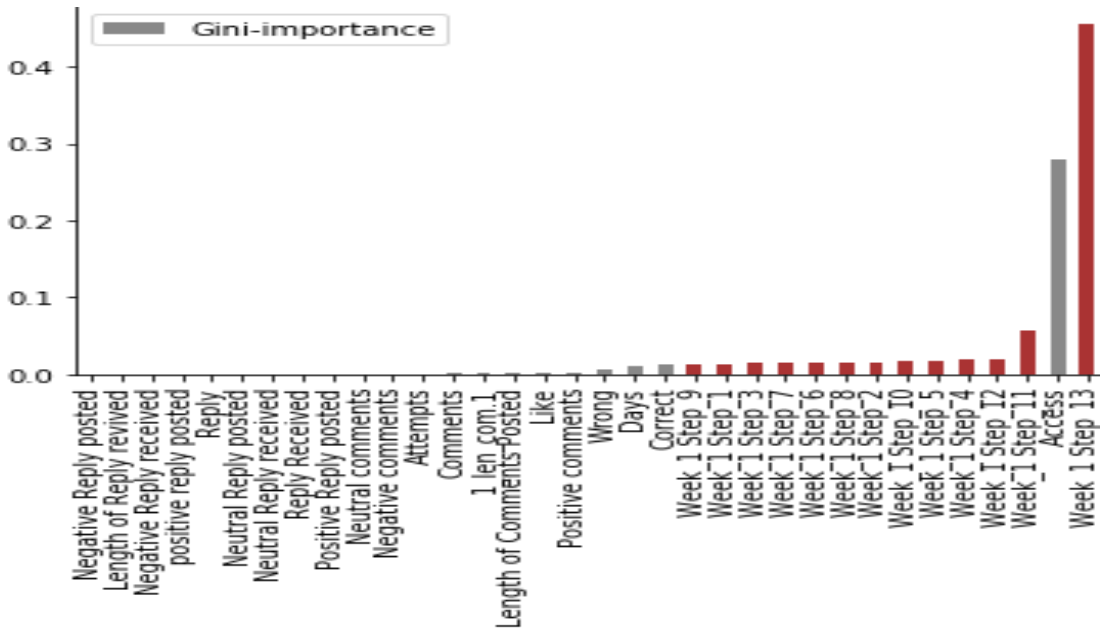
b) SLC


Gini-importance
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Negative Reply posted
Reply Received
1 len com.1
positive reply posted
Negative comments
Reply
Comments
Neutral Reply posted
Like
Week 1 Step 3
Days
Positive comments
Week 1 Step 1
Week 1 Step 8
Week 1 Step 2
Week 1 Step 7
Week 1 Step 5
Week 1 Step 4
Week 1 Step 6
Length of Comments Posted
Week 1 Step 9
Neutral comments
Week 1 Step 12
Week 1 Step 10
Week 1 Step 13
Week 1 Step 16
Week 1 Step 11
Week 1 Step 14
Week 1 Step 15
Access

c) Java



d) LMPCC



e) EXL

Figure 7.7 Gini-importance for five courses (a-e)

The figure below (Figure 7.8) illustrates the mean of the time spent by completers and non-completers on the first step of the first week across all ten courses. Results show that non-completers spent 39% more time than completers in the Java course and 21% more time in EXC course. On the other hand, the completers spent between 4% to 25% more time than non-completers in the first step of eight courses.

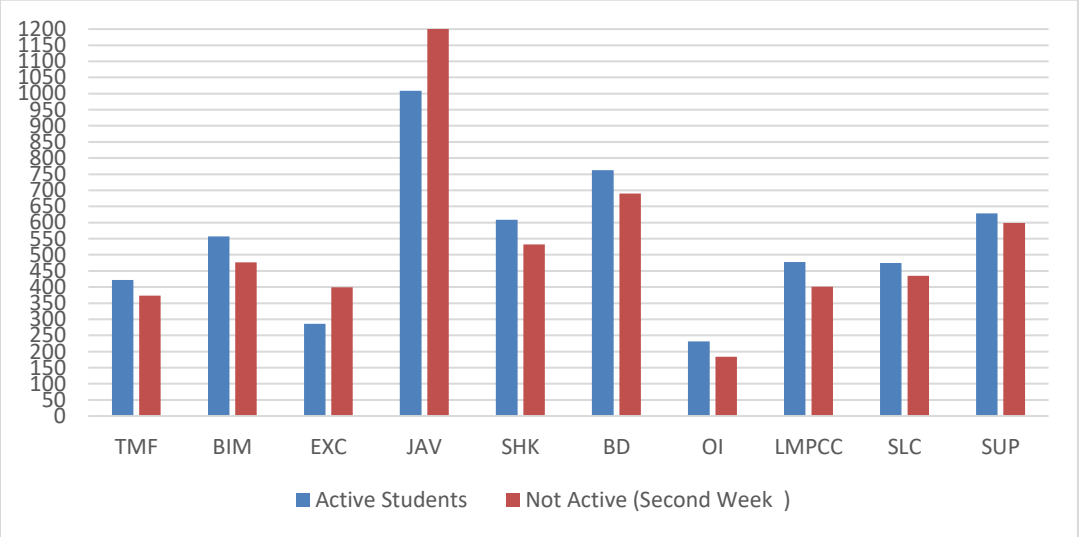


Figure 7.8 Time spent (in second) on the first step by the completers and non-completers.

Table 7.5 below shows the results for 10 courses. The prediction F1 score varied between 69% and 87%. We can see that the best-performing course across all models applied was the Self-Confidence (**SLC**) course. The worst-performing course, on the other hand, 'Babies in Mind', was the shortest (4 weeks). In general, the most robust model is random forest (RF), as it outperforms seven out of 10 courses.

Table 7.5 Prediction performance using the time spent and number of access

	AdB	**EX**	**GB**	**KNN**	**LR**	**RF**	**XGB**
SLC	87.44%	81.32%	87.31%	85.15%	82.86%	87.44%	87.44%
EXC	87.78%	86.81%	89.10%	75.02%	83.48%	89.21%	89.13%
JAV	68.84%	69.70%	71.18%	55.12%	62.97%	73.08%	70.12%

OI	82.10%	81.36%	80.70%	61.45%	81.17%	82.76%	77.93%
IMPCC	80.54%	80.23%	78.88%	60.11%	77.05%	80.62%	76.97%
BIM	68.71%	68.49%	64.36%	41.61%	66.86%	69.31%	63.88%
BD	78.45%	77.62%	77.98%	58.17%	75.07%	79.07%	77.36%
SHK	81.68%	80.44%	81.92%	66.11%	78.04%	81.68%	81.76%
TMF	80.51%	79.03%	79.90%	60.93%	76.00%	80.51%	80.63%
SUP	79.09%	75.95%	77.93%	62.50%	73.25%	78.76%	74.64%

The method of using only ***time spent*** and the ***number of accesses*** features showed competitive results compared to the ***WP*** method but utilised far fewer indicators to achieve success much earlier. This is due to the careful selection process of the two features, which are both generic and informative. One important reason the two early, first-week features were enough for such good prediction is the fine granularity of the mapping of these features – for each 'step' (or piece of content), we could compute both the ***number of accesses*** as well as the ***time spent***. Thus, the application of the features for the first week transformed into a multitude of features, which would explain the increased prediction power. Nevertheless, this method is widely applicable and does not detract from the generalisability of our findings.

7.3.4 Early prediction performance

In the previous three experiments, based on the first week of the course, we managed to predict only what the outcome would look like. For some courses, this represented a prediction based on a quarter of the course (e.g., for 'Babies in Mind'). For others, the prediction was based on one-tenth of the course data, which was a short time to draw conclusions.

A few further important remarks need to be considered. First, data pre-processing is vital. For such extremely skewed datasets as encountered when studying MOOC completion, where averages of 10% completion of the whole course are the norm, prediction can 'cheat'

easily: for example, by just predicting that all students fail, we would obtain a 90% completion rate. To avoid such blatant bias (e.g. using overall average accuracy), we used the F1 score, a commonly used metric for the binary classification of an unbalanced dataset.

As shown above, the difficulty in the problem we were tackling was the prediction of the completers; thus, it would be easy to hide the poor prediction in this ‘hard’ category by using overall accuracy.

Epilogue

This chapter presents three experiments to predict dropout students from an early stage, starting in week one. The first experiment compared the prediction of weekly dropouts and whole-course dropouts. The second experiment was to build a dropout predictive model based on students' jumping behaviours. Finally, the third experiment was to predict dropout students from an early stage based on two easily obtainable features (the ***number of accesses*** and the ***time spent***).

In the next chapter, we propose a concrete mapping between the tracking parameters and four of the most frequently used theories related to students' engagement in digital systems.

Chapter 8 : The Engage Taxonomy: SDT-based measurable Engage-ment Indicators for MOOCs and their Evaluation

Prologue

This chapter shows how Self Determination Theory (SDT) can be mapped onto concrete features extracted from tracking student behaviour on MOOCs. We map the dimensions of Autonomy, Relatedness and Competence, leading to methods to characterise engaged and disengaged MOOC student behaviours, and exploring what triggers and promotes MOOC students' interest and engagement.

8.1 Introduction

Recently, the dynamics of engagement and motivation in MOOC systems has been especially targeted (Ferguson et al., 2015). However, to the best of our knowledge, there are no works to date, which systematically employ motivational theories, mapping online student behaviour onto them, to analyse the drives and triggers promoting student engagement. Moreover, in the past, engagement theories have been created often based on theoretical findings from psychology, or small-scale experiments (Moreno-Murcia et al., 2013, Shen et al., 2009, Langdon et al., 2014). We advocate that, in addition, it is vital to provide numerical, tested engagement measures, for direct application in MOOCs and numerical comparisons. We consider the advent of 'big data' as a chance to evaluate these theories at scale.

To address these gaps, we use raw multimodal, multi-dimensional data from comprehensive tracking of student behaviour, as well as aggregate features - such as those extracted from natural language processing (NLP) - to cluster students, then analysing the engagement parameters of these clusters, based on solid motivational theories - starting with one of the most well-known, cited and used one, especially in the education domain - the Self-

Determination Theory (SDT) (Zhou, 2016, Duncan et al., 2020, Deci and Ryan, 2013). Many previous works focus only on some particular aspects of the student behaviour (Cristea, 2018, Shi and Cristea, 2018a). Instead, here, we triangulate multimodal tracking data at various granularity levels - temporal mode (time-stamp, day, week, course), action mode (where we compute frequencies of student actions for a given time interval), natural language mode (where we analyse the language exchange content, including its sentiment mode, etc.) - resulting in 17 indicators. To obtain significant, generalisable results, we perform this engagement analysis on a large longitudinal dataset (6 MOOC courses with 26 runs, spanning 2013-2018, delivering to 218,235 students).

This research targets the following research questions:

RQ4.1: Can engagement theories help in identifying student success on MOOCs?

RQ4.2: How are engagement theories applicable in MOOCs?

8.2 Engage Taxonomy: Mapping of MOOC indicators onto engagement theories

8.2.1 Extracting raw and computing aggregated MOOC indicators

As we have observed, most of the motivational theories discussed in Section 2.5 have relatively similar concerns about what triggers student engagement and motivation. The challenge is, then, to map their respective engagement concepts onto MOOC behaviour, available as tracked data, to extract concrete measures that address them. Thus, our next step is to have a collection of potentially relevant data that can be tracked from MOOCs in general. The point being that the more available the data is across different MOOC types and platforms, the more likely it is that this process is generalisable. Most MOOCs track access to the material, answers to any quizzes, and store chats of their students. Based on indicators used by prior MOOC research for clustering (Oyelade et al., 2010, Rana and Garg, 2016), prediction (Ding, 2019a, Gitinabard, 2018, Qiu, 2019, Mubarak, 2020, Doleck, 2020), data analytics (Moreno-Marcos et al., 2020, Shorfuzzaman et al., 2019), we gather the following indicators:

1. Number of accesses steps per week	7. Number of positive comments per week	13Number of negative replies posted per week
2. Number of correct answers per week	8. Number of negative comments per week	14.Number of neutral replies posted per week
3. Number of wrong answers per week	9. Number of neutral comments per week	15. Number of positive replies received per week
4. Number of attempts per week	10Number of replies posted per week	16. Number of negative replies received per week
5. Number of comments per week	11. Number of replies received per week	17. Number of neutral replies received per week
6. Number of likes received per week	12. Number of positive replies posted per week	

Please note that feature extraction has been introduced in section 4.5.2.

8.2.2 Mapping Indicators to Engagement Theories

Next, the mapping between the engagement and motivation concepts and the potential indicators within MOOCs was done independently by three experts (see details on the procedure in Section 8.3.2). Table 8.1 presents the mapping performed by our experts between the engagement and motivation concepts and the potential indicators within MOOCs that can be tracked, resulting in the Engage Taxonomy.

As can be seen from the table, due to the generic nature of the MOOCs indicators selected, our Engage Taxonomy is available to the research community for further exploitation of other motivational theories. In fact, Table 8.1 shows mapping onto four popular motivational theories, SDT, Drive, Engagement Theory and Process of Engagement. Thus, this research can be used to showcase how to tackle this exploitation and continue using both SDT and the other motivational theories.

Analysing the expert mapping, for example, experts found that MOOC learners' independent, voluntary activities, such as #Accessed Steps, could be related to their 'Autonomy' (AUT). Any activity showcasing users' skills, such as #Correct Answers, could be related to the 'Competence' (COM) construct in the SDT theory, and the 'Mastery' (MAS) construct in the Drive theory, respectively. User's skills in MOOCs could be tracked by storing numbers - such as the number of quizzes answered. Similarly, the users' social interactions address the 'Relate' (REL) concept in the Engagement Theory, the 'Relatedness' (REL) concept in SDT and other social attributes within the 'Period of Engagement' (PER);

this may be represented in MOOCs by indicators, such as the number of # Main Comments posted and interactions within those comments, or replies and likes.

Please note that similar constructs, such as AUT in SDT and Drive, have a similar mapping. 'Create' (CRE) in the Engagement Theory is somewhat related to autonomy, but it is more concerned with the process of creating a students' own path, or, even more interestingly, creating new information via comments. On the other hand, none of the indicators were considered appropriate for the 'Point of Engagement' (POI). All constructs regarding comments and replies were considered to be a good mapping for 'Relatedness' (REL) in both SDT and Engagement Theory. However, 'Competence' (COM) in SDT was only showcased in Quiz by # Correct Answers, # Wrong Answers, Comments by the #Likes Received and in Replies by positive replies posted or received – although the #Replies Posted also was considered a measure of competence, possibly as the ones feeling ready to answer other learners' questions would show a degree of competence.

Still, this research proposed mapping (according to our experts), whilst potentially of use to the research community, evaluated in various ways in the rest of the research and showed to be performant, is not claiming to be optimal, but can be further improved upon. Moreover, the process of obtaining and testing it are, we believe just as useful as the final product.

Next, we explain why and how we select SDT, for the evaluation of the expert mapping.

Table 8.1 *Engage Taxonomy*: Motivational Theories Mapped onto students' Activities (indicators) in MOOCs

Theories:		SDT			Drive			Engagement Theory			Process of Engagement			
MOOC features (indicators): Activities per week		AUT	COM	REL	AUT	MAS	PUR	CRE	REL	DON	POI	PER	DIS	REE
Steps	# Accessed Steps	x			x			x				x		
Quiz	# Correct Answers		x			x				x		x		
	# Wrong Answers		x			x				x		x	x	
	# Attempts to answering questions	x			x							x		
Comments	# Main Comments posted	x		x	x		x		x	x		x		x
	# Like Received		x	x		x			x					
	# Positive comments posted	x		x	x		x	x	x			x		x
	# Negative comments posted	x		x	x		x	x	x			x	x	
	# Neutral comments posted	x		x	x		x	x	x			x		x
Replies	# Replies Posted	x	x	x	x		x	x	x			x		x
	# Replies Received			x					x					
	# Positive Replies posted		x	x			x		x			x		x
	# Positive replies received		x	x		x			x			x		x
	# Neutral replies posted			x			x		x			x		x
	# Negative replies received			x			x		x				x	
	# Negative replies posted			x			x		x			x	x	
	# Neutral replies received			x			x		x					

8.2.3 SDT as illustrator of the Engage Taxonomy

We only work with Self-Determination Theory (SDT) (Deci and Ryan, 2013), as it is arguably the most well-known engagement theory, and is well-supported by socio-psychological literature (Gerber and Anaki, 2021). SDT is a macro-theory linking personality, human motivation, and optimal functioning. It stems from research on two main

types of motivation —intrinsic and extrinsic— that are further thought to shape human behaviour (Deci and Ryan, 2013). SDT posits that humans become self-determined when their needs for Competence, Relatedness and Autonomy are fulfilled. Self-determined individuals believe they are in control of their lives, take responsibility for their behaviours, are self-motivated and determine their actions based on internal values and goals. SDT has led to various sub-theories, such as organismic integration theory and causality orientation (Hagger and Hamilton, 2021). In education, students are more likely to learn and succeed when they are intrinsically motivated by their need for Competence, than when extrinsically motivated (Standage et al., 2005). Studies within SDT provide strong psychological evidence to support a more interactive, multidimensional picture of human nature in various sociocultural contexts (Chirkov, 2009). SDT has been used in musical education (Evans, 2015), physical education (Vasconcellos et al., 2020), science education (Lavigne et al., 2007), medical education, amongst others. It is worth mentioning that SDT is further connected with the self-regulated learning (Littlejohn et al., 2016) method – which is the predominant approach in MOOCs, and thus we consider SDT an excellent choice of a first analysis of motivational theory application and testing for MOOCs.

Whilst SDT is, as mentioned, well-known and frequently applied. Nevertheless, SDT and other motivational theories have not yet been evaluated on large-scale data. This opportunity is given to us in the context of online learning and MOOCs. Finding out to what extent SDT is really applicable is thus a useful endeavour. Hence, SDT represents a good starting point to experiment with mapping MOOC features onto motivational theories.

In this rsearch, we propose and use early measurable indicators of engagement (the Engage Taxonomy for SDT) from the first week activities (see Sections 8.2.2), as mapped by experts (see Section 8.3.2), and apply these onto the concrete data from our MOOCs to all student clusters. These early behavioural clusters are analysed in terms of their semantics derived from SDT, on the axes of Autonomy, Competence and Relatedness. The idea being that, if we find semantically relevant clusters, build based on SDT variables, and then further find that they are correlated to the success parameters, this confirms that the motivational theory-rooted methodology can be applied to characterise students' success. After performing the SDT mapping, we need to establish how to measure its success. This is further explained in Section 8.2.5. Next, we tackle how to compute the engagement measures.

8.2.4 Engagement Measures: Computing SDT aggregate constructs

As mentioned, we proceed in our further analysis with the SDT model and mapping. Thus, we analyse the following SDT-related constructs, proposing hereby also numerical ways of computing them:

Autonomy (per week): we created this aggregate construct, based on a generalised version of the data in Table 8.1, computed as a normalised value, $Aut \in [0,1]$, as follows:

$$\boldsymbol{Aut}(\boldsymbol{s},\boldsymbol{w}) = \frac{\sum_{c \in C_{Aut}}^{w} (\boldsymbol{we}_c * \frac{\boldsymbol{c}_w^s}{\boldsymbol{max}_{ss \in \mathbf{S(w)}} \boldsymbol{c}_w^{ss}})}{\sum_{c \in C_{Aut}}^{w}} \quad \textbf{(8.1)}$$

Here C_{Aut} are all constructs (extracted from tracking data) usable for establishing the Autonomy of students; where c_w^s is the value of construct $c \in C_{Aut}$ for student $s \in S(w)$ in week w, normalised by dividing it by the maximum of all values of c in that week w, for all students $ss \in S(w)$ in week w; we_c is the weight of construct c in the computation of the Autonomy, and should be a value between [0,1]; this weight allows to have different constructs to influence the result in a different way; currently, we used $we_c = 1$, although further experimentation could render more exact results. Finally, to ensure Aut(s,w) $\in [0,1]$, we normalise the result by dividing it by the number of constructs in C_{Aut}. E.g., if in week w1 there are 6 steps in total, 3 quizzes, and students have posted together 5 comments (out of which, 2 are positive, 2 neutral and 1 negative), and there have been in total 2 replies; a student s1 has accessed 3 out of the max 6 steps, answered 1 of the 3 quizzes, and, for simplicity, 0 out of 5 comments in week w1, as $\sum_{c \in C_{Aut}}^{w} = 7$ as we have 7 features on Autonomy for SDT in Table 8.1. Then Aut(s1,w1) = (3/6+1/3+0/5+0/2+0/1+0/2+0/1)/7= 0.119. A student s2 performing the maximum number of activities related to autonomy constructs would have a total of Aut(s2,w1) = (6/6+3/3+5/5+2/2+1/1+2/2+1/1)/7= 1, representing the maximum value of the autonomy computable in that week.

Competence (per week): is also our created aggregate construct, on Table 8.1, defined similarly to the Autonomy construct above (Eq. (8.1)), but summing only over C_{Com}, the Competence-related constructs (formed of tracked data). For example, if a student s3 has 1

correct answer out of a maximum of 3, 2 wrong answers out of 3, 1 like received out of 2, 2 replies posted out of 6, 1 positive reply posted out of 3 and 2 positive replies received out of a maximum 2, in week w1, as $\sum_{c \in C_{com}}^{w} = 6$, as we have 6 features on competence for SDT in Table 8.1, Com(s3,w1) = (1/3+2/3+1/2+2/6+1/3+2/2)/6= 0.53.

$$\boldsymbol{Com}(\boldsymbol{s},\boldsymbol{w}) = \frac{\sum_{c \in C_{Com}}^{w} (we_c * \frac{c_w^s}{max_{ss \in S(w)}\, c_w^{ss}})}{\sum_{c \in C_{Com}}^{w}} \quad (8.2)$$

To further illustrate the usefulness of the weights, it is possible that, instead of using the same weight overall, we would consider in a further iteration of this research (not further explored here beyond this section) that, for the Competence construct, quiz results are much more important than comments and replies. Thus, with the rest of the data as above, instead of $we_c = 1$ for all, we could have :

$\boldsymbol{we}_{\#\ Like\ Received}$ =$\boldsymbol{we}_{\#\ Replies\ Posted}$ = $\boldsymbol{we}_{\#\ Positive\ Replies\ posted}$ = $\boldsymbol{we}_{\#\ Positive\ replies\ received}$ =0.5;
$\boldsymbol{we}_{\#\ Correct\ Answers} = \boldsymbol{we}_{\#\ Wrong\ Answers} = \mathbf{1}$

Thus, if student s4 would have the maximum number of correct answers and no other competence-related accomplishments in week w1, whereas student s5 would have the maximum number of comments liked during the same week but no other accomplishments, in this case, we would have $Com(s4, w1) > \mathrm{Com(s5, w1)}$, as:

$$Com(s4,w1) = \frac{\left(\frac{3}{3}\right)*1+\left(\frac{0}{3}\right)*1+\left(\frac{0}{2}\right)*.5+\left(\frac{0}{6}\right)*.5+\left(\frac{0}{3}\right)*.5+\left(\frac{0}{2}\right)*.5}{6} = 0.167 > \mathrm{Com(s5, w1)} = \frac{\left(\frac{0}{3}\right)*1+\left(\frac{0}{3}\right)*1+\left(\frac{2}{2}\right)*.5+\left(\frac{0}{6}\right)*.5+\left(\frac{0}{3}\right)*.5+\left(\frac{0}{2}\right)*.5}{6} = 0.083$$

Relatedness (per week): is our final aggregate construct, again based on Table 8.1, and defined similarly to the Autonomy and Competence constructs (Eq. (8.1) (8.2)), summing here over C_{Rel}, the Relatedness-related constructs (formed of tracked data):

$$\boldsymbol{Rel}(\boldsymbol{s},\boldsymbol{w}) = \frac{\sum_{c \in C_{Rel}}^{w} (we_c * \frac{c_w^s}{max_{ss \in S(w)}\, c_w^{ss}})}{\sum_{c \in C_{Rel}}^{w}} \quad (8.3)$$

From the generic way we have created these formulas, one could see that they are immediately applicable onto the other theoretical engagement and motivation approaches analysed in Table 3.3, such as Drive, Engagement Theory, and Process of Engagement. Please note that we do not claim this model to be optimal; it is, however, a simple one, thus an Occam-razor based approach.

The overall connection between the definitions above, the mapping in Section 8.2.2 (Table 8.1) and the indicators in Section 8.2.1 are further shown in Figure 8.1. For instance, raw data, such as course contents, maps, via pre-processed data #Accessed_steps (number of steps accessed by the current student), to the construct on Autonomy in the SDT mapping. Similarly mapped to Autonomy are #Attempts to answer questions, #Main Comments posted, etc.). Based on this mapping, the Aut(s,w) measurable, SDT-related construct, is defined. Thus, Figure 8.1 illustrates how we can obtain measurable motivational theory constructs from raw student tracking data. It shows the sequence of operations for extracting these features and turning them into the input data used for clustering and machine learning.

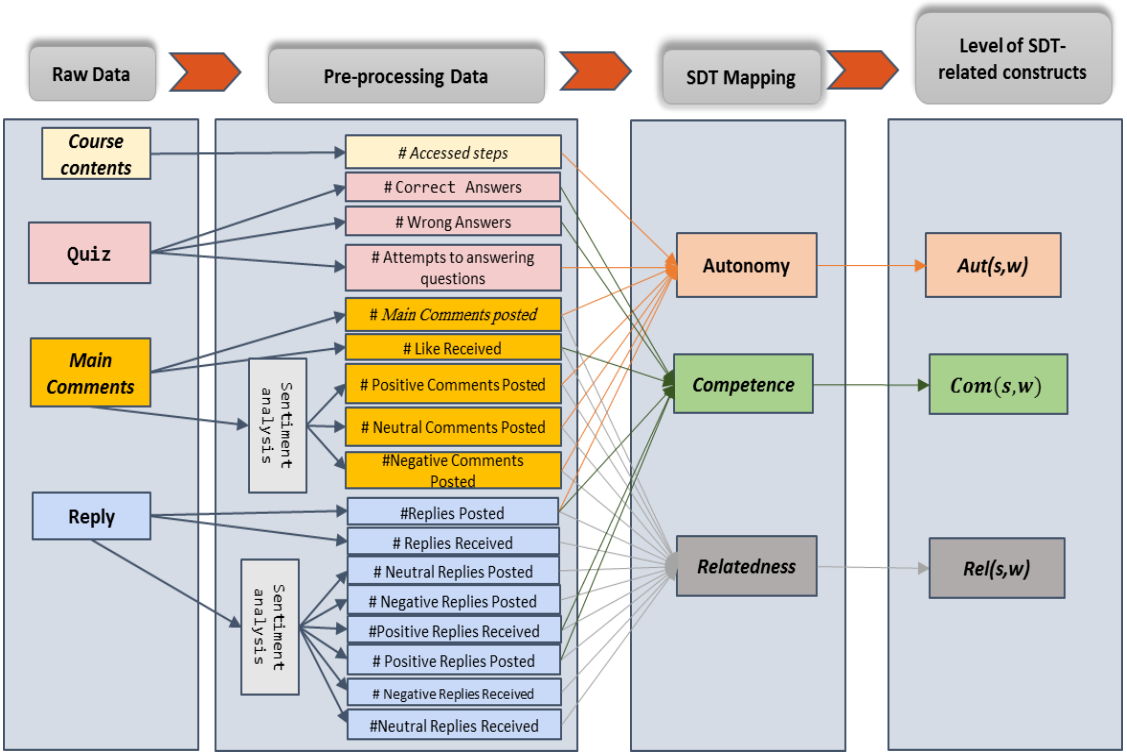


Figure 8.1 SDT Theory, mapped to students' activities in MOOCs

8.2.5 Computing SDT Success

Thus, besides completion, it is useful to look at other measures of success. We do this here in two ways: one, by using a new, rigorous methodological approach based on engagement and motivation theory - and mapping SDT constructs onto track data (raw or derived) of the students, from 17 features, including lower used features, such as sentiment-related ones (as described in Sections 8.2.1,8.2.2). The second way is in measuring the success in terms of the other ways a student can interact with a MOOCs, beside reading (and completing) pages: by answering quizzes and posting comments. Thus, we measure the student's success additionally via the Correct Answer ratio and the Reply ratio from week 2 to the last week (Figure 8.2 and see also Section 8.3.5). In terms of completion itself, various studies proposed different formulas to estimate the completion (Sunar, 2017). We use here the 80% threshold to define Active students in the following week as in (see Section 8.3.6).

