The application of Machine Learning for Early Detection of At -Risk Learners in Massive Open Online Courses

Raghad Al-Shabandar

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

August 2018

DECLARATION

I, Raghad ALshabandar, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm this has been indicated in the thesis.

Raghad AL-shabandar

Word count (excluding acknowledgement, appendices and references): 40,600 **words** (excluding the Appendixes and References).

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful. Praise and Great Thanks to Allah for providing me blessing, strength, and the opportunity to complete this research project.

I would like to thanks my supervisory team, Professor Abir Jaafar Hussain and Mr. Andy Laws for their unfailing assistance and patience throughout my PhD study. I would like to thank Abir for her exceptional supervision and invaluable guidance. Prof. Abir has always dedicated time to discuss my project and gave me valuable suggestions over the three years' period of my PhD study, without Prof. Abir's support; I would not have been able to complete the PhD. My special thanks go to Mr. Andy Laws for reviewing my work and enhancing my writing skills, which has helped me to publish the work in top conferences. You both have my sincerest gratitude

I am greatly indebted to my family for this project. In particular, my father and my brothers deserve the greater appreciation for their financial support. My family had extensive understanding and endless patience during my PhD.

I am also grateful to the technical team at the Department of Computer Science at LJMU for their help and cooperation throughout my PhD; I would like to thank Mrs. Tricia Waterson and Mrs. Carol Oliver for all their support over the three-year period at LJMU.

ABSTRACT

With the rapid improvement of digital technology, Massive Open Online Courses (MOOCs) have emerged as powerful open educational learning platforms. MOOCs have been experiencing increased use and popularity in highly ranked universities in recent years. The opportunity to access high-quality courseware content within such platforms, while eliminating the burden of educational, financial and geographical obstacles has led to a growth in participant numbers. Despite the increasing participation in online courses, the low completion rate has raised major concerns in the literature.

Identifying those students who are at-risk of dropping out could be a promising solution in solving the low completion rate in the online setting. Flagging at-risk students could assist the course instructors to bolster the struggling students and provide more learning resources. Although many prior studies have considered the dropout issue in the form of a sequence classification problem, such works only address a limited set of retention factors. They typically consider the learners' activities as a sequence of weekly intervals, neglecting important learning trajectories.

In this PhD thesis, my goal is to investigate retention factors. More specifically, the project seeks to explore the association of motivational trajectories, performance trajectories, engagement levels and latent engagement with the withdrawal rate. To achieve this goal, the first objective is to derive learners' motivations based on Incentive Motivation theory. The Learning Analytic is utilised to classify student motivation into three main categories; Intrinsically motivated, Extrinsically motivated and Amotivation. Machine learning has been employed to detect the lack of motivation at early stages of the courses. The findings reveal that machine learning provides solutions that are capable of automatically identifying the students' motivational status according to behaviourism theory.

As the second and third objectives, three temporal dropout prediction models are proposed in this research work. The models provide dynamic assessment of the influence of the following factors; motivational trajectories, performance trajectories and latent engagement on students and the subsequent risk of them leaving the course. The models could assist the instructor in delivering more intensive intervention support to at-risk students. Supervised machine learning algorithms have been utilised in each model to identify the students who are in danger of dropping out in a timely

manner. The results demonstrate that motivational trajectories and engagement levels are significant factors, which might influence the students' withdrawal in online settings. On the other hand, the findings indicate that performance trajectories and latent engagement might not prevent students from completing online courses.

Table of Contents

List of Tables

List of Figures

List of Algorithms

THESIS ACRONYMS

Acronyms	Meaning
LA	Learning Analytics
ICT	Information Communication Technology
LMS	Learning Management System
IRC	Internet Relay Chat
ITSs	Intelligent Tutoring Systems
OER	Open Educational Resources
MOOCs	Massive Open Online Courses
cMOOCs	cnnectivist Massive Open Online Courses
xMOOCs	eXtended Massive Open Online Course
EDM	Educational Data Mining
LAM	Learning Analytic Model
FAM	Factor Analysis Model
VLE	Virtual Learning Environment
CMA	Check My Activity
DKT	Deep Knowledge Tracing
RNN	Recurrent Neural Networks
GA	Genetic Algorithm
SOM	Self-Organised Map
PSL	Probabilistic Soft Logic
NSSE	National Survey of Student Engagement
SDT	Self-Determined Theory
IM	Incentive Motivation Theory
SIMS	Situational Motivation Scale
SRL	Self-Regulated Learning
XGBoost	Extreme Gradient Boosting
IMMS	Instructional Materials Motivation Survey
CNN	Convolutional Neural Networks
GBM	Generalised Boosted regression Models

HMM	Hidden Markov Model
GAN	Generative Adversarial Network
EBL	Explanation-Based Learning
L1	Lasso Regularization
L2	Ridge Regularization
OOB	Out-of-bag
MSE	Mean Square error
LR	Logistic Regression
FCM	Fuzzy C-means clustering
MCMC	Markov Chain Monte Carlo
BIC	Bayesian Information Criterion
LEC	Laplace Empirical Criterion
MML	Minimum Message Length
SVM	Support Vector Machine
AIC	Akaike's Information Criterion
FMD	Finite Mixture Model
MDA	Mixture Discriminant Analysis
EDDA	Eigenvalue Decomposition Discriminant Analysis
OULAD	Open University Learning Analytic Dataset
TMA	Tutor Marked Assessment
CMA	Computer Marked Assessment
EDA	Exploratory Data Analysis
PCA	Principal Component Analysis
RFE	Recursive Feature Elimination
RMHC	Random Mutation Hill Climber
SMOTE	Synthetic Minority Oversampling
ANCOVA	Analysis of Covariance
MDI	Mean Decrease Impurity
MDA	Mean Decrease Accuracy
ICL	Integrated Completed likelihood

PDF	Probability Density Function
EM	Expectation-Maximization
ROC	Receiver Operator Characteristic
AUC	Area Under Curve
RF	Random Forest
Rpart	Decision Tree
Nnet	Feedforward Neural Network with Single hidden layer
MLP	Multiple Layer Perceptron with two hidden layers
BFGS	Broyden-Fletcher-Goldfarh-Shanno

List of publications

International Journal:

Al-Shabandar, R., Hussain, A.J., Liatsis, P. and Keight, R., 2018. Analyzing Learners Behaviour in MOOCs: An Examination of Performance and Motivation Using a Data-Driven Approach. *IEEE Access*.

International Conference:

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J. and Radi, N., 2017, May. Machine learning approaches to predict learning outcomes in massive open online courses. In *Neural Networks (IJCNN), 2017 International Joint Conference on* (pp. 713-720). IEEE.

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R. and Lunn, J., 2017, August. Towards the Differentiation of Initial and Final Retention in Massive Open Online Courses. In *International Conference on Intelligent Computing* (pp. 26-36). Springer, Cham.

Alshabandar R, Hussain A, Keight R, Laws A, Lunn J, Shamsa T., 2018. The Application of Gaussian Mixture Models for the Identification of At-Risk Learners in Massive Open Online Courses. In *World Congress on Computational Intelligence* (WCCI), 2018.IEEE

Al-Shabandar, R., Hussain, A., Laws, A., Jumeily D., 2018, August. Supervised and Unsupervised Machine Learning Approaches for the Prediction of Learner Outcomes in Massive Open Online Courses**.** In *Interactive Digital Media (ICDIM), Springer,* Cham (accepted).

Chapter 1: Introduction

1.1 Introduction

Online education has become an area of continuing growth within both industrial and academic settings. There were more than 6 million students enrolled in online courses in 2012 (Qiu *et al.*, 2016). The new bellwether of online educational platforms is Massive Open Online Courses (MOOCs) (Coffrin *et al.*, 2014). MOOCs are open educational platforms that deliver learning resources through digital platforms (Hew, 2016).

MOOCs provide the same quality of learning as the traditional classroom without the time and geographical restrictions. As a result, learners are able to understand and learn coursework content at their own pace(Khalil and Ebner, 2014). In MOOC platforms, learners are connected with an array of learning resources, including video lectures, weekly quizzes, regular assessments and even PDF documents. Additionally, the learners can interact asynchronously with the instructors via discussion forum posts (Liyanagunawardena, Parslow and Williams, 2014).

MOOCs have seen dramatic increases in popularity over the last few years within the higher education sector (Qiu *et al.*, 2016). The highest ranking universities have developed and delivered hundreds of courses, including HarvardX, Khan Academy, and Coursera (Qiu *et al.*, 2016). In 2012, the Open University cooperated with more than 20 universities and educational institutions to deliver online courses. It offered the courses in different subjects such as health, business and management.

One of the distinctive features of MOOCs is their instant accessibility coupled with the elimination of financial, geographical, and educational obstacles. Consequently, the proportion of participants engaging in such courses could increase quickly (Qiu *et al.*, 2016)(Yang and Rose, 2013). For example, the number of participants has rapidly expanded in Harvard online courses with 1.3 million unique learners engaged in online courses reported at the end of 2014.

Despite the lowering of barriers to high-quality education, the ability of students to enrol and withdraw from courses freely often results in high rates of attrition (Yang and Rose, 2013). As such, during 2012, the University of Duke offered a bioelectricity course, attracting around 12,175 registered participants, of which only 315 learners

continued to undertake the final exam. At the end of 2012, It was reported that 93% of participants withdrew (Yang and Rose, 2013).

The low completion rate is a major issue related to MOOCs; research investigations reveal on average that out of each one million participants in MOOCs, an overwhelming majority of them withdraw from MOOCs prior to completion (Yang and Rose, 2013). Due to a lack of face-to-face interaction between instructors and learners in such courses it is understandably difficult for instructors to maintain direct awareness of the reasons of individual learner withdrawals (Hone and El Said, 2016).

The early identification of students who are at-risk of withdrawal from a course is one of the strategies that can be used to overcome the low completion rate in MOOCs. Detecting at-risk students in a timely manner could help the educators to deliver instructional interventions and improve the structure of the courses (Hung *et al.*, 2017). With a timely intervention solution, the instructors can provide real-time feedback to the students; hence, retention rate could be improved.

In order to build an accurate 'at-risk' students prediction model, the researchers investigate the reason behind the withdrawals. This has been attributed to a number of factors by the literature. The main reasons for students dropping out in the online courses include:

- **Lack of motivation:** Students' lack of motivation is the most influential factor that prevents them from completing the online course. Learning Analytics (LA) have been utilised to investigate the reasons for participants' motivation in online classes through the analysis pattern of students' engagement (Clow, 2012). The findings demonstrate that students engage and participate in the online course for two main reasons, namely feeling immediate satisfaction or attaining formal recognition by receiving a certificate. As such, 87% of learners who enrol in online courses tend not to complete the full courses they undertake for enjoyment and interest. Only 13% of students decide to participate in formal assessments (Dalipi, Imran and Kastrati, 2018). In addition, they have found the levels of student engagement differ from intrinsically motivated students to extrinsically motivated students.

- **Low student performance**: The trajectory performance is shown to be another reason, which affects the students' decision to quit the online course. Within certain subjects, students have been allowed to engage in multiple courses. Therefore, the

student performance on a previous course can be used as a predictor of their following course completion. In terms of the traditional classroom setting, the researchers conclude that students' prior GPA with a low mark is considered as a significant factor of withdrawal from the next course (Meier *et al.*, 2016).

- **Lack of Time:** The amount of time is another factor that causes students to withdraw from the course. The surveys show that students might need more time in the online setting course than in the traditional classroom course. This is because MOOC requires that the students watch video lectures and participate in a set of quizzes. Furthermore, a large number of MOOCs participants start late so it could be hard to follow up with the course schedules (Hone and El Said, 2016) (Dalipi, Imran and Kastrati, 2018).

- **Inadequate Skills:** The online course requires the students to have technical skills and a high degree of autonomy. Insufficient academic skill is another reason for disengagement. The studies show that if students do not have the knowledge base in the specific subject, their engagement level could decrease rapidly (Dalipi, Imran and Kastrati, 2018). References (Dalipi, Imran and Kastrati, 2018)(Kloft *et al.*, 2014) state that students who do not have the reading, writing and IT skills feel frustrated and struggle to engage with the course in particular if the course requires the synchronous 'chat' interaction.

1.2 Problem Statements

Although the literature suggests many factors for students' withdrawal, few researchers' efforts have been dedicated to investigating how such factors influence at-risk students. To this end, the link between the level of engagement, performance trajectories and motivational trajectories should be investigated from various perspectives. This is a significant challenge. In the following, highlights of the limitations of existing research works will be provided:

Evaluating learner motivation In MOOCs

One of the main shortcomings of existing research is the lack of approach for the evaluation of motivation in the online setting. The majority of studies employ both quantitative and qualitative methods to measure motivation within MOOCs, relying on the analysis of transcripts, interviews and survey data. Consequently, learner motivation is evaluated from a narrow perspective, which does not account

for learner interaction patterns within the MOOCs environment (Barak, Watted and Haick, 2016).

Student motivational status changes over time; for instance, the literature review demonstrated that students' motivational level during the online courses either decrees or increases according to social, cognitive and environmental factors. The motivational trajectory is an important indicator of students' dropout rate. As a consequence, motivational trajectories can be used as reliable predictors to detect which student is in danger of withdrawal from courses (Turner and Patrick, 2008). Motivational trajectories can be measured by exploring the changes in the subsequent of learners' behaviours across various courses. Until now, most researchers have not given attention to examining the association between motivational trajectories and at-risk students.

Investigate the association of student engagement levels with the withdrawal rate

In order to examine how the above reasons could influence at-risk students, the dynamic link between the level of engagement, performance and withdrawal rate should be considered (Hew, 2016). LA is utilised to analyse the students' engagement levels by tracking students' historical clickstream data within a single online course. With timely dropout prediction, educators could deliver timely intervention support to at-risk students at any given time (Dalipi, Imran and Kastrati, 2018).

Although many previous studies have identified at-risk students in a timely manner, such works address only a limited set of dynamic behaviours; typically students activities have been analysed according to the sequence of the weekly intervals. Assessment deadlines have not been taken into consideration. However, it has been argued that the assessment deadlines are an essential component of students engagement that might influence students performance in the online context (Wolff *et al.*, 2014).

In addition, existing works have been unable to flag at-risk students early and accurately as only limited types of learning activities have been analysed. Namely, pdf files viewed, videos watched, quizzes viewed and posts in discussion forums; they did not involve other significant features such as the number of clicks that

the student performs per homepage and subpage in addition to the students' scores in previous assessments.

Furthermore, the lack limitation of features engineering techniques is shown in literature review since features are extracted only from the number of clicks in a particular activity or the number of times that the student engaged in activity within the entire week (Fei & Yeung 2015).

To the best of our knowledge, none of the current work has paid attention to constructing features from the number of times that students launched the specific activity per single session for predicting at-risk students.

Students can participate in multiple courses within a certain subject. The student completion rates across various courses can be related to the level of student engagement and performance at the previous session (Huang and Fang, 2010). Insufficient attention has been paid by literature to evaluate whether the level of student engagement and achievement in the prior course could affect the at-risk student in the next course.

Inferring Latent Engagement in MOOCs

To explore how the patterns of engagement could influence the students who are at-risk of withdrawal, latent engagement should be investigated (Ferguson and Clow, 2015). The inferring of the students' hidden engagement would help educators to understand the intention of students to participate in certain activities (Ramesh *et al.*, 2014).

Categorising the latent engagement patterns of learners concerning the impact on their continuation within course activities remains a challenge (Wang and Chen, 2016) (Lan *et al.*, 2016). Few studies have been undertaken to investigate latent engagement as a sequential classification problem (Lan *et al.*, 2016). A notable limitation of the few existing studies is that behavioural features are distributed weekly. As a consequence, prediction models over time depend on a weekly basis without accounting for the assessment submission date as a factor of significant influence for student withdrawal. Within this approach, the estimating procedures which characterise at-risk students could be inaccurate due to their failure to

account for context-sensitive factors including the submission date of assessments.

Research Questions

1. Can the students' behavior activities be used to categorise the motivational status in MOOCs platform?
2. To what extent are the performance trajectories and motivational trajectories influencing students' dropout in MOOCs platforms?
3. Are the failing and successful learners different in their engagement style?
4. Is latent engagement a crucial factor that prevents students completing the assessments within a single online course?

1.3 Aims and objectives

There are two aims in this reserch. The first aim is to evaluate the learners' motivational statues based on the concept of motivation theory. The second aim is to build dynamic dropout predictive models that are capable of identifying the students who are in danger of withdrawal from the course at an early stage. Towards achieving the research aims, the following objectives are considered.

- Develop a motivational model for the prediction of student motivational categories in an online setting.
- Investigate the influence of motivational trajectories and performance trajectories on students' decision to quit the online course considering multiple courses.
- Evaluate the impact of latent engagement on at-risk students within a single course.
- Investigate the impact of engagement level on student performance with respect to single course and multiple courses.
- Compare various machine-learning techniques and select the appropriate techniques that are suitable for the project.
- Utilise different classification and regression evaluation metrics to demonstrate the predictive capability and generalizability of our results.

Table 1.1 Research workflow

Objective	Methodology	Chapter
Develop a motivational model for the prediction of students' motivational categories in an online setting.	(A)Conduct literature review in the behaviourist theory of motivation, evaluation of motivation in online courses. (B)LA to categorise the students' motivational statues	(A)Chapter 2 (B) Chapter4
Investigate the influence of motivational trajectories and performance trajectories on student decision to quit the online course considering multiple courses.	(A)Conduct literature review into the factors that influence learners' attrition in online courses. (B)Propose at-risk student framework in Harvard dataset.	(A)Chapter 2 (B) Chapter4
Evaluate the impact of latent engagement on at-risk students within a single course.	A) Conduct literature review to examine effect of latent engagement in online setting. (B)Propose at-risk student framework in OULAC dataset.	(A)Chapter 2 (B) Chapter4
Investigate the impact of engagement level on student performance with respect to single course and multiple courses.	A) Conduct literature review to explore effect of engagement on student performance in online setting. (B) Propose student performances models in both datasets.	(A)Chapter 2 (B) Chapter4
Compare various machine-learning techniques and select the appropriate techniques that are suitable for the project.	(A) Set of machine learning classifiers to be used in both datasets to predict students' performance and identify at-risk students.	(A)Chapter 5
Utilise different classification and regression evaluation metrics to demonstrate the predictive capability and generalisability of our results.	(A) The results of students' performance model and dropout predictive models are compared in terms of (Sensitivity, Specificity, Recall, F1 measures, AUC).	(A) Chapter 5

1.4 Anticipated Contributions

To overcome the limitations of existing work, this thesis proposes the prediction framework for detecting at-risk students early in online courses. The at-risk students has been investigated from various aspects including engagement level and motivational trajectories, performance trajectories, in addition to students' latent engagement. Two online datasets have been considered: Harvard dataset and Open

University Learning Analytic Dataset (OULAD). The research project provides four major novel contributions including:

- The first contribution of this research work is to employ a data-driven approach, a LA tool is utilised to characterise student motivational status based on Incentive Motivation Theory (IM). This has been accomplished using the Harvard dataset. According to the theory, the learners have been classified into three categories namely amotivation, extrinsic, and intrinsic. A set of machine learning models has been applied in the prediction of learners' motivational status. Hence, the predictors in the experiments depend on quantitative log data, rather than questionnaire response.
- The second contribution of this work is constructing dropout predictive models that are capable of delivering timely intervention support for at-risk students. LA is utilised to identify the at-risk students in a timely manner by tracking learners' historical behavioural records. In the first predictive model, the aim is to investigate whether the students' performance on the previous courses could influence their decision to quit the following courses. In the second predictive model, the engagement level in conjunction with students' behavioural status on the previous course have been examined to evaluate the effect of such factors on students persisting with participation in the following course.
- The third contribution of this project is using the novel and robust features engineering method. The student's activities are captured on a daily basis in the OULAD database; the behavioural features have been extracted from each activity according to assessments tradeoff date. With this method, students' records for each activity type are distributed across six-time slices leading to a multi-view of behavioural features. For each time interval, the features are derived from the number of sessions in which the student undertook the specific activity in addition to, the number of clickstreams that have been completed by students per single session.
- The fourth novel contribution that has been provided by this project is designing a temporal model for identifying at-risk students in the online course focused on the sequences of students' activities and latent engagement across a single course.

1.5 Thesis Structure

The remainder of this thesis is organised as follows. Chapter two displays the background of MOOCs. More specifically, an overview of students performance and the evaluation of performance in online courses is discussed in this chapter. Chapter two also gives a brief overview of motivation theory and summarises the methods that have been adopted by the literature for evaluating students' motivation in online courses. The earlier at-risk students models that have been developed by current researchers are also presented in this chapter.

Chapter three provides a brief history of machine learning. A comparison of supervised and unsupervised machine learning is explored in this chapter. The advantages and disadvantages of each algorithm are also explained. An overview of unsupervised machine learning algorithms is also presented in chapter 3 and, consequently, a discussion of the most popular unsupervised machine learning algorithms. Finally, a description of features selection method is included in this chapter.

A detailed discussion of research methodology is presented in chapter four. The description of two MOOCs datasets is discussed in this chapter. The Harvard and Open University Learning Analytic Dataset (OULAD) are compared in terms of potentiality and limitations followed by the data pre-processing procedure. Three sets of experiments are presented in this chapter. More specifically, the performance prediction model is proposed over two datasets.

The motivational prediction model is presented in the second experiment where machine learning is utilised to detect the lack of student motivation in the online context. In the second experiment, the temporal predictive models are proposed. In such models, three crucial factors are investigated; the influence of engagement levels, motivational trajectories and performance trajectories.

Features extracted from the OULAD dataset are discussed in detail followed by a description of the proposed algorithm. Finally, the evaluation of student engagement levels and latent engagement over a series of intervals on withdrawal rate is examined.

Chapter five highlights the results and discusses the proposed work. The results of supervised and unsupervised machine learning are also presented in this chapter. A discussion of the results of student performance predictive models across the individual course and multiple courses are presented in this chapter.

The results of the statistical analysis with respect to the association between the students' performance and engagement levels are presented in chapter five. The comparison between successful and failing students with respect to students' educational levels is also presented in this chapter. The summary of contribution is displayed in chapter six followed by the future work.

CHAPTER 2: Background and Literature Review

2.1 Introduction

The traditional teaching strategy is purely focused on the educator (teacher, lecturer, tutor or trainer); the students are expected to obtain the information dictated by the teacher. The first traditional educational approach depended on the oral recitation of students. In such an approach, the lessons are delivered by students themselves and the teacher's role was passive. The teacher only listens to students' recitations (Dimitrios *et al.*, 2013)

This approach focuses on students' verbal answers and the emphasis is on rote memorisation. The critical limitation of this approach is that it does not take into consideration the different levels of the student's education; it assumes that all students should be taught at the same Pace. To maintain positive classroom atmosphere, corporal punishment is used as the response to students' unacceptable behaviour(Dimitrios *et al.*, 2013).

The students would be compulsorily taught in school between ages 5 to 10 years. All schools in the UK were private schools belonging to religious institutions until the 1900 when reforms were introduced. The aim of the reforms was widespread public education around the world in the form of an input-output system in that period, local authorities established public primary and secondary schools around UK(Hargreaves, 1994).

By 1940, a tripartite system had been introduced to shape the education system of the UK. This organised schools into three categories: grammar schools, secondary technical schools and secondary modern schools. In 1980 the national curriculum was introduced in which all schools must deliver the same standards of teaching to students across the UK(Hargreaves, 1994).

The traditional teaching method has a critical flaw in the sense that is only suitable for young children and university students. Learners who want to increase their experiences by undertaking professional courses cannot attend regular classes (Wikramanayake, 2003)(Rovai, 2002).

Funding restrictions mean that class size has become a crucial learning barrier in professional courses. Educators are required to manage larger classes and as such are

less able to provide specific assistance to individual students. The problems of increased class size cause a twofold reduction in the educators' efficiency. For example, the educator may not be able to invest the time required for the needs of individual learners and might give poor feedback. This critical issue threatens to undermine the capabilities of educators and limit the potential of the students in such courses. (Rovai, 2002).

The geographical factor in relation to a student's ability to attend the physical location of course lectures, workshops and meetings can also play a role in how a student can engage their learning. Factors such as travel disruption, family life, travel time, work commitments, finances and even social life can form a barrier to students attending a geographical location, which can become a severe obstacle where important information is missed that can affect their understanding of the course subjects as well missing out on the guidance required to succeed(Buabeng-Andoh, 2012)(Pamuk, 2012).

2.2 Breaking Learning Barriers

Information Communication Technology (ICT) has become widespread to play a vital role in education. ICT has contributed to the support of the academic curriculum and allows for the creation of an interactive channel between students and instructors. ICT could improve student outcomes and enable the teacher to aid the student in solving exercises hence, high quality teaching would be delivered through advanced technology(Ghaznavi, Keikha and Yaghoubi, 2011).

ICT is capable of enhancing the learning resources and assisting the educators in the delivery of efficient teaching strategies. With advanced technology, educators and their teaching strategies can heavily influence courses. ICT can be implemented in a way that helps the educators to enhance the performance of students (Ghaznavi, Keikha and Yaghoubi, 2011)(Sarkar, 2012).

The issues of class sizes and geographical distances are overcome since, digital learning technologies open up opportunities for students, allowing them to attend or catch up on lectures and meetings through video and teleconferencing without the need to travel. Communications can also be distributed between students, their educators and their peers in non-real time. ICT allows the individual learners to respond to their learning requirements without taking into account the obstruction time, as opposed to

the traditional requirement for learners to fit themselves around a fixed learning schedule. Ultimately, new technologies allow the opportunity for the deployment of new education strategies that are the reverse of traditional teaching. For instance, instead of courses being designed around the educator dictating to their students, courses can now be designed around the learners. Therefore, the students' engagement could raise rapidly with such platforms(Ghaznavi, Keikha and Yaghoubi, 2011).

Educators and their requirements should still play a central role in virtual classes, as learners will not be able to achieve their full potential without the guidance of educators. Instead of new learning technologies replacing the role of the educator, they provide assistance to them. The new virtual class could provide better quality feedback between educators and their students. In addition, it could open new scope for researchers to utilise the AI for the purpose of investigation into the learner's progression. With educators becoming more aware of the needs of their students in an online setting, teaching strategy could become more effective online than the traditional methods(Sørebø *et al.*, 2009)(Wilson, 2004).

Table 2.1 Comparison Traditional Learning versus Online Learning

Traditional Courses	Online Courses
Course more suitable for learners who aim to improve their career opportunities.	Course more suitable for children and teenager students.
class is synchronous	class is asynchronous
More flexible as the student dose not regularly attend the class.	Less flexible as students are regular attend the class.
Small class size.	Large class size.
Lower financial cost.	Higher financial cost.
Self-regulated learner.	Educator control learning.
Face to face communication.	Social communication.

2.3Learning Management System

The rapid growth of digital technology has increased the growth of distance learning by providing different tools to deliver course content using multimedia such as animation, pictures, and figures, videos which are used to provide interactive content, control of the online activities and motivating the learner to build new cognitive skills(Sclater, 2008). To this extent, e learning offers to learners a flexible teaching approach that enables them to access information resources from anywhere. With the internet revolution, the virtual learning environment has introduced the virtual

classroom as an alternative to the traditional classroom. It provides facilities and breaks down the obstacles of the traditional teaching approach(Weaver, Spratt and Nair, 2008).

Learning Management System (LMS) is a web-based system that is used for distributed online courses via ICT. LMS is a rich application that has features to deliver online courses in digital form. The attractive learning material could lead to improved learners' attention, and allows them to learn at their own pace. Additionally, learning is more easily accessible to students. Educators also benefit by being able to better estimate the progression of their student's and help them deploy the learning strategy (Weaver, Spratt and Nair, 2008)(Mtebe and Raisamo, 2008).

The LMS uses social media such as Internet Relay Chat (IRC) and Skype; this is especially useful for communication where there is a barrier to Synchronous learning. The LMS allows the learner and educator to share the online lesson at the same time through computer technologies. As a consequence, providing a synchronous learning environment, the LMS has become an effective alternative to face-to-face traditional teaching methods(Dutton, H., Cheong, P., & Park, 2004).

The LMS such as Blackboard, Internet Relay Chat (IRC) and WebCT are widely used in higher education and educational institutes as a paradigm integrated e-learning platform. It has various levels of complexity with different types. Although, there are differences, they perform the same functions and have common characteristics(Weaver, Spratt and Nair, 2008).

The LMS can help the teacher to construct online assessments. With online evaluations the teacher can monitor students' answers and deliver immediate, dynamic feedback to students and measure the difficulty of tasks(Botički, I., Budišćak, I. and Hoić-Božić, 2008). The LMS presents a promising alternative to traditional student assessments, by aiming to overcome the inherent limitations of conventional approaches.

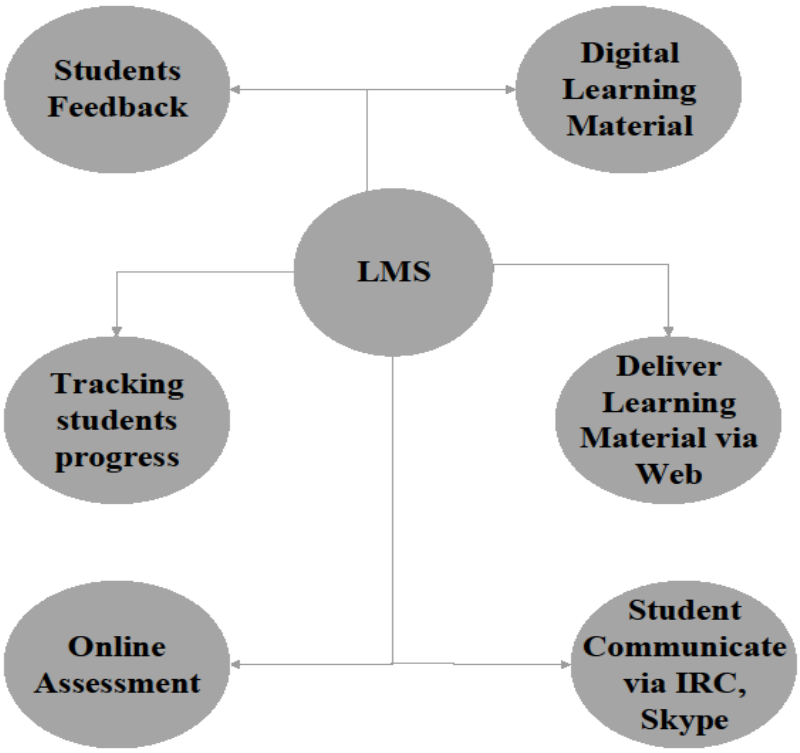


Figure 2.1 LMS Functions

2.4 Intelligent Tutoring systems

Intelligent Tutoring Systems (ITSs) are a computer-based formative assessment system designed to enhance the student's experience in e-learning platforms. It is based on the artificial intelligence concept that facilitates the learning process(Gharehchopogh and Khalifelu, 2011).

ITSs have the capacity to monitor the sequence of steps undertaken by students during interactive engagement with e-learning environments without the need for teacher assistance(VanLehn, 2006). ITSs are capable of tracking student behaviours during learning activities and providing immediate feedback to students(Koedinger *et al.*, 2013).

Moreover, The ITSs combine automated assistance with assessments to facilitate the learning process of students. They allow students to request a hint if they find the current task prohibitively difficult. Hints serve to guide the student towards successful problem resolutions and help educators to assess the student's educational level. In the

case of the student providing an incorrect answer to the current task, the assessment model responds by providing the student with a set of scaffold questions in which the original question is divided into multiple sub-questions. The model assists the student in the form of a sequence of hints, after which the student has the opportunity to practice the task multiple times. To avoid unnecessary practice questions and undue investment of student time, the tutoring system provides a bottom out hint that delivers the correct solution when the student reaches a trial threshold (Gharehchopogh and Khalifelu, 2011).

Cognitive Tutor is an example of an ITS that has been used by schools in the USA to assist students with learning maths. Researchers suggest that students performances increased by 80% when complex mathematical tasks were delivered through Cognitive Tutor. In ITS, structures and interfaces are adapted to the student's needs, while remaining based on educational theory(Ritter, S., Anderson, J.R., Koedinger, K.R. and Corbett, 2007).

2.5 Massive Open Online Courses

With progress in Open Educational Resources (OER) advancing from an emerging field towards an increasingly important learning modality Massive Open Online Courses (MOOCs) have become an alternative educational platform that allows learners from dispersed geographical locations to access digital learning material, regardless of the geographical and time obstructs(Shapiro *et al.*, 2017).

Massive Open Online Courses (MOOCs) is one of the most widespread e-learning platforms. The MOOCs present the course using digital tool materials in various forms such as visual, audio, video and plain text. Most students prefer using the video lectures to understand the contents of lessons over fully reading plain text documents. The interactive video in the MOOCs could reduce students' stress, help them to feel relaxed and learn quickly(Hew and Cheung, 2014).

MOOCs can be classified into two distinct types mainly, connectivist Massive Open Online Courses (cMOOCs) and eXtended Massive Open Online Courses (xMOOCs) The cMOOCs are a new learning model based on connectivist learning theory (Renz, Schwerer and Meinel, 2016)(Li, Tang and Zhang, 2016).With the connectivism approach the instructor would not provide the actual learning material, the students get

the course syllabus by asking the questions and sharing this information with other participants. References (Renz, Schwerer and Meinel, 2016)(Li, Tang and Zhang, 2016)(Zutshi, O'Hare and Rodafinos, 2013) posit the learning strategy of cMOOCs focused on a collaborative approach in which learning material combined remix, repurposable and provided, forwarded to other students. Students within cMOOCs control their learning process, have autonomy and create their own network .In addition the student can determine how much time they need to engage in the online course (Zutshi, O'Hare and Rodafinos, 2013)(Wang *et al.*, 2017). It could be hard to evaluate the Student's needs as the course content dynamically evolves as it is constructed. The cMOOCs do not include a formal assessment, hence universities do not consider them to be official courses(Wang *et al.*, 2017)(Gallego Arrufat, Gamiz Sanchez and Gutierrez Santiuste, 2015).

The online survey has been applied by reference (Zutshi, O'Hare and Rodafinos, 2013) to examine students' experiences in the cMOOCs environment. The results reveal that many factors occurring in such an environment could affect the students' interaction and connection negatively, for example, courses including a lot of information, lack of professional knowledge and unsuitable coursework content.

xMOOCs area learning paradigm based on the principles of cognitivist behaviorist theory(Kesim and Alt, 2015)(Zutshi, O'Hare and Rodafinos, 2013). The structure of the courses is similar to the traditional course where the syllabus consists of a set of video lectures and a set of multiple choice quizzes in addition to the final exam(Nkuyubwatsi, 2016). The video lectures featuring the course instructor reviewing the content of the previous online lesson are released weekly. The participants can watch and pause the video at their own pace. Moreover, the students can socially interact with other participants and the instructor through posting in discussion forums. The instructors usually post questions, provide task solutions and reply to student questions via these discussion forums; as a consequence the discussion forums play a vital role in enhancing the course quality and make online sessions collaborative and engaging(Adams *et al.*, 2014). Some xMOOCs tend to make courses more engaging by offering online simulation platforms such as serious games and live sessions(Staubitz *et al.*, 2014).

A comparison of xMOOCs and cMOOCs indicates there is a big difference between two types of courses in terms of their objectives. xMOOCs place emphasis on

delivering high-quality learning materials while cMOOCs focus more on the students' interaction and connection with each other and their instructor (Renz, Schwerer and Meinel, 2016)(Li, Tang and Zhang, 2016). With cMOOCs, it is impossible to involve expertise to assess the students' knowledge whereas in xMOOCs, university lecturers can evaluate the students' knowledge through the use of computer-marked assessment feedback (Gallego Arrufat, Gamiz Sanchez and Gutierrez Santiuste, 2015). In particular, the computer gives immediate feedback to the student when he completes the online assessment. The learner, upon successful completion, will be awarded their certification in xMOOCs (Gallego Arrufat, Gamiz Sanchez and Gutierrez Santiuste, 2015)(Jones, 2014).

2.6 Education Data Mining and Learning Analytics

Educational Data Mining (EDM) is an emerging field of research aimed at extracting knowledge from learning processes to support decision makers(Baker and Siemens, 2014)(Sachin, R. Barahate, 2012). Recently EDM has been used within the higher education setting to enhance teaching strategies (Baker and Siemens, 2014). EDM involves the use of statistics, visualization, and machine learning methods for the exploration and analysis of educational data (Baker and Siemens, 2014) (West, 2012). The possibility of capturing big data within MOOCs opens up new horizons to educational data mining researchers who can extract deeper insights from the analysis of the data (West, 2012). Although a prominent application of EDM is set within the online learning environment, the analysis and tracing of actionable data is challenging (West, 2012).

Learning Analytic Model (LAM) has been developed by reference (De Freitas et al. 2015) LAM is similar to LA dashboard, it aims to examine the factors that affect students' retention in higher education. Hypotheses were extracted from the dynamic interaction of students and instructors; hypotheses then iterate through implemented statistical methods. With LAM, students were split to groups of learners sharing common behavioural patterns hence; the stockholder would be able to capture students' requirements in real time.

The researchers are able to use the LA to investigate patterns of student engagement in MOOCS in further detail. As a result, such data can provide cohort information about the learning process, with LA researchers able to visualise and analyse the

information obtained from each tier of learning. Such analytical capacity may in turn enable the accurate prediction of student performance within MOOCs. There are various methods of LA utilised by researchers including Web analytics, Artificial Intelligence and Social Network Analysis (Baker and Siemens, 2014).

The author in reference (Yousef *et al.*, 2014) conducted a study to investigate the factors that influence the quality of MOOCs from the learners' and instructors' perspectives. There were 98 professors and 107 students participating in the survey. The result of this survey reveals that LA is the most critical factors that should be considered when design MOOCs. The LA was reported as the most important factor. There are other criteria, which should be taken into account for achieving successful MOOCs platforms such as video content, user interface and social tool.

Different approaches have been designed using both EDM and LA with the aim of understanding and analysing learners interaction in MOOCs efficiently(Baker and Siemens, 2014). LA has been used to identify dropout students (Kloft *et al.*, 2014) For example, the University of Michigan developed Michigan Tailoring System (E2Coach)(Mattingly, Rice and Berge, 2012). The E2Coach is an open source system, which aims to identify weaknesses and performance of physics students. E2Coach also delivers personalized learning by the customization of course materials. The LA tool was implemented in E2Coach to capture and collect data about students' progress from various resources .It provides indications to educators to reconstruct learning materials that match student ability and experience (Mattingly, Rice and Berge, 2012).

The association between the Virtual Learning Environment(VLE) data and student exam marks has been investigated at the University of Maryland , Baltimore County (UMBC) (Mullan, 2016). LA was used through the implementation of the Check My Activity (CMA) tool. CMA can be defined as an LA tool, which compares students VLE activities with other activities and provides lecturers frequent feedback of students' emotional status. The results showed the students who engage with the course frequently are more likely to earn mark C or higher than those who did not engage regularly(Mullan, 2016).

2.7 Latent Variable in Online Settings

The latent variable is an unobserved variable that is inferred from the observed variable through the statistical model. The mathematical model is called "Latent

variable modelling". It aims to measure the impact of the latent variable on the observed variable. Latent variables models are currently used in many domains such as Education, Psychology, Machine learning, Economics and Image processing (Loehlin, 1984).

Factor analysis is the most popular method used in education. Factor analysis is the statistical method used to reduce the number of variables into a lower number of factors. In this method, it is assumed that there is a linear relationship between the variables and factors(Tahar *et al.*, 2010). The pioneer of factor analysis goes back to psychologist Charles Spearman who was trying to measure human intelligence in 1904. He found that those students' scores in uncorrelated subjects were positively correlated with each other. His postulated that general mental ability (intelligence) is the main factor in achieving similar grades in unrelated subjects (Spearman, 1904).

The Factor Analysis Model (FAM) was proposed to predict the student's performance in ITS taking into consideration the difficulty level of assessments based on item IRT theory concept (Hao Cen, Kenneth Koedinger and Carnegie, 2006). The difficulty level of tasks can infer through measurement of correlation between the student's performances and assessment questions. To compute the probability of a student solving a task correctly, a set of predictor variables are defined in the LFA including the number of opportunities presented to the student at each task, the duration spent on each step and the difficulty level of each question or latent variable. The results reveal that incorporating the latent variables into estimates of student performance significantly enhances the model (Hao Cen, Kenneth Koedinger and Carnegie, 2006)(Koedinger, Mclaughlin and Stamper, 2012).

The Hidden Markov model (HMM) is another method used to measure the latent variable. The HMM model is a statistical model used to compute the probability for the sequences of observable events based on the Markov chain; The model assumes that the probability of events depends on the state that is not directly visible in the previous event (Jeong *et al.*, 2008).

In virtual learning, HMM is used to analyse the students' interaction with ITS. As such, the HMM is employed to discover the 'slip and guess' of students answers in ITS assessments. Bayesian Knowledge Tracing (BKT) is proposed by reference (Beal, Mitra and Cohen, 2007) to predict student performance. The authors define the

students' response into four categories; P (L0), P (T), P(S) and P(G). P (L0) represents the initial probability that the student knows the skill in specific tasks; P (T) is the probability that the student will shift from an unknown state to a known state about the skill in tasks. The P(S) is the probability of a student obtaining the incorrect answer, given a known level of skill; P (G) is defined as the probability of the student obtaining a correct answer in the absence of prior experience. The findings show that inferring the guessing and slipping latent variables lead to improvements in the accuracy of the prediction model.

The two-layer hidden Markov model (TL-HMM) was proposed by Geigle to infer the latent behaviour pattern of students in MOOCs platforms. The TL-HMM model is different from conventional HMM in its capacity to discover the micro-behaviour patterns of students in more detail and transition between latent states. For instance, when students undertake the quizzes, the student would tend to participate in the forum discussion. The model can deeply learn specific transitions between the quiz taking and quiz submission. The author concludes that high performing students have fewer latent behaviour states due to the fact that they have sufficient knowledge, they do not need much assistance (Geigle, 2017).

The expectation maximisation algorithm (EM) is widely used to measure the latent variable. The EM is the statistical model used to find maximum posterior estimates of observation when data has hidden variables (Fraley and Raftery, 2007). Regarding online courses, the Latent Dynamics Factor (LADfG) is proposed by (Qiu *et al.*, 2016) to predict the learning behaviour based on EM algorithms. The authors employ the student activities in forums, assignments and videos to infer their latent behaviour. The result shows notable behavioural heterogeneity in students learning pattern. For instance, the students who patriciate in the forum are more likely not earn a certificate.

2.8 Learner Performance in Online Course

Student performance is considered a key indicator of the effectiveness of the MOOCs platform. Researchers have adopted various methods to predict the performance of students in the online environment. In this section we will summarise the work of other researchers towards predicting student performnce in MOOCs.

Within the educational setting, machine learning is an effective technique that has been widely applied, primarily to the prediction of student performance in both traditional and virtual environments. Kabakchieva (Kabakchieva, 2013) applied supervised machine learning to predict student performance at a Bulgarian University. The author considers 20 predictive attributes extracted from personal information and the pre-university characteristics of students. The Bulgarian Score Level scale was used to categorise student performance into five distinctive classes, given as "Excellent","Very Good", "Good", "Average", "Bad". Several supervised Machine Learning techniques were used to predict student performance, including Decision Trees, Naive Bayes, Bayesian Networks, and k-Nearest Neighbours. The results demonstrated that all classifier models suffer from low performance, exhibiting an average accuracy range of 52-67% (Kabakchieva, 2013). The authors in reference (Asif *et al.*, 2017) applied data mining methods to predict the performance of undergraduate students at the Engineering University in Pakistan. Similar to (Kabakchieva, 2013), five levels of outcome are considered as targets, for which GPA is employed as a predictive feature. The result reveals that Naive Bayes achieves the highest accuracy, with a value of 83% (Asif *et al.*, 2017).

A technique called Deep Knowledge Tracing (DKT) is introduced in (Piech *et al.*, 2015).The authors implement Recurrent Neural Networks (RNN) on the Khan Academy online course to predict the future performance of students. RNN is a dynamic model with the ability to continually represent the latent knowledge status over time, while evaluating the level of the student's knowledge. A number of variables are considered for the DKT model, these include the student's previous knowledge, student clickstream features, the hidden variables, the factor difficultly that is associated with each task, and additionally the duration of task taken by the students during the online session. The result showed that the RNN model achieves good performance with an AUC value of 0.85 (Piech *et al.*, 2015).

Students' marks in the first assessment and quiz scores in conjunction with social factors are used to predict students' final performance in the online course (Jiang *et al.*, 2014). Two predictive models were introduced. In the first model, logistic regression was used to predict whether students gained a normal or distinction certificate. In the second predictive model, logistic regression was also used to predict if students achieved a certificate or not. The results indicated that the number of peer

assessment is the most effective feature for acquiring a distinction. The average quiz scores were considered the strongest predictor for earning a certificate. The accuracy of distinction and normal models were reported with the percentage of 92.6% for the first model and 79.6 % for second model respectively (Jiang *et al.*, 2014). The sensitivity and specificity were not reported in this study.

An online education web-based system was employed to predict student performance at Michigan state University (Minaei-bidgoli *et al.*, 2003). A number of features have been considered in this study such as how long students interact with the digital materials, when students submitted the assessments and the total number of attempts undertaken. In order to, enhance the classifier performance, the authors used the Genetic Algorithm (GA) to optimize the high ranking features.Four classifiers have been considered namely decision tree, neural network, Naïve Bayes and k nearest neighbour. The authors compared the classification accuracy in respect to GA versus non-GA. The results illustrated the accuracy improved 12% by considering GA features. As such, the binary classifier achieved the percentage of 83.87 % accuracy in the case of selecting all features while the performance accuracy acquired a proportion of 94.09% when considering GA selected features.

2.9 Engagement in Online Courses

Engagement is perceived as the sense or feeling that increases the level of student interaction with activities and the duration of such participation. Student engagement is considered as an important prerequisite for learning in the online context, impacting performance, motivation, and attrition (Trowler, 2010).Engagement can be classified into three main categories: behavioural engagement, emotional engagement, and cognitive engagement. Emotional engagement occurs when a student senses or feels emotionally engaged in a learning activity. Cognitive engagement refers to the student's feeling with regard to progress in the academic task, while behavioural engagement refers to the level of student participation in learning activity (Trowler, 2010).

Behavioural engagement concerns student behavioural activities. The absence of behavioural engagement could negatively influence student academic outcomes (Trowler, 2010). Behavioural engagement is considered the crucial factor in increasing

concentration, persistence, and social interaction, resulting ultimately in improved student performance.

Learner engagement has been widely investigated in online learning. Coffrin et al in reference (Coffrin *et al.*, 2014) employ a learning analytic technique to analyse the patterns of participant engagement within MOOCs. The number of video hits and dates of assessment submissions are used as features during the assessment of completion rates. The result showed that only 29% of participants completed their assessment, whereas more than 60 % viewed the video (Coffrin *et al.*, 2014).

Video lectures and assessments marks are used to describe the prototypical patterns of learners' engagement within the Coursera platform on a weekly basis. Four patterns of engagement are introduced namely completing, auditing, disengagement, and sampling (Kizilcec, Piech and Schneider, 2013). The completing class represents learners who submitted assessments on time. Auditing class represents learners who did not submit assessments but watched video lecture content; Disengagement represents learners who drop out from the course; Sampling represents learners who watch video on only a single occasion .k-means cluster algorithm is used to find sub-populations of this engagement pattern, with results indicating that most learners engage with the course for the purpose of watching video lectures (Kizilcec, Piech and Schneider, 2013). These taxonomies are suitable for any MOOC platform that considers only video lectures and assessment. Consequentially, the narrow focus on these features imposes a limit on the generality.

In reference (Alias, Ahmad and Hasan, 2015) ,the authors employ Self Organised Map (SOM) clustering to describe the learners behaviour in the e-learning management system. They have found SOM clustering is a powerful approach in terms of visualising the behavioural patterns of learners, due to its capacity to analyse highly dimensional data with different types of input variables.

Other researchers examine factors relevant to the structural aspects of MOOCs design that could raise the level of participants' engagement (Hew, 2015). Learner comments are used to validate how instructional design promotes students' engagement. The authors' findings indicated that courses materials, interaction, and persistent monitoring of participant progress are critical elements that increase the level of engagement (Hew, 2015).

Probabilistic Soft Logic (PSL) is used by reference (Ramesh *et al.*, 2013) to model students' engagement. The PSL can be defined as a paradigm for developing the probabilistic model. PSL uses first-order logic rules to represent variables in the graphical model. Three types of engagement have been defined in this study, namely active engagement, passive engagement and disengaged. The learners' activities are defined as active when learners demonstrated interaction with the course such as posting on the discussion forums and submitting assessments. Passive engagement is assigned to the learners who access the homepage of resources, without proceeding to further forms of interaction like voting on a post, watching the lectures and viewing the discussion forums. Disengaged learners were those who tended to quit from an online course. The authors investigated the link between the engagement and performance. The findings illustrated that latent engagement enhances the performance of the predictive model. As such, the PSL model that accommodated the latent engagement achieved better performance than the model without latent engagement. The AUC = 0.7492 for the PSL model with latent variable, while the AUC acquired a value of 0.7393 for the PSL model without latent variable (Ramesh *et al.*, 2013).

Jiye et al. in reference (Baek and Shore, 2016) examine the relationship between social engagement and performance in the MOOCs platform. In this study, the course provided by Boston University is delivered through the edX platform. To measure social engagement, the authors split people into large and small groups. The results indicated that students within the larger group interacted more with discussion forums and acquired better performance than the small group as a consequence.The MOOCs can be similar to crowdsourcing where a large number of students who have various experiences would provide different resources for solving a particular task (Baek and Shore, 2016).

Gamification is used to measure students engagement in Blended University programming courses (Tvarozek and Brza, 2014). Interactive badges are given to students who solve the exercises correctly. Hence, the high skill students may gain more badges. The authors conclude the badges are a more accurate method than a self-reported survey to determine students' engagement level as self-reported student motivation could be evaluated from a narrow perspective; students tend to claim that they are engaged in the course to earn high marks in exams (Tvarozek and Brza, 2014).

The findings indicated that students interact with interactive badges in different ways. It was notable only 13.8 % of students frequently engage with gamification while around 52.5% interact randomly. The authors also demonstrates a positive impact of interactive badges on students' performance (Tvarozek and Brza, 2014).

National Survey of Student Engagement (NSSE) is used by reference (Sinclair and Kalvala, 2016) to evaluate students' engagement in the online course environment; NSSE includes ten benchmarks: Higher order learning, Reflective learning, Learning strategies, Quantitative reasoning, Collaborative learning, Discussions with diverse others, Student-faculty interaction, Effective teaching practice, Quality of interactions and supportive environment. The findings reported that an average of 30% of learners collaborated with others (Sinclair and Kalvala, 2016). Approximately 30% of students discussed the idea with other participants. Conversely, 10% asked for assistance as a consequence, the students' engagement level for collaborative learning criteria is significantly low. The result also posits an average 30% of students never interacted with any learning activity (Sinclair and Kalvala, 2016).Table 2.2 list the literature review work on engagement in online setting.

Table 2.2 Overview of Researchers Work on Examination of learner Engagement Pattern in MOOC

Author	Year	Dataset Provider	Data Mining Method	Feature Set
ArtiRamesh, et al.	2013	Coursera	PSL	Discussion forums, assessment marks, number of video hits
René F. Kizilcec et al.	2013	Coursera	k-means cluster	Video hits, assessments marks
Coffrin et al.	2014	"Principles of Macroeconomics and Discrete Optimization" Course university of Melbourne	LA	Video hits, assessments marks
Jozef Tvarozek et al.	2014	Programing courses, Blended University	Gamification {Interactive badges}	Number of assessments solved
U. F.Alias et al.	2015	E-learning management system	SOM	Student's action behavioural activities
Jiye Baek et al.	2016	edX	Survey	Discussion forums post
J. Sinclair et al.	2016	Online course	NSSE	Questionnaire

2.10 Self-Determined Theory

One of the most empirical theories of motivation in education is Self-Determined Theory (SDT). The SDT is a contemporary theory of motivation which has been used to explore why human activity occurs and what is the goal of that activity(Hofer and Busch, 2011). SDT has been widespread in the educational domain. The theory confirm that intrinsic motivation of student propensity is considered as the main factor for student participation in the specific task (Zhou, 2016).

The SDT posits that students have a basic psychological need to engage in learning activities. Psychological needs are autonomy, competence and relatedness. Students' satisfaction of these psychological needs could raise intrinsic motivation, and in addition, students will tend to elaborate in more advanced learning resource in contrast to, deprivation of these Psychological needs could impact negatively on students achievement(Zhou, 2016)(Leal, Miranda and Carmo, 2012).

According to SDT theory, motivation falls into three main categories: intrinsic motivation, extrinsic motivation and amotivation. Extrinsic motivation can be further classified into four types which are external regulation, interjected regulation, identified regulation, and integrated regulation (Leal, Miranda and Carmo, 2012). External regulation is the lowest autonomous type of motivation because students are undertaking the task in order to obtain tangible rewards or to avoid punishment. The interjected regulation students engage in tasks for eschewal of self-derogation purposes. Identified regulation is associated with students who participate in the task with the goal of valuable consequence outcome or for future career prospects (Osborne and Jones, 2011).

A number of recent studies have been conducted suggesting that SDT is an efficient framework to evaluate motivation in the online environment . A notable limitation of SDT is that it is focused only on surveys to categories participant motivation as a consequence, the motivation could be measured from the learners' perspectives(Velazquez-Iturbide, Hernan-Losada and Paredes-Velasco, 2017). Researchers find learners' activities are critical factors that could influence motivation. Therefore, Incentive Motivation Theory is introduced which considers the learners' behavior to evaluate students' motivation in the educational setting.

2.11 Incentive Motivation Theory

Incentive Motivation Theory (IM) is the behaviourist theory of motivation survey developed by B.F. Skinner (Martimort, 1996). IM seeks to explain why human activities occur relative to goals. The IM theory introduces the notion of "ramifications", which are posited to be the basis for task focused incentives. In particular, ramifications are classified into main subtypes: tangible and intangible (Martimort, 1996). The motivation categories are further explained in terms of three main dimensions: intrinsic incentive motivation, extrinsic incentive motivation, and amotivation (Martimort, 1996)(Ryan and Deci, 2000). Intrinsic motivation is attained from a student's perception of a task as interesting, challenging, and enjoyable. In contrast, extrinsic motivation originates from the expectation of rewards that lie outside of the activity itself (Martimort, 1996). Intrinsically motivated students feel immediate satisfaction while undertaking a task. Conversely, extrinsically motivated students derive satisfaction from extrinsic reward mechanisms such as attaining favourable exam marks or social rewards. amotivation is an another category of motivation in which the lack of incentive represents a key factor in student dropout (Martimort, 1996) (Ryan and Deci, 2000).

The advantage of IM is that it provides an explanation of the student motivation process from the different perspectives including psychological and cognitive. Additionally, IM addresses the association of motivation types with student academic performance Therefore, IM could help the instructors to enhance learning strategies and guide them to identify student-learning styles.

2.12 Motivation in Online Courses

Motivation is defined as the process for achieving a goal, which provides energy and initiates to accomplish a specific task. In terms of education, motivation is described as a conceptual construct that directs and improves student behaviour towards a specific goal (Cho and Heron, 2015)**.** Although motivation plays an important role in online contexts, few contemporary studies have evaluated motivation in online setting.

Current studies have highlighted the importance of motivation as a factor in the learners' engagement. Much of the research reported in the literature focuses on the validation of motivational indicators within the setting of online courses. Osborne et al. (Osborne and Jones, 2011) found a strong correlation between motivation and

domain identification within MOOCs (e.g. job prospects, knowledge expansion, social development). The authors demonstrated that social factors play an important role in increasing student engagement and enhancing cognitive skill.

The researchers in (Tatiana, 2016)employ mediation analysis with logistic regression to evaluate the correlation between engagement, motivation and achievement. In this study, the data was collected from 20 online courses provided by Coursera during 2014-2015 at the higher school of economics. The database is derived from two resources mainly, actionable data and surveys. The results reveal that the level of engagement acts as a mediator between motivation and achievement. The results also indicate intrinsically motivated students achieve good performance only in the first week of the course ,while extrinsically motivated learners have the incentive to complete the entire online course successfully(Tatiana, 2016).

To validate motivation in MOOCs, several studies have designed questionnaire frameworks based on the Glynn scale ("Science Motivation Questionnaire II") (Glynn *et al.*, 2011). The authors employed the Glynn scale to evaluate four types of motivation: intrinsic motivation, self-determination, self-efficacy, and career motivation, comparing English with Arabic participants within the Coursera platform (Barak, Watted and Haick, 2016). The results reveal a similar pattern of motivation category for both English and Arabic participants within the studied setting (Barak, Watted and Haick, 2016). The Situational Motivation Scale (SIMS) has been adopted by (Hartnett, George and Zealand, 2011) to measure learner motivation on two "teacher education" courses delivered by New Zealand Tertiary Institution. Four subtypes of motivations have been assessed in these studies, namely Intrinsic Motivation, External Regulation, Identified Regulation, and amotivation. The students were asked to respond to 16 SIM questions relating to particular assessments. The results demonstrate that participants in both case studies score high level of Identified Regulation and Intrinsic Motivation (Hartnett, George and Zealand, 2011).

Instructional Materials Motivation Survey (IMMS) has been applied by (Huang and Hew, 2016) to evaluate the learners' motivation level in the MOOCs environment. Four factors have been considered to model motivation in an online course: attention, relevance, confidence and satisfaction (ARCS). There were around 27 learners participating in this experiment who were enrolled on various MOOCs platforms. The results demonstrated that overall motivation level was positive for all criterias, in

particular, most learners were satisfied with the learning material and course structure. However, the learners were not satisfied with the instructor feedback as a large number of learners engage at the same time with an online course; the instructor could find it difficult to respond to all students (Huang and Hew, 2016).