We use the measures of student success, as defined in Section 8.3.4, to evaluate the clustering (as explained in Section 8.3.3) and ML prediction based on the proposed SDT theory and the Engage Taxonomy, as illustrated in Section 8.3.6. Figure 8.2 shows the input values (SDT elements from Students' activities in the first week) and output results from both the Success measures from clusters analysis and Active students prediction from Machine learning.

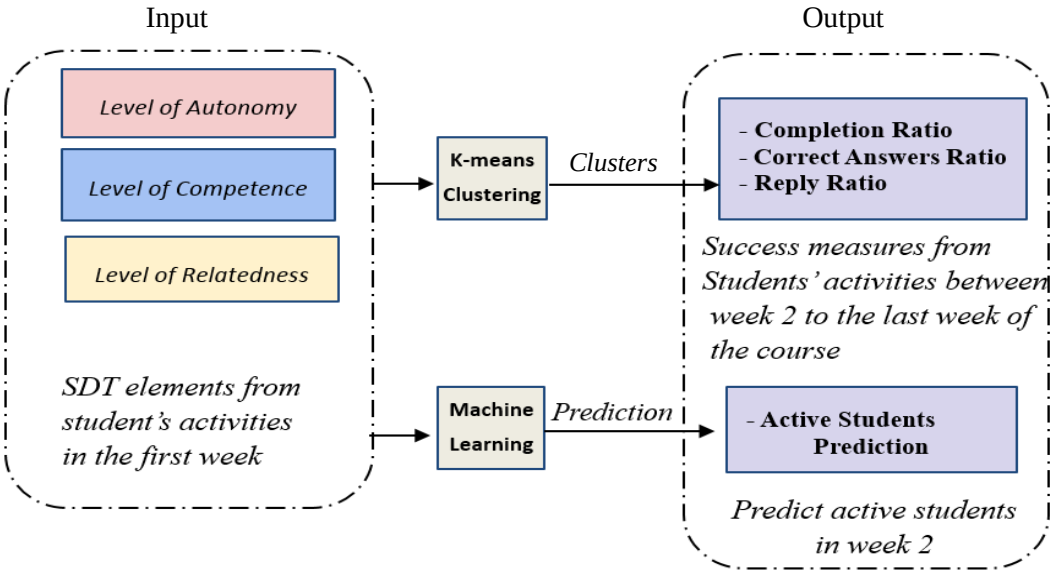


Figure 8.2: SDT constructs versus success measures

8.3 Methodology

8.3.1 Data Preparation and Pre-processing

We started with 6 courses with 26 runs with 2038 steps in total (an average of around 78.3 steps per run), and 136 quizzes (on average, about 5.2 quizzes per run). The courses are: Course 1, 'Ope n Innovation in Business (OI)'; Course 2, 'Leading and Managing People-Centred Change (LMPCC)'; Course 3, 'Babies in Mind (BIM)'; Course 4, ' Shakespeare (SHK)'; Course 5, 'Supply Chains (SUP)'; Course 6, 'The Mind is Flat (THM)' as shown in Table 8.2.

Table 8.2 FutureLearn courses' summary

Course	Enrolled	Accessed	Run
Open Innovation in Business (OI)	6071	2792	4
Leading and Managing People-Centred Change (LMPCC)	10417	6566	3
Babies in Mind (BIM)	48771	26175	6
Shakespeare (SHK)	63625	29432	4
Supply Chains (SUP)	5808	2912	2
The Mind is Flat (THM)	83543	39894	7
Total	218235	107,771	26

Originally, 218,235 students enrolled on these courses. The first step of the data preparation refined the raw data, by removing all students who had enrolled on one of the courses, but never viewed (accessed) any of the materials (steps). These students clearly never engaged and thus are irrelevant for our online course behaviour analysis with track data. As a result, we were left with analysing 107,771 students, which is still a large number, to extract behavioural patterns from (tracking data as above).

8.3.2 Expert Mapping

To map the features extracted (indicators) from MOOCs onto the set of chosen popular Engagement Theories and, importantly for this study, the SDT theory, we used expert labelling. We put forward some requirements to ensure the annotation quality. Therefore, we selected annotators who held at least a PhD degree and were experts in the domain of learning analytics (LA). The mapping between the engagement and motivation concepts and the

potential indicators within MOOCs was done independently by the experts (two professors and one Post-doctoral research assistant). In terms of our experts being LA experts, this was considered necessary, as LA experts understand the need for labelling data for any kind of neural network-based automatic machine learning. Additionally, they also had the educational expertise to understand what motivates students (as per the target of this study). In terms of their knowledge of the purpose of the study, the mapping was done independently of, and at a stage prior to the evaluation study. To further increase the quality of the categorical labelling, in the case where two experts disagreed on mapping a specific behaviour onto the theory's constructs, the mapping from the third expert was considered to determine the decision. Moreover, the inter-rater Fleiss' Kappa agreement test has been used to assess the inter-rater agreement between experts' mapping. The test resulted in $k = 0.72$, which is interpreted as a substantial agreement (Fleiss et al., 1981). The Engage Taxonomy constructed thus is described in detail in Section 8.2. Please note that, beside the efforts taken as described here to ensure the quality of the process, further validation of the experts' mapping is indirectly provided by measuring the success of students based on their SDT values (as further explained in Sections 8.3.4, 8.3.5 and 8.3.6, and evaluated in Sections 8.4.2 and 8.4.3). Additionally, the stronger, established and effective 'gold standard' measurement of criterion validity (Amirkhan, 1994) is provided by calculating the correlation between the results of the mapping and the results of the criterion measurement (here, the success measurement, as introduced in Section 8.3.4 and with results in Sections 8.4.2). The usage of the resulting 3 SDT constructs in practice for a prediction of active students in the following week provides further 'in-practice' proof of the usefulness of the SDT mapping (see Figure 8.2 and Sections 8.3.6 and 8.4.3). However, whilst we took great care with all steps in our and innovative process of evaluating motivational theories via data-driven approaches, and have had promising results (see Section 8.4), we do not claim each step is optimal; indeed, this process illustrated here is provided to the research community to further improve upon and explore, as also discussed in Section 8.5.

8.3.3 Clustering Students

As we wish to explore commonalities of students, Clustering seems like the most appropriate technique to employ first (Rana and Garg, 2016). In terms of student clustering, we analyse data at the level of individual students and all students for all runs of a course.

We have applied K-means clustering, K-Means clustering technique is an unsupervised machine learning algorithm and one of the most popular clustering techniques in data analytics (Xu and Wunsch, 2005). Previous research used this technique for related tasks - e.g., (Moreno-Marcos, 2018a) used sentiment analysis and k-means clustering to analyse discussion patterns on FutureLearn. K-means produces a pre-specified number k of clusters. To find the optimal k, we used the means silhouette coefficient, i.e., running clustering on a range of values of k (2 ~ 10, in our case). Thus, we used it to partition students based on their behaviour.

We use *raw data* and *aggregate data*, i.e. data composed from different raw data sources. We use data generated with various techniques: e.g., generated by 'simple' tracking of students, by applying motivational theories, by applying sentiment analysis on student information exchange, etc. Considering the multiple sources and complexity of the data processed, and to limit it somewhat for the current research, we have decided to perform a first aggregation step based on the weekly learning unit, which is used as a synchronisation point in instructor-led FutureLearn courses.

This approach further allows for early prediction (see Section 8.3.6) – starting by analysing clustering in *week 1*. Additionally, we ensure that tracking data covers all aspects of the motivational theories involved (see Section 8.2) - especially the SDT theory, which is studied here, as being the most widely used one (see Section 8.2.3).

8.3.4 Student Success Measure Definitions

Although a considerable amount of literature has been published on student success in a MOOC, there is no formal definition of student success. The concept of MOOC success is multidimensional and the researchers in the domain have been using a variety of definitions such as Course completion, Pass/Fail, certificate earners and final exam grade (Gardner and Brooks, 2018).

To measure student success in MOOCs, in the clusters we identify as explained in Section 8.3.3, we use an extended set of parameters (besides the 'basic' Completion ratio), as proposed by (Shi et al., 2020) (as explained at a generic level in Section 8.2.5).

- *Completion ratio*: this is the most often used success measure in learning in general and in online learning in particular: did the student complete the course? Here, instead of obtaining a binary value based on various criteria, we instead use the actual (normalised) proportion as a target; we normalised the *Completion rate* for each student by dividing the number of completed steps by the total course steps available, which had the effect of scaling all scores between 0 and 1.
- Correct Answer ratio: often, completion is not sufficient for estimating the success of a student. (Shi et al., 2020) have thus proposed to use other measures on a different 'axis', that of quizzes, and to explore how many answers have been correctly answered by a student, from all answers delivered by that student during the same period.
- *Reply ratio*: similarly, the social activity of a student may be considered another type of measure of success, which is here represented by the number of replies a student receives for their comments (Shi et al., 2020).

8.3.5 Analysing Student Clusters

To analyse the student subpopulations in the clusters obtained, we performed statistical analysis of the input parameters: the SDT constructs (Autonomy, Competence and Relatedness), defined via the Engage Taxonomy, see Section 9.3.1 In addition, the *success measures* (Completion ratio, Correct Answer ratio and Reply ratio, see Section 8.4.2) have been analysed in each cluster. For this purpose, we computed the mean and standard deviation of these parameters. Additionally, the highest two clusters in terms of the mean of the success measures (Completion Ratio, Correct Answers Ratio and Reply Ratio) are further compared pairwise via a Mann-Whitney U test, for each of the success measures. Moreover, the Pearson correlation coefficient test (PCC) (Benesty et al., 2009) has been employed to assess the relation between *SDT constructs* and *success measures*.

8.3.6 Machine Learning Prediction

To illustrate the power of the extracted SDT constructs, we evaluate if they can be used directly as early predictors of student activity. For this, we define *Active* students as those that did *access* 80% of the course material (80%=0.8 of the number of *#steps*; see also Section 8.2.5), and the rest as *Non-Active* students (Eq (8.4)):

$$\mathrm{NA(s,w)} = \begin{cases} 1, if\ TAS(s,w) < TS(w) * 0.8 \\ 0, rest \end{cases} \quad (8.4)$$

$$\boldsymbol{TAS(s,w)} = \sum_{\boldsymbol{j=0..TS(w)}} \boldsymbol{AS(s,w,j)}$$

$$\boldsymbol{AS(s,w,j)} = \begin{cases} \mathbf{1}, \boldsymbol{if\ student\ s\ accessed\ step\ j\ in\ week\ w} \\ \mathbf{0}, \boldsymbol{rest} \end{cases}$$

s: student, TAS(s,w): total steps accessed by student s in week w; TS(w): total course steps available in week w; AS(s,w,j): step j accessed by a student.Where TAS(s,w)<= TS(w), as the maximum number of steps a student could access in week w are all available ones.

As we aim at early prediction, we use machine learning to predict *Non-Active* Students (*NA*) in week 2, by using SDT constructs (defined via the Engage Taxonomy, see Section 8.2) as input, which were extracted from week 1. Thus our prediction problems is:

Given a student s, and student's SDT constructs from the current week ($\mathbf{week}\ \boldsymbol{w} = \mathbf{1}$) predict if the same student s is non-active (NA(s,w+1)) in the following week ($\mathbf{i.e.,}\ \mathbf{week}\ \boldsymbol{w} + \mathbf{1} = \mathbf{2}$).

For a comprehensive analysis, we employed several competing ML ensembles methods, as follows: Random Forest (RF) (Breiman, 2001), Gradient Boosting Machine (Gradient Boosting) (Friedman, 2001), Adaptive Boosting (AdaBoost) (Freund and Schapire, 1997), XGBoost (Chen and Guestrin, 2016), (ExtraTrees) (Geurts et al., 2006), Logistic Regression (LR) (Rawlings et al., 1998) and (K-nearest Neighbour) (Anchalia and Roy, 2014).

The current study used a balanced accuracy score (BA) to evaluate the performance of the models; this metric is widely used to calculate accuracy for imbalanced datasets, by preventing the majority of negative samples from biasing the result (Brodersen et al., 2010). Please note that, although we applied and compared various classifiers, our aim here was not to optimise the prediction of Active students in week 2, but to showcase how the SDT theory, and our mapping of indicators onto SDT constructs, can be used directly as a predictor.

As we have used a massive dataset for different courses, we have prepared the training and testing sets based on the last Run of the course. For example, in the Mind Is Flat course, we trained our ML models by using students' data extracted from early runs (Runs 1 to 6) for students who registered between 2013 to 2016. However, for testing the models, we used a new data set extracted from the last run (Run 7) that contains students' activities in 2017- see - which is similar to some extent to Transfer learning models (Getoor, 2020, Bote-Lorenzo, 2018).

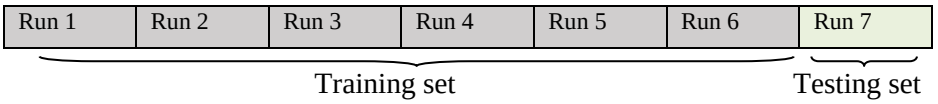


Figure 8.3 The Mind Is Flat course (Training and Testing set)

In addition, we combined all datasets (Runs) together for each course to predict Active and Non-Active Students in week 2 by using the 10-fold cross-validation, a widely used technique to evaluate a predictive models (An et al., 2007).

8.4 Results

8.4.1 Student Clusters

The indicators are aggregated, to obtain 6 datasets corresponding to the 6 courses. Further aggregation would not be applicable, as the structure of the courses varied in length, number of steps, quizzes, resources, etc., available, whereas within each course, these were (relatively) constant, thus progress would have been expected to be equal for each student - all other parameters being equal. As it is difficult to compare data from different courses, we normalised the indicators, by dividing each value by the highest value in the column (activity) within each course, which had the effect of scaling all scores between 0 and 1.

We first clustered the students in the 6 courses for the 26 runs and obtained the main clusters for students, based on the SDT variables from students' activities in the first week. The silhouette coefficient analysis showed that $k=3$ for K-Means is the most appropriate, when clustering the behavioural indicators. Hence, we obtain 3 clusters (Table 8.3).

Table 8.3 Number of students and percentages in each cluster for each course

Course	Cluster 1	Cluster 2	Cluster 3	Total
BIM	*2055 (8%)*	*10155 (39%)*	*13965 (53%)*	*26175*
SHK	*1454 (5%)*	*12247 (42%)*	*15731 (53%)*	*29432*
SUP	*96 (3%)*	*631 (22%)*	*2185 (75%)*	*2912*
THM	*1410 (3%)*	*12522 (31%)*	*25962 (65%)*	*39894*
OI	*164 (6%)*	*786 (28%)*	*1842(66%)*	*2792*
LMPCC	*314 (5%)*	*2255 (34%)*	*3997 (61%)*	*6566*

Table 8.3 shows an overview of students' distribution in each cluster. As can be seen from the table, cluster 3 contains the majority of the students: more than the other two clusters for all six courses. Approximately two-thirds of the students have been clustered in cluster 3 in Supply Chains and 66% of the students have been grouped in cluster 3 in Open Innovation in Business (OI). On the other hand, cluster 1 comprises the minority of the students (3%-8% of the students) and cluster 2 is mid-ranged (28% and 42%). Table 8.4 shows the number and percentage of the outlier students for each SDT element in each cluster.

Table 8.4 Number and percentage of the outlier students for each SDT element in each cluster

Course	***Cluster***	***AUT***	***COM***	***REL***
BIM	*Cluster 1*	*N=53-%2.5*	*N=127-%6.18*	*N=92-%4.4*
	Cluster 2	*N=1902-%18.7*	*N=439-%4.32*	*N=599-%5.8*
	Cluster 3	*N=0-%0.0*	*N=496-%3.55*	*N=484-%3.4*
SHK	*Cluster 1*	*N=18-%1.2*	*N=44-%3.02*	*N=47-%3.2*
	Cluster 2	*N=829-%6.7*	*N=579-%4.72*	*N=2750-%22.4*
	Cluster 3	*N=86-%0.5*	*N=573-%3.64*	*N=418-%2.6*
SUP	*Cluster 1*	*N=0-%0.0*	*N=3-%3.12*	*N=6-%6.2*
	Cluster 2	*N=68-%10.7*	*N=40-%6.33*	*N=120-%19.0*
	Cluster 3	*N=239-%10.9*	*N=57-%2.60*	*N=114-%5.2*
THM	*Cluster 1*	*N=14-%0.9*	*N=27-%1.91*	*N=45-%3.1*
	Cluster 2	*N=1221-%9.7*	*N=953-%7.61*	*N=2579-%20.5*
	Cluster 3	*N=191-%0.7*	*N=1069-%4.11*	*N=897-%3.4*
OI	*Cluster 1*	*N=0-%0.0*	*N=6-%3*	*N=4-%0.2.4*
	Cluster 2	*N=102-%13*	*N=22-%3*	*N=88-%11*
	Cluster 3	*N=0-%0.0*	*N=41-%2.2*	*N=44-%2.3*
LMPCC	*Cluster 1*	*N=4-%1.2*	*N=10-%3.18*	*N=15-%4.7*
	Cluster 2	*N=220-%9.7*	*N=104-%4.61*	*N=281-%12.4*
	Cluster 3	*N=227-%5.6*	*N=69-%1.72*	*N=76-%1.9*

Outliers: Values greater than the third quartile (Upper Bound (Q3 +(1.5 * IQR)) or less than the first quartile (Lower Bound (Q1 - (1.5 * IQR)) are considered outliers. Where Q1 is the middle number between the lowest and the median values of the dataset, Q3 is the middle number between the median and the maximum value in the dataset(Cho et al., 2008).

8.4.2 Student Cluster Analysis

To semantically analyse the clusters, Table 8.5 illustrates the three elements of SDT extracted from week 1 activities only, versus the success measures of student activities from week 2 to the last week (with highest values in bold). This cluster analysis thus allows us to estimate if the SDT values of week 1 would be a good predictor for success in the rest of the course.

Table 8.5 Mean and Standard Deviation for the 3 SDT construct-based clusters aggregated over the 6 courses, versus the success measures (highlighted in green)

Course			***Cluster 1***	***Cluster 2***	***Cluster 3***
			Mean/SD	*Mean/SD*	*Mean/SD*
Babies in Mind(BIM)	***SDT elements*** *from Students' activities in the first week*	*Autonomy*	***0.37 / 0.08***	*0.19 / 0.04*	*0.04 / 0.04*
		Competence	***0.2 / 0.09***	*0.09 / 0.03*	*0.0 / 0.01*
		Relatedness	***0.16 / 0.08***	*0.02 / 0.03*	*0.0 / 0.0*
	Success measures *from Students' activities between week 2 to the last week*	*Completion Ratio*	***0.55 / 0.4***	*0.40 / 0.4*	*0.02 / 0.09*
		Correct Answers Ratio	***0.39 / 0.36***	*0.27 / 0.34*	*0.01 / 0.05*
		Reply Ratio	***0.02 / 0.06***	*0.0 / 0.01*	*0.0 / 0.0*
Shakespeare(SHK)	***SDT elements*** *from Students' activities in the first week*	*Autonomy*	***0.39 / 0.07***	*0.22 / 0.03*	*0.04 / 0.04*
		Competence	***0.35 / 0.09***	*0.20 / 0.03*	*0.0 / 0.01*
		Relatedness	***0.19 / 0.09***	*0.01 / 0.03*	*0.0 / 0.01*
	Success measures *from Students' activities between week 2 to the last week*	*Completion Ratio*	***0.62 / 0.38***	*0.44 / 0.39*	*0.02 / 0.09*
		Correct Answers Ratio	***0.45 / 0.31***	*0.31 / 0.31*	*0.01 / 0.05*
		Reply Ratio	***0.02 / 0.07***	*0.0 / 0.01*	*0.0 / 0.0*
Supply Chains(SUP)	***SDT elements*** *from Students' activities in the first week*	*Autonomy*	***0.48 / 0.09***	*0.23 / 0.04*	*0.04 / 0.04*
		Competence	***0.33 / 0.12***	*0.19 / 0.05*	*0.0 / 0.01*
		Relatedness	***0.23 / 0.11***	*0.02 / 0.04*	*0.0 / 0.01*
	Success measures *from Students' activities between week 2 to the last week*	*Completion Ratio*	***0.63 / 0.43***	*0.49 / 0.42*	*0.03 / 0.13*
		Correct Answers Ratio	***0.4 / 0.33***	*0.32 / 0.31*	*0.0 / 0.02*
		Reply Ratio	***0.04 / 0.11***	*0.0 / 0.03*	*0.0 / 0.02*
The Mind is Flat(THM)	***SDT elements*** *from Students' activities in the first week*	*Autonomy*	***0.4 / 0.08***	*0.21 / 0.04*	*0.05 / 0.04*
		Competence	***0.34 / 0.1***	*0.18 / 0.03*	*0.0 / 0.01*
		Relatedness	***0.21 / 0.1***	*0.01 / 0.03*	*0.0 / 0.01*
	Success measures *from Students' activities between week 2 to the last week*	*Completion Ratio*	***0.56 / 0.36***	*0.45 / 0.37*	*0.01 / 0.08*
		Correct Answers Ratio	***0.41 / 0.3***	*0.33 / 0.3*	*0.0 / 0.04*
		Reply Ratio	***0.03 / 0.06***	*0.0 / 0.01*	*0.0 / 0.0*
Open Innovation in Business (OI)	***SDT elements*** *from Students' activities in the first week*	*Autonomy*	***0.37 / 0.08***	*0.18 / 0.04*	*0.04 / 0.03*
		Competence	***0.26 / 0.09***	*0.07 / 0.04*	*0.0 / 0.01*
		Relatedness	***0.24 / 0.09***	*0.02 / 0.04*	*0.0 / 0.01*
	Success measures *from Students' activities between week 2 to the last week*	*Completion Ratio*	***0.55 / 0.39***	*0.45 / 0.41*	*0.01 / 0.04*
		Correct Answers Ratio	***0.27 / 0.23***	*0.19 / 0.22*	*0.0 / 0.02*
		Reply Ratio	***0.02 / 0.11***	*0.0 / 0.0*	*0.0 / 0.0*

Leading and Managing People-Centred Change (LMPCC)	***SDT elements*** *from Students' activities in the first week*	*Autonomy*	***0.35 / 0.07***	*0.18 / 0.02*	*0.04 / 0.04*
		Competence	***0.2 / 0.09***	*0.1 / 0.03*	*0.0 / 0.01*
		Relatedness	***0.16 / 0.08***	*0.01 / 0.02*	*0.0 / 0.0*
	Success measures *from Students' activities between week 2 to the last week*	*Completion Ratio*	***0.68 / 0.37***	*0.53 / 0.38*	*0.02 / 0.12*
		Correct Answers Ratio	***0.4 / 0.24***	*0.33 / 0.28*	*0.0 / 0.05*
		Reply Ratio	***0.02 / 0.1***	*0.0 / 0.01*	*0.0 / 0.0*

The most striking result to emerge from the table is that there is a clear positive correlation between the SDT constructs (Autonomy, Competence and Relatedness) and the success measures (Completion Ratio, Correct Answers Ratio and Reply Ratio). This has further been proven statistically by using the Pearson correlation coefficient test. The results revealed a positive correlation (r >0), as shown in Table 8.6. The table shows that the Relatedness construct is the most correlated construct, strongly correlated with the Reply-Ratio measure in the six courses, whereas Autonomy and Competence constructs are less correlated. On the other hand, the Competence construct, in the Shakespeare, LMPCC, Supply Chains and THM courses, is the most correlated construct with the Answer Ratio measure, and the Autonomy is the most correlated construct with the Completion Ratio measure in four courses.

Table 8.6 Correlation between the SDT constructs and the success measures over the 6 courses

		Completion Rate	*Answer Rate*	*Reply Rate*
SHK	*Autonomy*	*0.60*	*0.56*	*0.17*
	Competence	*0.61*	*0.59*	*0.22*
	Relatedness	*0.29*	*0.29*	*0.36*
BIM	*Autonomy*	*0.57*	*0.50*	*0.24*
	Competence	*0.54*	*0.49*	*0.31*
	Relatedness	*0.30*	*0.29*	*0.32*
LMPCC	*Autonomy*	*0.68*	*0.62*	*0.15*
	Competence	*0.66*	*0.63*	*0.21*
	Relatedness	*0.26*	*0.23*	*0.30*
OI	*Autonomy*	*0.62*	*0.57*	*0.14*
	Competence	*0.57*	*0.54*	*0.24*
	Relatedness	*0.29*	*0.29*	*0.25*
SUP	*Autonomy*	*0.64*	*0.59*	*0.18*
	Competence	*0.63*	*0.65*	*0.18*
	Relatedness	*0.28*	*0.26*	*0.25*
TMF	*Autonomy*	*0.64*	*0.60*	*0.26*
	Competence	*0.66*	*0.64*	*0.29*
	Relatedness	*0.26*	*0.25*	*0.40*

Please note that this is even more important, as SDT are measured as early variables, potentially usable for prediction (week 1, as said) and success measures are collected week

2 till the last week. Thus, we can arguably claim that SDT motivational theory constructs can be used as early informer for success towards the end of the course. Therefore, it is an important result that we could confirm, via a data-intensive approach, that most motivated and engaged students, as defined by the SDT motivational theory, turn out to be the most successful. Interestingly, our clustering succeeded to showcase this, by grouping these students in cluster 1. Likewise, cluster 2 has naturally resulted in assembling the intermediate students, who have statistically significantly ($p<0.05$) lower results in terms of both SDT constructs as well as success (see Table 8.7). Cluster 3 gathers, on the other hand, a large number of users who are not very engaged, as per SDT parameters (which is not that surprising, considering clustering was done based on SDT parameters). A good example of this can be found in the 'Babies in Mind' course, as the mean of cluster 1 was 0.37 in Autonomy, 0.2 in Competence and 0.16 in Relatedness; the Completion rate was 0.55. On the other hand, cluster 2 reported lower values than cluster 1 in all SDT parameters (0.19 in Autonomy, 0.09 in Competence and 0.02 in Relatedness; with a Completion rate of 0.4 (0.15 less than cluster 1)). We can notice the same pattern for all other courses (see Table 8.5).

Table 8.7 Statistical significance analysis ($p<.05$) of the difference between the highest two clusters (cluster1 vs cluster2)

Course	Completion Ratio	Correct Answers Ratio	Reply Ratio
BIM	*4.30947e-54*	*1.97756e-46*	*0.0*
SHK	*1.35573e-69*	*7.74394e-53*	*0.0*
SUP	*0.004673*	*0.02695*	*2.45989e-30*
THM	*1.34367e-24*	*1.54078e-23*	*0.0*
OI	*6.440039e-4*	*5.668406e-06*	*8.801509e-63*
LMPCC	*4.544427e-15*	*0.000159*	*9.613372e-74*

Table 8.8 further shows the mean, standard deviation and maximum values for the three SDT construct over six courses. For all students, the Autonomy construct has the highest mean score (ranging from 0.133 to 0.096). The autonomy mean score in Table 8.5 for students in cluster 1 (ranging from 0.48 and 0.35) represents a high degree of autonomy, compared to students in cluster 2 and 3. The Competence construct ranked as the second highest, with a mean score ranging from 0.102 to 0.036. Finally, the Relatedness construct had the lowest mean score (from 0.021 to 0.01).