Other researchers adopt machine learning techniques to predict learners' motivation levels (Wen *et al.*, 2014). Three sets of features were considered in this work. The "unigram" feature, which represents the main features set. "Linguistic" features that only used student comments in post form. If student comments are positive, then the post is classified as motivated, otherwise, it is unmotivated. The third set of features is "Unigram+Ling" that combines the unigram features with linguistic features. Logistic regression was implemented to investigate motivation mode on two Coursera courses. The results showed the linguistic feature has not had a significant impact on motivation levels. It was reported only 1-3% that enhance the performance of predictive model over baseline feature set. The result of logistic regression also revealed that "Unigram+Ling" achieved the best performance with values of 73%, and 62% for "Accountable" and "Fantasy" courses, respectively(Wen *et al.*, 2014) .The sensitivity and specificity were not provided in this study.Other studies investigate how motivation can positively influence learner performance. For example, Barba et al. (de Barba, Kennedy and Ainley, 2016) demonstrated that motivation has a significant impact on a learner's participation. Motivated students are goal-oriented people who tend to expand their experiences and overcome challenges (Barak, Watted and Haick, 2016). In the online context, researchers indicated that most online learners are intrinsically motivated rather than extrinsically motivated.

LA is used to evaluate learners' participation and performance in Coursera. The authors utilized video hits and quiz attempts as features, serving as an indicator of learner participation. The results show that most successful participants tend to be intrinsically motivated (de Barba, Kennedy and Ainley, 2016). In another study, sentiment analysis of participants' interview transcript within the Coursera platform was adopted in order to investigate the association between motivation and engagement (Shapiro *et al.*, 2017). Acquired knowledge and work are reported to be the main factors of influence for learner motivation in online course participation. In this work, learner experiences were found to be the critical factors affecting engagement and motivation levels. Learners with higher levels of education are more

likely to engage than those with less formal education, as they are found to have the ability to overcome barriers including technical and subject difficulty (Shapiro *et al.*, 2017).

According to Cho et. al. (Cho and Heron, 2015), Self-Regulated Learning (SRL) is a key factor for the achievement of motivation for learning. The SRL framework identifies student control, autonomy in the learning process, and time management as factors for successful goal achievement. A highly autonomous approach towards learning is a distinctive characteristic of the self-regulated learner. Cho et al examined SLR regarding motivation and learning strategy in an online mathematics course. The results demonstrate that learning delivery strategies did not significantly influence motivation. The researchers concluded that self-regulated learners are goal orientated and therefore tend to adopt critical thinking strategies in order to solve difficult tasks and develop skills. Table 2.3 list the research work on motivation.

Table 2.3 Overview of Researchers Work on Evaluation of Motivation in MOOC

Author	Year	Method	Finding
M. Hartnett et al	2011	Questionnaire frameworks based Situational Motivational Scale (SIM) to evaluate four types of motivation	Most students had high level of identified regulation and intrinsic motivation
M. Wen et al.	2014	Machine Learning to predict student motivation in Coursera courses considering linguistic and baseline features	Result of logistic regression achieve the best performance with values of 73%, and 62%.
S. Tatiana et al.	2016	Mediation analysis with logistic regression to evaluate the relationship between engagement, motivation and achievement	Extrinsically motivated learners incentive to complete course successfully
M. Barak et al.	2016	questionnaire frameworks based on Glynn scale to evaluate four types of motivation	The language barrier does not impact on motivation
B. Huang et al	2016	Questionnaire frameworks based Instructional Materials Motivation Survey to assess the level of motivation	Learners were satisfied with material and course structure instructor feedback satisfaction reported low
de Barba et al	2016	LA was used to evaluate learners' motivation by tracking students video hits and quiz attempts	Most students are intrinsically motivated
Shapiro et al.	2017	The sentiment analysis of participants' interview transcript in Coursera to examine the factors impacting on motivation	Acquired knowledge, work and learner experiences are the most effective reasons

2.13 Learner Attrition in Online Courses

MOOCs have attracted the attention of many researchers, with an aim to provide an advantage over traditional classroom environments. Much existing work focuses on participant attrition in MOOCs. In this section, we will summarise the work of other researchers towards learner attrition in MOOCs.

The attrition phenomenon was described by (Clow, 2013) as a funnel of participation. The term funnel of participation emerges from the equivalent concept in marketing (marketing funnel). The LA is used to describe the funnel of participation approach.

LA classify learners' theoretical stages toward dropout from MOOCs according to four main stages. Such stages are defined as Awareness, Registration, Activity, and progress (Clow, 2013). The author concludes that the fluctuation of learners' behavioral activities leads to withdrawal from online courses.

LA was used by reference (Ye and Biswas, 2014) considering temporal features and behavioral features to identify early student dropout in MOOCs. The results implied that an average 60% of participants who only watched the lecture withdrew from the course Conversely, 20% of participated students who were watched the videos and undertook quiz (Ye and Biswas, 2014). The researchers in reference (Li *et al.*, 2014) propose a multi-view semi-supervised learning model to address the issue of the dropout prediction problem.With this approach the unlabelled data was driven from the student behaviour record as the result, the prediction performance of inadequate label could be improved. In this study, six behavioural features were considered, these features included undertaking assessments, watching videos, accessing other objects, posting on the forum, and closing the web page. Four types of classifier were used to train each feature separately.The findings reveal that accessing other objects feature is most effective features into withdrawal rate.The average value of F-measures around 83%-84% for all classifiers (Li *et al.*, 2014).

Discussion threads are used to measure the negative behaviors of learners that lead to demotivated engagement within MOOCs platforms. Two kinds of features have been considered, namely click stream events and discussion threads (Yang and Rose, 2013). Survival models have been used by (Yang and Rose, 2013) for measuring the likelihood of attrition events. Survival models can be described as predictive models that apply logistic regression to infer the probability of learners' survival in the course over time. The results indicated that social factors significantly impact the dropout rate.

The author in reference (Kloft *et al.*, 2014) applies support vector machine to predict the likelihood of learner dropout from MOOCs. Feature engineering over time was considered in order to obtain more accurate prediction rates (Kloft *et al.*, 2014). Results reveal the good accuracy found at the end of the course, which has improved the predictive accuracy by 15% whereas weak accuracy was observed at the beginning of the course.

Other researchers emphasise on forum posting as a prominent resource of information for dropout analysis in MOOCs. In such works, the author in reference (Wen, Yang and Rosé, 2014) adopts a sentiment analysis approach considering only posts on the forum as the main criteria for analysis. The work considers the daily data of user forum posts and undertakes analysis in order to evaluate participants' opinions regarding the quality of teaching, learning materials, and peer-assessments. The results show a significant association between learner sentiment and attrition rate.

Although forum posting acts as a major factor affecting attrition rates, it has been observed that around 5-10% of registrants participate in the discussion forums. Consequentially, the narrow focus on the forum post data imposes a critical limit on the generality of the approach, since other important factors such as behavioural activities are not taken into consideration(He, Bailey and Rubinstein, 2015).

Feedforward neural networks have been implemented in (Chaplot, Rhim and Kim, 2015) to predict student attrition rates in MOOCs, using student sentiments feature and click stream as baseline features. The data was collected from 3 million student click logs in addition to 5,000 forum posts via the Coursera platform at 2014. The imbalance data is one of the critical issues of this study. The researchers overcome this issue by considering Cohen's Kappa criteria instead of accuracy. The results of neural networks achieve 0.74 when considering the sentiment features while results drop 0.70 in the case of excluding the sentiment attributes.

The model called "ConRec Network" deep neural networks is proposed by(Wang, Yu and Miao, 2017). The authors of this work combined the Convolutional Neural Networks (CNN) and RNN to predict whether students are at-risk of dropping out from the online course "XuetangX" in the next ten days. The students' records are structured according to the sequence of time-stamps and contain various attributes such as time a particular event should happen, the event type and student enrolment date. The neural Network consists of two parts, namely, the lower part and the upper part. In the lower part, the hidden layer of CNN was utilised to extract the features automatically. In the upper part, RNN used to make a prediction by aggregate and combine extracted feature at each time point.

To evaluate the prediction accuracy of the deep dropout mode. The models have been compared with various baseline methods; the results indicated similar performance

across all models. The F1-score results were reported to range value of 90.74-92. 48. Although, there was a similarity in performance, the authors argue that "ConRec Network" model is more efficient than baseline methods as it has the ability to extract the features automatically from student record without need the help of feature engineering(Wang, Yu and Miao, 2017).

The author in reference (Cobos, Wilde and Zaluska, 2017) examines whether differences between MOOCs platforms with respect to structure and theme could impact learners' dropout rate. Various machine-learning algorithms were used to compare two different MOOCs platforms: "Future Learn dataset" and "edX MOOCs". The results show that extreme gradient boosting (XGBoost) classifier acquired the highest accuracy=0.91 for a course delivered by "Future Learn" platform while Generalised Boosted regression Models (GBM) obtained accuracy=0.90 for "edX "MOOCs. The author concluded learners who engaged socially with other peers were more likely persevere in the "Future Learn " course. Conversely, the learners who spend time reading digital material completed the "edX" course successfully hence, course structure is one important factor that significantly influences the student attrition in online courses (Cobos, Wilde and Zaluska, 2017). Table 2.4 list literature review work on evaluation of students withdraw in online setting.

Table 2.4 Overview of Researchers Work on Evaluation of Learners Attrition in MOOC

Author	Year	Features	Method	Finding
Clow.	2013	click stream features	LA to describe funnel of Participation	Fluctuation of learners' behavioral activities leads withdrawal from online courses
Yang & Rose.	2013	Click stream events discussion threads	Survival models	Social factors affect the withdrawal rate
Kloft, et al.	2014	click stream features	Support vector machine	Predictive accuracy improved by 15% at end of the course
We et al.	2014	Forum posting	Sentiment analysis	Sentiment analysis results show a significant association, learner sentiment and attrition rate
Cheng Ye et al.	2014	Temporal attributes, behaviour features	LA	Undertaking quiz could reduce the dropout rate by by 40%
Chaplot et al.	2015	sentiments features and click stream features	feedforward neural networks	Neural network gain higher performance when consider sentiment features
Wentao Li et al	2016	Behaviour attributes	semi-supervised learning model	Unlabeled data can enhance performance of model
Wang et al.	2017	Behaviour attributes	ConRec Network deep neural networks	Deep learning able to extract the features automatically

2.14 Chapter summary

This chapter has discussed the background of MOOCs. A brief description of the EDA and LA is provided. The researchers demonstrate that LA is more influential than EDA in it is ability to analyze, capture the data in a more precise way and monitor the learning process. The background of students' performance has been introduced; the current literature on prediction of student performance has been reviewed in this chapter. It also describes the method of evaluating student motivation in online courses. A brief description of engagement style has been defined. The work of literature regard learner engagement pattern in the online environment has been

presented. The extensive literature reviews show that the major issues relating to MOOCs is the low completion rate. This is considered a lack of person-to-person interaction between instructors and learners on such courses. Moreover, the ability of tutors to monitor learners is impaired, often leading to learner withdrawals. To address this problem, learner dropout frameworks have been proposed by researchers. The work of other researchers towards learner attrition in MOOCs has been introduced in this chapter.

CHAPTER 3:Machine Learning

3.1 Introduction

Machine Learning is an application of artificial intelligence that is capable of performing a task without explicit human intervention such as recognition, disease diagnosis and prediction. The key feature of machine learning is providing the computer ability to learn from data and make an accurate decision without the need for human assistance (Samuel, 1959).

Warren Mc Culloch and Walter pitts in 1943 proposed the first artificial neural network. The network was very simple but had significant computing capability(Daniels and Mascini, 1943). Neurophysiologists introduced the idea of the artificial neural network with electrical circuits at the end of 1950(Baxt, 1995). They described the workflow of the neural network as being similar to human neurones. Subsequently, the neural network was designed by computer scientists and mathematicians to eliminate echoes over a phone line. The researchers simulated neural network processing information to solve a real problem similar to the human neural system(Birkett and Goubran, 1995).

In 1958, Rosenblatt introduced the perceptron algorithm for image recognition. With single perceptron, the neural network makes a prediction based on linearly separable classes (Schmidhuber, 2015). By 1965, Ivakhnenko and Lapa constructed the first neural network with multiple layers. A few years later inductive algorithms called "Group method of data handling" was introduced. The algorithms were capable of selecting the optimal structure of the neural network and automatically finding interrelations between features (Ahmadi, Mottaghitalaband and Nariman-Zadeh, 2007).

Samuel proposed a prominent game program to predict the winner in a game of checkers by 1950. The program assists the players in enhancing their skills (Samuel, 1959). The author found that machine learning could evaluate the board positions of a player like the human.

By this time, AI researchers had examined the role of simple machine analogies in learning data. They tried to identify the problem as a mathematical model that simulates the workflow of biological neurons (Birkett and Goubran, 1995). In 1969,

Minsky and Papert found two limitations in the machines that process neural networks. Firstly, perceptrons were incapable of recognising all of the pixels of the image. Second, the limited capacity of the processor to handle the large neural network (Schmidhuber, 2015). The neural network became very popular when Paul J. Werbos utilised the back-propagation algorithm to train neural network feasibility. In that period the researchers described back-propagation as the revaluation of the neural network(Nawi, Ransing and Ransing, 2006).

By 1981, Dejong introduced Explanation-Based Learning (EBL). EBL is an approach used in machine learning to learn and analyse data through selecting the important features that significantly impact on the target (Dejong, 1986). In 1990, AI scientists changed their direction with respect to the role of machine learning from the symbolic approach in solving a particular problem to the data-driven approach. With the data-driven approach, a large amount of data could be analysed based on the statistical approach and probability theory (Langley, 2011). By 1994, the weightless neural network was constructed. The topology of the weightless neural network differs from the standard neural network as it does not have weight and learning depends on memories(Aleksander *et al.*, 2009).

Deep learning was proposed by Geoffrey Hinton in 2006 (Sadiku, Tembely and Musa, 2017). Deep learning is a family of machine learning that is capable of extracting complex features from high dimensional data. The ability to learn the various levels of data representations that match hierarchy elements of complex relational architecture is one of the distinctive features of deep learning (Sadiku, Tembely and Musa, 2017). Alex Krizhevsk built the convolutional neural network in 2012. The convolutional neural is a difference in the topology. As such, the layer of such a network is arranged in three dimensions, which are the width, height and depth. In addition, the neurons of one layer are partially connected to neurons of the next layers (Schmidhuber, 2015).

Behemoth's deep learning system was released in 2014 in Facebook. The system uses a deep neural network to recognise the human face in digital images on social media. The network shows a 27% improvement over the previous deep neural network (Schmidhuber, 2015). During the same period, Ian Goodfellow introduced the Generative Adversarial Network (GAN). The GAN type of neural network algorithms usually used unsupervised machine learning. The GAN consists of two neural

networks, generative and discriminative. Generative learns the data from the latent space while discriminative discriminates between examples of actual data and instances from the generative network (Zhao, Mathieu and LeCun, 2016).

Although statistical method and machine learning share the same goal, they are different. As such, the statistical method is a mathematics model focused on a hypothetical test, which requires human effort to make inferences about the relationship between variables compared with machine learning where the computer can learn without requiring a specific human intervention. Machine learning is focused on predictions; it is based on computational learning theory where different assumptions of probability are used to evaluate generalisation errors. The statistical method's emphasis is on human assumptions that need a good understanding of data whereas machine learning identifies the hidden patterns of data through iterations (Goldenberg, Kubica and Komarek, 2003)(Demšar, 2006).

A key feature of machine learning is adaptive learning; it can learn a task by adopting a particular learning algorithm. The learning mode can be classified into two categories; supervised learning and unsupervised learning (Lawrence and Giles, 2000). A brief explanation of each mode is displayed in the following section.

3.2 Supervised Learning

In supervised learning, machine learning can learn the task by mapping function from input to output. This approach assumes that training examples contain pairs of input and output targets (Tan and Gilbert, 2003). The learning algorithm is used to map the given examples with actual outputs and generalised new data. The main issue with this learning approach is the bias-variance trade-off. It is simultaneously bias and variance error that prevents the learning algorithm from making an accurate prediction and generalising beyond training examples (Lawrence and Giles, 2000)(Nilsson, 2005).

The Bias error occurs due to the erroneousness of learning algorithms. It measures the difference between the model's predictions and actual values. The high bias causes the learning algorithms to be incapable of discovering the association of the features with target class and leads to the underfitting issue (Lawrence and Giles, 2000)(Nilsson, 2005).

Underfitting happens when the size of the dataset is small and the model cannot train the data well enough as a result, the model then makes the wrong prediction and gives

a low performance. Underfitting also occurs when fitting the linear model to nonlinear data(Domingos, 2012). To overcome the problem of underfitting, more data must be trained.

On the other hand, the deviation of prediction is called the variance. In this case, the predictive model fits well for the training dataset but does not perform well on new data. The high variance can add random noise to the learning algorithm and cause the overfitting problem (Lawrence and Giles, 2000)(Nilsson, 2005). Overfitting occurs when a machine-learning model captures the noise of the dataset. More specifically, the model learns the detail in the training dataset. Therefore, it fails to train with more observation and negativity affects the generalisation of new data(Domingos, 2012) (Lawrence and Giles, 2000)(Nilsson, 2005).

Several methods are shown by literature to reduce the overfitting issue, the most popular solution for overfitting is regularisation. Regularisation adds penalties to simplify the model. There are two types of regularisation, namely, Lasso Regularization (L1) and Ridge Regularization (L2). In the neural network, the penalty is added to the error function. The sum absolute value of weights is used in the Lasso Regularization(L1) method while Ridge Regularization (L2) uses the sum of squared values of weights as a penalty (Piotrowski and Napiorkowski, 2013).

Cross-validation can prevent overfitting by partition data into two subsets namely train and test where only one subset allocated for test and use the remaining subsets as the train. An early stop is a very intuitive approach and can be used to avoid overfitting. With this approach, the dataset is trained until a certain number of iterations are reached and the performance of test dataset is monitored. Since new iteration has not improved the performance dataset, the learning procedure should be stopped(Kai *et al.*, 2008). The ensemble is another approach to overcome the problem of overfitting. In this approach, multiple models are trained and an average of these models is used to produce the final model. Thus, a sample set of data is trained at each iteration instead of for the whole dataset. Finally, overfitting can be reduced by removing the irrelevant features(Domingos, 2012).

Supervised machine learning can be further classified into two taxonomies: classification and regression. In classification, the target class is the discrete label and regression is used when outputs are continuous (Tan and Gilbert, 2003).

In the context of the educational setting, supervised machine learning is used to track students' activities, predict students' performance and identify students' learning styles. In addition, machine learning is an effective tool which has been used to provide better learning materials that match the student's educational level (Dalipi, Imran and Kastrati, 2018)(Kabakchieva, 2013)(Lykourentzou *et al.*, 2009).

Machine learning is a promising solution for the detection of patterns of learner attrition from course activities through the examination of learning behaviour features over time. As explained in the previous chapter, supervised machine learning has been effectively utilised to tackle the withdrawal issue in virtual learning settings (Lykourentzou *et al.*, 2009). The next section discusses various supervised machine-learning methods.

3.2.1 Decision Tree

A decision tree is a hierarchical subtype of directed acyclic graph (DAG), constructed by performing two steps; recursion and partitioning. The tree structure consists of three canonical components: a root node, a set of internal nodes, and a set of leaf nodes. Each node acts as a processing element that acts on a subset of the pattern space performing a logical test on a particular attribute, for which outcomes are propagated by outgoing edges (Zimmerman *et al.*, 2016). Each successive transfer from a parent to a child node is adapted such that the homogeneity of the resulting pattern is increased concerning the outcome classes, a property defined as purity. Attributes of the highest discriminative power are represented in the root node. With lessening power towards the leaf nodes, the overall objective is that all leaf nodes will be completely pure (Rounds, 1980). When the tree size becomes too complex, the generalization error increases although the training error keeps decreasing resulting in the reduction of tree performance (Pal and Mather, 2003).

The splitting of the training set into many subsets leads to duplication of the same subset within one tree (Phyu, 2009). In some cases, all the attributes on the right path are duplicated on the left path, resulting in creating a tree which has two copies of the subset; this is known as a replication problem, and negatively affects the tree's efficiency (Pal and Mather, 2003) (Phyu, 2009).

Let X_t represent a set of training examples relevant to node *t* and Y=$\{Y_1, \dots, Y_c\}$ is a set of target classes. The tree is constructed by spliting the observation feature X into the various groups. For continuous features, the tree is grown up based on a set of test conditions and questions with expected results in a terms of binary outcomes {yes,no}. Node *t* is partitioned into two branches as follows (Zimmerman *et al.*, 2016).

$$\begin{aligned} t_l &= \{t \in X : A < V\} \\ t_r &= \{t \in X : A > V\} \end{aligned} \tag{3.1}$$

where A is the test condition with outcome V $\in \{0,1\}$, $t_l and\ t_r$ represents the left and right nodes of new tree *t.*

To evaluate the best split in feature space, a variety of measures have been utilised including Entropy, Gini, and Classification error determined as follows(Zimmerman *et al.*, 2016).

$$\text{Entropy}(t) = \sum_{i=0}^{C-1} \rho(i|t) \log_2 \rho(i|t)) \tag{3.2}$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{C-1} [\rho(i|t)]^2 \tag{3.3}$$

$$\text{Classification error}(t) = 1 - \max_i [\ \rho(i|t)] \tag{3.4}$$

Where $\rho(i|t)$ is the probability of recodes that is associated with class i at a given node *t* and C is the number of classes.

The main advantage of the decision tree is that the output can be easily interpreted, even by non-professionals as it is represented in graphical form (Podgorelec, Kokol and Rozman, 2002). Another benefit is in the handling of nominal and numeric parameters; it is the nonparametric method, which does not require normalisation of data. In addition, the decision tree can handle databases that have missing and error values. As a consequence, it could easily be incorporated with other classification approaches (Podgorelec, Kokol and Rozman, 2002)(Rounds, 1980).

One of the main drawbacks of the decision tree is the overfitting phenomenon. As mentioned, the concept of creating a decision tree model depends on a split dataset, which leads to increasing the number of nodes (Pal and Mather, 2003).

3.2.2 Random Forest

Random forest is an ensemble method that constructs multiple decision trees during learning time where each tree is generated using random sample vector provided from input features. Random forest can be employed for the classification and regression problems (Liaw and Wiener, 2002)(Ham *et al.*, 2005). In terms of classification, Random forest uses the voting mechanism that selects the most popular classes to classify the target. In regression, the weight averages of trees are used for prediction (Biau and Scornet, 2015)(Liaw and Wiener, 2002).

In theory, the training algorithm of Random forest follows the bootstrap method. Given the training dataset consists of N samples and M features. The first step in training algorithms is based on a bootstrap technique where each tree is constructed by randomly selecting several N samples with replacements. Next, Trees are created by selecting the M predictor variables that give the best split. The procedure is repeated multiple times, and the tree governs the growth without pruning until stopping criteria are achieved (Bharathidason and Venkataeswaran, 2014)(Laboratories, Avenue and Hill, 1995).

The main difference between Bagging and Random forest is that Bagging considers all features when splitting nodes while Random forest chooses only a subset of features randomly. Features within the particular subset of predictors that give the best split are used to obtain nodes in trees (Banfield *et al.*, 2007).

There are two approaches that can be used to choose the features in Random forest namely, Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA). The MDI is based on decreased weight of impurity tree. Multiple nodes are created where each node corresponds to a single feature. The Gini impurity for classification and variance in regression should be computed for each node and averages this quantity across all trees to gain weight of tree. The best features selected should have lowest impurity weight(Louppe *et al.*, 2013).

The MDA relies on the Out-of-bag (OOB) error concept. As mentioned previously, trees are constructed using bootstrap samples, some of the observation set side from bootstrap samples and are not used in building trees (Louppe *et al.*, 2013). The prediction error of left-out observations is called OOB error. To evaluate the importance of a particular feature, the value of this feature permutes into OOB

observation. The MDA for this feature is computed by the average difference of OOB prediction errors prior and post permutation across all trees. The feature with the highest MDI is the most important feature (Biau and Scornet, 2015)(Louppe *et al.*, 2013).

Random forest is considered the most accurate machine-learning algorithm due to its capacity to discover the nonlinear association between the features and targets. Also, it can run efficiently in high dimensional data (Ham *et al.*, 2005) (Biau and Scornet, 2015).

Random forest can handle the numerical and categorical values without concern for the deletion of observation. When a large amount of data contains the missing value, it can deal with missing data by adopting an imputation algorithm that keeps enhancing accurate prediction results (Shah *et al.*, 2014).

The main drawback of Random forest is its huge computational cost. The computational complexity of training algorithms is high compared with other machine learning models. In a high dimensional dataset, Random forest builds thousands of trees. Therefore, it could take more time during the training phase as a result; computational efficiently of Random forest is significantly increased(Shah *et al.*, 2014).

3.2.3 Gradient Boosting

Gradient boosting is a sequence of decision trees adopting the ensemble technique used for classification and regression tasks. The trees train sequentially where early shallow trees fit the sample model of data. Later trees try to fix the error of the previous tree. As a consequence, the final prediction model builds in the form of boosting the weak classifier into a strong classifier (Natekin and Knoll, 2013)(Ridgeway, 2007).

The Mean Square error (MSE) and Logistic Regression (LR) in regression and classification are used as loss functions in the Gradient boosting model where the goal is to predict new value by minimizing the error between the predicted values and actual values (Friedman, 2002)(De'ath, 2007). The optimisation algorithm of gradient boosting utilises to a minimum the expected values of loss function$\Psi(Y, \mathrm{F}(X_i)$ as follows (Friedman, 2002).

$$\mathrm{F}^*(X_i) = \arg\min_{\mathrm{F}(X_i)} E_{X,Y}\Psi(Y, \mathrm{F}(X_i)) \quad (3.5)$$

Friedman (2001) developed the first gradient boosting algorithm. The algorithm given the training sample $\{X_{i,}\ Y_{i,}\}_1^N$ of N data point. At first iteration M, the algorithm assigns the initial loss function. The loss function is used to map the input X_i to response Y_i .The error of loss function can be reduced through utilising the optimisation algorithm gradient descent. The $h(X_i)$ is the sample function used to teach trees of weak learners ("base learners") which fit the preemptive predictive model.

During the iterative learning process, the weight of data corresponding to misclassified samples is increased while the weight of correct classified sample decreases. With this approach, the errors of the weak learners' model can reduce and be fixed by combining the sum weights of all trees. The final prediction model can be provided by average weights of all trees as described in the following equations.

$$\mathrm{F}_{M+1}(X_i) = \mathrm{F}_M(X_i) + h(X_i) = Y_i \tag{3.6}$$

$$h(X_i) = Y_i - \mathrm{F}_M(X_i) \tag{3.7}$$

Where $\mathrm{F}_M(X_i)$ is the boosting approximates function and $h(X_i)$ is the weak "base learner" function. $The\ Y_i$ is the output variable.

Friedman (2002) developed the stochastic gradient boosting algorithm which depends on the randomisation aspect (Friedman, 2002). The random subsample of the training dataset was chosen without replacement and then used to fit the base learners across each iteration of the learning process. The author concludes that randomisation significantly improves the performance of the predictive model (Friedman, 2002).

The main crucial features of stochastic gradient boosting are the ability to prevent Overfitting in the dataset. Using the smaller subsample helps to reduce the variance of combined trees over various iterations. Furthermore, the computational cost is less in stochastic gradient boosting than original gradient boosting (De'ath, 2007)(Nawar and Mouazen, 2017). The algorithm would teach and fit the subsample instead of the full sample of the dataset.

The critical limitation of gradient boosting is the complexity of tuning parameters. Gradient boosting builds series of trees. Within each tree, three hyper-parameters should be considered: learning rate, number of trees and depth of trees. In contrast, the random forest constructs the trees in parallel where only number and size of the trees

are taken into consideration during the tuning procedure (Olinsky, Kristin and Brayton, 2012).

3.2.4 Generalised Linear Model

The generalised linear model is a statistical method that is used for linear mapping between the observed variables and response variables through a specified link function. The Generalized linear model assumes that the observations follow a particular distribution, namely; Average, Binomial, Poisson and Gamma distribution (Kumar, Naughton and Patel, 2015)(Nelder, J.A. and Baker, 2014)(Czado and Tu, 2004)(Liang and Zeger, 1986).

In the Generalised linear model, we assume { $X_{1,\dots}X_n$ } is n observation with dependent variable η_i , each linear predictor η_i is generated from a particular distribution. The simple Generalised linear model can be described according to the following equation (Nelder, J.A. and Baker, 2014)(Czado and Tu, 2004).

$$\eta_i = \beta_0 + \beta_1 X_1 + \dots \beta_n X_n \tag{3.8}$$

Where X_i is the predictor variables with the coefficients value β_i . β_0 is intercept, it can be interpreted as the mean value of $\eta_i$when all predictor variables are set to value zero. The link function is used to transfer the mean of expected values of response into the linear model form. There are several link functions that can be used to fit the values to linear model scale such as Identity, Log, Reciprocal, Logit and Probit (Statistics, 1986). The basic formula of link function is defined as(Rodriguez, 2013).

$$\eta_i = g(\mu_i) \tag{3.9}$$

$$\mu_i = g^{-1}(X_i \beta_i) \tag{3.10}$$

Where $g(\mu_i)$ is the link function and η_i is the linear predictor. As can be seen, in equations 3.9 and 3.10 the linear predictor η_i equals the mean μ_i inverse expected value of predictor variables since the data follows exponential family density.