Table 8.8 Mean, Standard Deviation and maximum value for the 3 SDT constructs for all students

		Autonomy	Competence	Relatedness
SHK	*Mean*	*0.13*	*0.10*	*0.01*
	Std	*0.11*	*0.11*	*0.04*
	Max	*0.73*	*0.74*	*0.70*
BIM	*Mean*	*0.13*	*0.04*	*0.01*
	Std	*0.10*	*0.06*	*0.05*
	Max	*0.67*	*0.62*	*0.70*
LMPCC	*Mean*	*0.10*	*0.04*	*0.01*
	Std	*0.09*	*0.06*	*0.04*
	Max	*0.66*	*0.66*	*0.75*
OI	*Mean*	*0.09*	*0.03*	*0.02*
	Std	*0.10*	*0.07*	*0.06*
	Max	*0.56*	*0.59*	*0.61*
SUP	*Mean*	*0.09*	*0.05*	*0.01*
	Std	*0.11*	*0.10*	*0.05*
	Max	*0.74*	*0.73*	*0.68*
TMF	*Mean*	*0.11*	*0.06*	*0.01*
	Std	*0.10*	*0.10*	*0.04*
	Max	*0.82*	*0.79*	*0.82*

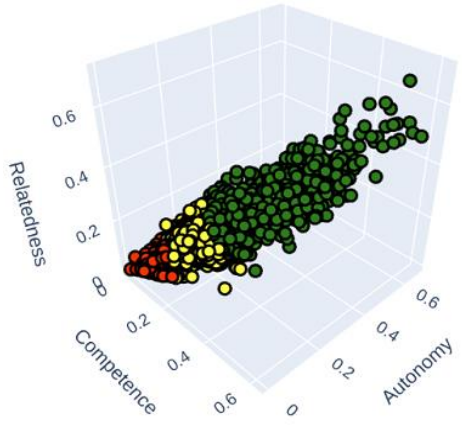
A follow-up analysis additionally shows that there is a significantly high correlation between Autonomy and Competence (ranging from 86 to 93), and a lower correlation between Relatedness and both Autonomy and Competence (ranging from 57 to 84) over the six courses (see Table 8.9). Section 8.5 further discusses these findings.

Table 8.9 Correlation between SDT constructs (Autonomy, Competence and Relatedness) over the 6 courses

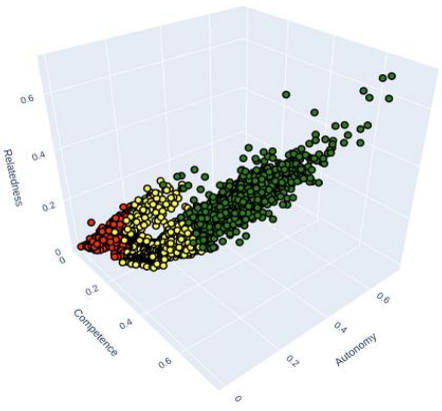
		Autonomy	Competence	Relatedness
SHK	*Autonomy*	-	*0.93*	*0.60*
	Competence	*0.93*	-	*0.57*
	Relatedness	*0.60*	*0.57*	-
BIM	*Autonomy*	-	*0.87*	*0.73*
	Competence	*0.87*	-	*0.75*
	Relatedness	*0.73*	*0.75*	-
LMPCC	*Auton omy*	-	*0.89*	*0.61*
	Competence	*0.89*	-	*0.62*
	Relatedness	*0.61*	*0.62*	-
OI	*Autonomy*	-	*0.86*	*0.75*
	Competence	*0.86*	-	*0.84*
	Relatedness	*0.75*	*0.84*	-
SUP	*Autonomy*	-	*0.87*	*0.69*
	Competence	*0.87*	-	*0.56*
	Relatedness	*0.69*	*0.56*	-
TMF	*Autonomy*	-	*0.92*	*0.62*
	Competence	*0.92*	-	*0.58*
	Relatedness	*0.62*	*0.58*	-

As previously mentioned, the clusters were created based on the SDT constructs (Autonomy, Competence and Relatedness) from students' activities in the first week (analysing basically if these newly proposed early engagement parameters would be good potential early predictors of success). Table 8.7 shows the results of the statistical analysis of the success measures for the highest two clusters in terms of SDT (cluster 1 versus cluster 2). We can see that a significant difference exists for all success measures between the highest two clusters for all six courses, meaning their differences are not due to chance.

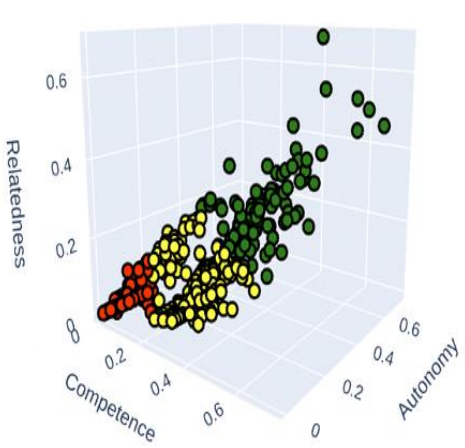
Figure 8.4 (a-f) allows for further visual analysis of the results, by showing the 3D-plots of the 6 courses, the relevance of SDT constructs being clearly visualised in the plots for each cluster. Clusters are well separated, with cluster 1 containing the higher SDT values (in green), cluster 2 the intermediate (in yellow) and cluster 3 the low values (in red). Cluster 1 contains the most motivated and engaged students whereas, cluster 3 identified the very low engagement students. Please note that cluster 3 contains students with very low SDT values and their data points are very close to each other. The visualisation for the students with similar SDT values are overlapping giving a 'feel' that the red cluster is smaller, whereas it is the largest.



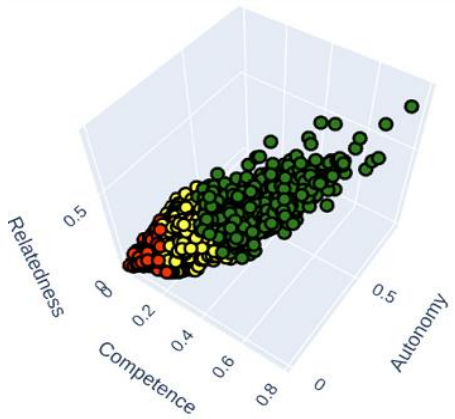
a) Babies in Mind (BIM)



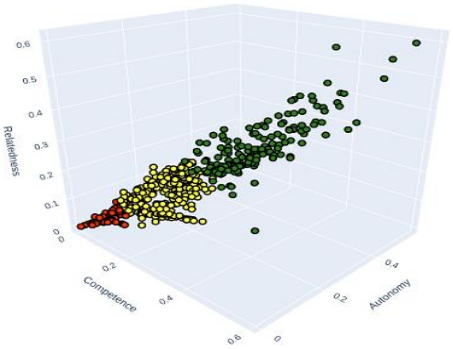
b) Shakespeare and His World (SHK)



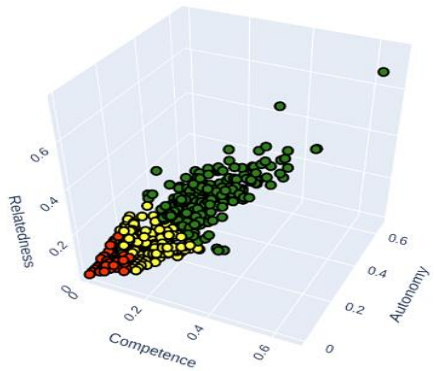
c) Supply Chains (SUP)



d) The Mind is Flat (THM)



e) Open Innovation in Business (OI)



f) Leading and Managing People-Centred Change (LMPCC)

Figure 8.4 (a-f) The three clusters mapped onto the Self-Determination Theory (SDT) (Cluster 1: green; Cluster 2: yellow; Cluster 3: red)

To better understand the usefulness of the SDT constructs with respect to the impact on the success measures, the centroids of each cluster are further represented as a point in the radar plot in Figure 8.5 (a-f); clusters with the same colouring convention as in Figure 8.4. It is clear that the students with higher values of the SDT features in the first week activities have a higher chance to be the most successful. Indeed, the wider spread of the green cluster 1 area shows for all SDT values (Autonomy, Competence and Relatedness) show a respective widespread for the success values (Answer Rate and Completion Rate). Similarly, the yellow cluster 2 is wider for all these when compared with the red cluster 3, for which all these values (SDT and success) are so low, it almost appears as a small dot.

a) ***Babies in Mind*** *b)* ***Shakespeare***

c) ***Supply Chains*** *d)* ***The Mind is Flat***

e) ***Open Innovation in Business (OI)*** *f)* ***LMPCC***

Figure 8.5 (a-f) Values of the SDT features versus success measures (Completion Ratio and Correct Answers Ratio (Cluster 1: green; Cluster 2: yellow; Cluster 3: red)

We have repeated the experiment for one course "The mind is flat" using only one dimension of the SDT constructs ("Relatedness") and compare it with the results obtained by using all SDT constructs. The clusters with one construct look worse than using all SDT constructs. The results showed that 93% of students were clustered in cluster 3 (was 65% when we used all SDT constructs), which means that a lot of high achievers' students in cluster 1 and 2 moved to cluster 3. Therefore, the mean completion rate for Cluster 3 increased from 0.01 to 0.15 please see Table 8.10.

Table 8.10 k-means clustering for The Mind is Flat course using only one dimension of the SDT constructs ("Relatedness")

			Cluster 1	***Cluster 2***	***Cluster 3***
The Mind is Flat (THM)	*Relatedness from Students' activities in the first week*		*Mean/SD*	*Mean/SD*	*Mean/SD*
			0.21 / 0.1	*0.01 / 0.03*	*0.0 / 0.01*
	Success measures *from Students' activities between week 2 to the last week*	*Completion Ratio*	***0.55 / 0.37***	*0.43 / 0.39*	*0.15 / 0.29*
		Correct Answers Ratio	***0.41 / 0.31***	*0.31 / 0.31*	*0.1 / 0.22*
		Reply Ratio	***0.04 / 0.08***	*0.01 / 0.03*	*0.0 / 0.0*

8.4.3 Machine learning prediction

Table 8.11 shows the performance of the predictive models, evaluated by the Balanced Accuracy score (Brodersen et al., 2010), a commonly used metric for binary classification of unbalanced datasets (see Section 8.3.6). Moreover, several measures, such as Precision, Recall and F1-Score have been also used to evaluate the prediction performance (full results provided in Appendix A, due to the extensive size of the table). In addition, Appendix B shows more results for using 10 fold cross-validation by combining all datasets (Runs) together for each course to predict Active and Non-Active Students in week 2.

In general, all algorithms achieved good results, indicating that, regardless of the employed model, the SDT constructs extracted from the first week in this study proved to be powerful in predicting Active and Non-Active students in the second week. Whilst all models' performances are generally relatively good; the most robust model is the ExtraTrees model,

as it outperforms in two courses: 'The Mind is Flat' 91.70%, and 'Leading and Managing People-Centred Change (LMPCC)' 90.13%.

Table 8.11 Prediction of Active and Non-Active Students in week 2 based on week 1 SDT constructs

Courses	AdaB	ExtrTr	GBoost	KNN	LR	RF	XGBoost
BIM	*82.87%*	*80.33%*	*62.48%*	*75.13%*	*83.01%*	*82.87%*	*67.50%*
SHK	*87.11%*	*87.29%*	*87.33%*	*80.57%*	*87.24%*	*87.26%*	*87.24%*
SUP	*91.35%*	*92.46%*	*92.73%*	*83.74%*	*90.74%*	*92.11%*	*92.73%*
THM	*91.11%*	*91.70%*	*91.16%*	*86.54%*	*91.31%*	*91.65%*	*91.17%*
OI	*90.09%*	*89.75%*	*84.70%*	*87.39%*	*83.35%*	*88.07%*	*89.08%*
LMPCC	*89.86%*	*90.13%*	*89.63%*	*87.12%*	*88.18%*	*89.57%*	*89.32%*

Thus, we have computed the Gini index (GI) (Dorfman, 1979) for the 'winning' ExtraTrees algorithm, to evaluate the feature importance of each feature used to predict non-active students in the following week (see Table 8.12). Briefly, results show that Competence is ranked as the most important construct in the classification (importance value ranging from 0.43 and 0.50). Autonomy is ranked as the second important construct (importance value ranging from 0.27 to 0.34). Finally, the Relatedness construct is ranked as the least important factor (importance value ranging from 0.20 to 0.23).

Table 8.12 The constructs importance values for the ExtraTrees algorithm

	Autonomy	Competence	Relatedness
SHK	0.31	0.48	0.20
BIM	0.34	0.43	0.23
LMPCC	0.29	0.48	0.23
OI	0.33	0.45	0.22
SUP	0.32	0.47	0.21
THM	0.27	0.50	0.22

Table 8.13 shows the prediction results of Active and Non-Active Students in week two based on one of the SDT constructs (the Relatedness); the prediction accuracies were considerably lower (between 50%-63%) compared to using all SDT constructs (between 67%- 92% (see Table 8.11)).

Table 8.13 Prediction of Active and Non-Active Students in week 2 based on one of SDT construct (Relatedness); evaluated by the Balanced Accuracy score

Courses	AdaB	ExtrTr	GBoost	KNN	LR	RF	XGBoost
BIM	50.00%	58.21%	55.53%	55.22%	58.85%	60.47%	55.45%
SHK	63.00%	61.44%	62.94%	61.89%	62.43%	63.00%	63.00%
SUP	55.26%	56.97%	55.77%	56.22%	57.13%	57.12%	54.90%
THM	59.81%	59.70%	59.51%	58.60%	61.55%	59.70%	59.75%
OI	50.68%	56.07%	54.55%	51.86%	56.24%	56.24%	51.35%
LMPCC	57.27%	56.40%	55.95%	55.68%	58.62%	56.38%	55.65%

8.5 Discussion

In this study, we propose the *Engage Taxonomy*, a mapping of 'bottom-up' MOOC data to 'top-down' high level concepts from motivational theories. To create it, we have mapped the engagement and motivation concepts and the potential indicators within MOOCs, with the help of three experts, onto these theories. Moreover, the Fleiss' Kappa agreement test has shown a high rate of agreement (substantial agreement) between experts (see Section 8.3.2 and Table 8.1). However, this process illustrated here is provided to researchers to further improve upon and explore. Indeed, a similar mapping can be extended to incorporate further MOOC variables, if other MOOCs provide additional behavioural meta-data. Additionally, our weighted model as in Section 8.2.4 can be further improved: weight optimisation sought - e.g. searching optimal values between [0,1]; or even proposing negative values. For instance, negative replies are considered here to influence positively the relatedness (in the sense of 'any news is good news', and any interaction and replies is affecting the relatedness). However, another model may consider this relation as a negative one. Our methodology can be seen as a shell to be applied to different MOOCs, or onto different motivational theories (as already started in Table 8.1, where we have not just SDT, explored further in-depth in this research, but also the Drive Theory, Engagement Theory and Process of Engagement Theory).

In terms of MOOC completion, although about 10% of participants complete their courses in MOOC, clearly many more students quit over time. Figure 7.2 presents the number of remaining students over time, showing that the curve of dropouts is itself dropping speedily. Hence, participants are most likely to drop out in the first few weeks. Therefore, identifying

those students at an early stage is important, to provide early intervention, to keep the engagement going. In this study, we used SDT constructs extracted from the first week as input features for machine learning. This provides an opportunity to deal with at-risk students at an early stage (week 2, which we identified as a critical period). Future work will explore how using week-by-week prediction affects the prediction accuracy and gain (in terms of number of students dropping out at later stages).

The correlation results between the SDT constructs and the success measures (Table 8.6) point to Relatedness being the best construct to measure the Reply Rate of students, as it shows higher correlation values. Both Autonomy and Competence constructs have similar correlation patterns with success measures. For the Completion Rate, the Autonomy was the most correlated construct in four courses and Competence was the highest in two courses. Finally, Competence was the most correlated construct to the Answer Rate in four courses.

This finding led us to further explore the direct correlation among the SDT constructs. The three SDT constructs are thought to represent different traits conceptually, so some independence is expected. This is confirmed by Figure 8.1, in that at least one feature uniquely maps to each SDT construct. However, the linear trend of the distribution of students in Figure 8.4 suggests that the Autonomy, Competence, and Relatedness dimensions from the expert mapping may be correlated. An additional analysis (Table 8.9), shows that a significantly high correlation between Autonomy and Competence, ranging between (.86-.93). Returning to Figure 8.1, this corresponds to experts allocating sometimes the same feature to more than one Engagement dimension. Whilst these findings correspond with the literature, which shows a positive correlation between Autonomy and Competence (Wangwongwiroj and Bumrabphan, 2021, Vlachopoulos and Michailidou, 2006, Qin, 2021, Gangire et al., 2021), data-driven approaches such as ours may be further explored to feed back to the theories and potential further improvement thereof.

Furthermore, it can be seen from the data in Table 8.5 that there is some consistency in the way the clusters appear, regardless of the course. As the bold, highest values for the means show, cluster 1, for all 6 analysed courses, tends to have the highest level of Autonomy, Competence and Relatedness (so SDT values), on one hand, as well as the best distribution of student success, taking into consideration the success measures from students' activities

(*Completion Ratio, Correct Answers Ratio* and *Reply Ratio)* between week 2 to the last week presented. We can observe that cluster 1 comprises the high achievers, cluster 2 contains the intermediate students, and cluster 3 comprises the students who probably end up dropping out. This is clearly seen in Figure 8.5 (a-f), which shows the centroids of each cluster as a point in the radar plot. In other words, the students with higher Autonomy, Competence and Relatedness in week 1 tend to be the students with higher success measures in later weeks.

Thus, we can say relatively confidently that our process extracts in cluster 1 the most motivated and engaged students, who also turn out to be the most successful. Interestingly, for all these parameters, the mean is statistically significantly higher than for the next best cluster ($p < .05$) see Table 8.7.

Cluster 3 gathers a large number of users who are not very engaged, as per SDT parameters, nor are they very successful (as per our three success measures). The mean silhouette coefficients in all courses range from 68 to 78, which shows that our early collected SDT features worked as expected.

Figure 8.4 further supports these findings and shows that our SDT features can be used to separate the students into three distinct clusters, with distinct success, as per Figure 8.5.

Our proposed SDT-based approach has thus been validated by leading to semantically relevant clusters. Indeed, we have clearly confirmed, with this method, facts known from literature (but from theory only, from small-scale studies, mainly using face-to-face data) via our *large-scale study on online data*, from different angles – such as the fact that engaged students have higher success (and have found them all as members of cluster 1). We have also identified the very low engagement students (cluster 3), who also were confirmed to be the least successful. Interestingly, we have identified via cluster 2 students who have some good results, but perhaps lower motivation. This is a very interesting find, because it may show students who would have the potential to complete, to succeed, but may fail, as being less motivated. Thus, some intervention towards motivating these students would have a better chance of an effect than on those in, e.g., cluster 3. Whereas cluster 1 students may come with intrinsic motivation and do not need much 'hand-holding'.

Analysing the outliers (see Table 8.4), we have noticed some nuances in the students' behaviours related to the SDT constructs and success measures. It turns out that, contrary to

the general trend, there are not only the high-achievers from cluster 1 who have higher Autonomy, Competency and Relatedness, the intermediate students from cluster 2 who have intermediate values of the SDT features and the non-completers from cluster 3 who have low values of the SDT features. In fact, there are also students who have intermediate or low Relatedness and were assigned to cluster 2 or cluster 3, but are high-achievers, according to the success measures. Such students belong to a group that are not engaged in participating in forums or commenting on the pedagogical materials, but are still committed to learning the course content and completing it. These trends can be seen for the other SDT features, that is, a student might not have a high Autonomy or Competence, but be a high- or intermediate achiever, as per our success measures. On the other hand, there are a few students who have high values for the SDT constructs, but do not achieve good results. Nevertheless, this is not the general pattern, but the exception.

We have two possible explanations about these outliers. First, we are collecting data from a very early stage (only the first week data) for each course and, hence, it is natural to have outliers, since student behaviour might change during the course. That is, a student might begin the course with positive attitudes and behaviours - however, end up failing or dropping out due to personal reasons or something that we cannot control. The opposite is possible as well, students might start with apparently bad attitudes and behaviours, e.g. due to personal problems, but may succeed in concluding the course, due some attitude or circumstances change after the first week.

The second potential reason is that we are dealing with big data and, thus, it is typical to find behaviours that do not follow the general trend. Indeed, this is a characteristic of the human-being, as reported by many authors (Hawkins, 1980, Rambo-Hernandez and Warne, 2015). This leads to the need of further adaptive systems in MOOCs, to consider such nuances of learning and engaging and how they can influence students' achievement. For further work, we can explore finding optimal subsets (or supersets) of features which would express the SDT (as currently we use the whole set of available features). We could also explore if there are essential features, which lead to a high drop in prediction power for the success variables, and optional features, which only lead to minimal increase in success.

It can be seen from the data in Table 8.11 that the SDT values (Autonomy, Relatedness and Competence) could be used directly, as good indicators for prediction to enable early interventions for students at-risk (*Non-Active student*) in the following week. Alternatively, different definitions of active students may lead to different results, and other predictions, such as completion of course, could be further attempted based on the SDT mapping proposed (Monllaó Olivé, 2020, Rawat, 2021, Alamri, 2019, Tóth, 2018, Kameas, 2021, Alamri, 2021); however, these are beyond the scope of this study. The seven classical machine learning algorithms achieved relatively good results. However, advanced data mining techniques, such as deep learning models have not been used in this study, due the low number of input features (three features), while deep machine learning models are used to find complex and hidden correlations in large input spaces and datasets (Wischmeyer and Rademacher, 2020).

Further to note in terms of the prediction, that whilst the Autonomy construct has the highest mean score (ranging from 0.133 to 0.096) compared to the Competence (ranging from 0.102 to 0.036) and Relatedness (ranging from 0.021 to 0.01) (see Table 8.8), the Competence construct is ranked as the most important construct in the classification to predict non-active students in the following week (see Table 8.12).

Indeed, mapping large-scale student behaviour onto motivational theories opens the way to inform student models and create appropriate pedagogical interventions to improve students' outcomes. For instance, students could be brought from cluster 3 to cluster 2, or the desirable cluster 1, by appropriate recommendations. This leads to further avenues of research, bringing together *measurable, data-driven metrics for engagement*, and *classical adaptive learning*.

Finally, however, any research on MOOCs and its engagement needs to note the caveat that MOOCs are not just for traditional students, and many working professionals use them to touch up on certain skills or to explore new areas of knowledge. Once that goal is accomplished, which may occur before the natural end of a registered course, these individuals may quit, having learned and met their goals. The balance between motivation and success would need to take this further into account. Indeed, conducting a pre-survey is one way to identify the students who do not intend to complete the whole course. However,

the response rates in MOOCs are generally lower for surveys. For example in (Mihalec-Adkins et al., 2016), only 1,624 completed responses, out of 22,000 students who were enrolled. Therefore, it is likely that MOOC statistics derived from surveys with low response rates would not accurately reflect the real population. Other methods may need to be devised to extract these ‘hidden agendas’.

Epilogue

This chapter has proposed mapping between the tracking parameters and four of the most used theories related to engagement in digital systems, generating the engage taxonomy. Finally, we showed how such a mapping can be put into practice by analysing the engaged and disengaged MOOC student behaviours in relation to SDT theory. The following chapter discusses the contributions of the studies presented in this thesis and their limitations.

Chapter 9 : Discussion

Prologue

This chapter discusses the contributions and overall findings of the studies presented in this thesis (Chapter 5, Chapter 6, Chapter 7, and Chapter 8). In addition, we will discuss the limitations of this thesis.

9.1 Introduction

As described in Chapter 2, the availability and use of MOOC platforms have increased dramatically during the last decade. MOOC platforms provide massive datasets that greatly aid in the advancement of knowledge in the area of learning analytics (LA). This kind of data can assist in gaining knowledge of a student's behaviour and provide insights into what works to help a learner improve (Watkins, 2017). Furthermore, LA is useful for ensuring students' progress more effectively, doing in-depth analyses of their activities and gauging the impact of these variables. LA is also useful for visualising students' activities to enhance education (Oliva-Cordova et al., 2021).

The poor completion rate is a fundamental problem associated with MOOCs. Researchers have shown that, most students who enrol in MOOCs, drop out before completing the course (Yang et al., 2013). One way to improve the low completion rate in MOOCs is to identify students who are at risk of dropping out at an early stage of the course. Detecting at-risk learners within a reasonable timeframe might support instructors in delivering educational interventions and improving course structures (Hung et al., 2015). Together, machine learning and LA can be used to identify potentially at-risk students (Al-Shabandar, 2019).

The research questions in this thesis aimed to exploit MOOC data to help early detection of at-risk students using several approaches, such as statistical data analysis, machine learning, data visualisation, and a concrete mapping between students tracking data and four of the most used theories. This includes exploiting the MOOC dataset to I) analyse the very first

interaction data with the MOOC system – the registration to predict students' completion, II) implement rules to automatically deliver personalised messages to students based on a statistical behaviour analysis of their registration date, III) propose visual multimodal graph analysis to discover linear or catch-up behaviours of completers and non-completers, IV) analyse completer and non-completer' learning paths, V) present students' learning paths with different granularity visualisations, VI) apply machine learning to compare the prediction of weekly and whole-course dropouts, VII) incorporate students' learning patterns, specifically jumping behaviours into the weekly predictive model, and demonstrate their effectiveness, VIII) provide a lightweight approach to predict dropout students based on two features (time spent and number of accesses), IX) map multimodal student behaviour over several motivational theories and conducting a large-scale evaluation of SDT for online learning and MOOCs based on success measures and X) cluster students and analyse the engagement parameters based on success measures.

9.2 Completion based on registration data

The first study of this thesis (Chapter 5) tackles the important and challenging issue of predicting student dropout and completion, which are the most targeted issues in research relating to MOOCs. However, most studies (rather predictably) analyse the course while it is running. We argue here that, in some cases, this might be too late. Thus, importantly, this work presents the results of a study aiming to discover if there are factors that can be identified before the students even begin the course, to predict which enrolled participants will not complete the MOOC and, possibly, take actions. This study is based on the analysis of a large dataset of FutureLearn MOOC users over several courses, each with several runs.

Our results show that completion can be predicted based on the date of registration. We performed a fine-grain analysis of this phenomenon based on our preliminary findings. Interestingly, we detected specific periods when it was likelier for the students registered (relatively) early to complete, as well as periods for which the opposite was true. We show that these periods are intrinsically linked to the course start date. We show how these findings can lead to personalisation strategies based on the earliest possible detection of potential issues. Additionally, this research is applied to a less explored MOOC platform, FutureLearn. Unlike many of its counterparts in other parts of the world, FutureLearn has arguably been

based, from the beginning, on solid pedagogical foundations, which makes it specifically interesting for education-related research. However, for this research, the results we obtained were founded on features shared by all MOOCs, such as the information on the date of student registration. Thus, we can claim that our results have a more generic impact. Furthermore, as we addressed the research question via a genuinely large-scale experiment involving several subjects in a truly longitudinal study, reaching over several iterations of all the courses considered, we further ensured the generality of our claims.