One example of a Generalised linear model is a logistic regression. The logistic regression can be used in a classification problem when the response variables are discrete values (Shaliz, 2012). To make a prediction, we assume that the input variables correspond to vector features that are denoted as X_i and response variable represented in the target class Y_i.

To predict class label, the maximum likelihood estimation is used. The gradient ascent is utilised to select the best parameter (Czado and Tu, 2004) (Czepiel, 2012). The gradient ascent can be defined as, hill climbing algorithm move small step to direction of optimal point. The weight of the previous step combines with the weight of the current step during the learning process until the optimal value is achieved. Due to maximum likelihood estimates, the class probabilities of the sigmoid function is derivative to convert estimated probabilities to discrete value. The sigmoid function is an S-shaped curve that transfers any real number in to value between 0 and 1.

Logistic regression is an efficient technique. It can be implemented easily since scale of the features and tuning parameters are not required. It shows an advantage over the General linear model. As such, in Logistic regression, the response variables could be generated from different distributions while the response variables should normally be distributed in the General linear model (Shaliz 2012). Another advantage of Logistic regression is that the cost with respect to computational complexity is low. It takes a small amount of time during the learning process (Shaliz 2012). The critical limitation of Logistic regression is that it is unable to solve nonlinear problems since it is a generalised linear model.

3.2.5 Neural Network

The human brain contains nearly one hundred billion brain cells, known as neurons. These neurons exist to pass information to individual target cells, with communication signals being sent through synapses – these are structures that connect the cell's plasma membrane to the membrane of the target cell, playing an important role in the nervous system (Howard-Johnes,2010).

Neural Networks are a problem solving methodology grounded in the connectionist paradigm, comprising networks of interconnected simple units whose adaptive parameters may be tuned to form an emergent solution. In particular, neural networks are modelled as a cannibalized abstraction of the biological neural networks found in the mammalian brain, aiming to capture the information processing capability of such structures(Perner, Zscherpel and Jacobsen, 2001).

The connected nodes are called artificial neurons; these neurons are connected via edges. The synapses can transfer the signal from one node to another similar to

synapses in the biological nervous system. Each neuron and edge has a weigh. The activation function is used to determine the output of neuron by computing the weighted sum of input nodes and adding the bias. There are two types of activation function, namely; linear activation function and nonlinear activation functions. The linear functions have the limited capacity to learn the complex mapping between the input variables and the target. Therefore, nonlinear functions are utilised that able to represent and learn complicated tasks(Ahmadi, Mottaghitalaband and Nariman-Zadeh, 2007).

Most popular activation function used in neural networks are; Linear threshold, Sigmoid (logistic) and Tanh(hyperbolic tangent). Since Linear threshold gives the discrete output, it is only working with binary classification with the output (Active (yes) / not active (no)). It could be hard to train and converge the neural network for multiclass tasks. The nonlinear sigmoid and Tanh functions consider probability that can be used in multiclass tasks. Nonlinear sigmoid and Tanh are differentiable, meaning the slop of functions can be found at any data points. The output of Sigmoid and Tanh functions are in the ranges (0,1) ,(-1,1) respectively(Vecci, Piazza and Uncini, 1998).

3.2.5.1 Feed-Forward Neural Network

The most familiar type of artificial neural network is a Feed Forward neural network. The information transfer is in one direction without cycles; in this network, neurons belong to layer (i), receive input from layer (i-1) and transmit output to layer (i+1) through one hidden layer. The hidden layer contains the number of hidden neurons that enhance the sufficiency of the neural network(Hosseini, Luo and Reynolds, 2006)(Jadhav, Urmi, 2016).

The simplest type of artificial neural network is single layer perceptron network where the information transfers directly from the input layer to output layer via the weights. The activation function used in the single layer perceptron network is a linear threshold. The Delta rule-learning algorithm is utilized to train single layer perceptron network. For Delta rule, gradient descent calculates the error between actual and predicted output then choses the lowest error to adjust the network weight (Lambert, Johnson and Xue, 1998)(Marcialis and Roli, 2005).The activation function can be defined as follow.

$$g(x)=\begin{matrix}1\ if\ z>\theta\\-1\ otherwise\end{matrix} \quad (3.11)$$

$$Z=w_1x_1+\cdots+w_mx_m=(x+a)^n=\sum_{j=1}^{m}w_mx_m \quad (3.12)$$

Where $x_i$the input is value and w_i is weight. The Z is activation function based on the threshold (θ), the neuron is active if the out values of activation function is above the threshold.

A Multilayer perceptron (MLP) is a type of feed-forward neural network that is able to learn none linearly separable data. It consists of multi-layers of units. Usually, the MLP comprises three layers; one input layer, one output layer and at least one hidden layer. Each node fully connects to the other nodes in the following layer through a sequence of weighted edges(Hosseini, Luo and Reynolds, 2006) (Bullinaria, 2015).

The basic architecture of MLP is shown in figure 3.1. The MLP formally consists of a number of *L* layers where each layer has a number of nodes. The collection of units in the input layer can be described as$\{(L^i)\}\ _{i=1}^{N-1}$. The$\{(L^h)\}\ _{h=1}^{m-1}$ is the vector represented by the complete set of units in the hidden layer. The $\{(L^o)\}\ _{o=1}^{U-1}$ is also the vector represented neurons in the output layer. The collection of weight can be represented by two matrices { ${W_{ij}}^1$, ${W_{kj}}^2$} .The weight matrix that connects the input neuron to the hidden layer can be represented as ${W_{ij}}^1$ and the weight that link hidden neuron to out layer is ${W_{kj}}^2$.*B* is the collection $\{(B^i)\}\ _{i=1}^{L-1}$, where B^i denotes the column vector of biases for layer *i* + 1. Assuming the training dataset as the pair of input and output $\{(X_{1,}Y_{1,}),\ldots(X_{n,}Y_{n,})\}$, the input $X_{i,}$ transfers to input node in the input layer and the value of input nodes are multiplied by weight. The equation (3.13) compute adjusted weight. The output of input layer can be gain by fed weight (u_j) into activation function (σ) as follow(Bullinaria, 2015).

$$u_{j=}\sum_{i=1}^{n}{W_{ij}}^1X_i^n+\mathrm{b} \quad (3.13)$$

$$d_j=f(u_j) \quad (3.14)$$

The similar producer undertakes with output layer. The output of hidden layer is the input to output layers. The weight ${W_{kj}}^2$ is adjusted and weight sum (v_j) also fed into transfer function. The output of link function is represented the predicted outcome(Bullinaria, 2015).

There are different types of learning rules utilised to train MLP, the most common is Backpropagation. Backpropagation is widely used to train the MLP based on delta rules. The Backpropagation algorithm is used to compute neural network weight through gradient descent. More specifically, the optimisation algorithm, Gradient descent is utilised to find the optimal set of weights by computing the gradient of the loss function. The cost function computes the error between the actual inputs and the predicted outputs then calculated errors are propagated backwards to the previous layer. During the learning, the gradient descent adjusts weight iteratively by computing the derivative of cost function until it reaches the lowest error of cost function. (Chu *et al.*, 2007) (Schmidhuber, 2015).

Various factors could affect the performance of MLP such as; the number of hidden layers, the number of hidden neurons, the type of activation function and learning rates. The researchers demonstrate that increases in the number of hidden layers could significantly improve the performance of MLP. In terms of the number of hidden neurons, the researchers argue that the nonlinearity relationship between the features and the target can be improved by increase the number of hidden neurons. There is any assumption made by literature on how to select the optimal cost function(Jadhav, Urmi, 2016).

The learning rate is another important factor that could affect the performance of the neural network. If the learning rate is small, the training process becomes slow and may lead to a local minima problem. Nevertheless, if the learning rate is high, it might cause divergent behaviour of a cost function and might lead to a global minima issue (Kwak, Hanock, 2018). The local minima occur when Gradient descent, an adjusting weight involves taking steps toward the positive gradient that leads to it getting stuck in an undesirable point or local minima(Kwak, Hanock, 2018) .In addition, the learning rate of the network structure can influence the local minima. Deterministic and probabilistic approaches are used to handle the problem of overfitting (Atakulreka and Sutivong, 2007).

In the deterministic approach, the learning algorithm is global descent rather than gradient descent. Global descent computes the error of cost function at each iteration. Although global descent can allocate its local minimisers in neighbourhood points that give the better estimation, the computational complexity is the main limitation of this approach(Doria, Freire and Basilio, 2013).

The Probabilistic approach is based on the weight initialisation concept. In particular, the neural network learning with a random set of weights, the best neural network is selected with the lowest error. Although, this method is efficient, it requires a large amount of time for training (Kwak, Hanock, 2018).

The simultaneous training method proposed by(Atakulreka and Sutivong, 2007).Multiple neural networks train in parallel and all networks have set of initial weights. All networks simultaneously start learning from the first epoch until max epoch. The removal criteria stop running the neural net with the worst error. In this approach, the poor networks are eliminated which can reduce the probability of acquiring the local minima. The authors have shown the effectiveness of this method to avoid the local minima in comparing with the conventional method.

MLP can learn and model complex relationships between features. Therefore, they have been used to find accurate solutions for complex problems. Another advantage of MLP is that they can quickly make the correct prediction upon unseen data. The new data can be generalised even if it has a high degree of noise(Tu, 1996).The main drawback of MLP is their black-box nature; it could be hard to understand the features that affect the prediction. The interpretability of results could be hard to explain. It requires huge computational resources. The training of neural networks can take more time than traditional algorithms (Tu, 1996).

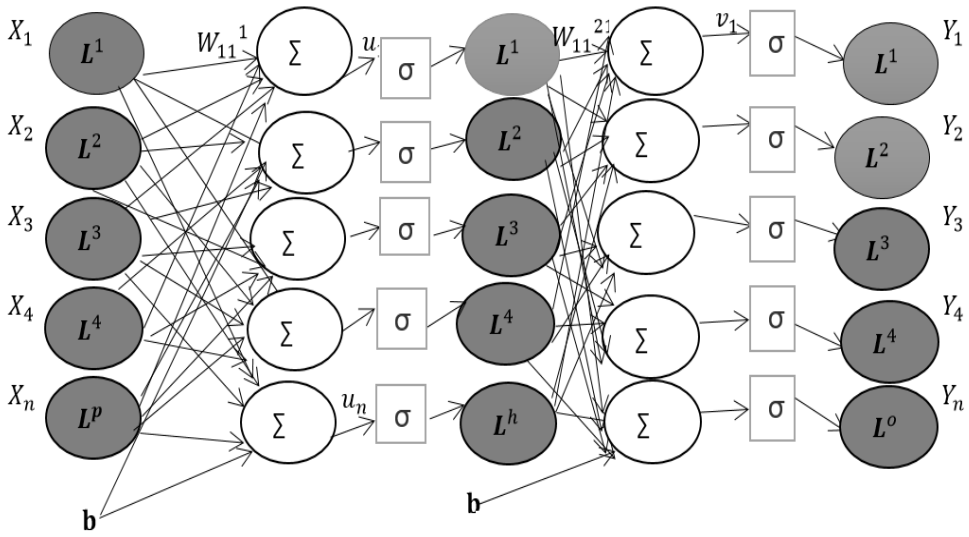


Figure 3.1 Multilayer Perceptron Neural Network Architecture

3.2.6 Support Vector Machine

Support Vector Machine(SVM) is a supervised machine learning algorithm that is capable of data analysis and can be used for classification and regression tasks. SVM classifies data by constructing the separating hyperplane that splits the data into two non-overlapping classes (Burges, 1998). More specifically, datasets split into training and test. The training set can be represented as a pair of input and output $\{(X_{1,}Y_{1,}), \ldots (X_{n,}Y_{c,})\}$ where X_i is n dimensional vector that denotes the features and $Y_{c,}$ is the class label where$Y_i \in \{class-1, class+1\}$. In the case where the dataset is linearly separate, the SVM predicts the target by finding the optimal hyperplane that maximises the margin distance between two classes. With this method, the SVM can differentiate the two classes very well. The hyperplane can be described according to following formulas (Moro, 2008).

$$\begin{array}{ll} \text{H} & W_i X_i + \text{b} = 1 \quad \forall X_i \in \text{class}+1 \\ & W_i . X_i + \text{b} = -1 \quad \forall X_i \in \text{class} - 1 \end{array} \qquad 3.15$$

Where W_i is the weight that gives information about the association of observation X_i with the target$Y_i$and b is the bias.

The distance between the two hyperplanes is described in equation (3.16). The following objective function is utilised to find optimal hyperplanes by minimising the weight and increasing margin width (Moro, 2008).

$$\begin{aligned} &\text{D} = \frac{2}{||W||} \\ &f(x) = \text{Min}\, \frac{1}{2} ||W||^2 \end{aligned} \qquad 3.16$$

If the data is not linearly separable, the hyperplanes' margins become soft. The kernel function (kernel trick) is used to transform the training data into high dimensional features space. The non-linear function kernel performs the linear separation by maximising the hyperplane margin in a transformed space (Chang and Lin, 2001). There are different types of kernel function, such as; Sigmoid, Linear, Non-linear, Polynomial and Gaussian radial basis function. Although the higher dimensional space could increase the generalisation error of SMV, the learning algorithm can learn well. The kernel function is described as follows (Moro, 2008).

$$\text{K}(X_{i,}X_j) = \varphi(X_i) . \varphi(X_j) \qquad 3.17$$

Where $(X_i).\varphi(X_j)$are transformation data point for observation X. The classification vector (W) in the transformed feature space can be defined as (Moro, 2008).

$$W=\sum_{i=1}^{n} C_i Y_i \varphi(X_i) \quad 3.18$$

Where C_i is the tuning parameter that controls the generalisation error of SVM. The good classifier should achieve the low weight of C and a wider hyperplanes margin. Classification of data using kernel trick is as follows (Moro, 2008).

$$Z=\sum_{i=1}^{n} C_i Y_i K(X_i, X_j) + b \quad 3.19$$

There are two main advantages of SVM. Firstly, in a nonlinearly separable dataset, it works effectively using the kernel trick. The kernel function implicitly performs the non-linear transformation without the need for expert human intervention. Secondly, the SVM is a powerful classifier that works well even when the dataset has some bias; it can give good generalisation if the tuning parameter is appropriately chosen(Karamizadeh *et al.*, 2014).

The main limitation of SVM is in how it chooses the kernel function. The user must tune the number of different parameters in order to gain the best classification result such as the kernel type, SVM type and gamma. A common disadvantage of SVM is the complexity. As such, in high dimensional space, the SVM requires extensive memory. With a greater number of samples, SVM becomes very slow. Therefore, it is not suitable for a high dimensional dataset (Karamizadeh et al., 2014).

3.3 Unsupervised Machine Learning

Unsupervised learning is a type of machine learning that has the capability to infer the pattern within a dataset without providing any corresponding output. Unsupervised learning is more complicated than supervised learning as it can learn tasks from unlabelled data without any response variables. The purpose of unsupervised machine learning is to discover and draw inference about a similar group of input observations. Various algorithms have been used in unsupervised learning including; Clustering, Mixture model and Self-organised map (Lloyd, Mohseni and Rebentrost, 2013).

Unsupervised learning has been widely used in a range of domains such as medicine, bioinformatics data, speech recognition, image processing, and finance. Researchers

have adopted unsupervised machine learning to compare the sub-populations of learners who engage in different learning activities (Lloyd, Mohseni and Rebentrost, 2013). The following section introduces the unsupervised machine learning methods that have been used in this research.

3.3.1 Fuzzy Cluster

The fuzzy cluster is a method of the cluster where each data point belongs to more than one cluster. The data points within each group should have similar measures while they are dissimilar in different clusters (Irani, Pise and Phatak, 2016)(Pal and Bezdek, 1995). A number of similarity measuring techniques can be used to measure the relationship between data such as; Euclidean distance, Manhattan distance and Minkowski distance (Jyoti Bora and Kumar Gupta, 2014). The Fuzzy cluster can be defined as the soft cluster where each data point assigns a different partial membership degree to all groups. The membership value is between 0 and 1 in contrast to hard cluster where each data point fully belongs to one cluster (Irani, Pise and Phatak, 2016)(Pal and Bezdek, 1995).

The most popular algorithm of the fuzzy cluster is Fuzzy C-Means clustering (FCM). The FCM was proposed by Dunn in 1973 and improved by Bezdek 1981(Pal *et al.*, 2005). It depends on minimising the objective function, in particular, the algorithm increases the similarity of data points within one cluster; however, the similarity of the data point is minimised among various clusters. The objective function can be defined as (Jyoti Bora and Kumar Gupta, 2014).

$$J_{m=} \sum_{i=1}^{N} \sum_{J=1}^{C} w_{ij}^{m} \left\| X_i - C_j \right\|^2, 1 < m < \infty \tag{3.20}$$

Where w_{ij}^{m} is degree membership of data point X_i belongs to cluster J. C_j is the centre of the cluster and X_i the data point measured. At each iteration, the membership degree for each data point is measured. This can be achieved by computing the distance between the data point X_i and cluster centreC_j. The probability of data point X_i across all clusters should equal one. The membership value of each data point is updated by selecting the cluster, which is nearest to it. The cluster centre $C_j$is also updated by recomputing the mean of all data points that belonged to it. The iterative optimisation

of the objective function continues until the cluster centre cannot be changed (Jyoti Bora and Kumar Gupta, 2014).

.

3. 3.2 Gaussian Finite Mixture Model

Mixture model is a probabilistic model that infers groups of observations within a population without prior knowledge of sub-group memberships. Mixture model has been widely applied in various domains such as; Statistical inference, Machine learning, Clustering, Classification, and Hidden variable modelling. The estimation of the parameters is based on cluster analysis, where the components represent a probability distribution across cluster memberships (Fraley and Raftery, 2007).

Different approaches have been used in the literature to determine the number of clusters. Such approaches can be classified into two categories, namely; Stochastic and Deterministic (Bouguila and Ziou, 2007). In stochastic approaches, the Markov Chain Monte Carlo (MCMC) method is employed. Deterministic approaches can be categorised into two main categories. In the first category, Bayesian criteria are employed such as the Bayesian Information Criterion (BIC) and Laplace Empirical Criterion (LEC) (Fraley and Raftery, 2007). In the second category, coding theory is considered for selecting the number of clusters, for instance using Minimum Message Length (MML) and Akaike's Information Criterion (AIC) (Figueiredo and Jain, 2002).

Gaussian finite mixture model is a popular type of mixture model. The key feature of this approach is the capacity to model complex data by mixing the properties of a density function of sub-populations into finite mixtures of components. In the finite Gaussian mixture, BIC and Integrated Completed likelihood (ICL) criterion are used to determine the number of clusters (Figueiredo and Jain, 2002).

Let X=$\{X_1,\ldots. X_n\}$ a sample of n univariate observations. The probability of X_i can be derived from the probability density function (PDF) as follows (Figueiredo and Jain, 2002).

$$\mathrm{P}\,[X_i] = \int_a^b p(x)\,dx \tag{3.21}$$

In mixture models, we assume observations are denoted by$X_n = (X_1, \ldots, X_n)$, where each observation belongs to g components. The empirical estimate of the PDF of X_i can be computed as (Fop, Murphy and Raftery, 2016).

$$\sum_g = \lambda_g \, D_g A_g D_g^T \tag{3.22}$$

$$f\,(X_i) = \sum_{g=1}^{G} \mathcal{T}_g \, f_g(X_i; \theta_g) \tag{3.23}$$

Where G is number of components and $\mathcal{T}_g$ is mixing weight of observation X_i associated with components of the g_{th} $(\sum \mathcal{T}_g = 1; \mathcal{T}_g > 0)$. $f_g(X_i; \theta_g)$ is the density of g_{th} component with estimated parameter θ_g in mixture model.

If the observation data follows a normal distribution, the Gaussian density function is considered to characterise the finite mixture model (FMD). In this case, within each cluster, the data is centred by the mean μ_g and the covariance$\sum_g$. The density of observation X_i takes the following form (Russell, Cribbin and Murphy, 2012).

$$f\,(X_i) = \sum_{g=1}^{G} \mathcal{T}_g \, \theta_g \, (X_i|\mu_g, \sum_g) \tag{3.24}$$

The covariance $\sum_g$ is used to specify the Geometric characteristics {shape, volume, orientation} of each cluster. Reference (Russell, Cribbin and Murphy, 2012) applies constraints on the covariance $\sum_g$ to represent the various models of elliptical clusters. The authors proposed the eigenvalue decomposition framework. The eigenvalue decomposition can be describe as follows (Russell, Cribbin and Murphy, 2012).

Where D_g is an orthogonal matrix and A_g is a diagonal matrix. The D_g, A_g parameters control the shape and orientation of g_{th} components in the mixture model while λ_g is constant which governs the volume of the g_{th} components.

3.3.3 Mixture Discriminant Analysis

The Mixture Discriminant Analysis (MDA) is a predictive model used for the supervised classification problem based on the mixture model. The model aims to assign the observation data belonging to the unknown class, to one of the true classes. The density of each class in the MDA model follows a finite Gaussian mixture distribution. The MDA can be described according to (Fop, Murphy and Raftery, 2016) formally defined as:

$$f(X_c) = \sum_{g=1}^{G_c} \mathcal{T}_{gc} \, \theta \, (X|\mu_{gc}, \sum_{gc}) \tag{3.25}$$

Where $\mathcal{T}_{gc}$ is the mixing weight of class c associated with the g_{th} component, such that $(\sum \mathcal{T}_{gc} = 1; \mathcal{T}_{gc} > 0)$. Accordingly, $\mu_{gc}, \sum_{gc}$ represents the mean and covariance of components g for class c respectively.

The MDA model, which assumes the number of components associated with each class, is known and the covariance matrix within each class is similar. In another study, Eigenvalue Decomposition Discriminant Analysis (EDDA) has been proposed by (Bensmail *et al.*, 1996), assuming that each class belongs to a single Gaussian component.

The Expectation-Maximization (EM) algorithm is typically used to estimate the model parameters in EDDA (Bensmail *et al.*, 1996); The EM algorithm consists of two steps, namely; Expectation (E) step and Maximization (M) step (Fraley and Raftery, 2002). During the E step, the conditional probability that an observation x_i associated with the g_{th} component is computed. In the subsequent M step, further parameter estimates are computed to maximise the expected log-likelihood obtained during the E step. The estimated parameters are then used to initiate further E-M steps iteratively until convergence. The ML procedure, therefore, continues until all observations are assigned to a cluster corresponding to the highest posterior probability (Fraley and Raftery, 2002).

3.4 Feature Selection

Feature selection has been used to reduce noise components and improve the performance of the prediction model. In terms of machine learning, feature selection selects a subset of features by eliminating redundant and irrelevant features (Guyon and Elisseeff, 2003)(Chandrashekar and Sahin, 2014). Applying the features selection approach to classification problems has been proved to enhance the predictive accuracy, decrease the training time and reduce computational complexity (Guyon and Elisseeff, 2003)(Chandrashekar and Sahin, 2014). There are various feature selection methods, namely; Wrapper approach, Filter approach and embedded method.

In the Wrapper approach, the classifier model is employed to evaluate the subset of the feature. Search algorithms are used to find an optimal number of features heuristically. In particular, the dataset is split to train and cross validation, backwards algorithm runs with a different number of features on each set, the set with the lowest validation error is selected as the final set. Although, the potential of the wrapper approach is in enhancing the predictive model's accuracy, the wrapper approach acts as a black box in the high dimensional dataset since the number of features have been

increased. In this case the features selction method could become computationally expensive (Chandrashekar and Sahin, 2014).

With regards to the filter approach, the optimal number of features are selected according to heuristic criteria without considering the classifiers process. There are various heuristic criteria which rank features in such a method, these including; Correlation coefficient, Chi-Square, Information Gain, Cross Entropy. More specifically, the weight is assigned to features based on these heuristic statistical data and features below the threshold are eliminated (Chandrashekar and Sahin, 2014).

As mentioned, the filter approach is independent of any type of classifier consequence. The machine learning algorithms that rely on the filter approach might achieve lower performance than the wrapper approach. Nevertheless, the filter approach could select the optimal number of features that might exist in the redundant subset. One of the advantages of the ranking method is the low computational cost (Guyon and Elisseeff, 2003). The brief description of features selction that has been used in this reserch project is presented in the following section.

3.4.1 Recursive Feature Elimination

The Recursive Feature Elimination (RFE) is one of most popular wrapper feature selection approaches. The RFE can be defined as an optimisation algorithm based on backwards selection and resampling techniques (Yun *et al.*, 2007). It keeps recursively creating the model until it gets a small number of features. The data set is partitioned into train and bootstrap samples with the different elements. At each iteration, the algorithms are chosen as the most important features. To assess probability of ranking features, the new model that includes the most important predictors is retained until all features are exhausted.

3.4.2 Hill Climbing

The Hill climbing is the search algorithm used in the wrapper selection method. It performs a partial exploration of features to find a candidate that is close to optimal(Nunes *et al.*, 2004). The algorithms perform sequential backwards selection to select the subset of features. The two subsets of features are compared to evaluate whether the new subset enhances the performance of the classifiers. The most popular algorithm of hill climbing is Random Mutation Hill Climber (RMHC). At each

iteration, the RMHC chose the random sample of observation, call it "best evaluated" then select a subsample of observation that nearest neighbour to the best-evaluated sample.

The algorithm compares the best-evaluated sample with selected subsample. If the prediction model of simple gives the better performance than best evaluated, then swap the samples until get the best features(Chandrashekar and Sahin, 2014). The hill climbing is described as the anytime algorithm which can give the optimal number of the features in all situations even if interrupted prior it ends.

3.5 Application of machine learning to identify at-risk-students in online setting

The low completion rate in MOOCs is the main concern of researchers. To tackle this problem machine learning has been used to identify the at-risk student at an early stage of the course. In the following sections, we summarise the other research work towards detection of at-risk students.

The authors in reference (He, Bailey and Rubinstein, 2015) identify at-risk students by applaying various machine learning algorithms including; Regularised logistic regression, Support vector machine, Random forest, Decision tree and Naïve bays. A set of features have been extracted from behavioral log data such as the number of times students visit a home page and the length of the session. The results illustrated that regularised logistic regression acquired the best AUC (He, Bailey and Rubinstein, 2015).

The VLE dataset is also employed to identify the students who are at-risk of failure (Wolff *et al.*, 2014). The authors select only three types of activities, namely; Resource, Subpage, and Forum to represent behavioral features. The input variables consist of three set of behavioral features followed by demographic features and students' previous assessment grades. The K nearest Neighbours, Classification & Regression tree and Bayes network are used for the prediction of an at-risk student. The result shows the first assessment has been shown to be a strong predictor of final success or failure. The sensitivity and specificity are not provided in this study.

The at-risk students were also identified on a weekly basis by reference (Jakub Kuzilek, Martin Hlosta, Drahomira Herrmannova, Zdenek Zdrahal, 2015) using the Virtual Learning Environment(VLE) dataset of the open university. Two sets of

features have been considered in this study namely behavioral attributes and demographic features. The results of machine learning indicates the proportion of at-risk students increased overtime. As such, the precision value dramatically increased from 0.50 at the beginning of the course to 0.90 at the end of the course while recall average value was stable with the range of 0.50-0.30. Again, the sensitivity and specificity are not provided in this study.

Other researchers consider the student dropout issue in the form of a time series classification problem (Balakrishnan and Coetzee, 2013)(Wang and Chen, 2016) (Taylor, Veeramachaneni and O'Reilly, 2014) (Fei and Yeung, 2015) (Li *et al.*, 2014). The Hidden Markov Model (HMM) has been applied on data collected from edX's platform to predict the student retention over time (Balakrishnan and Coetzee, 2013). In this work, the author proposed both a composite and an individual HMM. The study reported satisfactory performance for the composite HMM, obtaining an AUC value of 0.71, for which multiple behavioural features were considered as a source of input. Subsequently, the individual HMM provided insight into patterns of student activity, for instance, participants who do not check the course progress frequently were found to be more likely to withdraw following the fourth week of the course (Balakrishnan and Coetzee, 2013).

Clickstream data has been considered from which a series of features are extracted, such as the number of lecture videos viewed, the number of threads posted in online forums, and the number of quizzes attempted. The author in (Taylor, Veeramachaneni and O'Reilly, 2014) employs logistic regression to predict student dropout events within 6.002x platform. The authors split the course into fifteen time slices based on the weekly interval. The results show a best predictive performance of AUC 0.95, obtained from a week situated around the approximate midpoint of the course duration, with the lowest AUC of 0.77 obtained at the end of course (Taylor, Veeramachaneni and O'Reilly, 2014).

The Long Short Term Memory Neural Network (**LSTM**) has been applied by (Fei and Yeung, 2015) to predict student dropout in two MOOCs platforms, Coursera and edX. The results show that LSTM is the best classifier that is capable of discovering the nonlinear latent representation of the model.The accuracy, sensitivity and specificity are not provided in this study.