The results in this work are interesting and, to the best of our knowledge, have not been tackled before. However, as in any research, they come with caveats that need to be mentioned. First and importantly, the variances for the five periods (P1 to P5) in Table 5.2 (especially for P1 and P5), as well as in Table 5.1, were huge. This was consistent with the data spread, as can be seen in Figure 5.2. In the latter, it can be clearly seen that, especially for non-completers, the spread of the registration date is quite wide. However, this is less so for the completers, as can be seen on the left side of Figure 5.2.

However, only Warwick University data was used in this study, as this was the first dataset acquired in 2017, and it was available when the research was carried out. The other two datasets, obtained from Durham University and the Rawaq platform in 2019, were unavailable for use in this study.

9.3 Different granularity visualisations for learning patterns

The second study of this thesis (Chapter 6) was to visualise and compare different learning paths of completers and non-completers across four MOOCs and explore the learning theme from which learners tend to drop out.

We have shown how different granularity visualisations (*fish eye*, *bird eye*) allow both researchers, and potentially teachers, to understand where issues occur and where patterns emerge, backed up by statistical analysis. Specifically, we have shown that course completers are likelier to learn *linearly*, while dropout learners are likelier to jump forward to a later activity, which we dubbed here as a '*catch-up*' learning pattern. Moreover, we show how this type of analysis can generate fine-grained ideas for instructors and course designers; for

example, to improve retention, instructors (and online course designers) should enable more discussion on support mechanisms.

Only the University of Warwick dataset was taken into consideration for analysis in this study. This is because the study's main objective was to examine how students interacted with the course material, and this dataset was accessible for four courses via the FutureLearn website. Due to their lack of availability at the time the study was conducted, the additional datasets from Durham University and the Rawaq platform were not used.

9.4 Weekly prediction of at-risk students

Chapter 7 presents three different experiments. The First experiment is to compare two methods for predicting dropout students from an early stage (**CP** vs **WP**). Second, the students' jumping activities were incorporated into the weekly completion predictive model. In the third experiment, predictive models were implemented based on only two independent variables (the ***number of accesses*** and the ***time spent*** on a page).

The first experiment's results showed that the **WP** method outperformed the **CP** method across all courses. One possible explanation for this disparity was that students who dropped out in subsequent weeks behaved similarly to completers in the first week. Therefore, it is challenging to predict for these students from an early stage. According to the results of our analysis, we found that the early prediction models should primarily emphasise predicting students who would withdraw from their courses soon (next week).

In the second experiment, we used student jumping activities as a feature to predict student dropout. According to the data visualisation presented in Chapter 7, a learning route is an informative feature because successful learners will follow the guided path and display behaviours that are referred to as 'linear learning behaviours'. We showed that using students' jumping behaviour as an input feature to predict at-risk students enhanced the prediction performance. For example, after considering the jumping activity, the F1 score rose by 12.25% – from 78.99% to 90.91% – in SUP. This might be used to enhance the learning environment of MOOCs and provide suitable early intervention for students who are at risk of dropping out. Finally, in the third experiment, we have shown that we can provide reliable, very early (first week) prediction based on two easily obtainable features

only, using a lightweight approach for prediction, which allows for easy and reliable implementation across various courses from different domains. Such an early and accurate predictive methodology does not yet exist beyond our research, and as such, this is the first in this class of models. We have shown that these two features can provide a 'good enough' performance. The advantage of such an approach is clear: it is easier and faster to implement across various MOOC systems and does require the existence of only a limited amount of information. The implementation itself is lightweight, is much more practical when considering an on-the-fly response and has limited expenditure in terms of implementation resources and, more importantly, time.

9.5 Engage Taxonomy

Chapter 8 presented a way to confirm SDT theory (and potentially any motivational theory) via practical experimentation, which we believe is groundbreaking, as it has not been done before in the MOOC online context at this scale. For this purpose, we have proposed a novel, systematic way of analysing engagement, starting from multimodal tracking parameters and following established engagement and motivational theories. We proposed a concrete mapping between the tracking parameters and four of the most used theories of, or related to, engagement in digital systems, generating an engaged taxonomy.

We have also showcased how such a mapping procedure can be put into practice by analysing the engaged and disengaged MOOC student behaviours in relation to the SDT theory. We clustered students based on their engagement and analysed them via the connection to their success. This connection showed that the results support the SDT theory, along with its dimensions of autonomy, relatedness and competence. Thus, it validates the fact that mapping onto concrete features extracted from tracking student behaviour provides reliable, measurable (and thus directly comparable) variables tested against independent success variables for the student.

These findings are based on a large-scale, data-driven study where similar consistent results were obtained over several runs of the courses. This clearly supports a theoretically rooted approach on how to characterise engaged and disengaged MOOC student behaviours and explore what triggers and promotes MOOC students' interest and engagement. Finally, we

used these extracted SDT constructs directly as early predictors of active versus non-active students, showing successful results with several machine learning methods.

9.6 Beneficiaries of the Research Findings

The research effort in this thesis focused on identifying indicators for the early identification of dropout students in MOOCs. The findings of this research hold considerable advantages for different people involved in the field of education. For example**,** teachers may derive advantages from the research findings presented in Chapter 5 by implementing personalised approaches that rely on early predicting. Therefore, teachers can proactively intervene with at-risk students to promote their engagement and retention in the MOOC by sending them personalised recommendations.

To find problems and spot patterns in learning routes, platform providers may also employ other granularity visualisations, such as fish eye and bird's eye, as shown in Chapter 6. Students would gain advantages from a more engaging and supportive learning environment, which would contribute to higher success and completion rates.

The results from Chapter 7 might also help teachers increase retention rates by enabling more discussion. The utilisation of machine learning techniques by MOOC platform providers to predict student attrition based on their initial week activities is a feasible option. MOOC platform providers can initiate this process by integrating a monitoring mechanism that gathers information on student actions, including login frequency, jumping activities, duration of platform usage, and interaction with course materials.

Chapter 8 showed how the mapping between the students' tracking parameters and motivation theories could be put into practice by grouping engaged and disengaged students in different clusters. This could potentially facilitate teachers in promptly intervening with students who are at risk so that they can motivate these students.

9.7 Students emotion analysis

In this thesis, we aimed to extract more insights from the messages left by students by using sentiment analysis techniques. Through the identification of nine distinct features related to the sentiments of students' posts and mapping these to constructs of motivation theories by

three experts, we were able to evaluate the level of students' motivation and group them into three different clusters. The incorporation of sentiment analysis yielded valuable insights for understanding the driving forces that impact student engagement and attrition.

With respect to the prediction of student dropout, our results showed that the characteristics extracted from students' comments did not significantly contribute to dropout prediction. However, it is noteworthy that positive comment posts were identified as one of the most important indicators for predicting attrition in certain instances, as in the Shakespeare and His World and Excel courses (see Appendix C). This highlights the varied influence of sentiment analysis on student attrition prediction in different contexts.

9.8 Limitations for head-to-head Comparison

The thesis did not incorporate a head-to-head comparison of prediction models with other study outcomes for a number of valid reasons. The first problem is that much of the published work on MOOC prediction models is based on private datasets that are inaccessible to academics. This restriction prevents a direct comparison of the findings with those of other research since the datasets utilised may have different sample sizes, demographics, and data-gathering techniques.

Second, comparing prediction models among MOOC providers is complicated by the fact that their dataset structures may differ. For example, some researchers have utilised the actions of students in videos as a predictive feature, but such information is not available in the dataset used for this thesis. This difference in data format makes it even more difficult to make meaningful head-to-head comparisons with the findings of other investigations.

Moreover, the lack of openly accessible datasets from the FutureLearn and Rawaq platforms constrains the feasibility of employing the devised methodology to examine such data for comparative purposes.

9.9 Limitations

As in any research, especially when breaking new ground, some limitations need to be highlighted that may help enlighten future research paths.

Despite the fact that MOOC platforms were designed to enhance educational outcomes, MOOC data is not usually stored in the database expressly for research purposes. As a result, the data is often irregular, incomplete, noisy and unbalanced. Therefore, addressing these challenges requires significant time for data extraction, understanding and preparation. In addition, data standardisation is a critical issue, as discussed in Chapter 2. Problems with data standardisation in MOOCs might be an obstacle to the validation of completion prediction models.

Although we used a large dataset with more than 340,000 students, the data did not contain all students' activities on the platform, such as video-related activities (e.g. video pause, video speed, video stops, rewatch records, video backward jumps) and mouse behaviours (e.g. move, scroll and click). Exploring this kind of data may enhance the prediction performance.

For the registration data analysis, the large variance could diminish the value of the statistical significance of the results obtained; this could be further indirectly affected by the large size of our sample. However, as Figueiredo Filho et al. (2013) recommended, we also visualised the data (as in Figure 5.2), and saw that completers were less spread than non-completers. What this means is that statements about completion are likelier to be statistically significant than statements about non-completers. Possible further research could look into eliminating the outliers; however, this needs to be done with care, as important information should not be lost in the process. For the latter reason, and to avoid sampling errors, we opted for this research to keep all students in.

It is possible that the date of registration was not the cause of the completion or non-completion of the students. For instance, it could be that a certain type of student, more inclined to complete the course, tends to register at a certain time. Thus, suggesting that students alter their registration behaviour might not be enough.

This research included the ***time spent*** (a calculated value) as a feature to predict dropout students. This feature defined the difference between the access time to a certain step and the point at which the student clicked a button labelled 'Mark as Completed'. However, this did not entirely reflect learners' true engagement time (as some students may click the completion button before they complete the step).

Additionally, we considered only the first access to the course content (first_visited_at) due to our data limitations, which may have underestimated the jumping patterns of dropout learners. Other learning features could be considered.

The mapping of student behaviour onto the theory has been done here also for the first time, to our knowledge. Whilst we were careful on checking our mapping with the help of several experts in education and motivational theories, it is possible that we may have missed something (either a construct not being included where needed, or a construct not needed but included). The current results seem to point to our mapping being successful. However, as this large-scale evaluation of the SDT, and mapping of metrics over engagement theories, is a completely new direction of research, this opens the way for further analysis and possible extension of these findings, including increasing the accuracy of the prediction, or looking into ways in which adaptation or intervention built on the motivation-based prediction outcomes. Additionally, on a final note on SDT, only three coarse-grained constructs were considered and mapped. Further mappings and evaluations can look into more refined model mapping.

Epilogue

This chapter has discussed the outcomes of the previous four chapters. In addition, we highlighted the limitations and challenges of using the MOOCs dataset as well as the limitation of each study. In the next chapter, we will provide a general summary and conclusion of this thesis. Finally, we conclude by offering suggestions for further research in this area.

Chapter 10 Conclusion and Future Work

This thesis presented some factors underlying student disengagement in online learning (MOOC platforms). The primary focus of all the efforts in this work was on students who participated in massive open online courses (MOOCs) but ultimately disengaged from the course. This work aims to contribute to the development of efficient intervention strategies to deal with at-risk students at an early stage.

This research attempted to use data from MOOCs to help in the early detection of at-risk students. This was achieved through the use of several methods, such as statistical analysis, data visualisation, machine learning and mapping between the tracking parameters and motivational theories.

In fact, data quality is considered an essential element in building successful predictive models (Hall and Smith, 1998). We collected student behaviour data from the MOOCs of 344,783 students, which is not accessible to the general public. The collected data contained the activities of the students, such as their social interactions, the topics they accessed, the quizzes they attempted and their correct/incorrect responses. The duration of the courses ranged from 3 to 10 weeks, and they were offered at various times between 2013 and 2019.

The first three core chapters (Chapter 5, Chapter 6 and Chapter 7) successfully achieved the first aim posed at the beginning of this thesis, which was to develop a continuous predictive model that can be used in real-time to identify the students at risk of dropping out of MOOCs.

The findings showed that students' completion can be predicted even before the course starts based on the registration date. This can help implement some rules to automatically deliver personalised messages to the students based on a statistical analysis of their registration date.

Also, the visualisation results showed that students who completed the course were likelier to learn linearly, whereas students who dropped out were likelier to engage in the 'catch-up' learning pattern. This type of analysis can generate fine-grained ideas for instructors by providing a clearer view to monitor students' activities and highlight at-risk students. Also,

the visualisation can provide important feedback for the instructor to redesign the course for the optimum participation of students.

Moreover, the comparison results of the two weekly prediction methods (**CP** vs **WP**) showed that the models' performance in predicting students who will withdraw from their courses (next week) outperformed the models that predicted dropout students in subsequent weeks. In addition, the performance of weekly prediction models improved by utilising students' jumping behaviour as an input feature. In the third experiment, we proposed a lightweight approach for early prediction that can be reliably applied to different types of MOOCs with little effort. The results of this prediction show that the two features (the ***number of accesses*** and the ***time spent***) provided a satisfactory performance level.

The second aim was to employ motivational theories and map online student behaviour onto them. In this thesis, Chapter 8 proposed a mapping methodology of engagement, including designing, mapping, measuring and evaluating, which can be further applied not only to MOOCs and e-learning systems when exploring engagement along the SDT constructs but also in terms of the mapping of other engagement theories, such as Drive (Pink, 2011), Engagement (Kearsley and Shneiderman, 1998), and Process of Engagement (O'Brien and Toms, 2008) onto data-intensive applications. This research shows how this kind of mapping can be used by analysing the engaged and disengaged MOOC student behaviours with respect to SDT theory.

This thesis contributes significantly to the early detection of students at risk in MOOCs. The research addresses the dropout issue in an online environment by providing useful insights and suggestions. Developing accurate predictive models as an alerting tool might help education service providers proactively predict at-risk students in ongoing courses. This may assist educators and decision-makers in the process of planning for measures to be taken in advance to help these students. Moreover, it may assist education providers in developing plans for the subjects that will be offered in the future.

10.1 Future work

Potential future work has several dimensions; for example, another dataset from different platforms with new longitudinal behaviour over extended periods could be used to predict dropout students.

In addition, more reliable and advanced machine learning classifiers (e.g. Explainable Artificial Intelligence (Speith, 2022)) can be used. This kind of prediction technique is not only accurate at making predictions but also has the ability to describe the reasoning behind such predictions.

In addition, we have opted to analyse motivation from the point of view of the SDT theory. There are many other motivational theories out there. We have opted for SDT as being one of the most well-known and the use of SDT has become commonplace in the educational domain (Zhou, 2016).

The mapping of metrics over engagement theories is a new direction of research; this opens the way for further analysis and possible extension of these findings, including increasing the accuracy of the prediction. This approach can also help in validating theories from a data-intensive point of view, which is interesting for the future. However, for future research, we will explore others as well.

Appendix

Appendix A Prediction of Active and Non-Active Students in week 2 measured over the 6 courses

BA: Balanced Accuracy, ACC: Accuracy, PR: Precision, Re: Recall, F1-Score

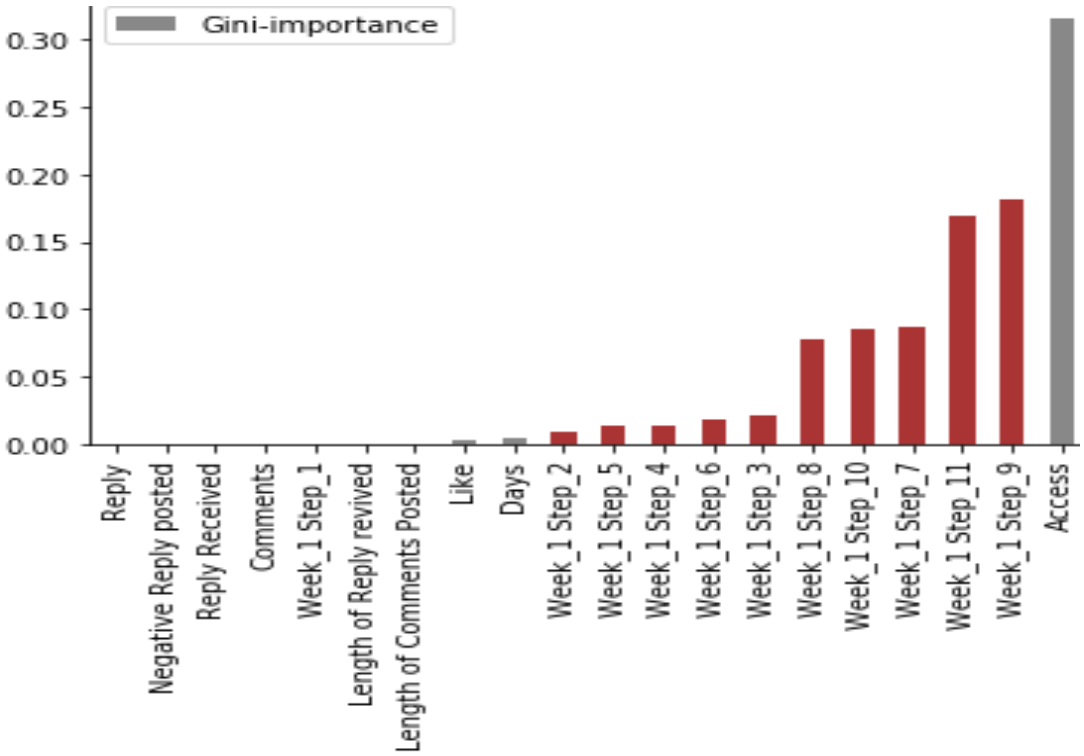
Course	Classifier	Active 1/ Non-Active 0	Precision	Re	F1	**ACC**	**BA**
The Mind is Flat	KNeighbors	0	94.60%	91.69%	93.12%	**89.43%**	**86.54%**
		1	73.39%	81.40%	77.19%		
	LogisticRe	0	98.01%	89.05%	93.31%	**90.04%**	**91.31%**
		1	70.64%	93.57%	80.51%		
	XGBClassif	0	97.65%	90.04%	93.69%	**90.53%**	**91.17%**
		1	72.29%	92.31%	81.08%		
	AdaBoostCl	0	97.84%	89.21%	93.33%	**90.04%**	**91.11%**
		1	70.82%	93.01%	80.41%		
	GradientBo	0	97.69%	89.88%	93.62%	**90.44%**	**91.16%**
		1	72.00%	92.45%	80.96%		
	ExtraTrees	0	98.14%	89.41%	93.57%	**90.41%**	**91.70%**
		1	71.41%	93.99%	81.16%		
	RandomFore	0	98.06%	89.60%	93.64%	**90.50%**	**91.65%**
		1	71.73%	93.71%	81.26%		
Supply Chains	KNeighbors	0	94.25%	93.73%	93.99%	**90.16%**	**83.74%**
		1	71.95%	73.75%	72.84%		
	LogisticRe	0	98.47%	87.74%	92.80%	**88.81%**	**90.74%**
		1	62.50%	93.75%	75.00%		
	XGBClassif	0	98.81%	90.46%	94.45%	**91.28%**	**92.73%**
		1	68.47%	95.00%	79.58%		
	AdaBoostCl	0	98.22%	90.19%	94.03%	**90.60%**	**91.35%**
		1	67.27%	92.50%	77.89%		
	GradientBo	0	98.81%	90.46%	94.45%	**91.28%**	**92.73%**
		1	68.47%	95.00%	79.58%		
	ExtraTrees	0	98.80%	89.92%	94.15%	**90.83%**	**92.46%**
		1	67.26%	95.00%	78.76%		
	RandomFore	0	98.52%	90.46%	94.32%	**91.05%**	**92.11%**
		1	68.18%	93.75%	78.95%		
Open Innovation in Business (OI)	KNeighbors	0	94.98%	88.93%	91.85%	**88.16%**	**87.39%**
		1	72.03%	85.86%	78.34%		
	LogisticRe	0	92.58%	87.92%	90.19%	**85.64%**	**83.35%**
		1	68.42%	78.79%	73.24%		
	XGBClassif	0	96.03%	89.26%	92.52%	**89.17%**	**89.08%**
		1	73.33%	88.89%	80.37%		
	AdaBoostCl	0	98.43%	84.23%	90.78%	**87.15%**	**90.09%**
		1	66.90%	95.96%	78.84%		
	GradientBo	0	92.78%	90.60%	91.68%	**87.66%**	**84.70%**
		1	73.58%	78.79%	76.10%		
	ExtraTrees	0	96.39%	89.60%	92.87%	**89.67%**	**89.75%**
		1	74.17%	89.90%	81.28%		
	RandomFore	0	95.34%	89.26%	92.20%	**88.66%**	**88.07%**
		1	72.88%	86.87%	79.26%		
Shakespeare	KNeighbors	0	87.88%	84.06%	85.93%	**81.72%**	**80.57%**
		1	70.99%	77.09%	73.91%		
	LogisticRe	0	96.58%	80.09%	87.57%	**84.90%**	**87.24%**
		1	70.58%	94.39%	80.77%		
	XGBClassif	0	96.17%	80.83%	87.84%	**85.14%**	**87.24%**
		1	71.20%	93.65%	80.89%		
	AdaBoostCl	0	96.45%	80.02%	87.47%	**84.78%**	**87.11%**

		1	70.47%	94.19%	80.62%		
	GradientBo	0	96.33%	80.74%	87.85%	**85.17%**	**87.33%**
		1	71.17%	93.92%	80.97%		
	ExtraTrees	0	96.47%	80.38%	87.69%	**85.02%**	**87.29%**
		1	70.84%	94.19%	80.86%		
	RandomFore	0	96.49%	80.30%	87.65%	**84.98%**	**87.26%**
		1	70.76%	94.22%	80.82%		
Leading and Managing People-Centred Change (LMPCC)	KNeighbors	0	95.36%	85.76%	90.31%	**86.48%**	**87.12%**
		1	69.22%	88.47%	77.67%		
	LogisticRe	0	96.95%	83.63%	89.80%	**86.05%**	**88.18%**
		1	67.21%	92.73%	77.94%		
	XGBClassif	0	97.73%	84.04%	90.37%	**86.85%**	**89.32%**
		1	68.20%	94.61%	79.27%		
	AdaBoostCl	0	99.07%	81.86%	89.64%	**86.11%**	**89.86%**
		1	66.13%	97.87%	78.93%		
	GradientBo	0	98.24%	83.40%	90.21%	**86.71%**	**89.63%**
		1	67.64%	95.86%	79.32%		
	ExtraTrees	0	98.66%	83.40%	90.39%	**86.98%**	**90.13%**
		1	67.87%	96.87%	79.81%		
	RandomFore	0	97.99%	83.90%	90.40%	**86.91%**	**89.57%**
		1	68.16%	95.24%	79.46%		
Babies in Mind	KNeighbors	0	91.26%	81.88%	86.31%	**79.19%**	**75.13%**
		1	48.35%	68.37%	56.64%		
	LogisticRe	0	97.51%	73.59%	83.88%	**77.34%**	**83.01%**
		1	46.47%	92.43%	61.85%		
	XGBClassif	0	86.96%	88.45%	87.70%	**80.12%**	**67.50%**
		1	50.00%	46.55%	48.21%		
	AdaBoostCl	0	97.76%	72.43%	83.21%	**76.58%**	**82.87%**
		1	45.64%	93.32%	61.30%		
	GradientBo	0	84.70%	91.77%	88.09%	**80.12%**	**62.48%**
		1	50.00%	33.18%	39.89%		
	ExtraTrees	0	95.60%	74.48%	83.73%	**76.80%**	**80.33%**
		1	45.58%	86.19%	59.63%		
	RandomFore	0	97.76%	72.43%	83.21%	**76.58%**	**82.87%**
		1	45.64%	93.32%	61.30%		

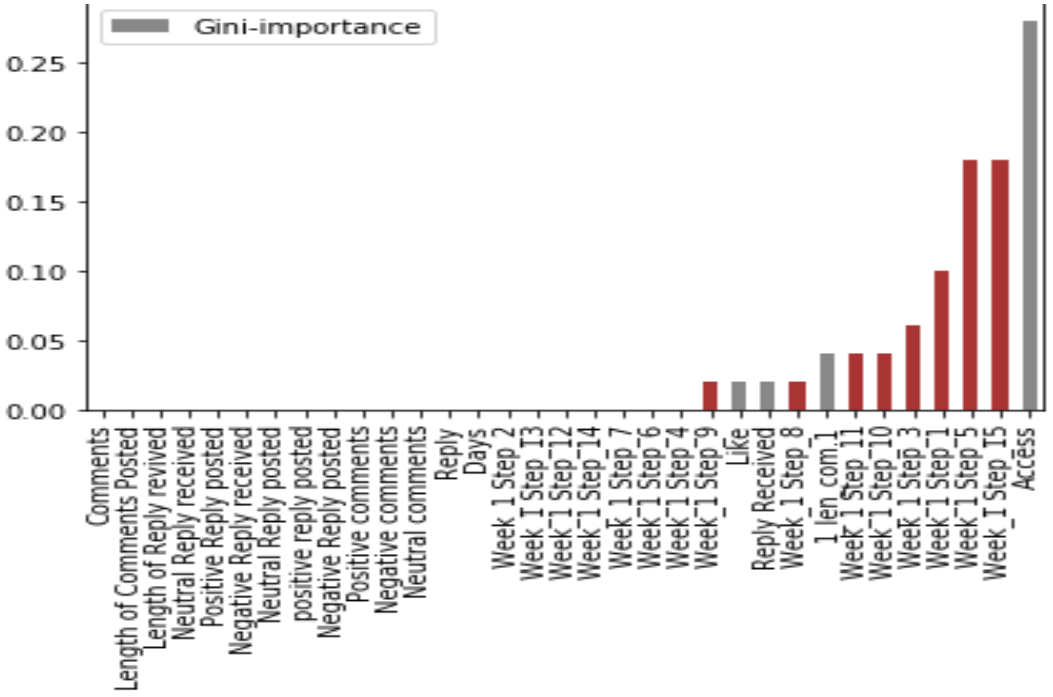
Appendix B Prediction of Active and Non-Active Students in week 2 based on week 1 SDT constructs (10 fold cross validation); evaluated by the Balanced Accuracy score

Courses	**AdaB**	**ExtrTr**	**GBoost**	**KNN**	**LR**	**RF**	**XGBoost**
Babies in Mind	83.93%	81.90%	70.13%	70.34%	84.17%	83.93%	72.69%
Shakespeare	86.76%	86.67%	87.01%	83.81%	86.84%	86.87%	86.96%
Supply Chains	91.30%	91.12%	88.88%	86.93%	91.22%	91.32%	90.69%
The Mind is Flat	90.08%	89.99%	90.13%	85.36%	90.23%	90.07%	90.51%
Open Innovation in Business (OI)	90.74%	90.14%	87.52%	87.24%	88.04%	90.27%	91.36%
(LMPCC)	89.55%	89.34%	88.97%	84.89%	87.59%	89.36%	89.57%