In a further study, at-risk students within online course settings have been investigated by (Wang and Chen, 2016). To understand the student motivation in relation to a particular activity, hidden latent engagement was analysed through application of a Nonlinear Status Space Model (NSSM) (Wang and Chen, 2016). The NSSM model was compared empirically with several other models, namely; Logistic regression (LG), Simultaneously smoothed logistic regression (LR-SIM) and Long-short-term memory (LSTM).Experimental results showed NSSM to acquire the highest performance, exhibiting an AUC value of 0.9 at the beginning of course. In contrast, the lowest AUC appeared at the end of the course with an AUC of 0.7. The results obtained indicate that the latent engagement patterns under analysis are time varying in nature (Wang and Chen, 2016).

Table 3.1 Overview of Researchers Work on identification of at- risk students in MOOC

Author	Year	Features	Method	Finding
Girish Balakrishnan et al.	2013	Temporal attributes Behaviour attributes	Hidden Markov Model(HMM)	HMM provided insight latent characteristics of students
Colin Taylor et al.	2014	Temporal attributes Behaviour attributes	Logistic regression	The AUC decrease over time and achieve value of 0.77 at the end of course
Wolff et al.	2014	Behaviour attributes (OULAD dataset)	(CART) classifier (k-NN) Bayes Network	The first assessment is a strong predictor of final success and failure
Jiazhen He et al.	2015	Click stream features	Set of Machine Learning models	Regularized logistic regression acquired the best AUC
Jakub Kuzilek et al.	2015	Demographic and Behaviour attributes (OULAD dataset)	Set of Machine Learning models	Machine learning gives a better result at the end of course
Fei Mi et al.	2015	Behaviour attributes	Long Short-Term (LSTM) netwrok	LSTM can find nonlinear representation of data
Feng Wang et al.	2016	Behaviour attributes	Nonlinear State Space Model (NSSM)	NSSM to acquire the highest performance, exhibiting an AUC value of 0.9 at the beginning of course

3.6 Summary

The chapter has presented an overview of machine learning algorithms with emphasis on the comparison between Supervised and Unsupervised algorithms. The machine learning that has been used in the project is introduced in this chapter. Learning algorithms have been described briefly. The section has highlighted the advantages and limitations of each classifier. Unsupervised machine learning has also been presented followed by the explanation of Cluster analysis and Mixture model. Mixture Discriminant Analysis (MDA) can be used as supervised classifiers. The Features selection method has been discussed, the difference between Wrapper and Filters approaches also explained followed by brief definitions of RFE and Hill climbing algorithms. At the end of chapter three, the application of machine learning for identification of at-risk students in the online context was introduced.

CHAPTER 4: Proposed Methodology

4.1 Introduction

LA is an effective tool for tracking student knowledge, precisely analysing behaviour, and measuring how such factors can affect 'at-risk' students (Siemens, 2013). Machine learning (ML) is capable of detecting potential patterns of learner attrition from course activity data, through the examination of learning behaviour features over time (Lakkaraju *et al.*, 2015). Moreover, machine learning has the scope to infer the underlying emotional status of learners, by discovering latent patterns of learner behaviour (Altrabsheh, Cocea and Fallahkhair, 2016). In the present study, machine learning, in conjugation with LA is applied to detect at-risk students. The proposed framework for the early detection of at-risk students has been explained in this chapter. The purpose of the proposed framework is to help educators flag at-risk students in their early stages and deliver timely intervention assistance to those students.

Further towards the above goal, both Harvard and Open University datasets are discussed, followed by an explanation of the data pre-processing procedure. The chapter also illustrate a detailed discussion of the set-up procedure of three experiments. The first two experiments applied to the Harvard dataset aim to predict students' performance during an independent course, and further evaluate the dynamic link between learners' educational background, engagement level and performance. The influence of crucial factors such as motivation and performance on students who are at-risk from withdrawal from such courses is investigated in the second experiment.

The LA is also used to characterise learner engagement patterns in MOOCs platform in first experiment. The student engagement pattern has been divided into two styles: {active} and {passive and active}. The unsupervised machine learning is used to discover the group of learners who share similar characteristics. The third experiment considers the Open University dataset. A detailed explanation of the extraction and selection features for VLE behavioural attributes is provided in this chapter. The analysis is performed according to the assessment submission date. The influence of performance trajectories on a given student's outcome is highlighted in this chapter, by considering former assessment grades as input to predict final student performance.

Our work differs from the prior research works as it concentrates on the analysis of various factors affecting the learners' outcome in MOOCs. As such, student performance and student behavioural activities at previous time interval consider as the input predictors. The regression analysis was used to predict student assessment grade based on the history of the student. Feature selection techniques were employed to discover if students performance in the previous assessments would impact his performance in next assessment.Furthermore, the impact of latent engagement on students and their subsequent risk of leaving a given course has been evaluated.

4.2 Database Introduction

In order to answer the research questions, two datasets were used. The first dataset was obtained from Harvard University, and contained information about 597,692 students during the first year of their courses. The second database was obtained from the Open University, which comprised of VLE data for a single course, namely, "Social Science". A total of 4,000 individual students were available for consideration. The two datasets differ in both scope and attributes. Table 4.1 compare datasets.

Table 4.1 Comparison of Harvard and OULAD Datasets

Harvard dataset	OULAD dataset
Multi-course scope The Harvard dataset is an aggregate dataset, comprising of course records over a set of differing topics. As a result, student behavioural activity for multiple courses is available to use during predictive inference, namely the use of information from a previous course to infer the trajectory of an independent course.	**Single Course scope** The OULAD dataset is a relational dataset that contains student records relating to the behavioural activity only for a single course. As a result, inference is limited to in-course analysis, where generalisation to additional courses cannot be evaluated.
Optional Assessment Students may choose whether to work towards a formal certification; the course material is made available, but assessments (including the final exam) are optional. Additionally, the student can access learning material after the course end date. An analysis of the participant's activity can therefore be used to identify intrinsically motivated learners.	**Mandatory Assessment** The students are required to undertake assessments (including a final exam) if they wish to remain on the course. Additionally, the students cannot access learning material after the course ends. Consequently, the intrinsic motivation of students cannot be evaluated for this dataset.
Video Engagement is Available The dataset includes features documenting the count of video views for each student in connection with their courses, providing a basis to examine the relationship between such activities and student risk factors.	**Video is not delivered** Videos content cannot be delivered through the course platform, meaning video engagement cannot be evaluated.
Course totals only	**Daily activity values**

The database does not provide a granular record structure for student activity over time; summary values are provided that incorporate totals, with the intermediate structure discarded. Consequently, intra-course dynamic engagement patterns cannot be evaluated.	Daily learning activities are provided, such that the evolution of student activity may be evaluated over the duration of the course. As a result, dynamic patterns in student activity can be evaluated.

4.3 Data Description

4.3.1 First Dataset Description

The Harvard University collaborates with The Massachusetts Institute of Technology (MIT) to deliver high quality MOOCs (Massive Open Online Courses). During the first year of providing MOOCs, 16 courses have been offered by Harvard and MIT (Ho *et al.*, 2014). The courses cover a variety of subjects, such as Computer Science, Mathematics, Humanities, History, Health, and Social Sciences (Ho *et al.*, 2014). Across all courses, only 30% of registrants succeeded in achieving certification (Ho *et al.*, 2014). The approximate percentage of learners who viewed the main course content and then subsequently dropped out from the courses is reported to be around 25%. (Ho *et al.*, 2014). The number of overall participants has markedly increased, with 1.3 million unique learners engaged in multiple courses reported at the end of 2014 (Ho *et al.*, 2015). Two sets of features are considered in the dataset - learner behavioural features, followed by demographic attributes (Ho *et al.*, 2014)(Ho *et al.*, 2015).

The primary feature of the dataset is the 'Click stream', which represents the number of user events relating to video lecture views, course content interaction, access to assessments, and posts in discussion forums (Ho *et al.*, 2014)(Ho *et al.*, 2015). The participants' demographic information is also considered in the dataset ('age', 'gender', and 'educational background'). Additionally, the date of learner registration in the course and the last learner activity was also captured (Ho *et al.*, 2014)(Ho *et al.*, 2015). The features denoting user exploration and viewed content are binary features that discretise the percentage of exploration and course content viewing, respectively (Ho *et al.*, 2014)(Ho *et al.*, 2015). If participants access more than half of the course content (chapter), the explored feature is encoded as 1, or 0 if otherwise. The viewed content is encoded as 1 when the participants access the home page of assessments and related videos, or 0 if otherwise (Ho *et al.*, 2014)(Ho *et al.*, 2015). The researchers have used these aforementioned features to measure what kinds of behavioural data could affect

the likelihood of certification gain. As such, the results show that during the first year there was a certification rate of 40%, where around 60% of the certificated learners fulfilled the criteria for explored participants (Ho *et al.*, 2014)(Ho *et al.*, 2015). A brief description of the dataset attributes is explained in Table 4.2.

Table 4.2 Harvard Database Description

Features	Type	Description
User-Id	Demographic	Learner identification number
YOB	Demographic	Learner date of birth
Gender	Demographic	Learner sex
LOE_DI	Demographic	Learner educational level
final_cc_cname_DI	Demographic	Learner continent area
Start_time_DI	Temporal	First date of learner activity
last_event_DI	Temporal	Last date of learner activity
ndays_act	Temporal	Number of unique days that learner interacts with course
Course Id		Course identification code
Nevent	Behavioural	Number of click stream
nplay_video,	Behavioural	Number of video viewed by learner
Nchapters	Behavioural	Number of chapter read by learner
nforum_post	Behavioural	Number of forum post by learner
Viewed	Behavioural	Discrete value describing if user accesses home page of videos
Explored	Behavioural	Discrete value describing if user accesses home page of chapter

4.3.2 Second Data Set Description

The second database in this study was obtained from the Open University, an institution located in the UK (Kuzilek, Hlosta and Zdrahal, 2016). The Open University delivers various online courses for undergraduate and postgraduate students. During 2013-2014, the Open University released a dashboard known as the Open University Learning Analytic Dataset (OULAD) (Kuzilek, Hlosta and Zdrahal, 2016). Two kinds of features have been considered in the database - namely demographic and behaviour. Here we consider the “Social Science” (“BBB”) course, which launched in October during 2013. The “BBB” module ran over 268 days.

The database is structured according to a relational schema, where all tables are joined to form a central composite table. The central table is designated “studentInfo”, and contains information relating to student demographic characteristics, such as gender, age, geographic area, and educational level (Kuzilek, Hlosta and Zdrahal, 2016).

In particular, the database contains fields relating to student performance and assessments, in addition to student interaction with online courses. In terms of behavioural features, a Virtual Learning Environment (VLE) system was used to capture student interaction within the online course setting. Each VLE is represented as an activity type, indicating the type of learning resources that the students are required to engage with within each module. There are various types of learning resources, such as reading PDF files, access to the home and sub-pages, and taking part in quizzes (Kuzilek, Hlosta and Zdrahal, 2016).

The table "StudentVle" includes information relating to student activities in particular modules. A series of student activities were collected on a daily basis and recorded in this table. The database captures daily information relating to student behaviour within an online course, in addition to the number of clicks that correspond to the students' interaction with the learning material on each day. The students are identified within both the "Vle" and "studentInfo" tables through unique numbers, providing consistent access to records (Kuzilek, Hlosta and Zdrahal, 2016).

The table labelled "Assessments" contains information about the number, weight and the type of assessments required for each module. In general, each module involves a set of assessments, followed by the final exam. There are two types of assessments, namely, the Tutor Marked Assessment (TMA) and the Computer Marked Assessment (CMA). The final average grade is computed with the sum of all assessments (50%) and final exams (50%). The "Student Assessment" table involves information relating to student assessment results, such as the date of the submitted assessment and the assessment mark. The student will succeed in the module if s/he gains an overall grade greater than 40%. To gain further information regarding the assessment, the Student Assessment table is linked to the Assessments through assessment identification number attributes (Kuzilek, Hlosta and Zdrahal, 2016).

The "Student Registration" table contains information about the date the students registered and unregistered in a particular module. The overall date is measured by counting numbers of unique days that students interact with courses until the course ends. In Open University online courses, students are able access a module even before being a student of the course; however, it is not possible to access the course post course close date.

4.4 Data Pre-Processing

Data Pre-Processing is an important step that enhances the performance of classifier models (Kotsiantis, Kanellopoulos and Pintelas, 2006). To achieve a more accurate analysis of the data, different data pre-processing methods have been applied over the two datasets.

4.4.1 Data Pre-Processing for the Harvard Dataset

Due to the large size of the dataset, a sample of 9,857 log file entries was sampled for each experiment. The log file records represent completed activities undertaken by learners on the respective MOOC platforms, where each entry corresponds to a single user session. The data Pre-Processing is divided into two distinct phases, implemented during the course of the procedure, namely, data cleaning and data transformation. Data cleaning was used to remove missing values, reduce noise, and remove inconsistencies within the data. On inspection, approximately 15% of the observations were missing for several behavioural variables, namely “Nevent”, “nplay_video”, “Nchapters” and “nforum_post”. The “YOB”, “Gender” and “LoE_DI” attributes are also included in the missing values. As a result, each incomplete observation was excluded from the candidate dataset. Subsequently, the dataset duplicate rows were also removed.

The Harvard dataset features have skewed distributions. Consequently, the data could suffer from the presence of non-normality. To overcome this issue, the Box-Cox transformation was used. This is a member of the class of power transform functions, which are used for the efficient conversion of variables to a form of normality, e.g., the equalization of variance, and to enhance the validity of tests for linearly correlated variables(Osborne, 2010). The data are furthermore standardised through scaling and centering, such that a mean value of 0 and standard deviation of 1 is obtained. The result of the transformation is shown in figure 4.1. The Box-Cox transformation was applied to only 10 of the features, as shown in Table 4.3.

Table 4.3 Box-Cox Transformation Harvard dataset

Features	Sample Skewness	Estimated Lambda
userid_DI	0.0135	0.7
final_cc_cname_DI	-0.569	1.2
LoE_DI	-0.163	0.7
YoB	-1.4	2
start_time_DI	-0.107	0.7
last_event_DI	0.0376	0.7
nevents	3.18	-0.1
ndays_act	1.76	0
nplay_video	6.21	0.1
nchapters	1.07	-0.4

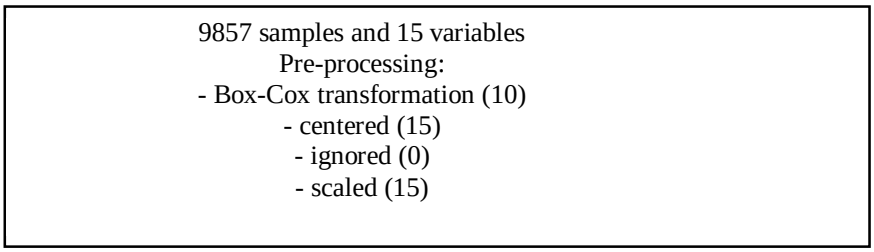


Figure 4.1 Data Pre-Processing transformation Harvard dataset

4.4.2 Data Pre-Processing for the OULAD Dataset

Because features are extracted in the OULAD dataset, the Pre-Processing procedure for OULAD is explained in section (4.8.3).

4.5 Predictive Model Evaluation Parameters

The Confusion matrix is used to evaluate the predictive model's performance. Furthermore, sensitivity, specificity, the F1-Measure, and accuracy are also utilized as quality measures, which are defined as (Oruç and Kanca, 2011)(Sing *et al.*, 2009).

Sensitivity = True Positive Rate (TPR)

$$\text{TPR} = \rho(\hat{C} = \oplus | C = \oplus) \simeq \frac{TP}{P} \quad (4.1)$$

Specificity = True Negative Rate (TNR)

$$\text{TNR} = \rho(\hat{C} = \ominus | C = \ominus) \simeq \frac{TN}{N} \quad (4.2)$$

False Positive Rate (FPR)

$$\text{FPR} = \rho(\hat{C} = \ominus | C = \oplus) \simeq \frac{FP}{N} \quad (4.3)$$

False Negative Rate (FNR)

$$\text{FNR} = \rho(\hat{C} = \oplus | C = \ominus) \simeq \frac{FN}{P} \quad (4.4)$$

Accuracy (ACC)

$$(\hat{C} = C) \simeq \frac{\mathrm{TP+TN}}{\mathrm{P+N}} \tag{4.5}$$

precision(p)

$$\mathrm{p}=\frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{4.6}$$

recall (r)

$$\mathrm{r}=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{4.7}$$

F_1-Measure (F1)

$$F_1 = \frac{2}{\frac{1}{r} + \frac{1}{p}} \tag{4.8}$$

Where, $\widehat{\boldsymbol{C}}$ and $\boldsymbol{C}$ are random variables that define class probability distributions for the prediction response and the actual class, respectively. The class outcomes are denoted as $(\oplus)$ for positive class and $(\ominus)$ for negative class outcomes. The empirical quantities P and N represent the number of positive and negative observations

The Receiver Operator Characteristic (ROC) and Area Under Curve (AUC) are also considered. ROC is a graphical representation in which TPR is plotted against FPR to generate a parametric curve that may subsequently be used to select appropriate cut-off values. AUC is defined according to (Vuk, 2006):

$$\mathrm{AUC}=\int_0^1 \frac{\mathrm{TP}}{\mathrm{P}} \quad d\frac{FP}{N} = \frac{1}{PN}\int_0^N TP\ dFP \tag{4.9}$$

AUC is used to measure the probabilistic classifier, which with the perfect classifier has a value close to 1. The probabilistic classifier randomly assigns a score for positive instances higher than the negative instances (Vuk, 2006). The scoring is computed based on the MaNnet Wilcoxon test (w) rules. The MaNnet Wilcoxon test is non-parametric, and is used to detect if observations in two different populations are identical. The MaNnet Wilcoxon test rules are described as (Hanley and McNeil, 1982):

$$\boldsymbol{s}(X_{\boldsymbol{p}}, X_{\boldsymbol{n}}) = \begin{cases} 1, & \text{if } X_p > X_n \\ 0.5, & \text{if } X_p = X_n \\ 0, & \text{if } X_p < X_n \end{cases} \tag{4.10}$$

The AUC is equivalent to the MaNnet Wilcoxon test (w), and can be computed as:

$$AUC \ = \mathrm{W} = \frac{1}{PN}\sum_{X_p \in pos}\sum_{X_{n \in neg}} s(X_p, X_n) \tag{4.11}$$

$$AUC = \rho(X_p > X_n) + \frac{1}{2}\ \rho(X_p = X_n) \tag{4.12}$$

Where $\boldsymbol{s}(X_{\boldsymbol{p}}, X_{\boldsymbol{n}})$ the score for probabilistic classifier is, X_p, X_n are probability-ranking examples that belong to positive and negative class, respectively.

With regard to the regression problem, Root Mean Square Error (RMSE), and relative square error (RSE), R-Square (R^2)are used to measure the performance of the regression model. The regression performance metrics are defined as follows (Huang and Fang, 2010).

$$\text{RMSE} = \sqrt{\sum_{i}^{n}\left(\widehat{Y}_{\iota}\right.} \tag{4.13}$$

$$\text{RSE} = \frac{\sum_{\text{i}}^{\text{n}}\left(\widehat{\text{Y}}_{\text{ı}} - \text{Y}_{\text{i}}\right)^2}{\sum_{\text{i}}^{\text{n}}(\overline{\text{Y}}_{\text{ı}} - \text{Y}_{\text{i}})^2} \tag{4.14}$$

$$\text{R}^2 = 1\text{- RSE}$$

Where, $\mathbf{Y_i}, \widehat{\mathbf{Y}}_{\mathbf{ı}}$) are vectors of actual values ($\mathbf{Y_i}$) and predicted values ($\widehat{\mathbf{Y}}_{\mathbf{ı}}$) for N observation. The $\overline{\mathbf{Y}}_{\mathbf{ı}}$ is the mean of actual values $\mathbf{Y_i}$.The difference between these two values is called a residual.

4.6 Experiment One Introduction

There are two sets of case studies presented in this experiment. The first case study examines the effectiveness of LA and machine learning approaches for the analysis and prediction of student outcomes within MOOCs. Behavioral features were used in conjunction with demographic features to predict whether learners gained a certification in MOOCs. LA are utilised to analyse the actionable data in greater detail. Machine learning is an effective technique that can be applied to LA, and has the capacity to discover patterns of student interaction with the MOOCs.

In this case study, machine learning in conjunction with LA is applied to predict if learners will achieve certification or not at the end of the respective course. The results of this experiment will assist educators in drawing inferences about students' performance and offer deeper insights. In the second case study, we suggest the use of unsupervised machine learning to discover prototypical learner engagement behaviour. We describe engagement patterns in two main categories: {active} and

{passive and active} engagements. The fuzzy clustering technique has been used to group learners who have similar prototypical engagement patterns. The Figure 4.2 shows the flowchart of experiment 1.

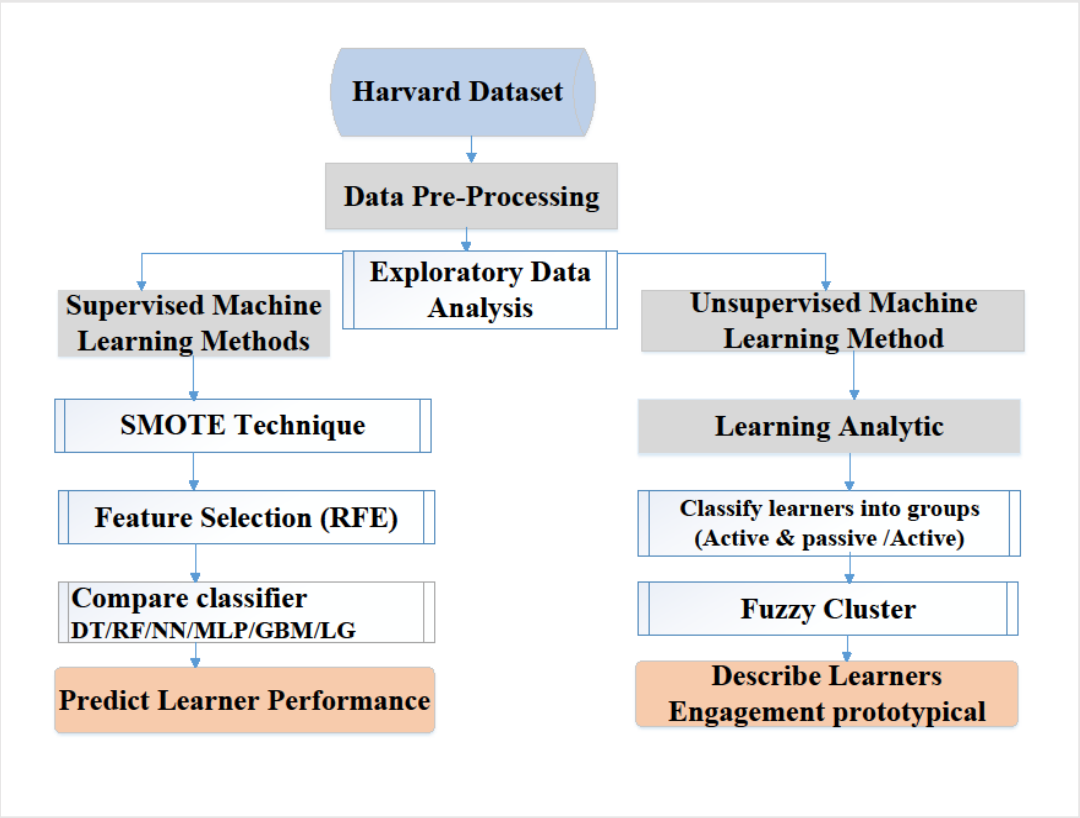


Figure 4.2 Experiment One Flowchart

4.6.1 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) is implemented in this study in order to gain an insight into the learners' behaviour, in conjunction with their performance. EDA is an important step within the machine learning approach, providing an intuition of the structure and relationships within the dataset (Kraska, T., Talwalkar, A., Duchi, J.C., Griffith, R., Franklin, M.J. and Jordan, 2013)(Leban, 2006). The objective of data visualisation is to provide information and an understanding of which type of features are more relevant to students' performance. The correlation matrix is applied to measure the dependency between the behavioural data and learners' certification. A heat map is used to visualise the correlation matrix more intuitively.

The Principal Component Analysis (PCA) is adopted to reduce the dimensionality and variance of a dataset. The PCA procedure facilitates a mapping from the original

feature space to a lower dimensional space, which is more conducive to learning (Chu *et al.*, 2007)(Pahor, 2009). To determine the number of principal components, the Kaiser method is used. The Kaiser approach is based on σ^2 to detect the number of optimal components, and retains components that have $\sqrt{\sigma} > 1$(Ferré, 1995).

4.6.2 Case Study One

4.6.2.1 Student Performance Prediction Model

The performance prediction model is designed to estimate learner certification rates in MOOCs. The learner must register in a course prior to accessing the coursework content. In order to be certified, the participant needs to achieve greater than 40% of the average grade. The average grade is calculated based on coursework, two mid-exams, and a final exam (Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., Chuang, I. and Ho, 2014)(D Seaton, J Reich, S Nesterko, T Mullaney, 2014). The coursework should be handed in on a weekly basis, and has a weight of 10% of the average grade, the two mid exams weighting is 40% and the final exam mark weighting is 50% of the average grade (D Seaton, J Reich, S Nesterko, T Mullaney, 2014). The certification is considered an inaccurate indicator of learning within the MOOCs. Due to free enrolment, a large number of learners interact with the course without aiming to undertake the final exam. Moreover, the participants who register after the course end date are precluded from obtaining a certificate. However, certificates are a good indicator to identify at-risk students as registrants who persisted in completing the entire course (Ho *et al.*, 2014).

Various linear and non-linear Machine Learning models have been used in the present study. The data is segmented into a number of subsets, with records of 8,000 learners in each subset. All dataset features have been considered, including both behavioural features and demographic variables.

4.6.2.2 Feature Selection

To identify the most important factors that impact student performance, a feature selection is used. Feature engineering could improve the performance of the predictive model by eliminating redundant variables (Granitto *et al.*, 2006).

The most important features that influence learners' performance are investigated. Two methods of wrapper approaches has been implemented namely Feature Eliminator algorithm (RFE) and hill climbing algorithm. RFE was used to select the

most important features. Each feature is ranked based on importance, using the random forest model (Yun *et al.*, 2007).The Hill climbing algorithm was applied to search for the optimal subset of features.

4.6.2.3 Synthetic Minority Oversampling Technique (SMOTE)

In binary classification, the number of instances should be equal for each class. The Harvard database is an unbalanced dataset, since 90% of the records are not certified (majority class) and 10% are certified (minority class). In this case, the predictive model will be more sensitive in predicting the majority class than the minority class; this leads to a bias problem.

To overcome this issue, the training data set should be re-sampled. In this work, the Synthetic Minority Over-Sampling Technique (SMOTE) is applied. SMOTE equalizes the class proportions by generating additional minority class examples. In particular, SMOTE applies a K nearest neighbours algorithm to interpolate new instances of each minority class through an evaluation of its nearest neighbors, according to a specific distance metric. Using this approach, the decision region of a minority class in the feature space becomes larger and more specific, and as a result, the training algorithm will obtain more results for the minority class. samples(Fernández *et al.*, 2018). The figure 4.3 offers a visualisation of the distribution of certified and non-certified students, using the SMOTE oversample approach.

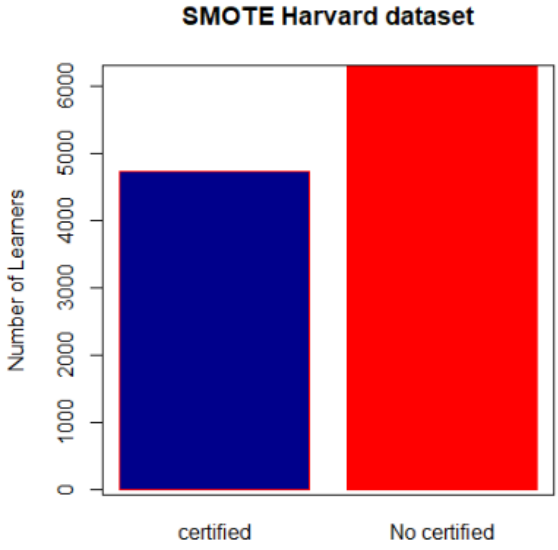


Figure 4.3 Smote Harvard Dataset

4.6.2.4 Evaluating Predictive Model

A ten-fold cross-validation involves five replications that are applied to assess the performance of classifier models. Cross-validation is capable of overcoming the problem of over-fitting by randomly partitioning the original sample of data into folds based on re-sampling (Pereira, Mitchell and Botvinick, 2009). Accordingly, 70% of the original dataset was allocated to the cross-validation training set; the subset elements of the training set were randomly partitioned into 10 equal-size subsets. For each round of cross-validation, 9-fold subsets are used as the training set, and the single subset is used as a test sample. A further 30% of the data is disjointed from the cross-validation set, and was used to evaluate the generalisation errors for each classifier.

To measure the predictive capabilities of classifiers over the test data, both ROC analysis and confusion matrix values were computed, forming the basis for comparing model responses to ground truth labels over each model. The details of ROC and the confusion matrix are explained in the "Predictive Model Evaluation Parameters section".