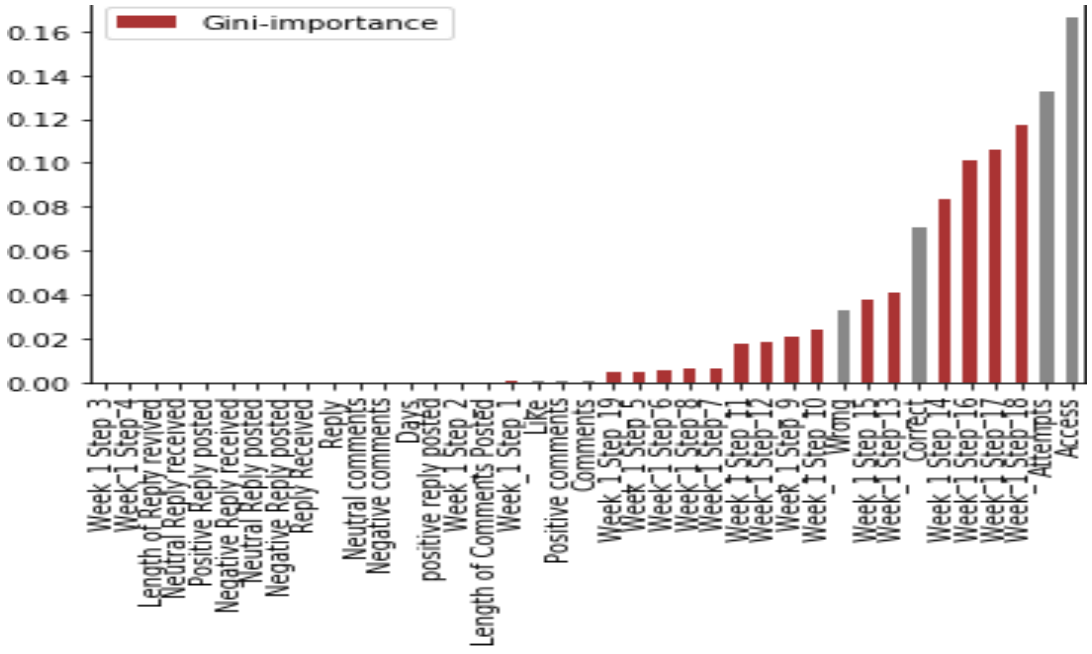
Appendix C Gini-importance for all courses (a-j)



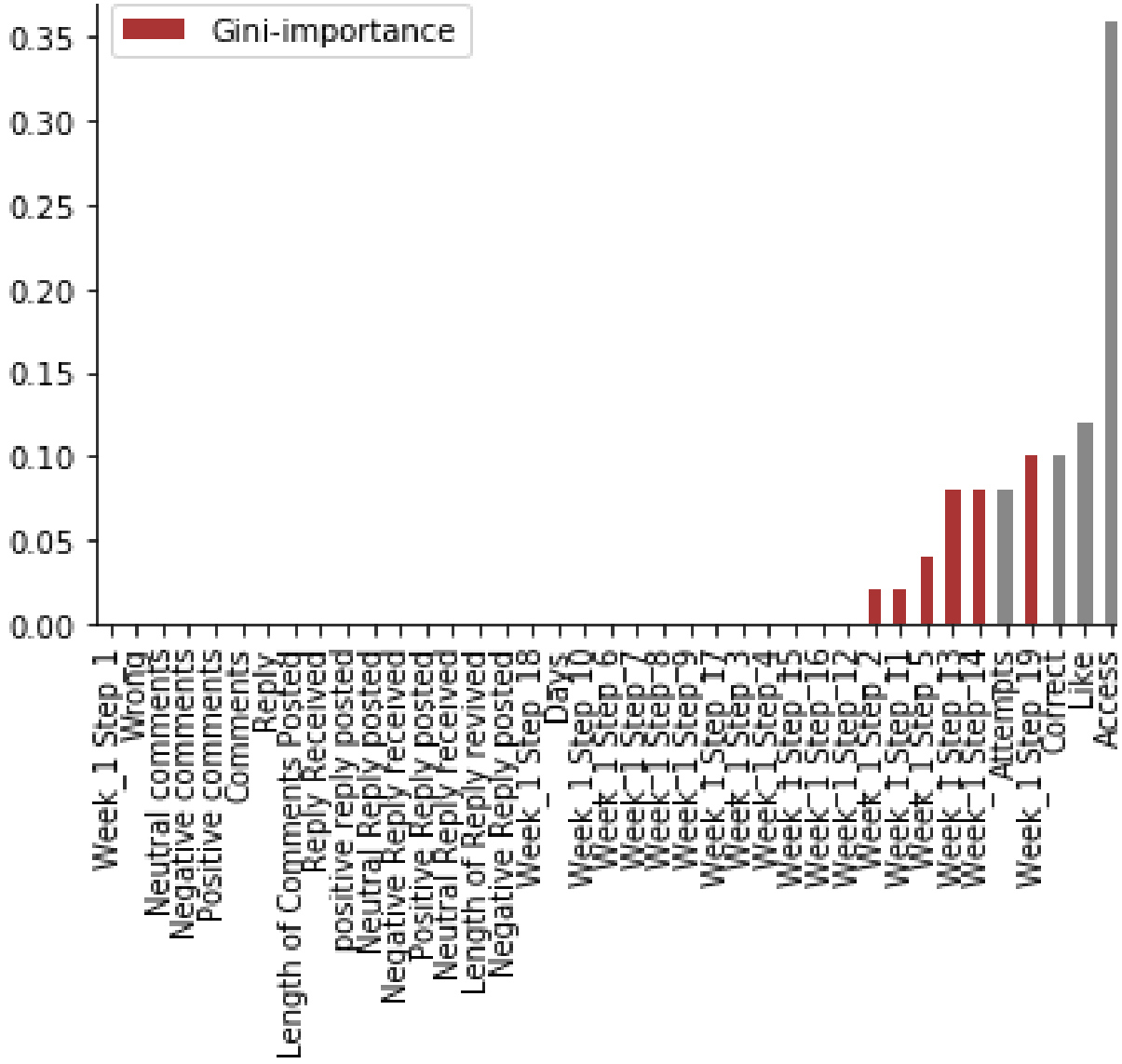
a) Big data


Gini-importance
0.25
0.20
0.15
0.10
0.05
0.00
Comments
Length of Comments Posted
Length of Reply revived
Neutral Reply received
Positive Reply posted
Negative Reply received
Neutral Reply posted
positive reply posted
Negative Reply posted
Positive comments
Negative comments
Neutral comments
Reply
Days
Week 1 Step 2
Week 1 Step 13
Week 1 Step 12
Week 1 Step 14
Week 1 Step 7
Week 1 Step 6
Week 1 Step 4
Week 1 Step 9
Like
Reply Received
Week 1 Step 8
1 len_com.1
Week 1 Step 11
Week 1 Step 10
Week 1 Step 3
Week 1 Step 1
Week 1 Step 5
Week 1 Step 15
Access

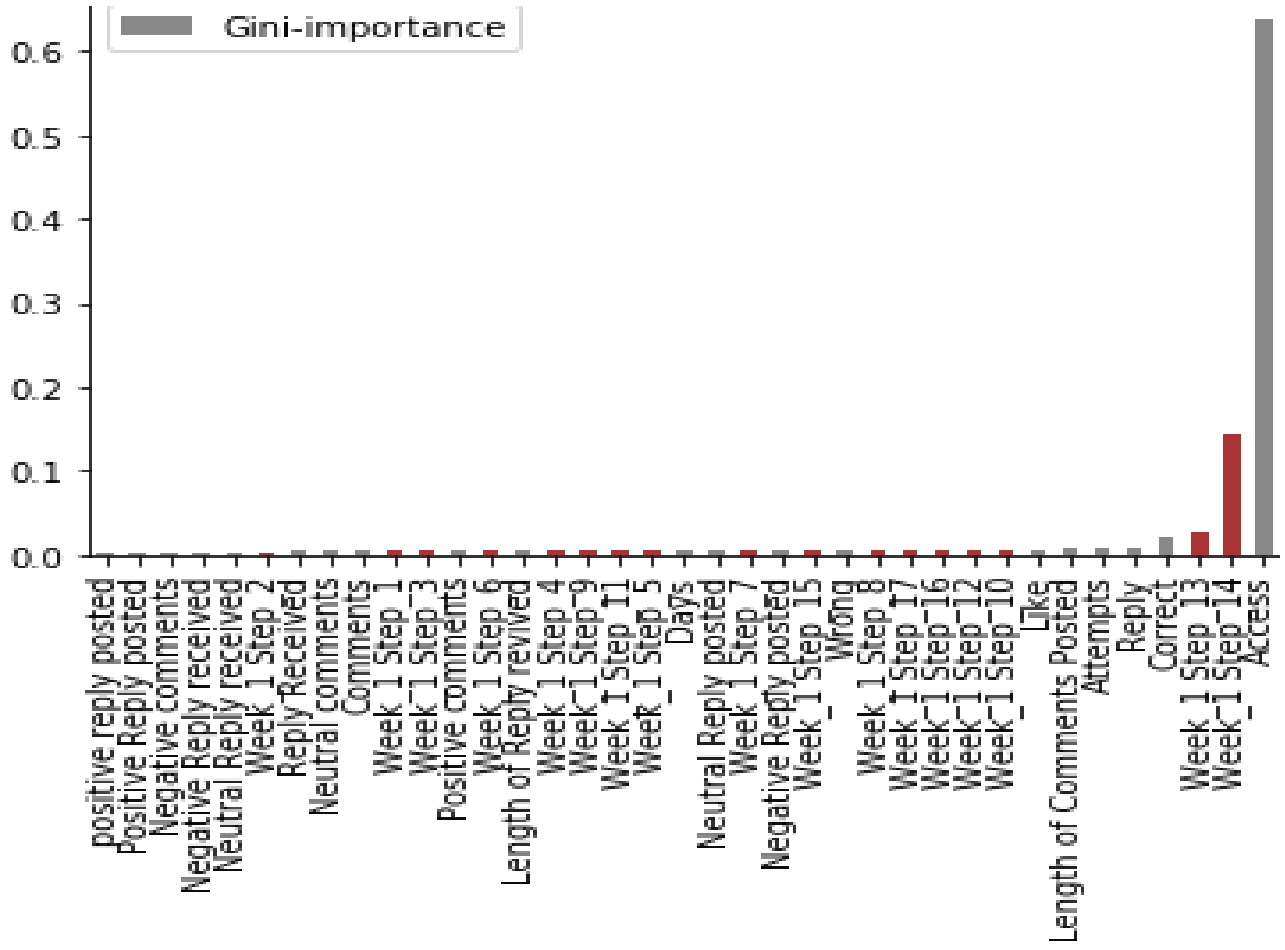
b) SLC


Gini-importance
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Week 1 Step 3
Week 1 Step 4
Length of Reply revived
Neutral Reply received
Positive Reply posted
Negative Reply received
Neutral Reply posted
Negative Reply posted
Reply Received
Reply
Neutral comments
Negative comments
Days
positive reply posted
Week 1 Step 2
Length of Comments Posted
Week 1 Step 1
Like
Positive comments
Comments
Week 1 Step 19
Week 1 Step 5
Week 1 Step 6
Week 1 Step 8
Week 1 Step 7
Week 1 Step 11
Week 1 Step 12
Week 1 Step 9
Week 1 Step 10
Wrong
Week 1 Step 15
Week 1 Step 13
Correct
Week 1 Step 14
Week 1 Step 16
Week 1 Step 17
Week 1 Step 18
Attempts
Access

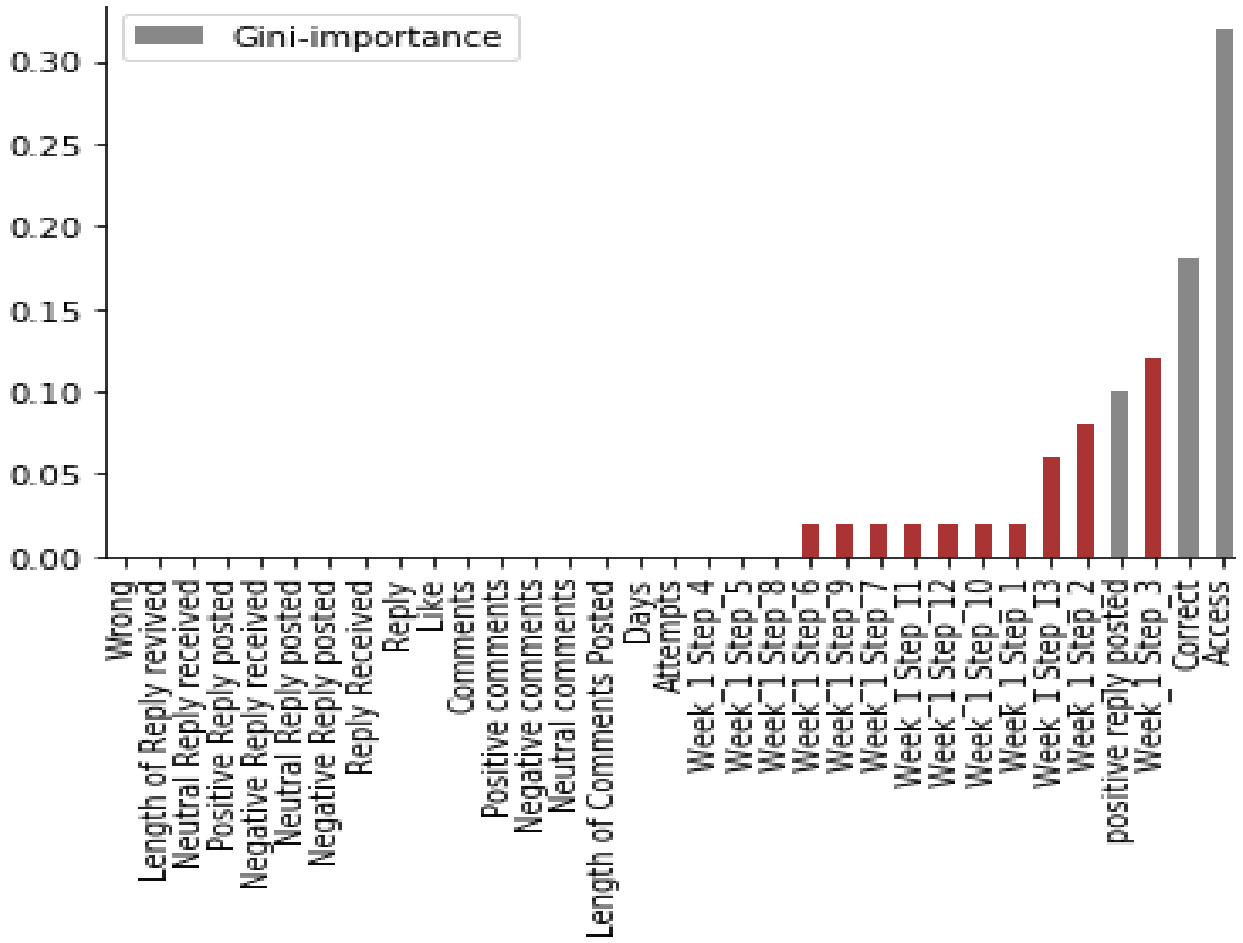
c) BIM


Gini-importance
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Week_1 Step_1
Wrong
Neutral comments
Negative comments
Positive comments
Comments
Reply
Length of Comments Posted
Reply Received
positive reply posted
Neutral Reply posted
Negative Reply received
Positive Reply posted
Neutral Reply received
Length of Reply revived
Negative Reply posted
Week_1 Step_18
Days
Week_1 Step_10
Week_1 Step_6
Week_1 Step_7
Week_1 Step_8
Week_1 Step_9
Week_1 Step_17
Week_1 Step_3
Week_1 Step_4
Week_1 Step_15
Week_1 Step_16
Week_1 Step_12
Week_1 Step_2
Week_1 Step_11
Week_1 Step_5
Week_1 Step_13
Week_1 Step_14
Week_ Attempts
Week_1 Step_19
Correct
Like
Access

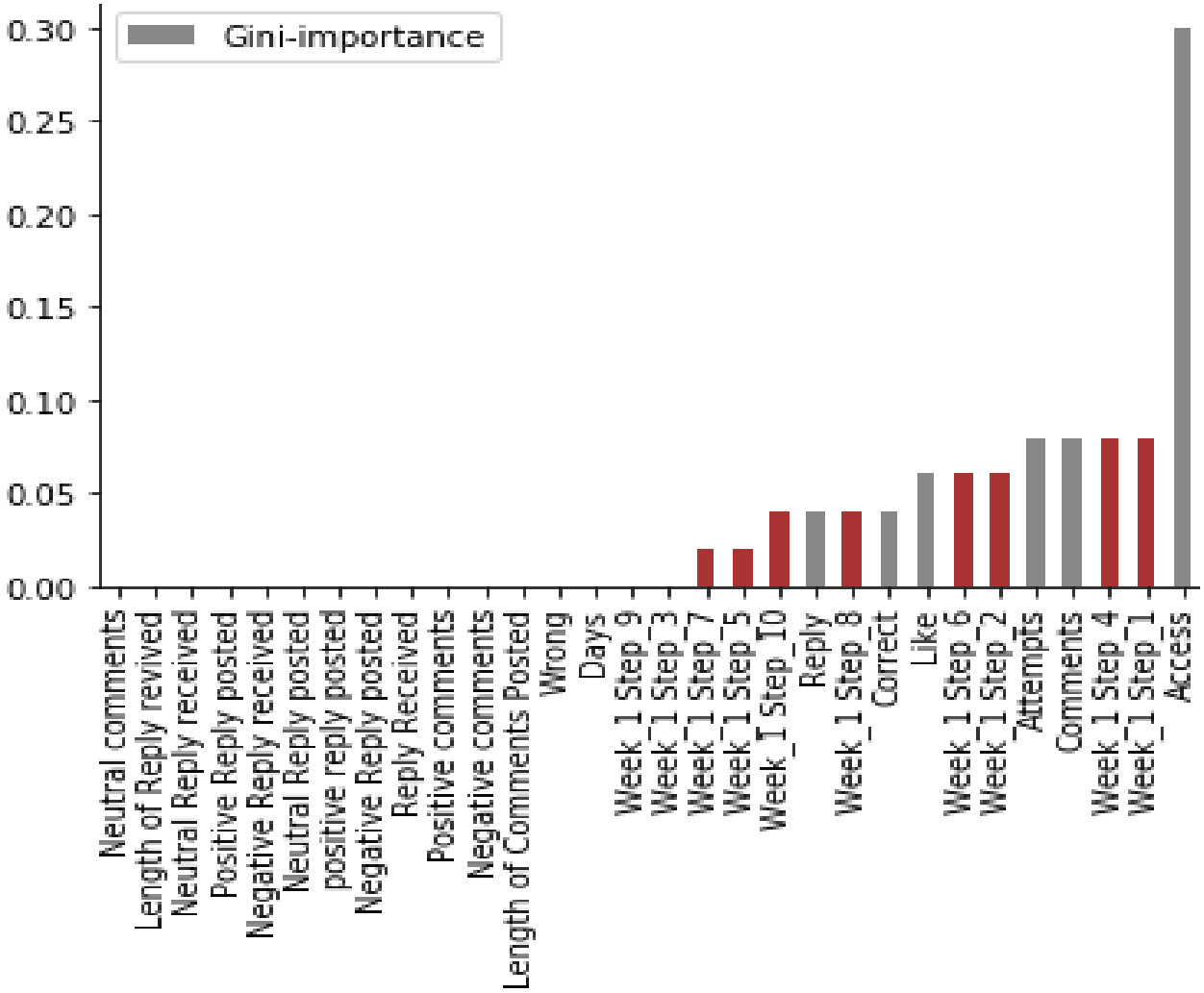
SUP


Gini-importance
0.6
0.5
0.4
0.3
0.2
0.1
0.0
positive reply posted
Positive Reply posted
Negative comments
Negative Reply received
Neutral Reply received
Week_1 Step_2
Reply Received
Neutral comments
Comments
Week_1 Step_1
Week_1 Step_3
Positive comments
Week_1 Step_6
Length of Reply revived
Week_1 Step_4
Week_1 Step_9
Week_1 Step_11
Week_1 Step_5
Days
Neutral Reply posted
Week_1 Step_7
Negative Reply posted
Week_1 Step_15
Wrong
Week_1 Step_8
Week_1 Step_17
Week_1 Step_16
Week_1 Step_12
Week_1 Step_10
Like
Length of Comments Posted
Attempts
Reply
Correct
Week_1 Step_13
Week_1 Step_14
Access

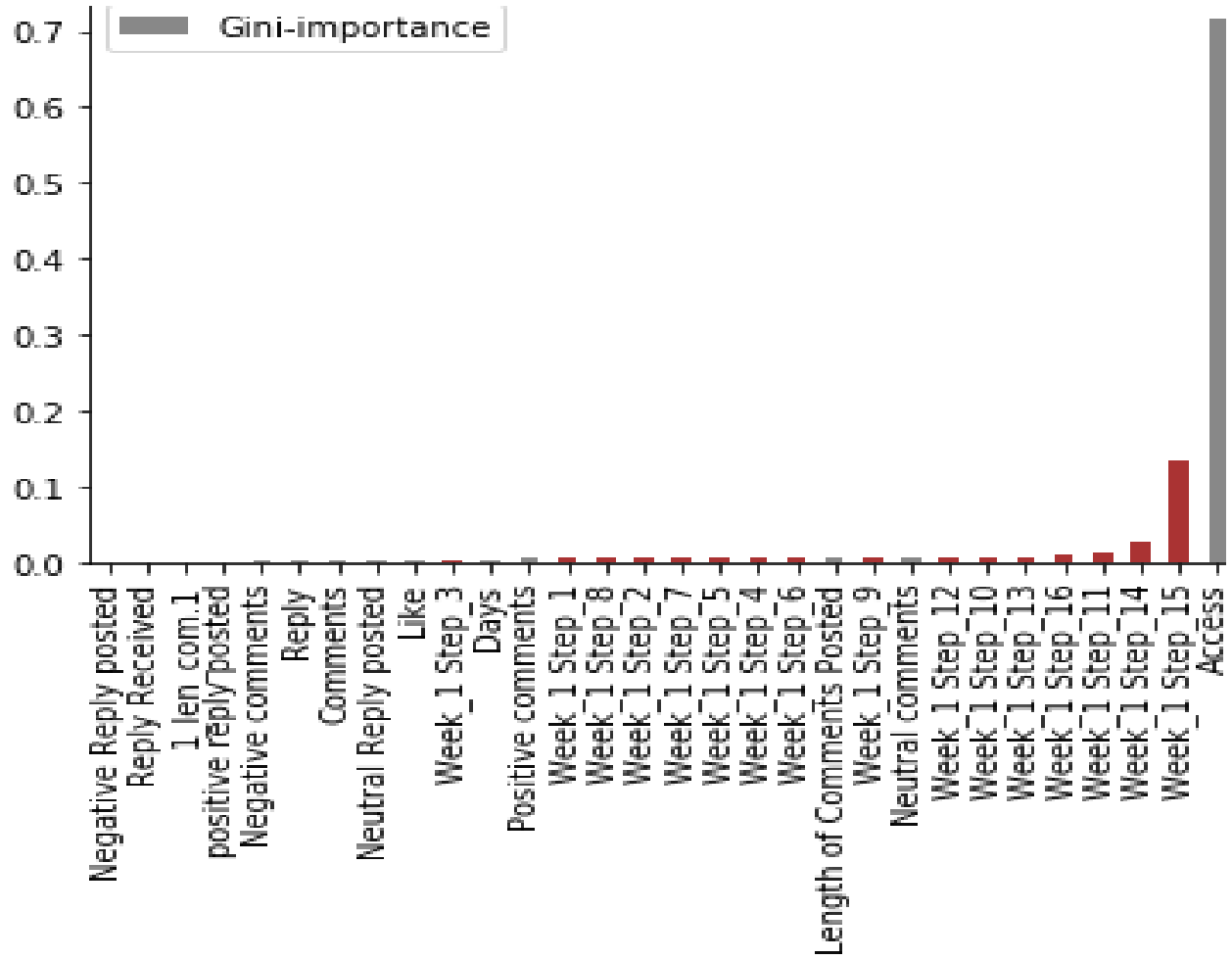
SHK


Gini-importance
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Wrong
Length of Reply revived
Neutral Reply received
Positive Reply posted
Negative Reply received
Neutral Reply posted
Negative Reply posted
Reply Received
Reply
Like
Comments
Positive comments
Negative comments
Neutral comments
Length of Comments Posted
Days
Attempts
Week 1 Step_4
Week_1 Step_5
Week_1 Step_8
Week_1 Step_6
Week_1 Step_9
Week_1 Step_7
Week_I Step_I1
Week_1 Step_12
Week_1 Step_10
Week_1 Step_1
Week_I Step_I3
Week_1 Step_2
positive reply posted
Week_1 Step_3
Correct
Access

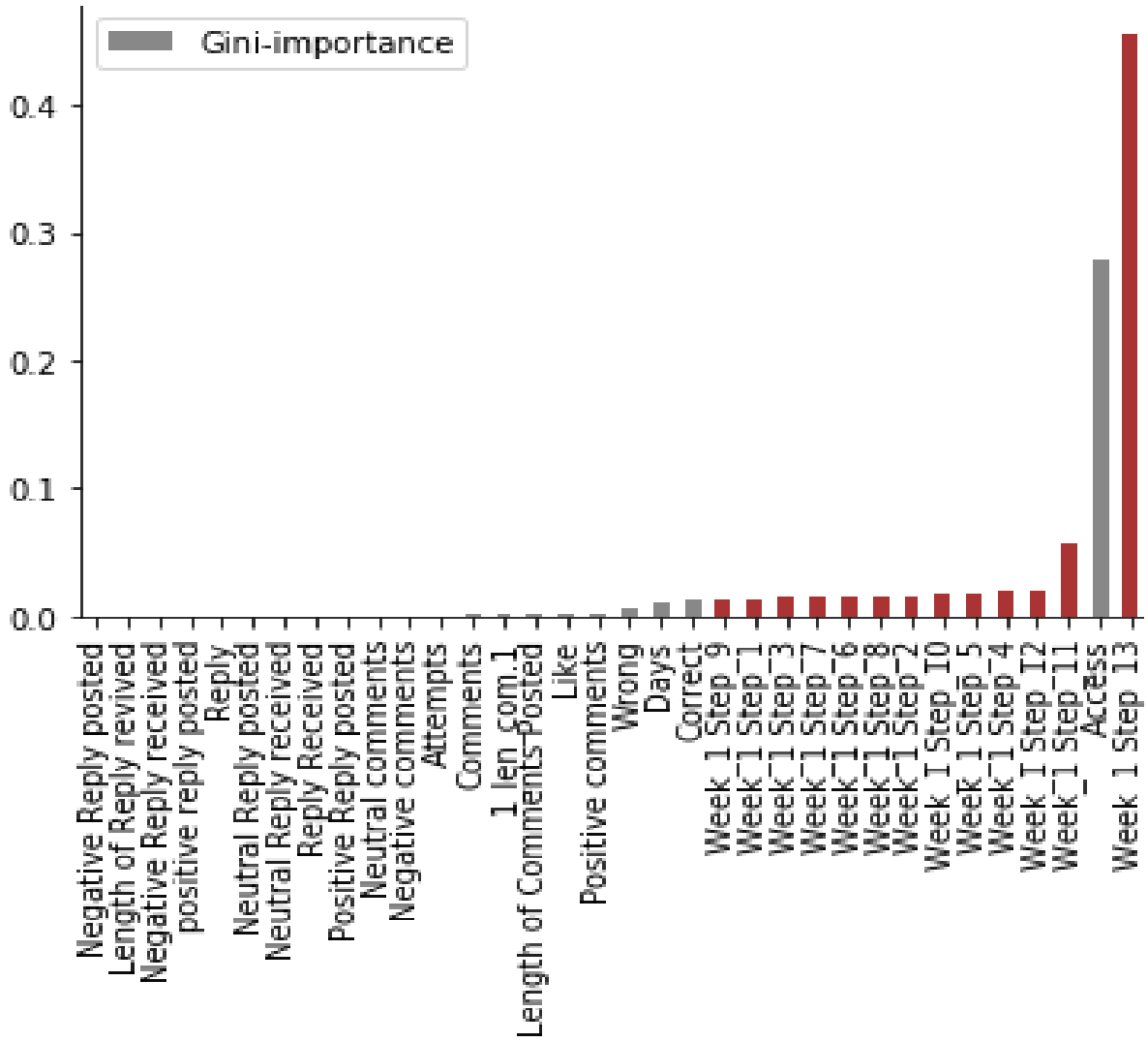
oi


Gini-importance
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Neutral comments
Length of Reply revived
Neutral Reply received
Positive Reply posted
Negative Reply received
Neutral Reply posted
positive reply posted
Negative Reply posted
Reply Received
Positive comments
Negative comments
Length of Comments Posted
Wrong
Days
Week_1 Step_9
Week_1 Step_3
Week_1 Step_7
Week_1 Step_5
Week_1 Step_10
Reply
Week_1 Step_8
Correct
Like
Week_1 Step_6
Week_1 Step_2
Attempts
Comments
Week_1 Step_4
Week_1 Step_1
Access