4.6.3 Case Study Two

4.6.3.1 Categorization Learners Based on Engagement Type

The LA approach is used to describe prototypical user engagement patterns. Only behavioural features were considered. Our hypothesis is based on the feature descriptions, as explained in the previous section. Here, members of the set of behavioural features have been combined to categorise the learner engagements within MOOCs in a meaningful way. We define the construction of the derived features according to logical aspect. The learners have been categorised based on engagements, into two main categories: {active} and {passive and active}. Below, a brief explanation of each class is provided.

Let V represents a set of students records, $|V| = N$ which is the number of students.

Let $R_i \in V$ represents the i_{th} student record, given as:

$$R_i = <v_i, g_i, s_i, e_i, c_i, l_i, w_i d_i u_i \mathrm{p_i}\ \mathrm{o_i}>$$

where v_i - Identity of the student for the i^{th} record

g_i - Grade of the i^{th} student record

s_i - Start date of the associated student interact with course

e_i - End date of the associated student interact with course

c_i - Identity of the course associated with the i^{th} entry

l_i - Launch date of the course referred to by c_i

w_i - Wrap date of the certification is issued by c_i

d_i - Number of videos viewed by i^{th} student

u_i - Number of chapters read by i^{th} student

p_i - Number of forum post by i_{th} student

h_i - Learners access home page of course content (chapter)

• Active Learners: Learner activity is defined as an active engagement activity, wherein learners demonstrate interaction with the course platform, such as interaction with a particular chapter, watching a video, and posting in forum, as defined in Equation 4.15.

$$ALg = \{\forall v \in V \| [\mathrm{d} > 0] \vee [U > 0] \vee [p > 0] \ \wedge [\mathrm{h} = 0] \| \quad (4.15)$$

• Active & Passive Learners: Learner activity is described as comprising of both active and passive characteristics, wherein learners explore both the content of the home pages and subsequently continue to make use of the chapters. This group of learners are described in Equation 4.16.

$$APLS = \{\forall v \in V | \left[[\mathrm{d} > 0] \vee [U > 0] \vee [p > 0] \ \wedge [\mathrm{h} = 1] \right] \quad (4.16)$$

4.6.3.2 Unsupervised Machine Learning to Describe Prototypical Engagement

In order to analyse the structure of the data, the use of unsupervised clustering was considered by applying the Fuzzy C-means clustering (FCM)(Suganya and Shanthi, 2012). Euclidean distance is used as a metric to compute dissimilarities between observations. The features associated with the target are removed when implementing the fuzzy clustering algorithm. The aim of the procedure is to establish a comparison between the engagement types established in the literature, as described previously, in contrast with the evidence represented by our empirical dataset. Such a data driven

approach provides a viable means to test the learner engagement type's hypothesis, which features in the work of other researchers.

The use of an unsupervised data analysis exposes the intrinsic structure of the data, such that the correspondence between engagement types and localized data structure can be reviewed. In the aforementioned procedure, we derived four distinct engagement types to serve as training patterns, while a total of four clusters were defined as a convergence imperative. To obtain the four features, the two original engagement types that were discussed previously were further subdivided by learner success outcome, namely, a binary value comprising of a certification/no certification dichotomy. The derived features used as input for the fuzzy cluster algorithm are shown in Table 4.4.

Table 4.4 Features Definition Based on Engagement Type

Feature Id	Definition
1	Active & cert
2	Active &no cert
3	Passive-Active & cert
4	Passive-Active & no cert

4.7 Experiments Two Introduction

Two sets of case studies are conducted in this experiment, with the aim of offering key decision makers the opportunity to intervene to assist at-risk students. In the first case study, the relationship between students' performance and engagement is investigated with a view to consider behavioural features.

The statistical techniques has been applied to find the association between learner engagement level and performance in the context of the learner educational background and geographical location. The associated statistical analysis identifies the key discriminative features between the successful and failing groups and provides a segmentation of the outcomes in the context of learners' educational background and geographical location, which can facilitate educators in future MOOCs design.

In the second case study, the impact of motivation and performance on students who are considered at-risk of not completing the courses has been examined. The Harvard dataset did not explicitly define the students' motivational label. Therefore, LA is used to derive learners' motivations, based on IM theory.

Two temporal predictive models are built in this case study. The first model is designed to investigate the impact of students' performance in prior courses on students' decisions to drop out in the following courses. The second model is constructed to examine the influence of changes in students' motivational status on at-risk students. The figure 4-4 displays the flowchart of experiment two.

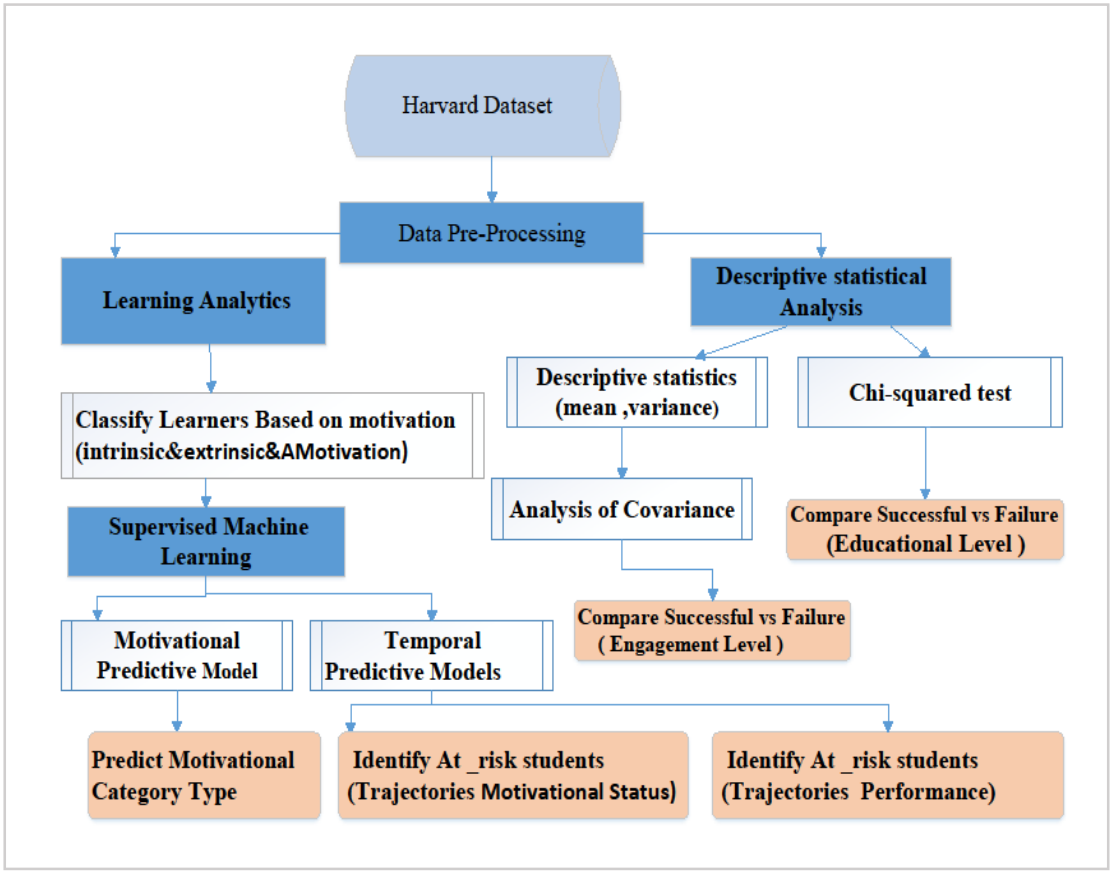


Figure 4.4 Experiment Two Flowchart

4.7.1 Course Description

In the present experiment, four courses are selected for analysis - "Introduction to Computer Science", "Circuits and Electronics", "Health in Numbers: Quantitative Methods in Clinical & Public Health Research" and "Human Health and Global Environmental Change".

In "Introduction to Computer Science", the course is focused on teaching students the use of computation in task solving (Guttag, 2014). The "Circuits and Electronics" course is an introduction to lumped circuit abstraction. The course was designed to serve undergraduate students of the Massachusetts Institute of Technology and was available online to learners worldwide (Mitros *et al.*, 2013). The "Health in Numbers:

Quantitative Methods in Clinical & Public Health Research" is a health research course that is designed to teach students adopting the quantitative method to monitor the health records of patients. In the "Human Health and Global Environmental Change" course, students learn to investigate how changes in the global environment could impact individual health. The reason why these particular four courses were selected is that the temporal information was only available for these courses.

All courses run in two semesters: Fall and spring. Fall courses were delivered in the fall of 2012, and the spring courses were covered in the spring of 2013. The courses differ in their structures and length. As such, all MITx courses run over a 15-week period, including a final exam and two examination periods. The HarvardX courses run over approximately 12-13 weeks. The MITx courses are entitled: "Circuits and Electronics Fall", "Circuits and Electronics Spring", "Introduction to Computer Science and d Programming Fall." In addition, "Introduction to Computer Science and d Programming Spring". The "Health in Numbers: Quantitative Methods in Clinical & Public Health Research" course provided by HarvardX launched by the end of 2012. Harvard also delivered "Human Health and Global Environmental Change" in the spring of 2013.

Table 4.5 Course Acronym

Course	Course Acronym
Circuits and Electronics Fall	Electronics Fall
Circuits and Electronics Spring	Electronics Spring
Introduction to Computer Science and Programming Fall	Computer Science Fall
Introduction to Computer Science and Programming Spring	Computer Science Spring
Health in Numbers: Quantitative Methods in Clinical & Public Health Research	Health Fall
Human Health and Global Environmental Change	Health Spring

4.7.2 Case Study One

4.7.2.1 Statistical Analysis Methods for Analysis Students' Activities

Various statistical methods have been employed in this case study to understand the patterns of participants' behaviour and explore how behavioural engagement can influence performance in MOOCs courses. A statistical analysis is capable of tracing and tracking learning activities in online courses hence, it could assist decision makers

in identifying the most effective learning tool (Brinton *et al.*, 2014). Brief descriptions of the statistical methods explored in our experiments are offered below.

- **Descriptive Statistics**: Descriptive statistics considers the utilisation of the mean and the standard deviation method (μ, σ). Those parameters are used in our case study to compare successful completion learners and non-successful completion learners in terms of location and engagement level. The students are distributed into 5 geographical areas, and 2 behavioural features are considered: the "nplay_video" and the "Nchapters". Learners are allowed to reattempt activities frequently, meaning that there is no specified boundary on the number of recorded attempts for each student per activity. Therefore, it is not possible to set a specific threshold relating to the number of click events for users watching the video and reading PDF files. Descriptive statistics help educators highlight the reason behind students' success and failure. The (μ, σ) descriptive statistics are described as follows (Friedman, 2001).

$$\mu_j = \left(\frac{1}{N_j} \sum_{i=1}^{N_j} X_{ji} \right) \tag{4.17}$$

$$\sigma_j = \sqrt{\frac{1}{N_j} \sum_{i=1}^{N_j} \left(X_{ji} - \mu_j \right)^2} \tag{4.18}$$

Where j is the location parameter, N_j is the total number of students at location j and X_{ji} is a students' interact with online courses from location j.The μ_j is the mean of i^{th} per particular location.

- **Analysis of Covariance**: To evaluate the results of descriptive statistics, the analysis of covariance (ANCOVA) is used. It is a statistical test used to test the mean of the independent variable across two groups. In our experiment, ANCOVA is used to determine whether the μ of successful and failing learners are identical, regarding their engagement level. The ANCOVA variable is defined as (Gribble, 2014):

$$Y_j = \sum_j^m \mu + T_j + \beta(C_j - X\overline{C}_j) + \epsilon_j \tag{4.19.1}$$

Where m is the number of geographical locationsG_1, G_m and n is the number

of success and failure students. In this case, μ is the population mean and $\overline{C}_j$ is a group mean. T_j is the effect of j^{th} group on the independent variable, and ϵ_j is the error term per j^{th} geographical location and X is the observation under the jth group. β is the slope of regression line. C_j is the covariate values of success and failure students in j^{th} geographical location. The C_j is defined according to the following equation.

$$C_j = \#\{Students\ S_1,..,S_n \in G_j\} = \sum_{i=1}^{n} \rho\left(S_i \in G_j\right) \quad (4.19.2)$$

Where $\rho(S_i \in G_j)$ is the probability of student S_i belong to particular geographical area.

- **Chi-Squared Test:** The Chi-squared Test is a statistical hypothesis test, which has been used to examine the difference between failure and success groups per course, with respect to learners' academic level. The Chi-squared test summarises differences between observed frequency values and expected frequency values for educational levels. The revealed results of the Chi-squared test help educators to determine if educational level factors can impact a learner's performance. A Chi-squared is defined as follows(Agresti, 2007).

  Let r represent the levels of educational background L_1 ,…, L_r and n represents the total number of success and failure students.

$$\chi_j^2 = \sum_{j}^{r} \frac{\left(O_j - E_j\right)^2}{E_j} \quad (4.20)$$

Where O_j is the number of success and failure students per j^{th} educational level described as (Agresti, 2007).

$$O_j = \#\{Students\ S_1,..,S_n \in L_j\} = \rho(S_i \in L_j) \quad (4.21)$$

E_j is the expected frequency of the number of success and failure students per j^{th} educational level and $\rho(S_i \in L_j)$ is probability of student S_i belong to j^{th}. E_j is given as (Agresti, 2007).

$$E_j = \#\sum_{i=1}^{n} E\ (S_i \in L_j) = \sum_{i=1}^{n} \rho\ (S_i \in L_j) \quad (4.22)$$

4.7.3 Case Study Two

4.7.3.1 Machine Learning Techniques

To evaluate the influence of engagement levels and motivations for students who are at-risk of quitting their courses, three predictive models were carried out in the present study namely, motivational predictive model and two temporal dropout predictive models.

The purpose of a motivational predictive model is to predict students' motivational categories in online courses. The problem is defined as a multi-class classification problem in such a model. Two temporal predictive models are proposed to predict at-risk students. The first temporal predictive model analysis is conducted according to student performance, while the analysis of the second predictive model depends on students' motivational status.

Machine Learning is applied over three predictive models. It represents a powerful data-intensive approach, which we apply within our proposed LA framework. Machine Learning is appropriate for the detection of students who are at-risk of not completing their next courses. In the following section, a brief description of the predictive models has been presented.

4.7.3.2 Categorization Learners Based on Motivational Status

Motivation and engagement are crucial factors that affect at-risk students. The Harvard database does not include the categories of student motivation, therefore, LA is employed in this experiment to understand the patterns of participants' engagement and motivation, and further explore how engagement could influence their performance in a MOOCs course. With LA learning, the students' performance trajectories can be examined in greater depth therefore. The decision makers should be able to acquire a deeper insight into the ground truth behind learner success and failure within MOOC platforms across various courses (del Blanco *et al.*, 2013). In addition, decision makers will be able to provide more attention to students who lack the motivation to persevere in their courses (Fei and Yeung, 2015).

To examine how such factors, influence students who are at-risk of dropping out, the taxonomy of learners is constructed, which relies on the Incentive Motivation Theory (IM) aspect. The following categories are defined based on IM:

Assume V represents a set of students records, $|V| = N$ which is the number of students. The explanation of students record R_i has been already described in section (**4.6.3.1**). Let $R_i \in V$ represents the i_{th} student record, given as:

$$R_i = <v_i, g_i, s_i, e_i, c_i, l_i, w_i d_i u_i >$$

Retention, completion and attrition learners are defined in detail below.

- **Retention Learners** (intrinsically motivated) are defined as those who engage in a given course without aiming to earn a certification, as defined in Equation (4.23)

$$RL = \{\forall v \in V |\ g = 0 \ \wedge\ [[\ l < s] \ \vee [w \ < e]]\} \quad (4.23)$$

Where V is the student's records, g is the grade, s is the join day, l is the course launch day, w is the leaving date, and e is the course end day.

- **Completion Learners** (extrinsically motivated) undertake courses with the expectation of obtaining a certification. Group is further divided into two subsets: those who pass and achieve certification, and those who do not pass. Pass Completion learners are defined in Equation (4.24), whereas Failure Completion learners are defined in Equation (4.25).

$$CLsc = \{\forall v \in V \mid\ g \geq 40 \ \wedge s \leq l \quad (4.24)$$
$$CLsn = \{\forall v \in V \mid 0 < g < 40 \ \wedge s \leq l \quad (4.25)$$

- **Attrition Learners** (amotivation) withdrew from the course within the same week, as expressed in Equation (4.26).

$$AL = \{\forall v \ \in V \mid g = 0 \wedge e - s < 8 \quad (4.26)$$

Algorithm 4.1 shows the groups of learners according to IM theory. Three groups were defined by considering the students' exam grades, course start and end dates, in addition to the first and last date that students interacted with the course. In both the RL and AL groups, students did not undertake the assessment, however, in the RL group, they engaged in the course longer than the AL group. Completion learners can be further classified into {CLsc, CLsn}. The assessment cut-off grade was used for distinguishing between these two groups. Due the records of extrinsically motivated

who engage after the course start date contain NA values of behaviours features, theses group of learners have been excluded.

Algorithm 4.1 Learners group according to IM Theory

1. $\forall V \in R^P: Ri = < v_i, g_i, s_i, e_i, c_i, l_i, w_i, d_i, u_i >$
2. $R_i \in RL \leftrightarrow g_i = 0\ ;\ l_i < s_i,\ w_i < e_i$
3. $R_i \in Al \leftrightarrow g_i = 0\ ;\ e_i - S_i < 8$
4. $R_i \in CLsc \leftrightarrow g_i \geq 40\ ;\ s_i \leq l_i$
5. $R_i \in CLsn \leftrightarrow 0 < g_i < 40;\ s_i \leq l_i$

4.7.3.3 Motivational Prediction Model

Students' motivation is the most important aspect of students' learning process. According to the previous definition of learning categories, the motivation predictive model is built in this case study. The multiclass classification is used where the set of label 1… L represents the target classes. In this experiment, learner motivation is classified into three distinct categories: amotivation, extrinsic, and intrinsic. The training dataset is represented as the pair$(\boldsymbol{F_i}, \boldsymbol{T_i})$, where $\boldsymbol{F_i} \in \mathbb{R}^{\boldsymbol{p}}$, denotes features of i_{th} observation and $\boldsymbol{T_i}$ are the targets. $\boldsymbol{T_i} \in \{1\ldots, \text{L}\}$.

Our training set consists of 3,373 data points, randomly sampled specifically from a subset of the courses considered, namely "Health Fall", "Electronic Fall" and "Computer Fall". Subsequently, a further 1,424 data points are randomly sampled from a separate subset of courses, comprising of "Health Spring", "Electronic Spring" and "Computer Spring", which is then used to evaluate the generalisation errors of each classifier model. The classifier model has been trained in one course, and the predictive model has been tested in another course. Such a scheme enables the generality of the features learned by the classifiers to be examined beyond the specific differences of the individual courses. The data has been selected with a balanced class over training and cross-validation. The proportion of each motivation category represented within the data for amotivation, Extrinsic and Intrinsic was 29%, 35%, and 36 %, respectively.

4.7.3.4 Temporal Models for Identifying At-risk Students

It would be impossible to track the temporal intervention behaviour of learners over a single course in the Harvard dataset. However, we can build a temporal model by adopting the course trajectories mechanism. In this case, LA is used for the students'

temporal records over their previous courses, with a view to investigate whether or not the students are at-risk of dropping out in the proceeding courses.To capture how students' performance and motivation could influence students' decision to abandon a course or not, the following two definitions of at-risk students are introduced.

- **At-Risk Student Definition(1):**Here, we consider the students who participated in the fall and spring courses within the same topic. In this case, if students engaged in fall courses and did not interact in the spring courses, they are defined as withdrawal students.
- **At-Risk Student Definition(2):**The learners who were engaged in both Fall and Spring courses are considered. As mention earlier, the students who withdrew from the course within one week are considered "amotivation" students. If a student's motivational status is "amotivation" during the Spring courses, then the student can be defined as withdrawal. Using this approach, LA could help course instructors provide the timely intervention to assist at-risk students. Figure 4.6 illustrates the at-risk student framework.

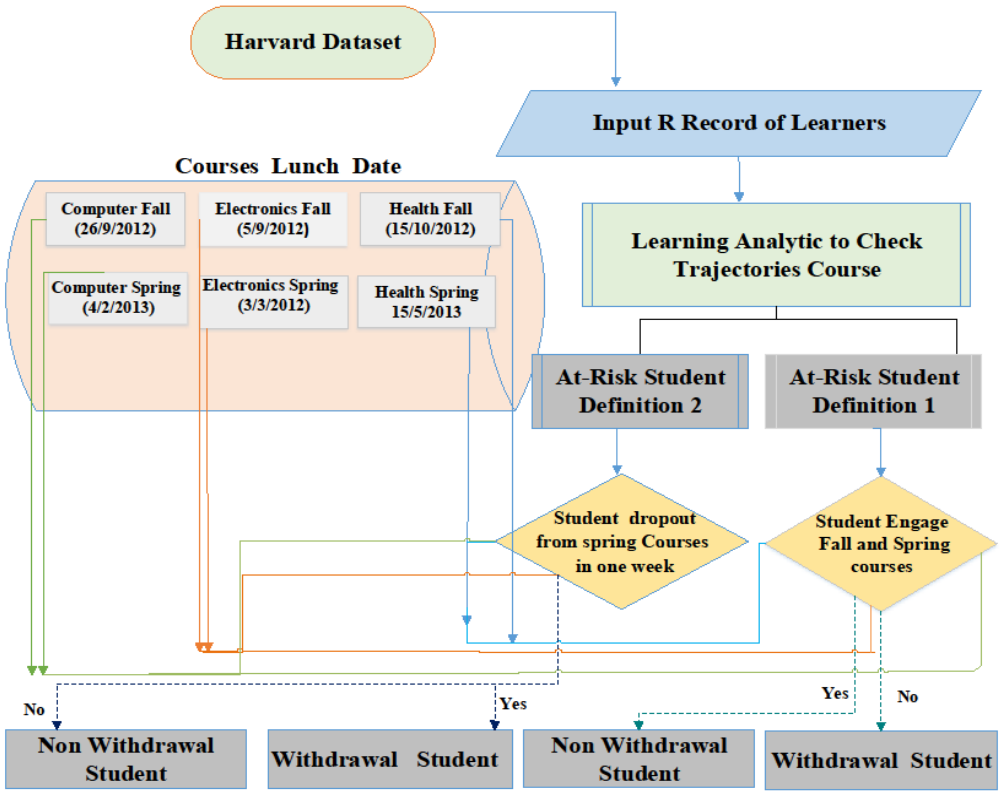


Figure 4.5 AT- RISK Student Framework in Harvard Dataset

4.7.3.5 Dropout Prediction Model Based on Student Performance

The first prediction model is based on the first at-risk student definition, and aims to investigate the impact of failure factors that influence students' decisions to persist in participating with another course. Therefore, only completion learners' groups who participated in the same topic are considered. for example, the students who failed the "Introduction to Computer Science" course, which was run in fall 2012. In the event that they were to re -enrol in the next "Introduction to Computer Science "course which was delivered in spring 2013, the students' record and their performance in previous courses are factors used to predict whether they were at-risk of dropping out in future courses. The trajectory analysis is based on course temporal launch data. The courses are split into two intervals t, where $t \in \{1,2\}$. The label withdrawal students is derived from students' trajectories records. It can be represented as a vector $Y_{(C)}$ where $Y_{(C)} \in \{0,1\}$. If students participated in previous and current courses at time *t-1* and t respectively, $Y_{(C)}$ denoting 0, and 1-otherwise.Withdrawal students are compared with non-Withdrawal students according to demographic features and students' grades. The students' records contain 6 demographic attributes that are driven from 2,175 records of Withdrawal students and 870 non-Withdrawal students.

The behavioural features have not been considered as there were insufficient behavioural records for at-risk students on the second time interval. A series of machine learning algorithms are applied over two sets of features. In the first set, we consider all demographic features, including students' grades (GPA), whereas in the second set of features, GPA is excluded from the analysis.

Class imbalance is an issue, which occurs in this subset. In this case, 74% of class instances occurred with the class "Withdrawal", whereas 26% of the data occurs with the class "not Withdrawal"; to solve this problem, Smote has been used.

4.7.3.6 Dropout Prediction Model Based on Student Motivational Status

The second temporal dropout predictive model is based on the second definition of at-risk students, and aims to examine the influence of motivational trajectories and engagement level on students' decisions to quit their courses. The motivation status is considered a high value factor that could impact at-risk students. As such, students with low motivation achievements in the current cohorts are more likely to withdraw from courses in the future cohorts.

To deliver timely intervention for at-risk students ,we consider only students who engage in the same Fall and spring courses. The students who lack motivation to persist in spring courses are classified as withdrawal students. The course is divided into two intervals t, where t $\in [1,2]$. Here student records can be described as $X_{i,j}{}^{(t)} = [X_{i,1}{}^{(t-1)}, X_{i,1}{}^{(t)}, \dots, X_{i,n}{}^{(t)}]$ where $X_{i,1}{}^{(t-1)} \in \{RL, Al, CLsc, CLsn\}$ is dimensional vector that represent student motivational statues in Fall courses. The $X_{i,1}{}^{(t)}$ is an activity undertook by student S_i at time t. The target class 'at-risk' student can be described as $Y_{(C)}$. The $Y_{(C)}$ takes value of 1 when student motivational status is reported as "amotivation" on the following courses, and 0 otherwise.

An analysis of motivational trajectories will provide new insight into the motivation behind at-risk students. As a result, the course instructors could immediately provide support for these students, by improving their motivation and increasing their learning. The correlation analysis is undertaken in this study with the aim to examine the relationship between the response variable (target) class and independent variables.

Various machine-learning algorithms are used to predict whether a student is at-risk or not. Machine Learning is capable of detecting changes in students' motivational status over time. The dataset contains 4,800 records for non-withdrawal students and 6,500 records for withdrawal students. The student's behavioural features of following courses, along with student motivational categories features at the pervious courses are used in the prediction of at-risk students.

4.7.3.7 Evaluating Temporal Prediction Models

For each predictive model, we split the original dataset into 50% for a cross-validation training set. A ten-fold cross-validation is considerd with five repetitions where training dataset is further divided into 10 different sets, 9 sets are used to train the classifier, and one is used as a test set. A further 50% of the data is used as an external test dataset to validate generalisation errors for each model. The training set can be described as $\{(X_{1,}Y_{1,}), \dots (X_{N,}Y_{C,})\}$ where $X_i \in \mathbb{R}^p$ *can* be represented by the i^{th} observation and $Y_{C,}$ is the target where $Y_i \in \{0,1\}$.

The empirical results of both temporal predictive models has been compared in terms of performance metrics comprising accuracy, specificity and sensitivity, precision,

recall, ROC, and AUC. A detailed explanation of performance metrics can found in the “predictive model evaluation parameters” section.

4.8 Experiment Three Introduction

Due to the student behavioural attributes represented as daily activities, the feature extraction procedures applied. Each student activity type aggregates into a single action according to the assessment cut-off date. Two features have been extracted over each activity, the number of sessions that a student engages in individual activity type and number of clickstreams that the student performs in each activity. As mention previously, the OULAD data set consists of several relation tables. We defined various feature sets such as static behavioral features, dynamic behavioral features, demographic features and assessment grades features. According to these features set we carried out two set of experiments.

In the first set of experiments, student performance in OULAD dataset is considered. For the prediction of students’ assessment scores, the regression analysis is implemented. The student past and current activities in addition to past performance are employed to predict student outcome. Tracing student performance over time will assist the educator to monitor the progress of the student in more detail.

The Final students’ performance prediction model is also proposed in this work. It It is computed based on the six TMA assessment, five CMA and final exam. The supervised machine learning method have been utilized to predict the long-term student performance. Three type of types of candidate predictors have been considered firstly behavioural features, followed by the temporal and demographic features. The Performance prediction offers new insight into determining the most important learning activity and assist the educators in keeping tracking of timely student performance. To best of our knowledge, the student performance has been evaluated in online course consider only two targets: success and fail. The long-term student performance predicts the performance with three-class labels success, fail and withdrew.

The influence of latent engagement on at-risk students is investigated in the second case study. The Gaussian Mixture Models is applied which aims to capture such important dynamics, providing an analytical assessment of the influence of latent engagement on students and their subsequent risk of leaving the course. Additionally, a set of machine learning models are used to provide a performance comparison. The

features used in the study were constructed from student behavioural records, capturing activity over time, which were subsequently organized into six time intervals, corresponding to assessment submission dates. Figure 4.6 describes the flowchart of the experiment.