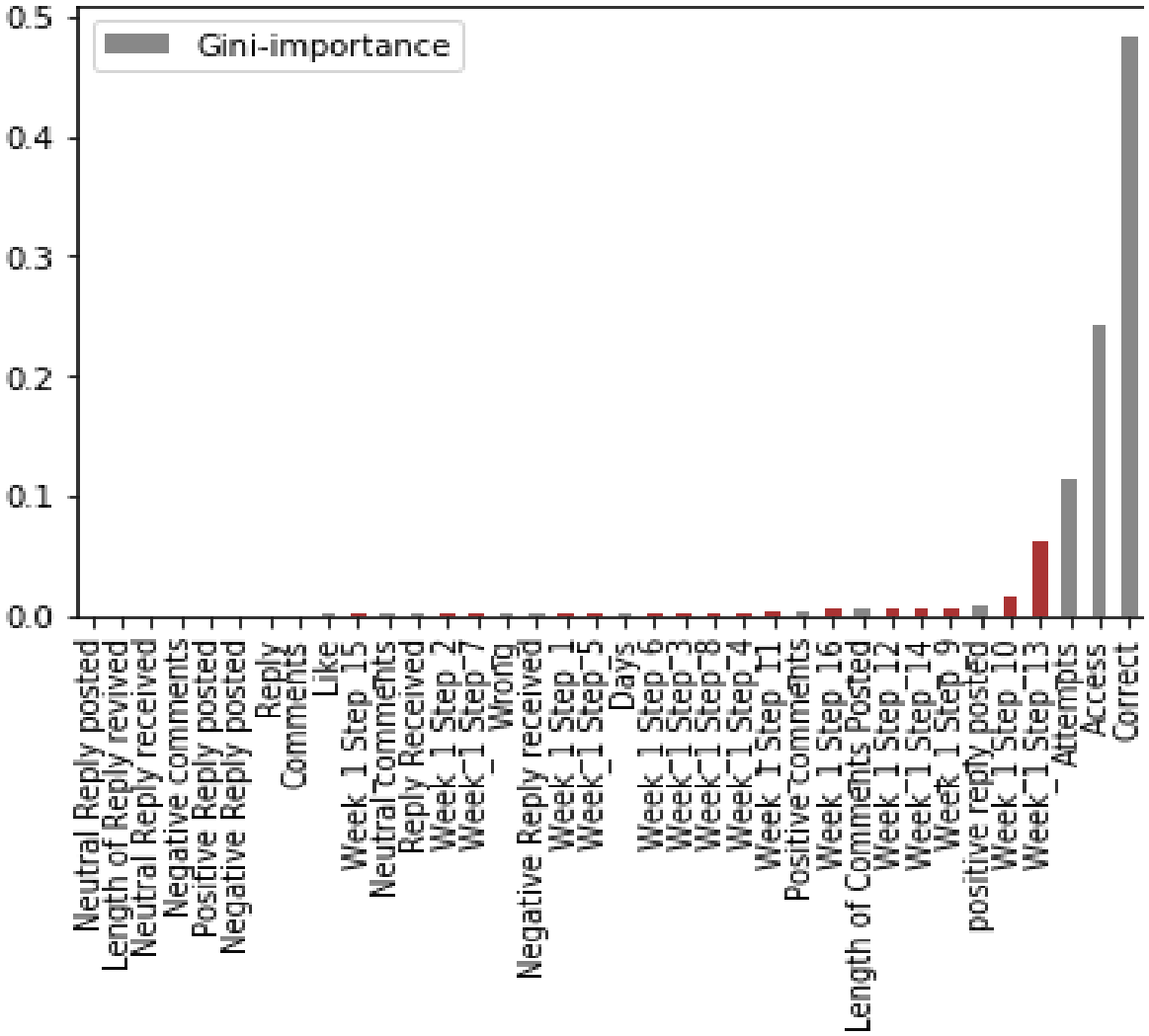
LMPCC


Gini-importance
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Negative Reply posted
Reply Received
1 len com.1
positive reply posted
Negative comments
Reply
Comments
Neutral Reply posted
Like
Week_1 Step_3
Days
Positive comments
Week_1 Step_1
Week_1 Step_8
Week_1 Step_2
Week_1 Step_7
Week_1 Step_5
Week_1 Step_4
Week_1 Step_6
Length of Comments Posted
Week_1 Step_9
Neutral comments
Week_1 Step_12
Week_1 Step_10
Week_1 Step_13
Week_1 Step_16
Week_1 Step_11
Week_1 Step_14
Week_1 Step_15
Access

Java


Gini-importance
0.4
0.3
0.2
0.1
0.0
Negative Reply posted
Length of Reply revived
Negative Reply received
positive reply posted
Reply
Neutral Reply posted
Neutral Reply received
Reply Received
Positive Reply posted
Neutral comments
Negative comments
Attempts
Comments
1 len_com_1
Length of Comments Posted
Like
Positive comments
Wrong
Days
Correct
Week 1 Step 9
Week 1 Step 1
Week 1 Step 3
Week 1 Step 7
Week 1 Step 6
Week 1 Step 8
Week 1 Step 2
Week 1 Step 10
Week 1 Step 5
Week 1 Step 4
Week 1 Step 12
Week 1 Step 11
Access
Week 1 Step 13

EXC


Gini-importance
0.5
0.4
0.3
0.2
0.1
0.0
Neutral Reply posted
Length of Reply revived
Neutral Reply received
Negative comments
Positive Reply posted
Negative Reply posted
Reply
Comments
Like
Week_1 Step_15
Neutral comments
Reply Received
Week_1 Step_2
Week_1 Step_7
Wrong
Negative Reply received
Week_1 Step_1
Week_1 Step_5
Days
Week_1 Step_6
Week_1 Step_3
Week_1 Step_8
Week_1 Step_4
Week_1 Step_11
Positive comments
Week_1 Step_16
Length of Comments Posted
Week_1 Step_12
Week_1 Step_14
Week_1 Step_9
positive reply posted
Week_1 Step_10
Week_1 Step_13
_Attempts
Access
Correct

a) TMF

Reference List

ADAM, S. P., ALEXANDROPOULOS, S.-A. N., PARDALOS, P. M. & VRAHATIS, M. N. 2019. No free lunch theorem: A review. *Approximation and Optimization: Algorithms, Complexity and Applications*, 57-82.

ADAMOPOULOS, A., NTASI, M., MAVROUDI, S., LIKOTHANASSIS, S., ILIADIS, L. & ANASTASSOPOULOS, G. Revealing the Structure of Childhood Abdominal Pain Data and Supporting Diagnostic Decision Making. 2009 2009. Springer, 165-177.

ADAMOPOULOS, P. 2013. What makes a great MOOC? An interdisciplinary analysis of student retention in online courses.

AGRAWAL, N. 2018. OER and Management Education in India: Managing Strategy in ODL System. *Open and Distance Learning Initiatives for Sustainable Development.* IGI Global.

AHMAD, M. S. A., A. H.; MOHAMMED, A. A Machine Learning Based Approach for Student Performance Evaluation in Educational Data Mining. 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference, MIUCC 2021, 2021. 187-192.

AI, D. Z., T.; YU, G.; SHAO, X. A Dropout Prediction Framework Combined with Ensemble Feature Selection. PervasiveHealth: Pervasive Computing Technologies for Healthcare, 2020. 179-185.

AITKENHEAD, M. J. 2008. A co-evolving decision tree classification method. *Expert Systems with Applications,* 34, 18-25.

AL-SHABANDAR, R. 2019. *The application of Machine Learning for Early Detection of At-Risk Learners in Massive Open Online Courses*, Liverpool John Moores University (United Kingdom).

ALAMI, H. A. A. C. M. B. A. T. E. Prediction MOOC's for student by using machine learning methods. 2021 XI International Conference on Virtual Campus (JICV), 30 Sept.-1 Oct. 2021 2021. 1-3.

ALAMRI, A. A., M.; CRISTEA, A.; PEREIRA, F. D.; OLIVEIRA, E.; SHI, L.; STEWART, C. 2019. Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

ALAMRI, A. S., Z. T.; CRISTEA, A. I.; STEWART, C.; PEREIRA, F. D. 2021. MOOC Next Week Dropout Prediction: Weekly Assessing Time and Learning Patterns. ***INTELLIGENT TUTORING SYSTEMS (ITS 2021).***

ALARCON, G. M. & EDWARDS, J. M. 2011. The relationship of engagement, job satisfaction and turnover intentions. ***Stress and Health,*** **27, e294-e298.**

ALBREIKI, B. Z., N.; ALASHWAL, H. 2021. A systematic literature review of student' performance prediction using machine learning techniques. ***Education Sciences,*** **11.**

ALDOR, E. & HELLE, D. 2021. Interweaving AutoML and Data Science Method.

ALHARBI, H. & JACOBSEN, M. 2014. A proposed framework for designing MOOCs based on the learning sciences and the first principles of instruction. ***Thannual,*** **212-220.**

ALHASSAN, Z. & NASSER, H. 2021. ***Machine Learning for Diabetes and Mortality Risk Prediction From Electronic Health Records.*** **Durham University.**

ALJOHANI, T. & CRISTEA, A. I. Training temporal and nlp features via extremely randomised trees for educational level classification. Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17, 2021. Springer, 136-147.

ALJOHANI, T. & MUSLIH, M. 2022. ***Learner Profiling: Demographics Identification Based on NLP, Machine Learning, and MOOCs Metadata.*** **Durham University.**

ALLEN, I. E. & SEAMAN, J. 2011. ***Going the distance: Online education in the United States, 2011,*** **ERIC.**

ALSHABANDAR, R. H., A.; KEIGHT, R.; LAWS, A.; BAKER, T. The Application of Gaussian Mixture Models for the Identification of At-Risk Learners in Massive Open Online Courses. 2018 IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings, 2018.

ALSOLAMI, F. J. 2020. A Hybrid Approach for Dropout Prediction of MOOC Students using Machine Learning. ***INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY,*** **20, 54-63.**

ALTINPULLUK, H. & KESIM, M. The evolution of MOOCs and a clarification of terminology through literature review. 2016 2016. 220-231.

AMANE, M., AISSAOUI, K. & BERRADA, M. Big Data in E-learning: Literature

Review and Challenges. International Conference on Advanced Intelligent Systems for Sustainable Development, 2020. Springer, 89-102.

AMIRKHAN, J. H. 1994. Criterion validity of a coping measure. *Journal of personality assessment,* 62, 242-261.

AN, S., LIU, W. & VENKATESH, S. 2007. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognition,* 40, 2154-2162.

ANCHALIA, P. P. & ROY, K. The k-nearest neighbor algorithm using MapReduce paradigm. 2014 5th International Conference on Intelligent Systems, Modelling and Simulation, 2014. IEEE, 513-518.

ANDRES-BRAY, J. M. L. 2021. *Replication in Massive Open Online Course Research Using the MOOC Replication Framework.* University of Pennsylvania.

ARCHIBALD, D. & WORSLEY, S. 2019. The father of distance learning. *TechTrends,* 63, 100-101.

ARDCHIR, S. O., Y.; OUNACER, S.; JIHAL, H.; EL GOUMARI, M. Y.; AZOUAZI, M. 2020. Improving Prediction of MOOCs Student Dropout Using a Feature Engineering Approach. *ADVANCED INTELLIGENT SYSTEMS FOR SUSTAINABLE DEVELOPMENT (AI2SD'2019): VOL 1 - ADVANCED INTELLIGENT SYSTEMS FOR EDUCATION AND INTELLIGENT LEARNING SYSTEM.*

ARNAVUT, A., BICEN, H. & NURI, C. 2019. Students' Approaches to Massive Open Online Courses: The Case of Khan Academy. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience,* 10, 82-90.

ARROYO, I., FERGUSON, K., JOHNS, J., DRAGON, T., MEHERANIAN, H., FISHER, D., BARTO, A., MAHADEVAN, S. & WOOLF, B. P. Repairing disengagement with non-invasive interventions. AIED, 2007. 195-202.

ATAPATTU, T., FALKNER, K. & TARMAZDI, H. Topic-wise Classification of MOOC Discussions: A Visual Analytics Approach. EDM, 2016. 276-281.

ATENAS, J. 2015. Model for democratisation of the contents hosted in MOOCs. *International Journal of Educational Technology in Higher Education,* 12, 3-14.

AZAR, A. T., ELSHAZLY, H. I., HASSANIEN, A. E. & ELKORANY, A. M. 2014. A random forest classifier for lymph diseases. *Computer methods and programs in biomedicine,* 113, 465-473.

BABU, R. N. L. S. K. J. C. S. N. P. N. S. C. Implementation of learning analytics framework for MOOCs using state-of-the-art in-memory computing. 2017 5th National Conference on E-Learning & E-Learning Technologies (ELELTECH), 3-4 Aug. 2017 2017. 1-6.

BAKHARIA, A. Towards cross-domain MOOC forum post classification. Proceedings of the Third (2016) ACM Conference on Learning@ Scale, 2016. 253-256.

BALAKRISHNAN, G. & COETZEE, D. 2013. Predicting student retention in massive open online courses using hidden markov models. *Electrical Engineering and Computer Sciences University of California at Berkeley*, 53, 57-58.

BALLES, L., ZAPPELLA, G. & ARCHAMBEAU, C. 2021. Gradient-matching coresets for continual learning. *arXiv preprint arXiv:2112.05025*.

BARAK, M., WATTED, A. & HAICK, H. 2016. Motivation to learn in massive open online courses: Examining aspects of language and social engagement. *Computers & Education*, 94, 49-60.

BARBA, P. G. D., 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*, 32, 218-231.

BATOOL, S. R., J.; NISAR, M. W.; KIM, J.; MAHMOOD, T.; HUSSAIN, A. A Random Forest Students' Performance Prediction (RFSPP) Model Based on Students' Demographic Features. Proceedings of the 2021 Mohammad Ali Jinnah University International Conference on Computing, MAJICC 2021, 2021.

BELANGER, Y. & THORNTON, J. 2013. Bioelectricity: A quantitative approach Duke University's first MOOC.

BELL, R. & TIGHT, M. 1993. *Open Universities: A British Tradition?*, ERIC.

BENESTY, J., CHEN, J., HUANG, Y. & COHEN, I. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*. Springer.

BLANCO, Á. F., GARCÍA-PEÑALVO, F. J. & SEIN-ECHALUCE, M. A methodology proposal for developing adaptive cMOOC. 2013 2013. 553-558.

BOEHMKE, B. & GREENWELL, B. 2019. *Hands-on machine learning with R*, Chapman and Hall/CRC.

BORISOV, V., LEEMANN, T., SEßLER, K., HAUG, J., PAWELCZYK, M. & KASNECI, G. 2022. Deep neural networks and tabular data: A survey. ***IEEE Transactions on Neural Networks and Learning Systems*****.**

BORRELLA, I. C.-C., S.; PONCE-CUETO, E. Predict and intervene: Addressing the dropout problem in a MOOC-based program. Proceedings of the 6th 2019 ACM Conference on Learning at Scale, L@S 2019, 2019.

BOTE-LORENZO, M. L. G.-S., E. Predicting the decrease of engagement indicators in a MOOC. ACM International Conference Proceeding Series, 2017. 143-147.

BOTE-LORENZO, M. L. G.-S., E. 2018. An approach to build in situ models for the prediction of the decrease of academic engagement indicators in massive open online courses. ***Journal of Universal Computer Science,*** **24, 1052-1071.**

BOTHWELL, E. & HAVERGAL, C. 2016. Moocs can transform education–but not yet. ***Times Higher Education (THE)*****.**

BOTIČKI, I., BUDIŠĆAK, I. & HOIĆ-BOŽIĆ, N. 2008. Module for online assessment in AHyCo learning management system. ***Novi Sad J. Math,*** **38, 115-131.**

BOYER, S. & VEERAMACHANENI, K. Robust Predictive Models on MOOCs : Transferring Knowledge across Courses. EDM, 2016.

BREIMAN, L. 2001. Random forests. ***Machine learning,*** **45, 5-32.**

BRESLOW, L., PRITCHARD, D. E., DEBOER, J., STUMP, G. S., HO, A. D. & SEATON, D. T. 2013. Studying learning in the worldwide classroom research into edX's first MOOC. ***Research & Practice in Assessment,*** **8, 13-25.**

BRINTON, C. G., CHIANG, M., JAIN, S., LAM, H., LIU, Z. & WONG, F. M. F. 2014. Learning about social learning in MOOCs: From statistical analysis to generative model. ***IEEE transactions on Learning Technologies,*** **7, 346-359.**

BRODERSEN, K. H., ONG, C. S., STEPHAN, K. E. & BUHMANN, J. M. The balanced accuracy and its posterior distribution. 2010 20th international conference on pattern recognition, 2010. IEEE, 3121-3124.

BUCKLAND, M. & DYE, C. M. 1991. The development of electronic distance education delivery systems in the United States. Recurring and emerging themes in history and philosophy of education.

CAHYANA, N., KHOMSAH, S. & ARIBOWO, A. S. Improving imbalanced dataset

classification using oversampling and gradient boosting. 2019 2019. IEEE, 217-222.

CHASTNEY, R. 2019. *FutureLearn reaches 10 million learners* [Online]. Available: https://www.futurelearn.com/info/press-releases/futurelearn-reaches-10-million-learners [Accessed 10/12/2019 2019].

CHAUHAN, N. S. 2019. Decision Tree Algorithm—Explained. *Towards Data Science,* 24.

CHEN, J., FANG, B., ZHANG, H. & XUE, X. 2022. A systematic review for MOOC dropout prediction from the perspective of machine learning. *Interactive Learning Environments*, 1-14.

CHEN, J. F., J.; SUN, X.; WU, N. N.; YANG, Z. Z.; CHEN, S. S. 2019a. MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine. *MATHEMATICAL PROBLEMS IN ENGINEERING,* 2019.

CHEN, M. W., L. A dropout prediction method based on time series model in MOOCs. Journal of Physics: Conference Series, 2021.

CHEN, T. & GUESTRIN, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016. 785-794.

CHEN, W. B., C. G.; CAO, D.; MASON-SINGH, A.; LU, C.; CHIANG, M. 2019b. Early Detection Prediction of Learning Outcomes in Online Short-Courses via Learning Behaviors. *IEEE Transactions on Learning Technologies,* 12, 44-58.

CHENG, J., YUEN, A. H. & CHIU, D. K. 2022. Systematic review of MOOC research in mainland China. *Library Hi Tech*.

CHIANG, W. C. C. G. B. D. C. A. M.-S. C. L. M. 2019. Early Detection Prediction of Learning Outcomes in Online Short-Courses via Learning Behaviors. *IEEE Transactions on Learning Technologies,* 12, 44-58.

CHIRKOV, V. I. 2009. A cross-cultural analysis of autonomy in education: A self-determination theory perspective. *Theory and Research in Education,* 7, 253-262.

CHIU, Y. C. H., H. J.; WU, J.; YANG, D. L. 2018. Predicting student performance in MOOCs using learning activity data. *Journal of Information Science and Engineering,* 34, 1223-1235.

CHO, H., KIM, Y.-J., JUNG, H. J., LEE, S.-W. & LEE, J. W. 2008. OutlierD: an R package for outlier detection using quantile regression on mass spectrometry data. ***Bioinformatics,*** **24, 882-884.**

CHUA, S.-M., TAGG, C., SHARPLES, M. & RIENTIES, B. 2017a. Discussion analytics: Identifying conversations and social learners in FutureLearn MOOCs. ***MOOC analytics: live dashboards, post-hoc analytics and the long-term effects,*** **36-62.**

CHUA, S. M., TAGG, C., SHARPLES, M. & RIENTIES, B. 2017b. Discussion Analytics: Identifying Conversations and Social Learners in FutureLearn MOOCs.

CIOLACU, M., TEHRANI, A. F., BEER, R. & POPP, H. Education 4.0—Fostering student's performance with machine learning methods. 2017 2017. IEEE, 438-443.

COATES, A., CARPENTER, B., CASE, C., SATHEESH, S., SURESH, B., WANG, T., WU, D. J. & NG, A. Y. Text detection and character recognition in scene images with unsupervised feature learning. Document Analysis and Recognition (ICDAR), 2011 International Conference on, 2011. IEEE, 440-445.

COBOS, R., WILDE, A. & ZALUSKA, E. 2017. Comparing attrition prediction in FutureLearn and edX MOOCs.

COCEA, M. Learning engagement: what actions of learners could best predict it? AIED, 2007. Citeseer, 683-684.

CODISH, D., RABIN, E. & RAVID, G. 2019. User behavior pattern detection in unstructured processes–a learning management system case study. ***Interactive Learning Environments,*** **27, 699-725.**

COFFRIN, C., CORRIN, L., DE BARBA, P. & KENNEDY, G. Visualizing patterns of student engagement and performance in MOOCs. Proceedings of the fourth international conference on learning analytics and knowledge, 2014. 83-92.

COKLUK, O. 2010. Logistic Regression: Concept and Application. ***Educational Sciences: Theory and Practice,*** **10, 1397-1407.**

CRISTEA, A. I., ALSHEHRI, M., ALAMRI, A., KAYAMA, M., STEWART, C. & SHI, L. 2018. How is Learning Fluctuating? FutureLearn MOOCs Fine-Grained Temporal Analysis and Feedback to Teachers.

CRISTEA, A. I. A., A.; KAYAMA, M.; STEWART, C.; ALSHEHRI, M.; SHI, L. Earliest predictor of dropout in MOOCs: A longitudinal study of futurelearn courses. Proceedings of the 27th International Conference on Information Systems Development: Designing Digitalization, ISD 2018, 2018.

CUI, H., HUANG, D., FANG, Y., LIU, L. & HUANG, C. Webshell detection based on random forest–gradient boosting decision tree algorithm. 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), 2018. IEEE, 153-160.

CUI, Y. C., F.; SHIRI, A.; FAN, Y. 2019. Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Science,* 120, 208-227.

CUNNINGHAM, P., CORD, M. & DELANY, S. J. 2008. Supervised learning. *Machine learning techniques for multimedia.* Springer.

D'INVERNO, M. M. A.-R. M. Y.-K. M. Boolean prediction of final grades based on weekly and cumulative activities. 2017 Intelligent Systems Conference (IntelliSys), 7-8 Sept. 2017 2017. 462-469.

DALIPI, F., IMRAN, A. S. & KASTRATI, Z. MOOC dropout prediction using machine learning techniques: Review and research challenges. Global Engineering Education Conference (EDUCON), 2018 IEEE, 2018. IEEE, 1007-1014.

DASCALU, M.-D. R., STEFAN; DASCALU, MIHAI; MCNAMARA, DANIELLE S.; CARABAS, MIHAI; REBEDEA, TRAIAN; TRAUSAN-MATU, STEFAN 2021. Before and during COVID-19: A Cohesion Network Analysis of students' online participation in moodle courses. *Computers in Human Behavior,* 121, 106780.

DAVIS, D. C., G.; HAUFF, C.; HOUBEN, G. J. Gauging MOOC learners' adherence to the designed learning path. Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016, 2016. 54-61.

DAVIS, H., LEON, K. D. M., VERA, M. & WHITE, S. 2013. MOOCs for Universities and Learners. *An analysis of motivating factors*.

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,* 32, 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,* 46,

455-471.

DECI, E. L. & RYAN, R. M. 2013. ***Intrinsic motivation and self-determination in human behavior,*** **Springer Science & Business Media.**

DEEKSHATULU, B. L. & CHANDRA, P. 2013. Classification of heart disease using k-nearest neighbor and genetic algorithm. ***Procedia technology,*** **10, 85-94.**

DENG, R., BENCKENDORFF, P. & GANNAWAY, D. 2020. Learner engagement in MOOCs: Scale development and validation. ***British Journal of Educational Technology,*** **51, 245-262.**

DILLON, J., BOSCH, N., CHETLUR, M., WANIGASEKARA, N., AMBROSE, G. A., SENGUPTA, B. & D'MELLO, S. K. 2016. Student Emotion, Co-Occurrence, and Dropout in a MOOC Context. ***International Educational Data Mining Society*****.**

DING, M. H., E.; WANG, Y.; O'REILLY, U. M. Transfer learning using representation learning in massive open online courses. PervasiveHealth: Pervasive Computing Technologies for Healthcare, 2019a. 145-154.

DING, M. Y., D. Y.; YANG, K.; PONG, T. C. Effective feature learning with unsupervised learning for improving the predictive models in massive open online courses. ACM International Conference Proceeding Series, 2019b. 135-144.

DIPINTO, D. & PRINCIPI, F. 2015. THE IMPACTS OF MASSIVE, OPEN AND ONLINE COURSES ON UNIVERSITIES: AN EXPLORATORY FRAMEWORK.

DOLECK, T. L., D. J.; BASNET, R. B.; BAZELAIS, P. 2020. Predictive analytics in education: a comparison of deep learning frameworks. ***Education and Information Technologies,*** **25, 1951-1963.**

DORFMAN, R. 1979. A formula for the Gini coefficient. ***The review of economics and statistics,*** **146-149.**

DROUSIOTIS, E., PENTALIOTIS, P., SHI, L. & CRISTEA, A. I. Capturing Fairness and Uncertainty in Student Dropout Prediction–A Comparison Study. International Conference on Artificial Intelligence in Education, 2021. Springer, 139-144.

DROUSIOTIS, E. S., L.; MASKELL, S. 2021. Early Predictor for Student Success Based on Behavioural and Demographical Indicators. ***Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence***

and Lecture Notes in Bioinformatics).

DUMOUCHEL, G. 2015. University teaching versus the MOOC: Reports from the field, current trends, and future directions. *International Journal of Technologies in Higher Education,* 12, 1-2.

DUNCAN, N. J., STREVENS, C. & FIELD, R. 2020. Resilience and student wellbeing in higher education: A theoretical basis for establishing law school responsibilities for helping our students to thrive. *European Journal of Legal Education,* 1, 83-115.

DUTTA, D., PAUL, D. & GHOSH, P. Analysing feature importances for diabetes prediction using machine learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018. IEEE, 924-928.

DUTTA, R. To Find the Best-Suited Model for Sentiment Analysis of Real-Time Twitter Data. International Conference on Innovative Computing and Communications, 2021. Springer, 445-452.

EDALATI, M. I., A. S.; KASTRATI, Z.; DAUDPOTA, S. M. 2022. The Potential of Machine Learning Algorithms for Sentiment Classification of Students' Feedback on MOOC. *Lecture Notes in Networks and Systems.*

EKANAYAKE, I., MEDDAGE, D. & RATHNAYAKE, U. 2022. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials,* 16, e01059.

ELIZONDO-GARCIA, J. & GALLARDO, K. 2020. Peer Feedback in Learner‐Learner Interaction Practices. Mixed Methods Study on an xMOOC. *Electronic Journal of e-Learning,* 18, pp122-135.

ENGEL, H. A. 1936. *WHA, Wisconsin's pioneer*.

ER, E., GÓMEZ-SÁNCHEZ, E., BOTE-LORENZO, M. L., DIMITRIADIS, Y. & ASENSIO-PÉREZ, J. I. 2020. Generating actionable predictions regarding MOOC learners' engagement in peer reviews. *Behaviour & Information Technology,* 39, 1356-1373.