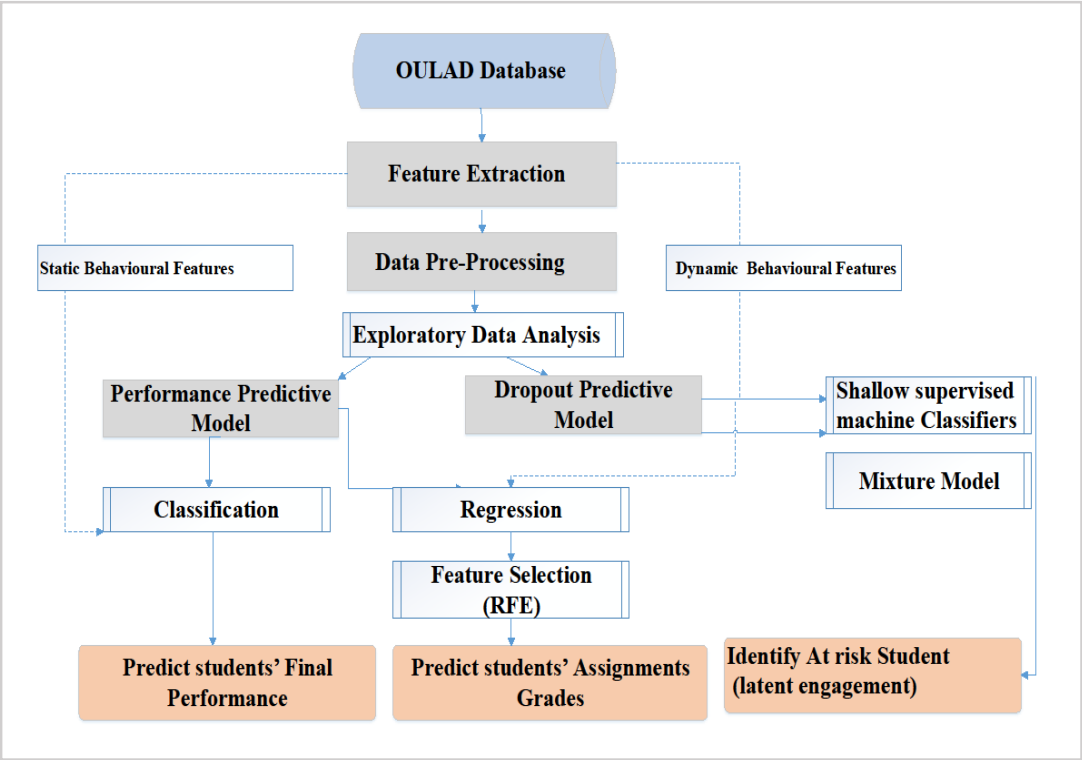


Figure 4.6 Experiment Three Flow Chart

4.8.1 Features Extraction

As the student VLE the data has been captured on daily basis, the feature extraction procedure has been undertaken. The VLE features were extracted according to the assessment submission dates. The course is split into seven time slices, where each time slice mapping is oriented around the final date of the TMA assessment submission. However, the first time slice captures student VLE information relating to learning activity prior to the course start, since students are permitted to enrol prior to the official course commencement. Our analysis of VLE features is undertaken to examine the association between student performances and the measured behavioural features, with respect to the assessment submission dates. For each student, there are a number of VLE learning activities at a specific time *t*.

The VLE activity types for each student are aggregated per time interval into single values. Hence, at each time interval, the students' VLE information records include 11 VLE activity types. Two features are extended, namely the number of sessions (o^t) and total number of clicks (c^t). The number of sessions is defined as the number of sessions wherein students engage in specific activity during the entire course. The number of clicks can be defined as the number of click-streams that students participate in per each activity during a single session. The procedure of feature extraction is described in Algorithm 4.2. Table 4.6 lists the set of extracted features.

Algorithm 4.2 Feature Extraction procedure

1. Split course into 7 interval t where t ∈ [1,7]
2. **For** each student S_i, i= 1…n, **do**
3. **For** each course activity type X_j,, i = 1…m, **do**
4. **For** each session $O_p$per activity type X_j, p = 1…w, **do**

$$O^t = \sum_{p=1}^{W} O_p$$

5. **For** each sum_click $C_e per$ activity type $X_{j,e}$ = 1…c, **do**

$$C^t = \sum_{e=1}^{c} C_e$$

6. **End**
7. **End**
8. **End**
9. **End**

Table 4.6 Set of Extracted Features

number of sessions (o^t)	number of clicks(c^t)
session.forumng	sum_click.forumng
session.glossary	sum_click.glossary
session.homepage	sum_click.homepage
session.oucollaborate	sum_click.oucollaborate
session.oucontent	sum_click.oucontent
session.ouellumi0te	sum_click.ouellumi0te
session.quiz	sum_click.quiz
session.resource	sum_click.resource
session.sharedsubpage	sum_click.sharedsubpage
session.subpage	sum_click.subpage
session.url	sum_click.url

4.8.2 OULAD Features

Behavioural features: The behaviour features have been extracted from the activity types. The OULAD dataset contains 11 VLE activity types. For each student at specific time t, two features are extracted: the number of sessions (o^t) and the number of clicks ((c^t). The behavioural features can be divided as either static or dynamic.

- **Static Behavioural features:** These are a set of behavioural features that are corresponding to student activities since the first time they engaged in the course till last day they quit the course. Let the tensor$X \in \mathbb{R}^{T \times n \times m}$, in which $X_{t,i,j}$ represents the j^{th} activity of the i^{th} student at time *t*. S is a set of students denoted as an n-dimensional vector [S_1 ….., Sn], where n is the number of students. Furthermore, M is defined as an m-dimensional vector that represents VLE learning activity types, M = [M_1 ….., M_m], where m is the number of learning activities that the i^{th} student is assigned.
- **Dynamic Behavioural Features:** These are a set of behavioural features that vary over time. Let t be a sequence of disjointed time intervals, where $t \in [1,6]$. To represent all student activities at time *t*, we define the type of student's activity records as the vector$X_{t,i,j} = [X_{t,i,1}, X_{t,i,2,} \dots, X_{t,i,m,}]$. Here the j^{th} denotes learning activity that is undertaken at time *t* by student S_i, such that $j = 1, \dots, m$; where m is given as the number of learning activities.

Demographic Features: These are given as$G \in \mathbb{R}^{n \times L}$, in which $G_{i,k}$ represents the k^{th} demographic feature for the i^{th} student, where the set of demographic features assigned to each student are considered constant over the course duration. The demographic features for the i^{th} student may therefore be given by the L dimensional vector $G_{i,k} = [G_{i,1}, G_{i,2}, \dots, G_{i,L}]$, where $k = 1 \dots L$. Table 4.7 describes the demographic attributes.

Temporal Features: The temporal features represented the date of student's registration and deregistration from the online course. Table 4.8 lists the temporal features.

Assessment Grades Features: These are a set of assessments submitted by student S_i at time *t*, we define the vector$A_{t,i,a} = [A_{t,i,1}, A_{t,i,2,} \dots, A_{t,i,a}]$. Here, $A_{t,i,a}$ denoted

the a^{th} assessments undertaken by student $S_{i\ at}$ time t. Additionally, $t \in [1,6]$ indexes course time intervals, where within each, students are allocated a single assessment.

Table 4.7 Demographic Features OULAD Dataset

Features	Description
id_student	Learner identification number
age_band	Learner age
Gender	Learner Sex
highest_education	Learner educational level
Region	Learner geographic area
studied_credits	The number of credits for the module that the learner is currently involve
disability	Indicator of student disability
num_of_prev_attempts	Number of time that student undertook the course
imd_band	Social-economic indicator measure student economic level

Table 4.8 Temporal Features OULAD Dataset

Feature	Description
id_student	Student identification number
date_registration	The date of learners registration in the course
date_unregistration	The date of learners quit the course

4.8.3 Data Pre-Processing

Data Pre-Processing is applied over the extracted behavioural features and demographic variables, with the aim to achieve the best performance from machine learning. As previously discussed, the analysis of the Open University database relies on the assessment submission dates, wherein we split the course into six-time intervals. At each interval, there are 11 learning activities. Two features have been extracted over each learning activity: number of sessions and number of clicks.

The first step in pre-processing the data is to investigate the highly correlated variables. We set a correlation cut off - if the correlation between two features >0.8, then these features are highly correlated. The highly correlated features are removed from the final model, given that the problem of feature redundancy could be solved. In addition, the issue of over-fitting might therefore be reduced. The zero and near-zero variance predictors are also investigated in this database; the features that same values that appear frequencies become zero variance predictors when the data is split into training and test. These features, which have a “near-zero-variance” are diagnosed

and eliminated during the pre-processing procedure. The Table 4.9 lists the near-zero predictors.

The Open University dataset is non-normally distributed; in order to address this problem, the transformation methods are applied. Yeo-Johnson is one of the data transformations methods, and performs a similar function to the boxCox transformation method, but applies a continuous variable that has a raw value equal to zero (Weisberg, 2001). In the present case, when a student did not participate in a particular activity, the value of the extracted features become zero. To this end Yeo-Johnson is more useful than boxCox. The figure 4.7 compares the results of both transformation methods over six time intervals. The results of all behavioural features are transformed by Yeo-Johnson; only two of features are processed by boxCox.

Table 4.9 Near Zero-Variance Predictors OULAD Dataset

Zero- and Near Zero-Variance Predictors
“ session.ouellumi0te”
“ sum_click.ouellumi0te”
“session.oucollaborate”
“sum_click.oucollaborate
“ sum_click.oucontent”
“ session.sharedsubpage”
“ sum_click.glossary”
“ session. glossary”

Pre-processing BoxCox
- Box-Cox transformation (2)
- centered (86)
- ignored (0)
- scaled (86)

Pre-processing Yeo-Johnson
- Yeo-Johnson transformation (85)
- centered (86)
- ignored (0)
- scaled (86)

Figure 4.7 Data Pre-Processing transformation OULAD dataset

4.8.4 Exploratory Data Analysis

The EDA is applied in this case study, with the aim to obtain insight into behavioural features in association with students’ performance. EDA helps educators gain

intuitions into the data and guides their decisions concerning teaching strategies. Such graphical tools could help educators fulfil the requirements of students. In the present case study, EDA is applied to the OULAD dataset as a precursor to the modelling phase. The objective of data visualisation is to provide insight into the correlation between extracted features and student performance.

The correlation matrix is used to evaluate the dependency between the behavioural, demographic variables and learners' outcome. A heat map is utilised to visualise two correlation matrices. The PCA is used to reduce dimensionality by eliminating the correlated variables. In this case, study, the PCA is applied only on behavioural features in order to remove any redundancy in the extracted features.

4.8.5 Case study One

4.8.5.1 Student Performance Prediction Model

The first case study in the OULAD dataset focuses on performance predictions. The problems are formulated as classification and regression problems. The regression setting is considered when we aim to predict students' assessments grades, whereas classification setting is utilised when we seek to predict final student performance in the entire course. It is considered a multi-class problem where the target class is whether students pass, fail or withdraw from courses.

Early grade prediction could help educators deliver timely intervention support and additional learning materials to help students who have low scores. As discussed previously, the student should participate in five CMA assessments and six TMA assessments, in addition to the final exam. The assessments should be handed in within a specific time period. Due to the TMA assessment weighing 45% of the final result, while the CMA assessment weighs only 5%, our temporal analysis is based on the submission date of the TMA. Students are allowed to submit after the deadline, but they might lose some marks. Furthermore, the student can access learning prior to the course, but are not allowed to engage with the course after it closes. Table 4.10 shows the TMA assessments submission date.

To predict student performance in a timely manner, as can be seen, in figure 4.8 the course is subsequently organised into six-time intervals, corresponding with assessment submission dates. The student behavioural records are distributed according to assessment date. With regards, to the regression analysis, the students'

performance during the early stages in conjunction with their interaction behaviour should be considered when predicting student assessment grades. For a specific student S_i across time interval *t*, the student record R_i , can be obtained from the student's learning activity and performance, which are described as an input sequence $(X_j^{(t-1)}, \ldots, X_n^{(t-1)}, A_a^{(t-1)}, X_j^{(t)}, \ldots, X_n^{(t)})$. Here X_j represents the j^{th} behavioural activity attempts by Si at time *t* and *t-1*. *The* A_a denotes the a^{th} assessments undertaken by Si at time t-1. The corresponding target can be represented as a sequence of output $(A_1^{t-1} \ldots A_a^{t})$.

In terms of classification analysis we aggregate the student's behavioural activities across the six time slices into a single time slice.Three sets of features are considered in this anlysis the behavoural features ,deomogrphic features and temporal features. We didn't account for past assessments grade and final exam mark as the final target class is computed based on these features. The dataset contains 4004 records where the proportion of "fail", "withdrawn" and "pass" classes are 28% ,40% and 32% respectively. Different linear and nonlinear regression algorithms and classifications have been used in this study.

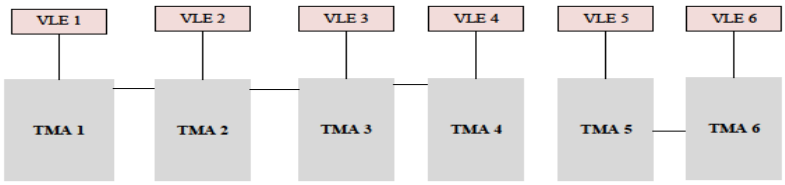


Figure 4.8 Student behavioral features based on TMA intervals

Table 4.10 TMA Assessments Submission Date

Module name	Weight	Day
2013B	5	19
2013B	18	47
2013B	18	89
2013B	18	124
2013B	18	159
2013B	18	187
2013J	5	19
2013J	18	47
2013J	18	96
2013J	18	131
2013J	18	166
2013J	18	208

4.8.5.2 Feature Selection

To detect the most important activity types that reflect on students' performance, feature selection is utilised. As previously discussed, the database contains many features, for instance, at each time interval, past student behavioural features were aggregated with current behavioural features, although some of these features could be considered irrelevant. The RFE is utilised to examine the high-ranking features that influence the learners' performance in particular assessments. Furthermore, the RFE is also used to discover behavioural features that affect students' final performance during the entire course.

With the RFE approach, irrelevant and redundant features are eliminated, and as a consequence of this, the predictive models will perform faster and more efficiently, in addition to reducing the over-fitting of the data and improving the generalisation of the learning algorithm.

4.8.5.3 Evaluating Student Performance Prediction Model

In order to evaluate the student performance predictive model, several metrics have been considered. In terms of the regression analysis, the RSME and R^2 are used to predict students' assessment grades. With regards, to the classification analysis accuracy, specificity and sensitivity, F-Measure, ROC, and AUC are employed to predict final student performance. Furthermore, a tenfold cross-validation is used for both a regression and classification analysis; 50% of the dataset is selected to train the model and 50% is selected to test the model.

4.8.6 Second Case Study

4.8.6.1 Temporal Model for Identify At-Risk Student in OULAD

In OULAD dataset, the timely behavioral intervention of learners can be traced by analyzing historical data. In this case study, LA is utilized to detect the students who participated in the current and past assessments using historical behavioral data. Based on the feature extraction procedure, the student records R are distributed across ssix-timeinterval according to assessment date tradeoff. We consider one definition of at-risk students in the OULAD dataset as explain in the following paragraph.

- **At- Risk Student Definition**: The students who have undertaken a sequence of assessments in a single course. The at-risk student is derived from assessments scores features which can be represented as $Y^t(i)$vector where $Y^t(i) \in \{0,1\}$, if

student Si undertook assessment$A_a{}^t$ at time t, then student is defined as non-withdrawal and denoting 0 , and 1 otherwise.

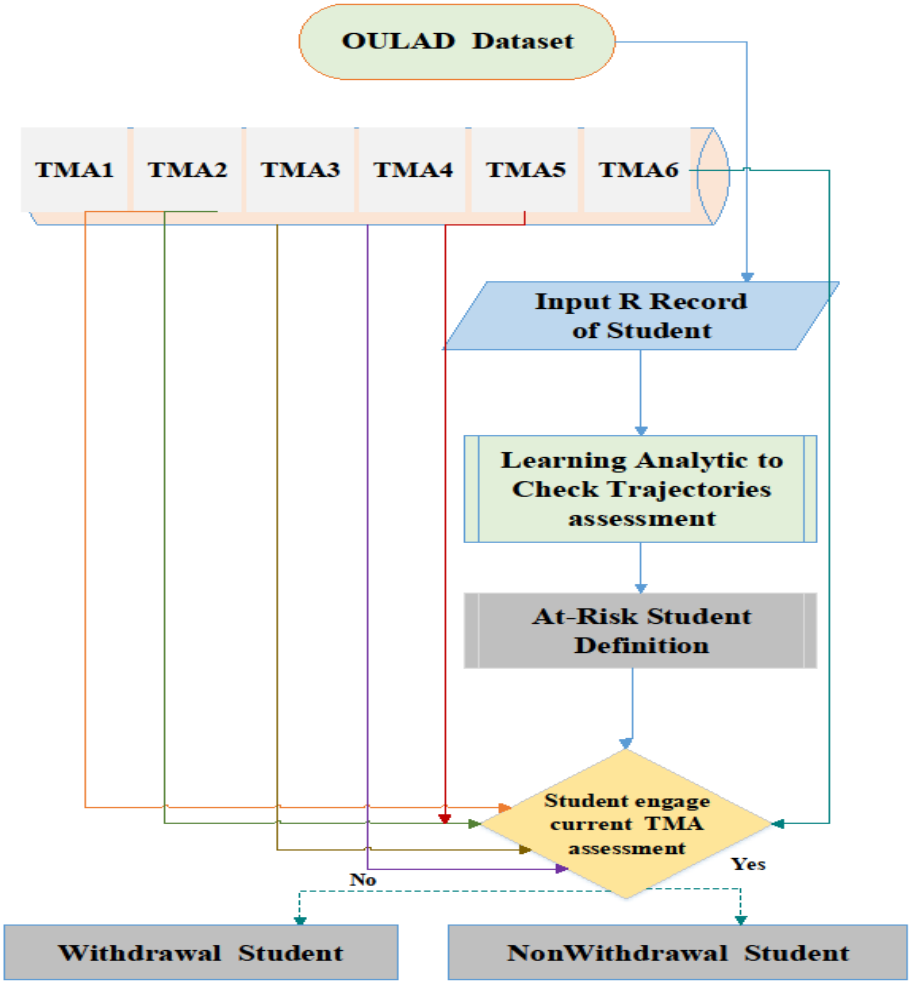


Figure 4.9 AT -RISK Student Framework in OULAD Dataset

4.8.6.2 Dropout Prediction Model Based on Latent Engagement

The temporal predictive model is based on the previous definition of at-risk students that aims to investigate how a student's engagement affects the at-risk student. Student engagement is a significant factor that influences the learning outcome. Behavioral records such as watching a video, undertaking assessments, accessing the home page, can dictate the student engagement and reading PDF documents .Categorizing the latent engagement pattern of learners concerning the impact on their continuation within course activities is crucially important to predict at-risk students.

The main challenge of a mixture model is how determined the number of the components. It has been attempted to increase the number of the component to be more than 10 but, the computational burden was the crucial issue as the higher number of clusters more time would be required to learn a model. In this project (1, 2, 3, 4, 5,10) components has been tested for each class. It has been found that selecting one cluster for each class would acquire the best accuracy result.

Eigenvalue Decomposition Discriminant Analysis (EDDA) model is utilised for the prediction of at-risk students within course environments, and considers two categories, "at-risk", or "not at-risk". The supervised classification algorithms did not take into consideration the impact of unlabelled data on one class (Bouveyron, 2014). Mixture model is capable autonomously of discovering unobserved latent engagement and assigns these unlabelled data to one of the classes. The mixture model is a powerful inference framework that can approximately represent high dimensional data as a linear combination of multiple Gaussian components (Bouveyron, 2014)(Moe and Fader, 2004).

To include all information about the learners past behaviour at each interval, we combine the student's behavioural features at the current time *t* with the student's learning behavioural attributes at the previous time *t-1*. A total of 30 behavioural features are considered across each time interval , denoted as the n-dimensional vector $X^t(i)$, producing a sample indexed over each student S_i , per each completed time interval *t*, where t ∈ [1,6]. Subsequently, we split data into 60% for use with model training and 40% for test evaluation. We consider the training dataset as the complete form of the variables, consisting of set observations, denoted$X^t(i)$, and a set of latent variables. The latent variables can be represented by$L^t(i)$, whose unknown labels can be described as $Z^t(i) = [Z_{t,i,1}, Z_{t,i,2,} \dots, Z_{t,i,m,}]$, such that $Z^t(i) \in \{0,1\}$. The class label for the i^{th} observation,$X^t(i)$, is given as $Y^t(i) \in \{0,1\}$. For example, if the i^{th} student submits the a^{th} assessment at the current time interval *t*, and his previous latent status at time *t*-1 was active, then $Y^t(i)$ = 0, else 1. The Algorithm 4.3 describes the learning procedure of EDDA model per each interval. To evaluate whether the latent engagement influences negatively or positively on the students who are likely to drop out from the course, the mixtures models are compared to a set of the supervised classifier.

Algorithm 4.3 Learning Procedure by Mixture Model

1:Given the incomplete training dataset $\{(X_{1,}Y_{1,}), \ldots (X_{N,}Y_{C,})\}$
2: **For** each X_i,i=1 … N **do**
3: Initialize model parameters θ
4: Initialize the hidden engagement status Z
5: Repeat
6: **Procedure E-Step**
7: Use the estimate parameters: $\widehat{\mathcal{T}_g}$, $\widehat{\mu_g}, \widehat{\sum_g}$
8: Compute initial expected value of latent engagement $\widehat{Z_{ig}}$ by using Eqn. 4.27
9: **End Procedure**
10: **Procedure M-Step**
11: Update expected value of latent engagement Z_{ig} via Eqn. 4.28 and Eqn. 4.29
12: **End Procedure**
13: Until Converged
End do

$$\widehat{Z_{ig}} = \frac{\widehat{\mathcal{T}_g}\theta(L_i|\widehat{\mu_g}, \widehat{\sum_g})}{\sum_{g=1}^{G}\widehat{\mathcal{T}_g}\theta(L_i|\widehat{\mu_g}, \widehat{\sum_g})} \tag{4.27}$$

$$L(\theta_{g,}\mathcal{T}_g, \sum_{g,}|L_{ig}) = \sum_{i=1}^{n} log\left[\sum_{g=1}^{G}\mathcal{T}_{g\theta\,(L_i|\mu_g,\sum_g)}\right] \tag{4.28}$$

$$Z_{ig} \begin{cases} 1\ if\ L_i \text{ belong to group g} \\ 0\ otherwise \end{cases} \tag{4.29}$$

To gain maximum likelihood for initial estimates parameters ($\widehat{\mathcal{T}_g}$, $\widehat{\mu_g}, \widehat{\sum_g}$), the function L($\theta_{g,}\mathcal{T}_g, \sum_{g,}|L_{ig}$) is applied. These estimates are calculated by considering latent variable L_i.The expected value of latent variable Z_{ig} are updated until latent variables L_i assign to component g that matching the highest probability.

4.8.6.3 Evaluating Student Dropout Prediction Model

To evaluate the performance of the student dropout predictive model, both ROC analysis and confusion matrix values were computed, forming the basis for comparing model responses to ground truth labels, over each model. Various performance summary metrics are considered, namely accuracy, sensitivity, specificity, the F1 measure, and the AUC.

4.9 Chapter Summary

This chapter has presented a detail explanation of the methodology used in this thesis. The detailed analysis of two MOOCs datasets have been discussed in this chapter. The Harvard and OULAD datasets. Harvard dataset contains the behavioural student activities for multiple courses while OULAD includes the behavioural student activities of a series of assessments within the single course.

Three definitions of at-risk students have been introduced in this research project. The student motivational categories have been evaluated only over Harvard dataset, as it is not possible to evaluate the student motivational statues in the OULAD dataset since all students were categorized as extrinsically motivated. Moreover, students' latent engagement cannot be inferred in the Harvard dataset because the latent variables depend on estimation of student activities in the prior time steps within a single course and this does not exist in such a dataset.

With these three definitions, factors that impact at-risk students from such as, students' performance, students' motivation status and students' engagement can be investigated. Statistical techniques have been applied to examine the relation between learner engagement level and performance in the context of the learner educational background and geographical location. As such, descriptive statistics, analysis of covariance and the chi-squared test are used to distinguish between failing and successful learners.

We proposed performance and dropout prediction models over each dataset. For each predictive model description of data pre-processing, exploratory data analysis, features selection, oversample and predictive models evaluations have been provided. The mixture Model has been used to predict at-risk student. The mixture model is a useful probabilistic model that can be applied to infer the student's latent engagement state over time. Mixture model provides advantages over traditional machine learning with the capability of automatically identify unlabelled data. The results of experiments are displayed in the following chapter.

Chapter 5: Results and Discussion

5.1 Introduction

This chapter produces the results of three experiments. The same set of supervised machine learning algorithms have been used. These are including Random Forest(RF), Decision Tree(Rpart), Feedforward Neural Network with single hidden layer(Nnet/NN), Multiple Layer Perceptron(Mlp), with two hidden layers, Gradient Boosting Machine (Gbm), Logistic regression(Glm/LR).

In the first experiment, a set of supervised machine learning algorithms is used to predict student performance. Moreover, the unsupervised machine learning is employed to explore whether students share similar characteristics. In the second experiment, the association between students' performance and engagement levels will be explored. The impact of temporal students' performance and motivational status on at-risk students have been also investigated.

Various statistical techniques and hypothesis tests are utilised including descriptive statistics, analysis of covariance and Chi-squared. As such, the successful students and failure students are compared according to descriptive techniques. Machine learning is utilised to identify the at-risk student in the early stage. The results of two dropout prediction models have been provided in this chapter.

In the third experiment, the regression and classification are considered to predict student performance. The dynamic behaviour features are employed to predict students' assessment scores. The static behavioural features are used to predict final t student performance. The purposes of these analyses are to investigate the influence of students' activities on students' performance. In addition, it could determine which interval contains the lowest number of students.

Furthermore, the mixture models and six sets of supervised machine learning used to identify the students who are at-risk to drop out from the courses.

5.2 Experiment One Results

In this section, we present the details results for the first experiment, which includes the results of EDA features selection, and machine learning.

5.2.1 Exploratory Data Analysis Results

A heat map is applied to visualize the correlation analysis. Figure 5-1 presents the

heat map of the Harvard dataset. The cell is colored based on the degree of correlation between the variables. The map shows that the attributes (ndays_act ,nchapters, Nevents) tend to be positively correlated with a target (certified attribute), showing coefficient values of 0.72,0.71 and 0.68 respectively. The remaining behavioral features display a weak positive correlation, such as (noforum_post) attributes, achieving a value of 0.09. It is notable that the demographic features are not highly correlated with student performance. The results of the PCA are shown in Figure 5.2. As can be seen, Harvard dataset exhibits high variance. The number of principal components was reported as 7 in this dataset. Figure 5.3 visualizes the result of the Kaiser method. The figure shows only first component Comp1 is chosen as an optimal component.

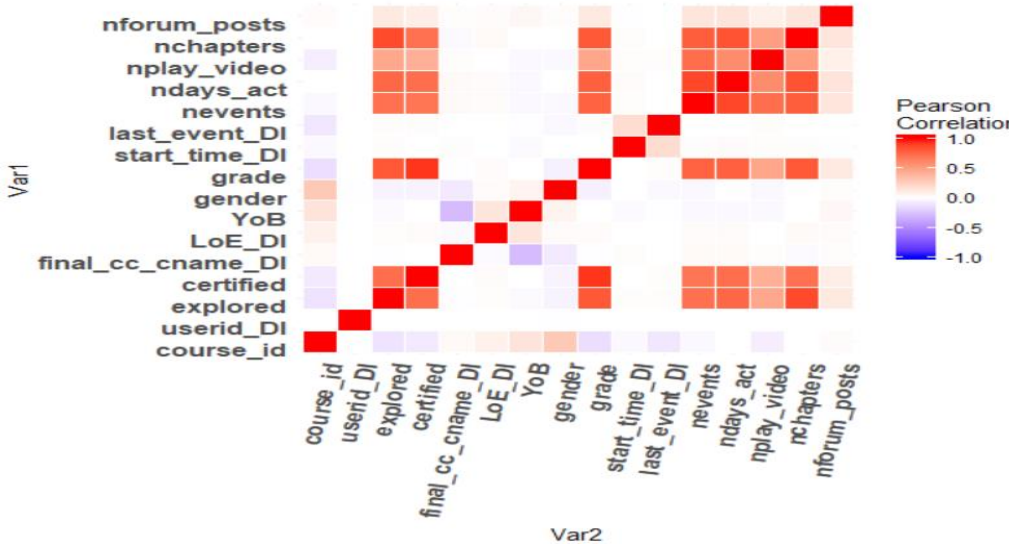


Figure 5.1 Heat Map For Harvard dataset

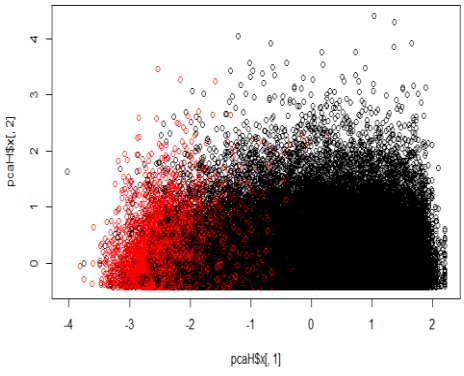


Figure 5.2 PCA for Harvard dataset

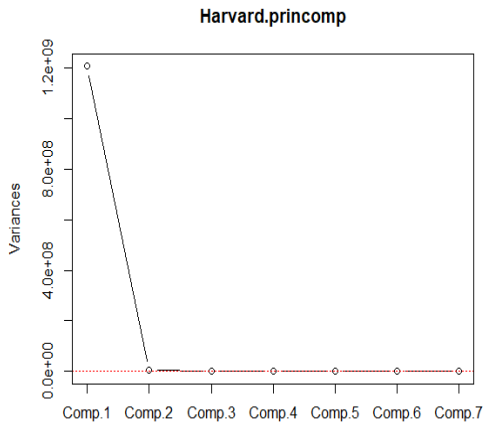


Figure 5.3 selected component Kaiser Method

5.2.2 Feature Selection Result

The results obtained by both RFE and Hill climbing algorithms show that both indicate the same subset of features. Figure 5.4 illustrates the result of RFE based on accuracy criteria. The features with higher accuracy correspond to the most important features. The top five features are "nchapters","nplay_video","ndays_act", "nevents" and “explored".