ER, E. G.-S., EDUARDO; BOTE-LORENZO, MIGUEL L.; DIMITRIADIS, YANNIS; ASENSIO-PÉREZ, JUAN I. 2020. Generating actionable predictions regarding MOOC learners' engagement in peer reviews. *Behaviour & Information Technology,* 39, 1356-1373.

ESSA, A. & AYAD, H. 2012. Improving student success using predictive models and data visualisations. ***Research in Learning Technology,*** **20.**

EVANS, P. 2015. Self-determination theory: An approach to motivation in music education. ***Musicae Scientiae,*** **19, 65-83.**

FAUZIATI, E. M. I. H. S. Dropout Prediction Optimization through SMOTE and Ensemble Learning. 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 5-6 Dec. 2019 2019. 516-521.

FEI, M. Y., D. Y. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015, 2016. 256-263.

FERGUSON, R., CLOW, D., BEALE, R., COOPER, A. J., MORRIS, N., BAYNE, S. & WOODGATE, A. Moving through MOOCS: Pedagogy, learning design and patterns of engagement. European Conference on Technology Enhanced Learning, 2015. Springer, 70-84.

FERREIRA, D. R. 2017. ***A primer on process mining: Practical skills with python and graphviz,*** **Springer.**

FIGUEIREDO FILHO, D. B., PARANHOS, R., ROCHA, E. C. D., BATISTA, M., SILVA JR, J. A. D., SANTOS, M. L. W. D. & MARINO, J. G. 2013. When is statistical significance not significant? ***Brazilian Political Science Review,*** **7, 31-55.**

FLEISS, J. L., LEVIN, B. & PAIK, M. C. 1981. The measurement of interrater agreement. ***Statistical methods for rates and proportions,*** **2, 22-23.**

FREUND, Y. & SCHAPIRE, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. ***Journal of computer and system sciences,*** **55, 119-139.**

FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine. ***Annals of statistics,*** **1189-1232.**

FU, Q. G., Z.; ZHOU, J.; ZHENG, Y. 2021. CLSA: A novel deep learning model for MOOC dropout prediction. ***Computers and Electrical Engineering,*** **94.**

FU, Y. Z. Z. G. Y. W. Q. 2020. MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series. ***IEEE Access,*** **8, 225324-225335.**

GALLAGHER, P. S. 2008. ***Assessing SCORM 2004 for its Affordances in Facilitating a***

***Simulation as a Pedagogical Model*, George Mason University.**

GAMAGE, D., PERERA, I. & FERNANDO, S. Evaluating effectiveness of MOOCs using empirical tools: Learners' perspective. 2016 2016.

GANGIRE, Y., DA VEIGA, A. & HERSELMAN, M. 2021. Assessing information security behaviour: A self-determination theory perspective. *Information & Computer Security*.

GARDNER, J. & BROOKS, C. 2018. Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction,* 28, 127-203.

GARDNER, M. W. & DORLING, S. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment,* 32, 2627-2636.

GERBER, Z. & ANAKI, D. 2021. The role of self-compassion, concern for others, and basic psychological needs in the reduction of caregiving burnout. *Mindfulness,* 12, 741-750.

GÉRON, A. 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, O'Reilly Media.

GETOOR, A. R. D. G. B. H. H. D. L. 2020. Interpretable Engagement Models for MOOCs Using Hinge-Loss Markov Random Fields. *IEEE Transactions on Learning Technologies,* 13, 107-122.

GEURTS, P., ERNST, D. & WEHENKEL, L. 2006. Extremely randomized trees. *Machine learning,* 63, 3-42.

GHAZNAVI, M. R., KEIKHA, A. & YAGHOUBI, N.-M. 2011. The Impact of Information and Communication Technology (ICT) on Educational Improvement. *International Education Studies,* 4, 116-125.

GITINABARD, N. K., F.; LYNCH, C. F.; WANG, E. Y. Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018, 2018.

GOEL, Y. G., R. 2020. On the Effectiveness of Self-Training in MOOC Dropout Prediction. *Open Computer Science,* 10, 246-258.

GREENBERG, G. 1998. Distance education technologies: Best practices for K-12 settings. *IEEE Technology and Society Magazine,* 17, 36-40.

GREENE, J. A., OSWALD, C. A. & POMERANTZ, J. 2015. Predictors of retention and achievement in a massive open online course. ***American Educational Research Journal,*** **52, 925-955.**

GRYLLOS, P., MAKRIS, C. & VIKATOS, P. Marketing campaign targeting using bridge extraction. Proceedings of the Symposium on Applied Computing, 2017. 1045-1052.

GUJJAR, J. P. & HR, P. K. 2021. Sentiment analysis: Textblob for decision making. ***Int. J. Sci. Res. Eng. Trends,*** **7, 1097-1099.**

GUO, P. J., KIM, J. & RUBIN, R. How video production affects student engagement: An empirical study of MOOC videos. Proceedings of the first ACM conference on Learning@ scale conference, 2014. 41-50.

GUO, P. J. & REINECKE, K. Demographic differences in how students navigate through MOOCs. Proceedings of the first ACM conference on Learning@ scale conference, 2014. 21-30.

HADWAN, M., AL-SAREM, M., SAEED, F. & AL-HAGERY, M. A. 2022. An Improved Sentiment Classification Approach for Measuring User Satisfaction toward Governmental Services' Mobile Apps Using Machine Learning Methods with Feature Engineering and SMOTE Technique. ***Applied Sciences,*** **12, 5547.**

HAFEZ, M. M., SHEHAB, M. E., EL FAKHARANY, E. & ABDEL GHFAR HEGAZY, A. E. F. Effective selection of machine learning algorithms for big data analytics using apache spark. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016 2, 2017. Springer, 692-704.

HAGGER, M. S. & HAMILTON, K. 2021. General causality orientations in self-determination theory: Meta-analysis and test of a process model. ***European Journal of Personality,*** **35, 710-735.**

HAIR, J. F., RINGLE, C. M. & SARSTEDT, M. 2011. PLS-SEM: Indeed a silver bullet. ***Journal of Marketing theory and Practice,*** **19, 139-152.**

HALL, M. A. & SMITH, L. A. 1998. Practical feature subset selection for machine learning.

HARASIM, L. 2000. Shift happens: Online education as a new paradigm in learning. ***The Internet and higher education,*** **3, 41-61.**

HART, P. 1968. The condensed nearest neighbor rule (corresp.). ***IEEE transactions on***

information theory, **14, 515-516.**

HASSAN, S. U. W., H.; ALJOHANI, N. R.; ALI, M.; VENTURA, S.; HERRERA, F. 2019. Virtual learning environment to predict withdrawal by leveraging deep learning. ***International Journal of Intelligent Systems,*** **34, 1935-1952.**

HASTIE, T., ROSSET, S., ZHU, J. & ZOU, H. 2009. Multi-class adaboost. ***Statistics and its Interface,*** **2, 349-360.**

HAWKINS, D. M. 1980. ***Identification of outliers,*** **Springer.**

HE, Y. C., R.; LI, X.; HAO, C.; LIU, S.; ZHANG, G.; JIANG, B. 2020. Online at-risk student identification using RNN-GRU joint neural networks. ***Information (Switzerland),*** **11, 1-11.**

HEW, K. F. 2016. Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS. ***British Journal of Educational Technology,*** **47, 320-341.**

HLIOUI, F. A., N.; GARGOURI, F. 2021. A withdrawal prediction model of at-risk learners based on behavioural indicators. ***International Journal of Web-Based Learning and Teaching Technologies,*** **16, 32-53.**

HOFER, J. & BUSCH, H. 2011. Satisfying one's needs for competence and relatedness: Consequent domain-specific well-being depends on strength of implicit motives. ***Personality and Social Psychology Bulletin,*** **37, 1147-1158.**

HOLMBERG, B. 2005. ***Theory and practice of distance education,*** **Routledge.**

HONE, K. S. & EL SAID, G. R. 2016. Exploring the factors affecting MOOC retention: A survey study. ***Computers & Education,*** **98, 157-168.**

HONG, B. W. W., Z. Q.; YANG, Y. Q.; IEEE, 2017. Discovering Learning Behavior Patterns to Predict Dropout in MOOC. ***2017 12TH INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND EDUCATION (ICCSE 2017).***

HU, W., HU, W. & MAYBANK, S. 2008. Adaboost-based algorithm for network intrusion detection. ***IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),*** **38, 577-583.**

HUNG, J.-L., WANG, M. C., WANG, S., ABDELRASOUL, M., LI, Y. & HE, W. 2015. Identifying at-risk students for early interventions—A time-series clustering approach. ***IEEE Transactions on Emerging Topics in Computing,*** **5, 45-55.**

IMRAN, A. S. D., F.; KASTRATI, Z.; ASSOC COMP, MACHINERY 2019. Predicting Student Dropout in a MOOC: An Evaluation of a Deep Neural Network Model. ***ICCAI '19 - PROCEEDINGS OF THE 2019 5TH INTERNATIONAL CONFERENCE ON COMPUTING AND ARTIFICIAL INTELLIGENCE.***

IPAYE, B. 2013. Opportunities and Challenges for Open Educational Resources and Massive Open Online Courses: The Case of Nigeria. ***Commonwealth of Learning*****.**

ITANI, A. B., L.; GARLATTI, S. 2018. Understanding Learner's Drop-Out in MOOCs. ***INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING - IDEAL 2018, PT I.***

JACKSON, G. T., GRAESSER, A. C. & MCNAMARA, D. S. What Students Expect May Have More Impact Than What They Know or Feel. AIED, 2009. 73-80.

JHA, N. I. G., I.; MOLDOVAN, A. N. OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques. CSEDU 2019 - Proceedings of the 11th International Conference on Computer Supported Education, 2019. 154-164.

JIANG, S., WILLIAMS, A., SCHENKE, K., WARSCHAUER, M. & O'DOWD, D. Predicting MOOC performance with week 1 behavior. Educational data mining 2014, 2014.

JIN, C. 2020. MOOC student dropout prediction model based on learning behavior features and parameter optimization. ***Interactive Learning Environments*****, 1-19.**

JIN, C. 2021. Dropout prediction model in MOOC based on clickstream data and student sample weight. ***SOFT COMPUTING,*** **25, 8971-8988.**

JOHNSON, R. B. & CHRISTENSEN, L. 2019. ***Educational research: Quantitative, qualitative, and mixed approaches*****, Sage publications.**

JOKSIMOVIĆ, S., POQUET, O., KOVANOVIĆ, V., DOWELL, N., MILLS, C., GAŠEVIĆ, D., DAWSON, S., GRAESSER, A. C. & BROOKS, C. 2018. How do we model learning at scale? A systematic review of research on MOOCs. ***Review of Educational Research,*** **88, 43-86.**

JORDAN, K. 2015. Massive open online course completion rates revisited: Assessment, length and attrition. ***International Review of Research in Open and Distributed Learning,*** **16, 341-358.**

KAMEAS, G. K. T. P. S. K. C. P. A. 2021. Interpretable Models for Early Prediction of Certification in MOOCs: A Case Study on a MOOC for Smart City Professionals. ***IEEE Access,*** **9, 165881-165891.**

KASHYAP, A. N., A. Different Machine Learning Models to Predict Dropouts in MOOCs. 2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018, 2018. 80-85.

KEARSLEY, G. & SHNEIDERMAN, B. 1998. Engagement theory: A framework for technology-based teaching and learning. ***Educational technology,*** **38, 20-23.**

KENTNOR, H. E. 2015. Distance education and the evolution of online learning in the United States. ***Curriculum and teaching dialogue,*** **17, 21-34.**

KHALIL, H. & EBNER, M. 2014. MOOCs completion rates and possible methods to improve retention-A literature review. ***EdMedia+ innovate learning,*** **1305-1313.**

KHAN, M. A. A. H. J. A. S. M. A. A. R. M. A. M. B. S. U. 2021. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. ***IEEE Access,*** **9, 7519-7539.**

KHAN, M. K., SALLAM, M. H., PRATT, C. B. & FARID, T. 2022. Stereotyping the Chinese in Arab Nations: Effects of media use, perceived realism, and perceived Chinese-Arab relations. ***The Social Science Journal,*** **1-21.**

KHODEIR, N. A. 2021. Bi-GRU Urgent Classification for MOOC Discussion Forums Based on BERT. ***IEEE Access,*** **9, 58243-58255.**

KIM, B. H. V., E.; GANAPATHI, V. GritNet: Student performance prediction with deep learning. Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018, 2018.

KJÆRGAARD, H. W., KJELDSEN, L. P. B., JELSBAK, V. A. & BENDSEN, T. 2013. MOOCs-perspektiver for UC-sektoren i Danmark. ***Tidsskriftet Læring og Medier (LOM),*** **6.**

KLOFT, M., STIEHLER, F., ZHENG, Z. & PINKWART, N. Predicting MOOC dropout over weeks using machine learning methods. Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 2014. 60-65.

KLOOS, P. M. M.-M. C. A.-H. P. J. M.-M. I. E.-A. C. D. Sentiment analysis in MOOCs: A case study. 2018 IEEE Global Engineering Education Conference (EDUCON), 17-20 April 2018 2018. 1489-1496.

KODIYAN, D., HARDEGGER, F., NEUHAUS, S. & CIELIEBAK, M. Author Profiling with bidirectional rnns using Attention with grus: Notebook for PAN at CLEF 2017. CLEF 2017 Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11-14 September 2017, 2017. RWTH Aachen.

KOENIG, A. E. 1969. The development of educational television. University of Wisconsin Press Madison.

KOENIG, A. E. & HILL, R. B. 1967. The farther vision: Educational television today.

KOLLER, D., NG, A., DO, C. & CHEN, Z. 2013. Retention and intention in massive open online courses: In depth. *Educause review,* 48, 62-63.

KONONENKO, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine,* 23, 89-109.

KŐRÖSI, G. F., R. 2020. MOOC Performance Prediction by Deep Learning from Raw Clickstream Data. *Communications in Computer and Information Science.*

KUH, G. D. 2003. What we're learning about student engagement from NSSE: Benchmarks for effective educational practices. *Change: The magazine of higher learning,* 35, 24-32.

KULSHRESTHA, T. & KANT, A. R. 2013. Benefits of learning management system (LMS) in Indian education. *International Journal of Computer Science & Engineering Technology (IJCSET),* 4, 1153-1154.

KUMAR, K. S., SEMWAL, V. B. & TRIPATHI, R. C. 2011. Real time face recognition using adaboost improved fast PCA algorithm. *arXiv preprint arXiv:1108.1353*.

LACKNER, E., KOPP, M. & EBNER, M. How to MOOC?–A pedagogical guideline for practitioners. 2014 2014.

LAI, S. Z., Y. X.; YANG, Y. Q.; ASSOC COMP, MACHINERY 2020. Broad Learning System for Predicting Student Dropout in Massive Open Online Courses. *ICIET 2020: 2020 8TH INTERNATIONAL CONFERENCE ON INFORMATION AND EDUCATION TECHNOLOGY.*

LAN, M. & HEW, K. F. 2020. Examining learning engagement in MOOCs: A self-determination theoretical perspective using mixed method. *International Journal of Educational Technology in Higher Education,* 17, 1-24.

LANGDON, J., WEBSTER, C., HALL, T. & MONSMA, E. 2014. A self-

determination theory perspective of student performance at the end of a volleyball unit in compulsory high school physical education. ***Sport Scientific & Practical Aspects,*** **11.**

LAVIGNE, G. L., VALLERAND, R. J. & MIQUELON, P. 2007. A motivational model of persistence in science education: A self-determination theory approach. ***European Journal of Psychology of Education,*** **22, 351-369.**

LEMAY, D. J. D., TENZIN 2020. Predicting completion of massive open online course (MOOC) assignments from video viewing behavior. ***Interactive Learning Environments,*** **1-12.**

LI, B., YU, Q. & PENG, L. 2022. Ensemble of fast learning stochastic gradient boosting. ***Communications in Statistics-Simulation and Computation,*** **51, 40-52.**

LI, X., WANG, T. & WANG, H. Exploring n-gram features in clickstream data for MOOC learning achievement prediction. International Conference on Database Systems for Advanced Applications, 2017. Springer, 328-339.

LI, Y. F., C.; ZHANG, Y. When and who at risk? Call back at these critical points. Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017, 2017. 168-173.

LI, Y. J. 2018. Feature Extraction and Learning Effect Analysis for MOOCs Users Based on Data Mining. ***INTERNATIONAL JOURNAL OF EMERGING TECHNOLOGIES IN LEARNING,*** **13, 108-120.**

LIANG, J., LI, C. & ZHENG, L. Machine learning application in MOOCs: Dropout prediction. 2016 11th International Conference on Computer Science & Education (ICCSE), 2016. IEEE, 52-57.

LIANG, J. L., C.; ZHENG, L. Machine learning application in MOOCs: Dropout prediction. ICCSE 2016 - 11th International Conference on Computer Science and Education, 2016. 52-57.

LIAW, A. & WIENER, M. 2002. Classification and regression by randomForest. ***R news,*** **2, 18-22.**

LIÑÁN, L. C. & PÉREZ, Á. A. J. 2015. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. ***International Journal of Educational Technology in Higher Education,*** **12, 98-112.**

LITTLEJOHN, A., HOOD, N., MILLIGAN, C. & MUSTAIN, P. 2016. Learning in MOOCs: Motivations and self-regulated learning in MOOCs. ***The internet and higher education,*** **29, 40-48.**

LIU, K. T., S.; KHONG, A. W. H. 2020a. A Weighted Feature Fixtraction Technique Based on Temporal Accumulation of Learner Behavior Features for Farly Prediction of Dropouts. ***PROCEEDINGS OF 2020 IEEE INTERNATIONAL CONFERENCE ON TEACHING, ASSESSMENT, AND LEARNING FOR ENGINEERING (IEEE TALE 2020).***

LIU, L. Q. Y. L. Y. 2018. An Integrated Framework With Feature Selection for Dropout Prediction in Massive Open Online Courses. ***IEEE Access,*** **6, 71474-71484.**

LIU, Y. W. Y. T. B. W. Q. Z. G. C. S. 2020b. Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs. ***Tsinghua Science and Technology,*** **25, 336-347.**

LIU, Y. Z. L. C. T. MOOCs Dropout Prediction Based on Hybrid Deep Neural Network. 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 29-30 Oct. 2020 2020c. 197-203.

LIVINGSTON, F. 2005. Implementation of Breiman's random forest machine learning algorithm. ***ECE591Q Machine Learning Journal Paper,*** **1-13.**

LIYANAGUNAWARDENA, T. R., LUNDQVIST, K. Ø. & WILLIAMS, S. A. 2015. Who are with us: MOOC learners on a F uture L earn course. ***British Journal of Educational Technology,*** **46, 557-569.**

LOHAR, P., XIE, G., BENDECHACHE, M., BRENNAN, R., CELESTE, E., TRESTIAN, R. & TAL, I. Irish attitudes toward COVID tracker app & privacy: sentiment analysis on Twitter and survey data. The 16th International Conference on Availability, Reliability and Security, 2021. 1-8.

LOHSE, J. J., MCMANUS, C. A. & JOYNER, D. A. Surveying the MOOC Data Set Universe. 2019 2019. IEEE, 159-164.

LOIZZO, J., ERTMER, P. A., WATSON, W. R. & WATSON, S. L. 2017. Adult MOOC Learners as Self-Directed: Perceptions of Motivation, Success, and Completion. ***Online Learning,*** **21, n2.**

LONG, P. 2011. ***LAK'11: Proceedings of the 1st International Conference on Learning Analytics and Knowledge, February 27-March 1, 2011, Banff, Alberta, Canada,*** **ACM.**

LU, X. W., S.; HUANG, J.; CHEN, W.; YAN, Z. 2017. What decides the dropout in MOOCS? ***Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).***

MACKENZIE, O. & CHRISTENSEN, E. L. 1971. The Changing World of Correspondence Study: International Readings.

MAHESH, B. 2020. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet],* 9, 381-386.

MARCZEWSKI, A. 2015. Even Ninja Monkeys like to play. *London: Blurb Inc,* 1, 28.

MARTÍN-MONJE, E., CASTRILLO, M. D. & MAÑANA-RODRÍGUEZ, J. 2018. Understanding online interaction in language MOOCs through learning analytics. *Computer Assisted Language Learning,* 31, 251-272.

MASSIMIANI, M., LACKO, L. A., SWANSON, C. S. B., SALVI, S., ARGUETA, L. B., MORESI, S., FERRAZZANI, S., GELBER, S. E., BAERGEN, R. N. & TOSCHI, N. 2019. Increased circulating levels of Epidermal Growth Factor-like Domain 7 in pregnant women affected by preeclampsia. *Translational Research,* 207, 19-29.

MAZINI, M., SHIRAZI, B. & MAHDAVI, I. 2019. Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms. *Journal of King Saud University-Computer and Information Sciences,* 31, 541-553.

MAZRAEH, S., GHANAVATI, M. & NEYSI, S. H. N. 2019. Intrusion detection system with decision tree and combine method algorithm. *International Academic Journal of Science and Engineering,* 6, 167-177.

MCDONALD, R. A., HAND, D. J. & ECKLEY, I. A. An empirical comparison of three boosting algorithms on real data sets with artificial class noise. 2003 2003. Springer, 35-44.

MIHALEC-ADKINS, B., HICKS, N., DOUGLAS, K. A., DIEFES-DUX, H., BERMEL, P. & MADHAVAN, K. Surveying the motivations of groups of learners in highly-technical STEM MOOCs. 2016 IEEE Frontiers in Education Conference (FIE), 2016. IEEE, 1-6.

MISHINA, Y., MURATA, R., YAMAUCHI, Y., YAMASHITA, T. & FUJIYOSHI, H. 2015. Boosted random forest. *IEICE TRANSACTIONS on Information and Systems,* 98, 1630-1636.

MOHAMED, M. H. & HAMMOND, M. 2018. MOOCs: a differentiation by pedagogy, content and assessment. *The International Journal of Information and Learning Technology*.

MOHAMED, U. U. E. & SALLEH, N. Measuring The Success of Massive Open

Online Courses: A Mixed-Method Case Study. 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2021. IEEE, 1-5.

MOLNAR, C. 2020. *Interpretable machine learning*, Lulu. com.

MONLLAÓ OLIVÉ, D. H., D. Q.; REYNOLDS, M.; DOUGIAMAS, M.; WIESE, D. 2020. A supervised learning framework: using assessment to identify students at risk of dropping out of a MOOC. *Journal of Computing in Higher Education,* 32, 9-26.

MOORE, M. G. & KEARSLEY, G. 1996. *Distance education: A system view*, Wadsworth.

MORENO-MARCOS, P. M., ALARIO-HOYOS, C., MUÑOZ-MERINO, P. J., ESTÉVEZ-AYRES, I. & KLOOS, C. D. Sentiment analysis in MOOCs: A case study. 2018 IEEE Global Engineering Education Conference (EDUCON), 2018. IEEE, 1489-1496.

MORENO-MARCOS, P. M., PONG, T.-C., MUNOZ-MERINO, P. J. & KLOOS, C. D. 2020. Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access,* 8, 5264-5282.

MORENO-MARCOS, P. M. A.-H., C.; MUNOZ-MERINO, P. J.; ESTEVEZ-AYRES, I.; KLOOS, C. D. Sentiment analysis in MOOCs: A case study. IEEE Global Engineering Education Conference, EDUCON, 2018a. 1489-1496.

MORENO-MARCOS, P. M. A.-H., C.; MUNOZ-MERINO, P. J.; KLOOS, C. D. 2019. Prediction in MOOCs: A Review and Future Research Directions. *IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES,* 12, 384-401.

MORENO-MARCOS, P. M. M.-M., P. J.; MALDONADO-MAHAUAD, J.; PÉREZ-SANAGUSTÍN, M.; ALARIO-HOYOS, C.; DELGADO KLOOS, C. 2020a. Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs. *Computers and Education,* 145.

MORENO-MARCOS, P. M. M.-M., PEDRO J.; ALARIO-HOYOS, CARLOS; ESTÉVEZ-AYRES, IRIA; DELGADO KLOOS, CARLOS 2018b. Analysing the predictive power for anticipating assignment grades in a massive open online course. *Behaviour & Information Technology,* 37, 1021-1036.

MORENO-MARCOS, P. M. P., T. C.; MUNOZ-MERINO, P. J.; KLOOS, C. D. 2020b. Analysis of the Factors Influencing Learners' Performance Prediction with Learning Analytics. *IEEE Access,* 8, 5264-5282.

MORENO-MURCIA, J. A., GIMENO, E. C., HERNÁNDEZ, E. H., BELAN-DO PEDREÑO, N. & MARÍN, J. J. R. 2013. Motivational profiles in physical education and their relation to the Theory of Planned Behavior. ***Journal of sports science & medicine,*** **12, 551.**

MORRIS, N. P., SWINNERTON, B. & HOTCHKISS, S. Can demographic information predict MOOC learner outcomes? Experience track: proceedings of the European MOOC stakeholder, 2015. Leeds.

MOURDI, Y. S., M.; EL KABTANE, H.; EL ABDALLAOUI, H. E. A. A Multi-Layers Perceptron for predicting weekly learner commitment in MOOCs. Journal of Physics: Conference Series, 2021.

MRHAR, K. D., O.; ABIK, M. 2020. A dropout predictor system in moocs based on neural networks. ***Journal of Automation, Mobile Robotics and Intelligent Systems,*** **14, 72-80.**

MUBARAK, A. A. C., H.; AHMED, S. A. M. 2021a. Predictive learning analytics using deep learning model in MOOCs' courses videos. ***Education and Information Technologies,*** **26, 371-392.**

MUBARAK, A. A. C., H.; ZHANG, W.; ZHANG, W. 2021b. Visual analytics of video-clickstream data and prediction of learners' performance using deep learning models in MOOCs' courses. ***Computer Applications in Engineering Education,*** **29, 710-732.**

MUBARAK, A. A. C., HAN; HEZAM, IBRAHIM M. 2021c. Deep analytic model for student dropout prediction in massive open online courses. ***Computers & Electrical Engineering,*** **93, 107271.**

MUBARAK, A. A. C., HAN; ZHANG, WEIZHEN 2020. Prediction of students' early dropout based on their interaction logs in online learning environment. ***Interactive Learning Environments,*** **1-20.**

MUHAMMAD, I. & YAN, Z. 2015. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. ***ICTACT Journal on Soft Computing,*** **5.**

MURRAY, B. 2001. What makes students stay? concern over quitters has online programs stepping up retention strategies. ***eLearn,*** **2001, 1.**

NAM, S., FRISHKOFF, G. & COLLINS-THOMPSON, K. 2017. Predicting students disengaged behaviors in an online meaning-generation task. ***IEEE Transactions on Learning Technologies,*** **11, 362-375.**

NARAYANASAMY, S. K. E., A. 2020. An Effective Prediction Model for Online

Course Dropout Rate. ***INTERNATIONAL JOURNAL OF DISTANCE EDUCATION TECHNOLOGIES,*** **18, 94-110.**

NIKOLAOU, N., EDAKUNNI, N., KULL, M., FLACH, P. & BROWN, G. 2016. Cost-sensitive boosting algorithms: Do we really need them? ***Machine Learning,*** **104, 359-384.**

NITTA, I. I., R.; SHINGU, M.; NAKASHIMA, S.; MARUHASHI, K.; TOLMACHEV, A.; TODORIKI, M. Graph-based massive open online course (MOOC) Dropout prediction using clickstream data in virtual learning environment. ICCSE 2021 - IEEE 16th International Conference on Computer Science and Education, 2021. 48-52.

NIU, Z. L., W.; YAN, X.; WU, N. Exploring causes for the dropout on massive open online courses. ACM International Conference Proceeding Series, 2018. 47-52.

O'BRIEN, H. L. & TOMS, E. G. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. ***Journal of the American society for Information Science and Technology,*** **59, 938-955.**

OKEREKE, J. E. M. F. B. B. R. N. G. Algorithms for the Development of Deep Learning Models for Classification and Prediction of Behaviour in MOOCS. 2020 IEEE Learning With MOOCS (LWMOOCS), 29 Sept.-2 Oct. 2020 2020. 180-184.

OLIVA-CORDOVA, L. M., GARCIA-CABOT, A. & AMADO-SALVATIERRA, H. R. 2021. Learning analytics to support teaching skills: a systematic literature review. ***IEEE Access,*** **9, 58351-58363.**

OLMOS, R. C. L. A Learning Analytics Tool for Predictive Modeling of Dropout and Certificate Acquisition on MOOCs for Professional Learning. 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 16-19 Dec. 2018 2018. 1533-1537.

ORTIGOSA, C. I. R. M. C. A. Dropout Detection in MOOCs: An Exploratory Analysis. 2018 International Symposium on Computers in Education (SIIE), 19-21 Sept. 2018 2018. 1-6.

OSHIRO, T. M., PEREZ, P. S. & BARANAUSKAS, J. A. How many trees in a random forest? , 2012 2012. Springer, 154-168.

OYELADE, O., OLADIPUPO, O. O. & OBAGBUWA, I. C. 2010. Application of k Means Clustering algorithm for prediction of Students Academic Performance. ***arXiv preprint arXiv:1002.2425*****.**

ÖZDAŞ, M. B., UYSAL, F. & HARDALAÇ, F. 2022. Retina Disease Classification in Optical Coherence Tomography Images Using Machine Learning and Firefly Algorithm. ***Available at SSRN 4240403*****.**

PANAGIOTAKOPOULOS, T. K., S.; BOROTIS, S.; LAZARINIS, F.; KAMEAS, A. 2021. Applying Machine Learning to Predict Whether Learners Will Start a MOOC After Initial Registration. ***IFIP Advances in Information and Communication Technology.***

PAPPAS, C. 2020. The best learning management systems (2020 update). ***Retrieved from eLearning Industry: https://elearningindustry. com/the-best-learning-management-systems-top-list.***

PARDO, A., HAN, F. & ELLIS, R. A. 2016. Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. ***IEEE Transactions on Learning Technologies,*** **10, 82-92.**

PEI, B. X., W. 2021. An Interpretable Pipeline for Identifying At-Risk Students. ***Journal of Educational Computing Research*****.**

PENG, X. & XU, Q. 2020. Investigating learners' behaviors and discourse content in MOOC course reviews. ***Computers & Education,*** **143, 103673.**

PERIWAL, N. R., K. An empirical comparison of models for dropout prophecy in MOOCs. Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2017, 2017. 906-911.

PHILLIPS, J. M. 2005. Strategies for active learning in online continuing education. ***The Journal of Continuing Education in Nursing,*** **36, 77-83.**

PILLI, O. & ADMIRAAL, W. F. 2016. A taxonomy for massive open online courses. ***Contemporary Educational Technology,*** **7, 223-240.**

PINK, D. H. 2011. ***Drive: The surprising truth about what motivates us,*** **Penguin.**

PRENKAJ, B., VELARDI, P., STILO, G., DISTANTE, D. & FARALLI, S. 2020. A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. ***ACM Computing Surveys,*** **53, 57-57:34.**

PRENKAJ, B. D., D.; FARALLI, S.; VELARDI, P. 2021. Hidden space deep sequential risk prediction on student trajectories. ***Future Generation Computer Systems,*** **125, 532-543.**

PULIKOTTIL, S. C. G., M. ONet - A Temporal Meta Embedding Network for

MOOC Dropout Prediction. Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020, 2020. 5209-5217.

QI, Q. L., Y.; WU, F.; YAN, X.; WU, N. Temporal models for personalized grade prediction in massive open online courses. ACM International Conference Proceeding Series, 2018. 67-72.

QIN, Y. 2021. Attractiveness of game elements, presence, and enjoyment of mobile augmented reality games: The case of Pokémon Go. *Telematics and Informatics,* 62, 101620.

QIU, J., TANG, J., LIU, T. X., GONG, J., ZHANG, C., ZHANG, Q. & XUE, Y. Modeling and predicting learning behavior in MOOCs. Proceedings of the ninth ACM international conference on web search and data mining, 2016. 93-102.

QIU, L. L., Y.; HU, Q.; LIU, Y. 2019. Student dropout prediction in massive open online courses by convolutional neural networks. *Soft Computing,* 23, 10287-10301.

QU, S. L., K.; FAN, Z.; WU, S.; LIU, X.; HUANG, Z. 2021. Behavior Pattern based Performance Evaluation in MOOCs. *Advances in Intelligent Systems and Computing.*

QU, S. L., K.; WU, B.; ZHANG, S.; WANG, Y. 2019. Predicting student achievement based on temporal learning behavior in MOOCs. *Applied Sciences (Switzerland),* 9.

RADI, R. A.-S. A. H. A. L. R. K. J. L. N. Machine learning approaches to predict learning outcomes in Massive open online courses. 2017 International Joint Conference on Neural Networks (IJCNN), 14-19 May 2017 2017. 713-720.

RADOVANOVIĆ, S. D., B.; SUKNOVIĆ, M. 2021. Predicting dropout in online learning environments. *Computer Science and Information Systems,* 18, 957-978.

RAJ, N. S. P., S.; HARISH, P.; BOBAN, M.; CHERIYEDATH, N. 2021. Early Prediction of At-Risk Students in a Virtual Learning Environment Using Deep Learning Techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

RAMBO-HERNANDEZ, K. E. & WARNE, R. T. 2015. Measuring the outliers: An introduction to out-of-level testing with high-achieving students. *Teaching Exceptional Children,* 47, 199-207.

RANA, S. & GARG, R. Application of hierarchical clustering algorithm to evaluate students performance of an institute. 2016 second international conference on computational intelligence & communication technology (CICT), 2016. IEEE, 692-697.

RASOULI, A., YAU, T., ROHANI, M. & LUO, J. Multi-modal hybrid architecture for pedestrian action prediction. 2022 IEEE Intelligent Vehicles Symposium (IV), 2022. IEEE, 91-97.

RAWAT, S. K., D.; KUMAR, P.; KHATTRI, C. 2021. A systematic analysis using classification machine learning algorithms to understand why learners drop out of MOOCs. *NEURAL COMPUTING & APPLICATIONS*, 33, 14823-14835.

RAWLINGS, J. O., PANTULA, S. G. & DICKEY, D. A. 1998. *Applied regression analysis: a research tool*, Springer.

REN, Y. H., S.; ZHOU, Y. Deep learning and integrated learning for predicting student's withdrawal behavior in MOOC. Proceedings - 2021 2nd International Conference on Education, Knowledge and Information Management, ICEKIM 2021, 2021. 81-84.

ROBINSON, C. Y., M.; REICH, J.; HULLEMAN, C.; GEHLBACH, H. Forecasting student achievement in MOOCs with natural language processing. ACM International Conference Proceeding Series, 2016. 383-387.

ROJAS, R. 2009. AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep*.

ROMERO, C. & VENTURA, S. 2020. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10, e1355.

RONG, X. L. H. Z. Y. O. X. Z. W. A Shallow BERT-CNN Model for Sentiment Analysis on MOOCs Comments. 2019 IEEE International Conference on Engineering, Technology and Education (TALE), 10-13 Dec. 2019 2019. 1-6.

ROSE, C. & SIEMENS, G. Shared task on prediction of dropout over time in massively open online courses. Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 2014. 39-41.

ROSÉ, C. P., CARLSON, R., YANG, D., WEN, M., RESNICK, L., GOLDMAN, P. & SHERER, J. Social factors that contribute to attrition in MOOCs. Proceedings of the first ACM conference on Learning@ scale conference, 2014. 197-198.

SAADATDOOST, R. S., ALEX TZE HIANG; JAFARKARIMI, HOSEIN; MEI HEE, JEE 2015. Exploring MOOC from education and Information Systems perspectives: a short literature review. ***Educational Review,*** **67, 505-518.**

SAHA, S., YADAV, J. & RANJAN, P. 2017. Proposed approach for sarcasm detection in twitter. ***Indian Journal of Science and Technology,*** **10, 1-8.**

ŞAHIN, M. 2021. A Comparative Analysis of Dropout Prediction in Massive Open Online Courses. ***Arabian Journal for Science and Engineering,*** **46, 1845-1861.**

SANCHEZ-GORDON, S. & LUJÁN-MORA, S. 2014. MOOCs gone wild. 2014 2014. 1449-1458.

SANTOS, J. L., KLERKX, J., DUVAL, E., GAGO, D. & RODRÍGUEZ, L. Success, activity and drop-outs in MOOCs an exploratory study on the UNED COMA courses. Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, 2014. 98-102.

SCHAPIRE, R. E. A brief introduction to boosting. 1999 1999. Citeseer, 1401-1406.

SEBBAQ, H. & EL FADDOULI, N.-E. 2022. Fine-Tuned BERT Model for Large Scale and Cognitive Classification of MOOCs. ***International Review of Research in Open and Distributed Learning,*** **23, 170-190.**

SEHABA, K. Learner performance prediction indicators based on machine learning. CSEDU 2020 - Proceedings of the 12th International Conference on Computer Supported Education, 2020. 47-57.

SHANG, L.-H., LUO, A.-L., WANG, L., QIN, L., DU, B., HE, X.-J., CUI, X.-Q., ZHAO, Y.-H., ZHU, R.-H. & ZHI, Q.-J. 2022. Objective Separation between CP1 and CP2 Based on Feature Extraction with Machine Learning. ***The Astrophysical Journal Supplement Series,*** **259, 63.**

SHEA, R. H. 2002. E-learning today. ***US News & World Report,*** **28, 54-56.**

SHELTON, K. & SALTSMAN, G. 2005. An administrator's guide to online education. Greenwich, CT. Information Age Publishing, Inc.

SHEN, B., MCCAUGHTRY, N., MARTIN, J. & FAHLMAN, M. 2009. Effects of teacher autonomy support and students' autonomous motivation on learning in physical education. ***Research Quarterly for Exercise and Sport,*** **80, 44-53.**

SHEN, L. & BAI, L. Adaboost gabor feature selection for classification. 2004 2004. Citeseer, 77-83.

SHENG, D. Y., J.; ZHANG, X. How MOOC videos affect dropout? A lightweight pipeline making student dropout interpretable from several levels. Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE, 2021. 189-194.

SHERNOFF, D. J., RUZEK, E. A. & SINHA, S. 2017. The influence of the high school classroom environment on learning as mediated by student engagement. *School psychology international,* 38, 201-218.

SHI, L. & CRISTEA, A. I. Demographic indicators influencing learning activities in MOOCs: learning analytics of FutureLearn courses. 2018a. Association for Information Systems.

SHI, L. & CRISTEA, A. I. In-depth exploration of engagement patterns in MOOCs. International conference on web information systems engineering, 2018b. Springer, 395-409.

SHI, L., CRISTEA, A. I., TODA, A. M. & OLIVEIRA, W. 2020. Revealing the hidden patterns: a comparative study on profiling subpopulations of MOOC students. *arXiv preprint arXiv:2008.05850*.

SHIH, J. 2019. Using Clickstream to Understand Learning Paths and the Network Structure of Learning Resources: Using MOOC as an Example. *the 27 th International Conference on Computers in Education.*

SHORFUZZAMAN, M., HOSSAIN, M. S., NAZIR, A., MUHAMMAD, G. & ALAMRI, A. 2019. Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. *Computers in Human behavior,* 92, 578-588.

SI, L. 2022. *Multi-modal Deep Learning.* Worcester Polytechnic Institute.

SIDDIQUE, S. A. 2020. Improvement of Online Course Content using MapReduce Big Data Analytics.

SIEMENS, G., GAŠEVIĆ, D. & DAWSON, S. 2015. Preparing for the digital university: A review of the history and current state of distance, blended, and online learning.

SIGRIST, F. 2018. Gradient and newton boosting for classification and regression. *arXiv preprint arXiv:1808.03064*.

SINHA, T., JERMANN, P., LI, N. & DILLENBOURG, P. 2014. Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. *arXiv preprint arXiv:1407.7131*.

SMITH, P. L. & DILLON, C. L. 1999. Lead article: Comparing distance learning and classroom learning: Conceptual considerations. ***American Journal of Distance Education,*** **13, 6-23.**

SORAYA, K., PURNAWARMAN, P. & SUHERDI, D. Revisiting Massive Open Online Courses Concept in The 21 st Century Era. 2019 2019. IEEE, 79-82.

SOUTHARD, S., MEDDAUGH, J. & FRANCE-HARRIS, A. 2015. Can SPOC (Self-Paced Online Course) live long and prosper? A comparison study of a new species of online course delivery. ***Online Journal of Distance Learning Administration,*** **18.**

SOZYKIN, K., PROTASOV, S., KHAN, A., HUSSAIN, R. & LEE, J. Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks. 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2018. IEEE, 146-151.

SPEITH, T. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022. 2239-2250.

STANDAGE, M., DUDA, J. L. & NTOUMANIS, N. 2005. A test of self-determination theory in school physical education. ***British journal of educational psychology,*** **75, 411-433.**

STONE, J. E. 2021. ***Self-Determination Theory and MOOC Enrollment Motivation: Validation of the Online Learning Enrollment Intentions Scale.*** **Oklahoma State University.**

SUN, D. M., Y.; DU, J.; XU, P.; ZHENG, Q.; SUN, H. Deep learning for dropout prediction in MOOCs. Proceedings - 2019 8th International Conference of Educational Innovation through Technology, EITT 2019, 2019. 87-90.

SUNAR, A. S., WHITE, S., ABDULLAH, N. A. & DAVIS, H. C. 2016. How learners' interactions sustain engagement: a MOOC case study. ***IEEE Transactions on Learning Technologies,*** **10, 475-487.**

SUNAR, A. S. W., S.; ABDULLAH, N. A.; DAVIS, H. C. 2017. How learners' interactions sustain engagement: A MOOC case study. ***IEEE Transactions on Learning Technologies,*** **10, 475-487.**

TAIT, A. 2003. Reflections on Student Support in Open and Distance Learning. 2003.

TANG, C. O., Y.; RONG, W.; ZHANG, J.; XIONG, Z. 2018. Time series model for

predicting dropout in massive open online courses. ***Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).***

TERUEL, M. A., L. A. Co-embeddings for student modeling in virtual learning environments. UMAP 2018 - Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, 2018. 73-80.

THAKKAR, S. R. & JOSHI, H. D. E-learning systems: a review. 2015 IEEE seventh international conference on Technology for education (T4E), 2015. IEEE, 37-40.

TÓTH, G. K. P. E. R. F. K. Clickstream-based outcome prediction in short video MOOCs. 2018 International Conference on Computer, Information and Telecommunication Systems (CITS), 11-13 July 2018 2018. 1-5.

TRIGUERO, I., MAILLO, J., LUENGO, J., GARCÍA, S. & HERRERA, F. From big data to smart data with the k-nearest neighbours algorithm. 2016 2016. IEEE, 859-864.

TUBMAN, L. H. Z. W. P. B. P. A Time Series Classification Method for Behaviour-Based Dropout Prediction. 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), 9-13 July 2018 2018. 191-195.

UDDIN, I. I., A. S.; MUHAMMAD, K.; FAYYAZ, N.; SAJJAD, M. 2021. A Systematic Mapping Review on MOOC Recommender Systems. ***IEEE Access,*** **9, 118379-118405.**

UMER, R. S., TEO; MATHRANI, ANURADHA; SURIADI, SURIADI 2017. On predicting academic performance with process mining in learning analytics. ***Journal of Research in Innovative Teaching & Learning,*** **10, 160-176.**

VASCONCELLOS, D., PARKER, P. D., HILLAND, T., CINELLI, R., OWEN, K. B., KAPSAL, N., LEE, J., ANTCZAK, D., NTOUMANIS, N. & RYAN, R. M. 2020. Self-determination theory applied to physical education: A systematic review and meta-analysis. ***Journal of Educational Psychology,*** **112, 1444.**

VEERAMACHANENI, K., DERNONCOURT, F., TAYLOR, C., PARDOS, Z. & O'REILLY, U.-M. Moocdb: Developing data standards for mooc data science. 2013 2013. Citeseer.

VEERAMACHANENI, S. B. B. U. G. B. S. K. Data science foundry for MOOCs. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 19-21 Oct. 2015 2015. 1-10.

VELU, N., MUTHU, S. R. U., NARASIMMALU, N. K. & KANMANI, M. 2023. Comparative Study of Machine Learning and Deep Learning for Fungi Classification. ***Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022.*** **Springer.**

VIGENTINI, L., LEÓN URRUTIA, M. & FIELDS, B. FutureLearn data: what we currently have, what we are learning and how it is demonstrating learning in MOOCs. Proceedings of the Seventh International Learning Analytics & Knowledge Conference, 2017. 512-513.

VITIELLO, M. G., C.; AMADO-SALVATIERRA, H. R.; HERNANDEZ, R. 2017. MOOC Learner Behaviour: Attrition and Retention Analysis and Prediction Based on 11 Courses on the TELESCOPE Platform. ***LEARNING TECHNOLOGY FOR EDUCATION CHALLENGES, LTEC 2017.***

VIVIAN, R., FALKNER, K. & FALKNER, N. 2014. Addressing the challenges of a new digital technologies curriculum: MOOCs as a scalable solution for teacher professional development.

VLACHOPOULOS, S. P. & MICHAILIDOU, S. 2006. Development and initial validation of a measure of autonomy, competence, and relatedness in exercise: The Basic Psychological Needs in Exercise Scale. ***Measurement in physical education and exercise science,*** **10, 179-201.**

VYAS, J. D. & HAN, M. Understanding the Mobile Game App Activity. Proceedings of the 2019 ACM Southeast Conference, 2019. 206-209.

WAHEED, H. H., S. U.; ALJOHANI, N. R.; HARDMAN, J.; ALELYANI, S.; NAWAZ, R. 2020. Predicting academic performance of students from VLE big data using deep learning models. ***Computers in Human Behavior,*** **104.**

WALLIN, D. L. 1990. Televised interactive education: Creative technology for alternative learning. ***Community Junior College Research Quarterly of Research and Practice,*** **14, 259-266.**

WANG, F. C., L. A nonlinear state space model for identifying at-risk students in open online courses. Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016, 2016a. 527-532.

WANG, J., XIONG, X., ZHOU, N., LI, Z. & WANG, W. 2016. Early warning method for transmission line galloping based on SVM and AdaBoost bi-level classifiers. ***IET Generation, Transmission & Distribution,*** **10, 3499-3507.**

WANG, L. W. H. Learning Behavior Analysis and Dropout Rate Prediction Based on MOOCs Data. 2019 10th International Conference on Information Technology in Medicine and Education (ITME), 23-25 Aug. 2019 2019. 419-

423.

WANG, W. Y., H.; MIAO, C. Deep model for dropout prediction in MOOCs. ACM International Conference Proceeding Series, 2017. 26-32.

WANG, X., YANG, D., WEN, M., KOEDINGER, K. & ROSÉ, C. P. 2015. Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. *International Educational Data Mining Society*.

WANG, X. L. L. X. H. Grade Prediction in MOOCs. 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 24-26 Aug. 2016 2016b. 386-392.

WANG, Z., ANDERSON, T., CHEN, L. & BARBERA, E. 2017. Interaction pattern analysis in cMOOCs based on the connectivist interaction and engagement framework. *British Journal of Educational Technology,* 48, 683-699.

WANGWONGWIROJ, T. & BUMRABPHAN, K. 2021. Self-Determination Theory: Statistical Correlations Between Motivational Regulations and Basic Psychological Needs. *International Journal of Higher Education Pedagogies,* 2, 53-58.

WARREN, J., RIXNER, S., GREINER, J. & WONG, S. Facilitating human interaction in an online programming course. Proceedings of the 45th ACM technical symposium on Computer science education, 2014. 665-670.

WATKINS, D. 2017. *Open educational resources movement gains speed* [Online]. Available: opensource.com/article/17/10/open-educational-resources-alexis-clifton [Accessed 2018].

WATTED, A. & BARAK, M. 2018. Motivating factors of MOOC completers: Comparing between university-affiliated students and general participants. *The Internet and Higher Education,* 37, 11-20.

WEAVER, D., SPRATT, C. & NAIR, C. S. 2008. Academic and student use of a learning management system: Implications for quality. *Australasian journal of educational technology,* 24.

WEI, X., LIN, H., YANG, L. & YU, Y. 2017. A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information,* 8, 92.

WEN, M. & ROSÉ, C. P. Identifying latent study habits by mining learner behavior patterns in massive open online courses. Proceedings of the 23rd ACM international conference on conference on information and knowledge management, 2014. 1983-1986.

WEN, M., YANG, D. & ROSE, C. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? Educational data mining 2014, 2014a. Citeseer.

WEN, M., YANG, D. & ROSÉ, C. Linguistic reflections of student engagement in massive open online courses. Proceedings of the International AAAI Conference on Web and Social Media, 2014b. 525-534.

WEN, M., YANG, D. & ROSÉ, C. P. Linguistic Reflections of Student Engagement in Massive Open Online Courses. ICWSM, 2014c.

WENQING, X. G. J. X. P. L. L. W. C. Are the Performance Prediction Models in MOOC General: Perspective from Big Data. 2020 IEEE Learning With MOOCS (LWMOOCS), 29 Sept.-2 Oct. 2020 2020. 84-89.

WHITEHILL, J. M., K.; SEATON, D.; ROSEN, Y.; TINGLEY, D. MOOC dropout prediction: How to measure accuracy? L@S 2017 - Proceedings of the 4th (2017) ACM Conference on Learning at Scale, 2017. 161-164.

WILLMS, J. D. 2003. Student engagement at school. *A sense of belonging and participation. Paris: Organisation for Economic Co-operation and Development.*

WISCHMEYER, T. & RADEMACHER, T. 2020. *Regulating Artificial Intelligence*, Springer.

WITCHEL, H. J. Engagement: the inputs and the outputs: conference overview. Proceedings of the 2013 Inputs-Outputs Conference: An Interdisciplinary Conference on Engagement in HCI and Performance, 2013. 1-4.

WOOD, D. N. & WYLIE, D. G. 1977. *Educational telecommunications*, Wadsworth Publishing Company.

WU, F. Z., J.; SHI, Y.; YANG, X.; SONG, W.; PENG, Z. 2020. Predicting MOOCs Dropout with a Deep Model. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

WU, N. Z., M.; ZHANG, L.; SUN, X.; GAO, Y.; FENG, J. CLMS-Net: Dropout prediction in MOOCs with deep learning. PervasiveHealth: Pervasive Computing Technologies for Healthcare, 2019.

XIA, X. Prediction of learning behavior based on improved random forest algorithm. Journal of Physics: Conference Series, 2020.

XING, W. C., X.; STEIN, J.; MARCINKOWSKI, M. 2016. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior,* 58, 119-129.

XING, W. D., D. 2019a. Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *Journal of Educational Computing Research,* 57, 547-570.

XING, W. T., H.; PEI, B. 2019b. Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of MOOCs. *Internet and Higher Education,* 43.

XU, R. & WUNSCH, D. 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks,* 16, 645-678.

Y. CHEN; Q. CHEN; MINGQIAN, Z. S. B. K. V. H. Q. DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction. 2016 IEEE Conference on Visual Analytics Science and Technology (VAST), 23-28 Oct. 2016 2016. 111-120.

YANG, D., SINHA, T., ADAMSON, D. & ROSÉ, C. P. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. Proceedings of the 2013 NIPS Data-driven education workshop, 2013. 14.

YANG, H. Chinese Sentiment Analysis of MOOC Reviews Based on Word Vectors. 2021 2nd International Conference on Artificial Intelligence and Education (ICAIE), 18-20 June 2021 2021. 68-71.

YANG, Z., MIAO, N., ZHANG, X., LI, Q., WANG, Z., LI, C., SUN, X. & LAN, Y. 2021. Employment of an electronic tongue combined with deep learning and transfer learning for discriminating the storage time of Pu-erh tea. *Food Control,* 121, 107608.

YARBROUGH, B. V. 1991. Trade and Investment Relations Among the United States, Canada, and Japan. JSTOR.

YE, C. & BISWAS, G. 2014. Early prediction of student dropout and performance in MOOCs using higher granularity temporal information. *Journal of Learning Analytics,* 1, 169-172.

YE, C. F., D. H.; KINNEBREW, J. S.; NARASIMHAM, G.; BISWAS, G.; BRADY, K. A.; EVANS, B. J. Behavior prediction in MOOCs using higher granularity

temporal information. L@S 2015 - 2nd ACM Conference on Learning at Scale, 2015. 335-338.

YOUNG, P. A. 2021. The ever evolving MOOC. *Educational Technology Research and Development,* 69, 363-364.

YU, C. H. W., J.; LIU, A. C. 2019. Predicting learning outcomes with MOOC clickstreams. *Education Sciences,* 9.

YU, C. H. W., J.; LIU, M. C.; LIU, A. C. 2021. Adopting software product lines to implement an efficient learning analytics framework in MOOCs. *Journal of Information Science and Engineering,* 37, 139-155.

YU, Z., HAGHIGHAT, F., FUNG, B. C. M. & YOSHINO, H. 2010. A decision tree method for building energy demand modeling. *Energy and Buildings,* 42, 1637-1646.

ZHANG, L. C. G. Prediction of MOOCs Dropout based on WCLSRT Model. 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 12-14 March 2021 2021. 780-784.

ZHANG, Y. C., L.; LIU, T. MOOCs Dropout Prediction Based on Hybrid Deep Neural Network. Proceedings - 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2020, 2020a. 197-203.

ZHANG, Y. Z., Q.; LIU, X. 2020b. Dropout Predictions of Ideological and Political MOOC Learners Based on Big Data. *PROCEEDINGS OF NINETEENTH WUHAN INTERNATIONAL CONFERENCE ON E-BUSINESS.*

ZHAO, Y., DAVIS, D., CHEN, G., LOFI, C., HAUFF, C. & HOUBEN, G.-J. Certificate achievement unlocked: How does MOOC learners' behaviour change? Adjunct publication of the 25th conference on user modeling, adaptation and personalization, 2017. 83-88.

ZHENG, J. L. J. Y. Y. W. C. L. L. Big Data Application in Education: Dropout Prediction in Edx MOOCs. 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), 20-22 April 2016 2016. 440-443.

ZHENG, Y. & YIN, B. Big data analytics in MOOCs. 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 2015. IEEE, 681-686.

ZHOU, M. 2016. Chinese university students' acceptance of MOOCs: A self-

determination perspective. *Computers & Education,* 92, 194-203.

ZHOU, Y. Experiment Report: Peripheral and Central Persuasion Name. 2022 International Conference on Social Sciences and Humanities and Arts (SSHA 2022), 2022. Atlantis Press, 355-358.

ZHU, M., SARI, A. & LEE, M. M. 2018. A systematic review of research methods and topics of the empirical MOOC literature (2014–2016). *The Internet and Higher Education,* 37, 31-39.

ZHU, M., SARI, A. R. & LEE, M. M. 2022. Trends and issues in mooc learning analytics empirical research: A systematic literature review (2011–2021). *Education and Information Technologies,* 27, 10135-10160.

ZHU, X. Y., Y.; ZHAO, L.; SHEN, C. 2021. MOOC behavior analysis and academic performance prediction based on entropy. *Sensors,* 21.