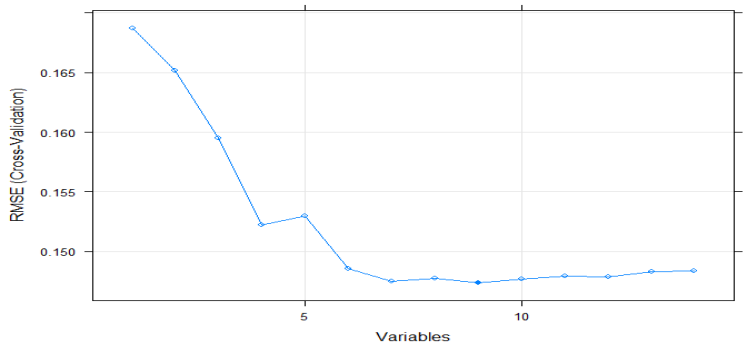


Figure 5.4 RFE Feature Ranking for Harvard Dataset

5.2.3 Student Performance Prediction Model Result

Supervised machine learning has been applied to two subsets of features, namely all dataset features and high weighted features, evaluated using the RFE. During the model training stage, cross-validation was used to evaluate the fit of classifiers. Figure 5.5 and Figure 5.6 show the classifier accuracy of the training set for all models over both subsets. The graphs show that both set of features have nearly the same accuracy. The RF acquired the best performance over the trained dataset for both sets of features.

The results are listed in Tables 5.1 and 5.2 respectively. As can be seen, simulation results show all features yield a slightly higher performance than the selected features. The Gbm achieves the highest accuracy, with a value of 0.967 in the first set of features, while both Mlp and RF gain the best performance, with a value of 0.963 in the second set of features. Glm, RF, and Mlp give albeit compelling results, with an accuracy of 0.963, 0.962, and 0.958 respectively in the first set of features. Conversely, Nnet and Gbm achieved similar accuracy, with values of 0.96 in the second set of features. In both sets of features, Rpart shows a lower performance than the other classifiers, acquiring values of 0.92.4 and 0.908 respectively.

Due to the Harvard dataset being imbalanced, an F1–Measure could be an effective metric to evaluate the performance of the classifiers. The F1–Measure accounts for both precision and recall (Bekkar, Djemaa and Alitouche, 2013). The F1–Measure metric produced a relatively similar result to accuracy. Over both sets of features, Glm, RF, Mlp, achieves the best F1–measure whereas Rpart obtained the lowest F1–measure value.

Specificity (true negative) results over all the classifiers are seen to be slightly higher in the second set of features than the first set of features. However, the RF model obtains similar specificities values over two sets of features, with a value of 0.974. The Nnet gains weakness specificities, with 83% over the first set of features. In contrast, the Nnet model obtains the highest sensitivity with a value of 0.99 over the same set of features.

In terms of true Positive, all classifier models over both sets of features obtained viable sensitivity values 0.99 to 0.91. However, Rpart obtains the lowest range of sensitivity, with a true positive percentage of 87% over second set of features.

The ROC and AUC were considered. Figures 5.7 and 5.8 show ROC for both experiments. The curves are shown to converge to roughly the same semblance of the plot, indicating similarities of performance across models.

In order to evaluate the feasibility of the classifier models for both sets of features, computational performance was considered. Figure 5.9 shows the speed run time measured in seconds for each learning algorithm. In general, the time required to train all features is longer than the selected features for all classifier models. The fastest algorithm speeds were Glm, and Rpart, which achieved 60 and 120 seconds respectively, for the first features while it takes 40 seconds for the second set of features.

The Nnet algorithm is not particularly affected by size of features. As can be seen, the average run time in Nnet achieves a value of 720 seconds when selecting all features, which slightly declines when selecting the highest-ranking features. Conversely, the Mlp model requires greater time to train in entire set of features than selected features. There is a gap in the training time between both sets of features for RF classifiers, in which approximately a third of the time declined when training high-ranking features. The RF is the slowest learning algorithm compared with the other

algorithms. A number of reasons could affect the speed of the RF. The most significant one is that it is a learning algorithm based on the bootstrap method (Koch, Konen and Hein, 2010).

Table 5.1 Classification Performances for First Set of Features

Classifier	Acc.	F-Meas.	Sens.	Spec.	AUC
Mlp	0.958	0.968	0.954	0.966	0.932
RF	0.962	**0.971**	0.955	**0.974**	0.994
Rpart	0.92.4	0.941	0.912	0.947	0.965
Glm	0.963	**0.971**	0.959	0.971	**0.995**
Gbm	**0.967**	0.974	0.965	0.971	**0.995**
Nnet	0.94	0.956	**0.992**	0.836	0.961

Table 5.2 Classification Performances for Second Set of Features

Classifier	Acc.	F-Meas.	Sens.	Spec.	AUC
MlP	**0.963**	**0.972**	0.952	**0.985**	**0.995**
RF	**0.963**	**0.972**	0.957	0.974	**0.995**
Rpart	0.908	0.926	0.87	0.981	0.980
Glm	0.959	0.969	**0.96**	0.958	0.994
Gbm	0.96	0.969	0.95	0.979	0.994
Nnet	0.96	0.969	0.952	0.974	0.994

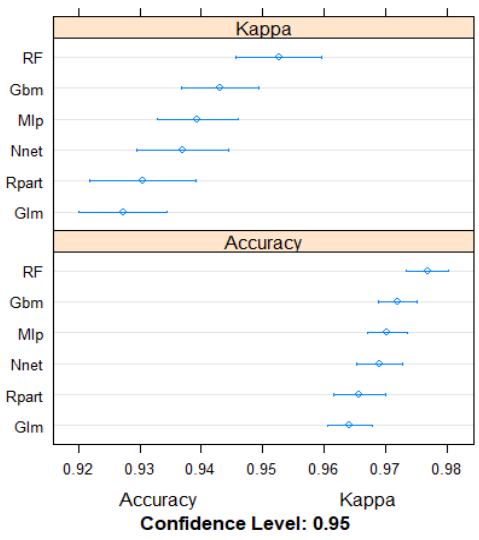


Figure 5.5 Estimation Accuracy Classifier First Set of Features

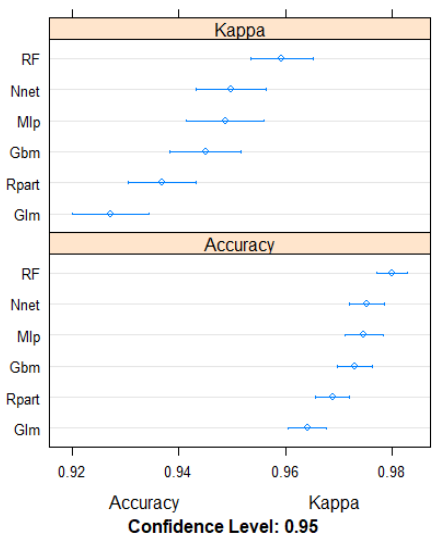


Figure 5.6 Estimation Accuracy Classifier Second Set of Features

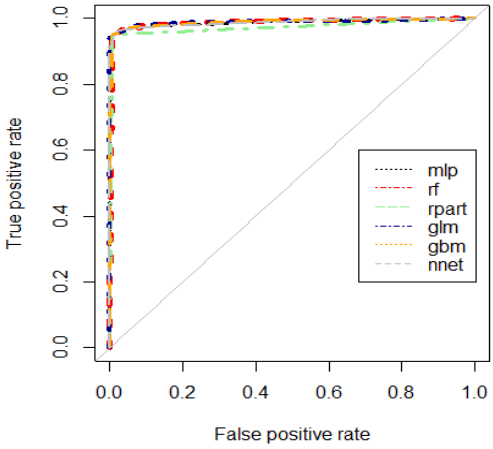


Figure 5.7 Roc Curve First set of Features

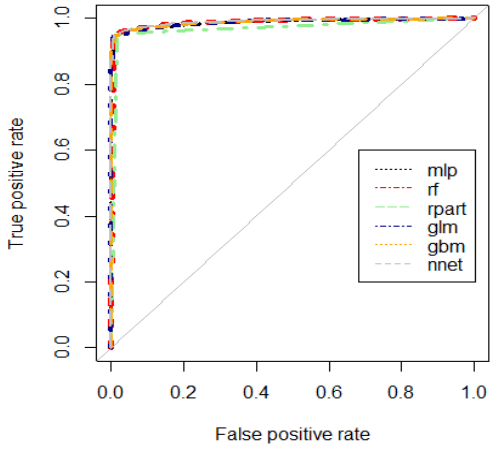


Figure 5.8 Roc Curve Second set of Features

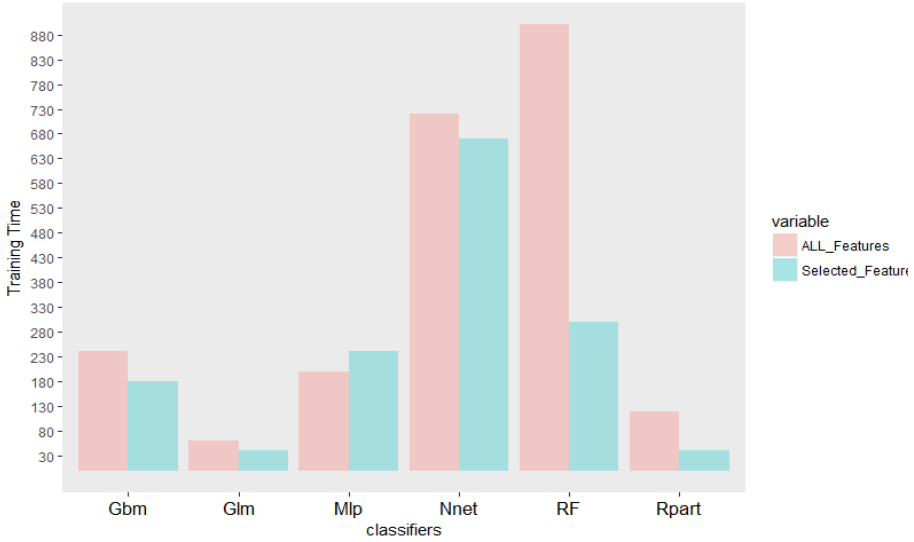


Figure 5.9 Comparing Computational Training Time

5.2.4 Unsupervised Machine Learning Results

The results for the fuzzy cluster are provided as follows. The Table 5.3 describes the descriptive statistics of the FCM algorithm. As can be seen, the values range of DuNnet's Fuzziness Coefficients, was reported at 0.25. According to the reference (Noordam *et al.*, 2000) ,if the value of DuNnet's Fuzziness Coefficients is close by 0, this indicates that the clusters are very fuzzy clusters.

The results also reveal that an association membership of observations across all clusters is equalised. Table 5.4 lists the number of learners across each cluster, based on engagement type. The table shows that only one student fits with cluster 1, while cluster 2 and cluster 3 contain the homogeneous data of students who were different in their prototypical engagement. The figure 5.10 shows that the cluster's boundaries are determined by the learning process overlap with a two-dimensional representation of the feature space. The FCM could provide an explanation in the patterns of students' behaviour. As such, the results demonstrate that learners share similar characteristics, although they do differ in their engagement categories.

Table 5.3 Fuzzy C-Means Cluster Result for Harvard dataset

DuNnet's Fuzziness Coefficients	0.25
Root Mean Squared Deviations (RMSD)	3.113661
Mean Absolute Deviation (MAD)	86.83641

Table 5.4 Learners' Distribution per Cluster Based on Engagement Type

Cluster	Active No Cert	Active Cert	Passive-Active Cert	Passive-Active No Cert
1	0	0	1	0
2	37	15	2394	292
3	4639	17	292	418

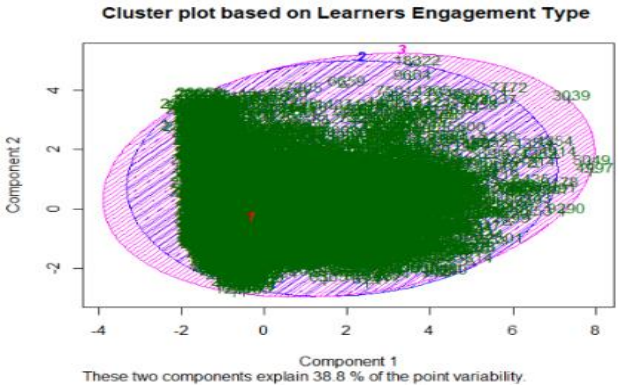


Figure 5.10 Cluster Plot

5.3 Experiment Two Result

The experiment two results are displayed in the following section. The first case study describes the statistical method result. The second study involves machine learning.

5.3.1 Engagement Level of Failing and Successful Learners Results

Descriptive statistics are computed and stratified according to the demographic region. The engagement levels of learning activities are determined. A comparison between failure groups with success groups was conducted, whilst accounting for geographical location ("Africa", "Asia"," Australia"," America"," Europe"), in conjunction with behavioural features (nplay_video, Nchapters).

The findings' results in Tables 5.5 and 5.6 indicate that there is a significant difference between the two groups for each course. The results also demonstrate that successful learners watch more videos than failure students. Europe dominated the top ranking of successful learners with $\mu($ 1376.21,1332.6, 1010.67, 734.74,560.85) for "Health Fall", "Electronics Fall", "Electronics Spring"," Computer Science Fall" ," Computer Science Spring" courses, respectively, during 2012 and 2013. However, the highest number of successful learners in "Health Spring" were living in America with $\mu(338.39)$.

As can be seen in appendix 2 .The result also reports that "Health Fall" is considere d the most watched course, with approximately 355,583 videos viewed by certified st udents. Conversely, "Health Spring" was the lowest viewed course, in which successf ul learners viewed only 41,710 videos. Within the successful group of learners, Europ ean students watched an average 30% of videos in" Health Fall", whereas the Americ an learners viewed around 52% of the videos in the "Health Spring" course.

The African and Australian learners viewed the lowest percentage of videos. In the "Computer Science Spring" and "Electronics Fall" course, European learners undertook once again the highest percentage of video usage, with approximately 50-70%, of the video resources used by this group of students. Please refer to Appendix 6.

For the the failure group of students, the largest proportion of videos is watched by the American participants in all the courses, who used 30-40% of the video resources. However, Asian students in "Electronics Spring" and "Electronics Fall" achieved the

highest proportion, and used around 35% of the videos. In all courses, the lowest rate of video usage was reported again for the Australian participants, except in "Health Spring" .African students showed the lowest percentage. All of these results can be found in the appendix 1 and 6.

The result shows that there is significant variability between the successful participants and the failed learners in respect to the number of chapters read. In general, successful learners read learning materials three times more than non-successful learners. For example in America the mean number of chapters read is reported as $\mu\,(16.30; 16.94;\ 14.82, 17.94;\ 16.54, 8.39)$ in "Electronics Fall", "Computer Science Fall", " Health Fall", "Electronics Spring" , "Computer Science Spring" and "Health Spring" courses for each successful group, in contrast to a reduction by one third, as seen in the failure peers $\mu(6.45, 4.68, 6.35, 6.42, 4.81, 3.69)$.

With respect to the number of chapters read, the most successful students are reported again as European, with an average of 40%- 46% of chapters viewed by the group across all courses. However, in the "Electronics Spring" the American students acquired the highest μ with a value of 17.94.

With regards, to non-successful students, students who engaged in "Health Spring" and "Electronics Spring" respectively reported the lowest rate viewed. Participants within this group read only a small proportion of the available materials. Moreover, the proportion of failure students who engaged in reading chapters rose to 90% in the "Computer Science" course, for which the percentage of reading material was greater than in the other courses. There was a similar average of chapters read by students who enrolled in "Health Fall" and "Electronics Fall" across all courses. The Australian students used the least reading material. Again, American participants in "Computer Science" courses and "Health Fall" achieved the highest rate, using around 40%-28 percentage of the Chapters.

Figures 5.11 and 5.12 show the box plot, in general, the engagement level of the successful group is higher than the failure group, over "nplay_video" and "Nchapters". Additionally, the boxplot shows successful learners in fall courses read more learning resources than those in the spring courses. Boxplot also shows that most successful learners in the "Health Fall" and "Electronics Fall" courses viewed an average of 1,000-1,200 videos and read approximately 15 chapters. The percentage of resource

usage drops slightly in “Health Spring” and “Electronics Spring” courses to 1000-400 videos. As can be seen, the number of videos viewed by the most successful group is slightly higher in the “Computer Science Fall” course rather than “Computer Science Spring” course. However, the percentage of reading documents is similar across both courses. In this study, ANCOVA is used to determine whether the mean of success and failure learners are identical, with regards, to their engagement level. The results reveal a notable difference between the two groups across all courses. The p-value was ($p<0.0002$) for all behavioural features. Hence, there is a significant difference between certified versus failure students. Table 5.7 lists the results of ANCOVA.

Table 5.5 Descriptive statistics of Analysis Failure Learners

Courses		Mean					SD				
"2012 Courses"	Africa	Asia	Australia	America	Europe		Africa	Asia	Australia	America	Europe
Electronics Fall											
nplay_video	876	275.6	155.12	200.81	**485.3**		3693	425.5	194.56	280.33	567.1
Nchapters	7.65	7.76	7.875	6.45	**8.30**		3.95	4.06	5.19	3.13	4.34
Computer Science Fall											
nplay_video	**371.5**	213.58	162.11	76.167	202.78		447.6	258.3	287.63	282.96	267.5
Nchapters	4 .97	5.22	4.67	4.68	**5.33**		3.57	3.58	3.48	3.20	3.58
Health Fall											
nplay_video	**531.7**	333.2	184.78	311.48	311.48		859.6	745.9	238.79	611.65	611.6
Nchapters	**8.68**	8.125	8.35	6.35	7.84		4.95	4.75	5.38	4.52	4.81
"2013 Courses" Electronics Spring											
nplay_video	231.17	144.25	**247.21**	209.72	220.53		575.10	256.97	202.88	312.27	315.21
Nchapters	6.36	5.90	**6.81**	6.42	5.92		4.91	3.60	4.47	3.86	3.41
Computer science spring											
nplay_video	123.1	134.20	105.14	130.83	**140.05**		174.90	266.48	173.74	203.19	217.6
Nchapters	5.08	5.21	4.64	4.81	**5.27**		3.40	3.53	3.09	3.27	3.46
Health Spring											
nplay_video	**74.54**	74.34	92.14	61.12	62.56		108.36	152.74	118.28	73.68	84.36
Nchapters	**4.59**	3.946	2.85	3.69	3.918		2.50	2.13	0.690	1.989	2.175

Table 5.6 Descriptive Statistics of Analysis Success Learners

Course		Mean					SD				
"2012 Course"	Africa	Asia	Australia	America	Europe		Africa	Asia	Australia	America	Europe
Electronics Fall											
nplay_video	383.57	576.08	62.66	1003.0	**1332.6**		399.2	713.88	84.29	1056.6	1496.1
Nchapters	16.34	15.16	14.3	16.30	**16.42**		. 09.1	2.12	3.62	1.63	1.64
Computer Scienc Fall											
nplay_video	538.6	499.78	197.8	634.12	**734.74**		579.90	759.37	189.9	509.38	753.15
Nchapters	16.41	16.11	16.36	16.94	**17.11**		2.34	2.62	2.54	1.69	1.61
Health Fall											
nplay_video	981.7	717.04	1357.8	1035.5	**1376.2**		2007.5	782.4	1725.0	1075.6	1203.2
Nchapters	14.27	14.27	15	14.82	**15.07**		1.425	1.44	0.816	1.28	1.24
"2013 Courses" Electronic Spring											
nplay_video	616.55	333.50	212.66	801.70	**1010.67**		704.34	505.91	328.35	609.45	1258.61
Nchapters	17.61	16.01	16.43	**17.94**	17.35		2.25	2.80	2	2.22	2.33
Computer Science Spring											
nplay_video	287.14	36.06	342.6	472.37	**560.85**		258	1.544	224.29	410.11	567.89
Nchapters	16.6	16.63	16.73	16.54	**16.83**		1.40	1.46	1.334	1.56	1.51
Health Spring											
nplay_video	171.0	171.02	157.16	**338.39**	242.6		382.93	380.9	71.51	606.72	208.49
Nchapters	7.78	7.620	8.333	8.39	**8.44**		1.999	1.862	1.63	1.32	1.64

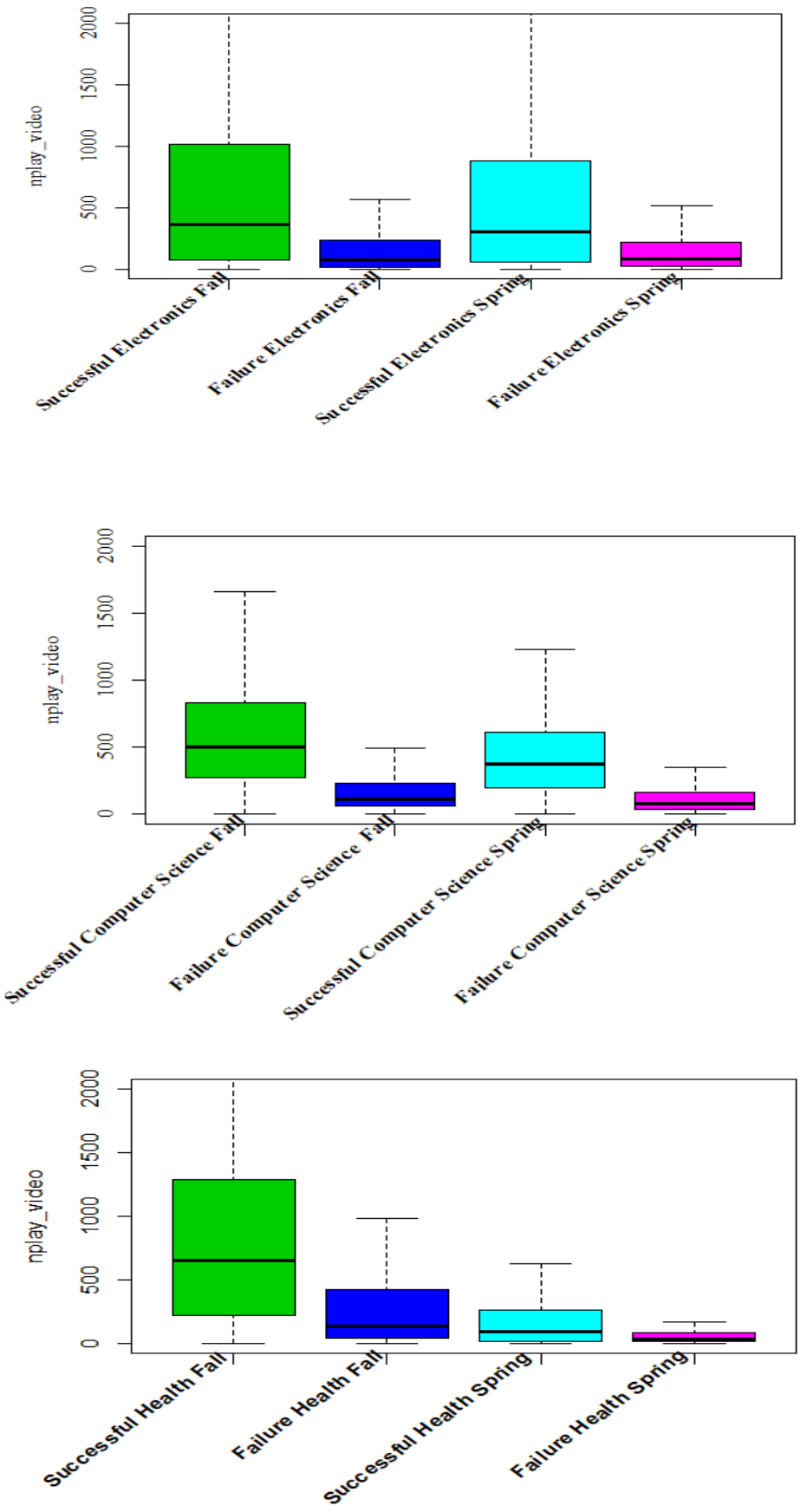


Figure 5.11 Mean values of failing and successful learners per video view

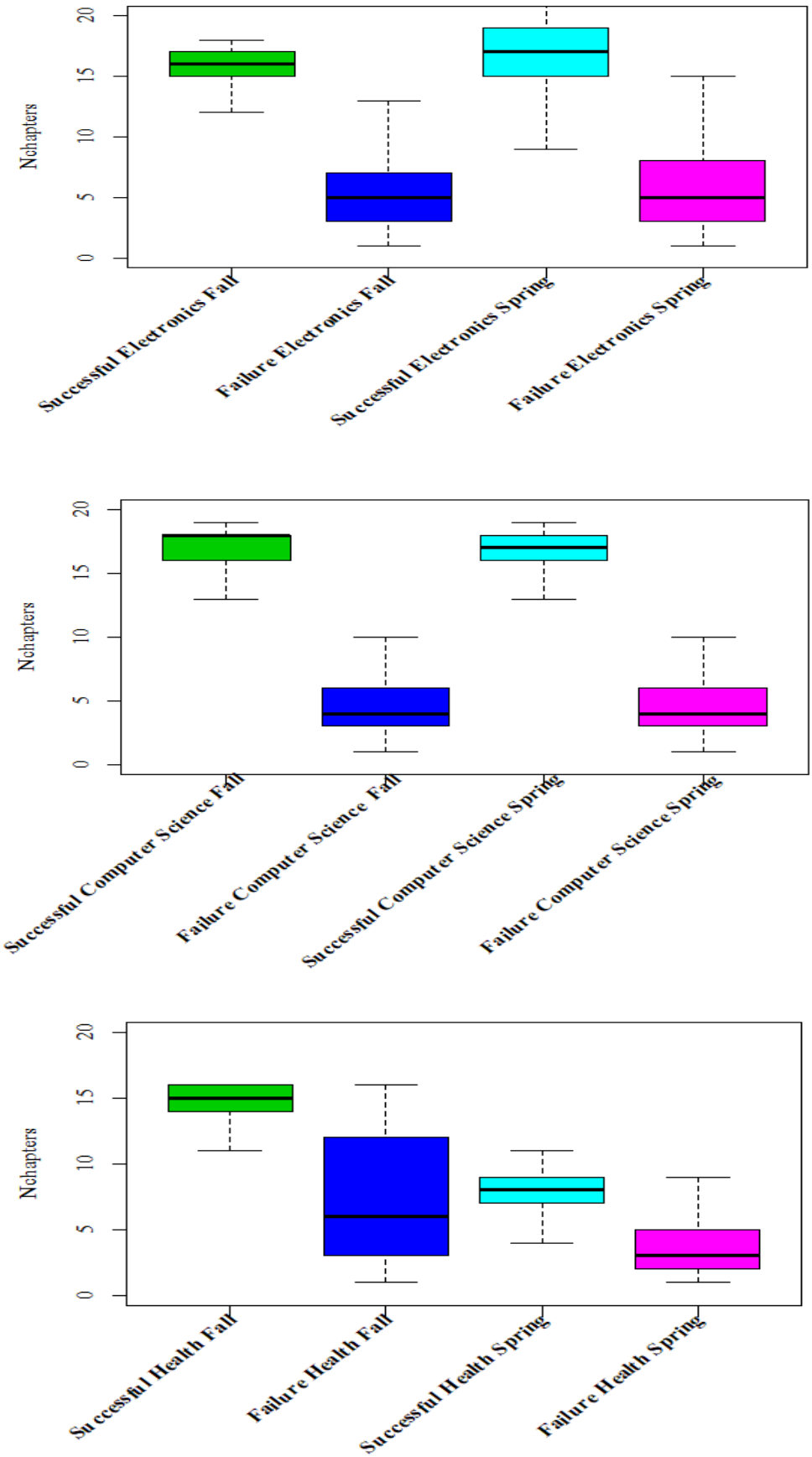


Figure 5.12 Mean values of failing and successful learners per chapters read

Table 5.7 ANCOVA Result

Courses	DF	Sum Sq	Mean Sq	F value	Pr(>F)
"Electronics Fall"					
nplay_video	1	8.91	8.91	97.92	< 2e-16
Nchapters	1	115.94	115.94	1273.74	< 2e-16
Residuals	953	86.75	0.09		
“Computer Science Fall”					
nplay_video	1	29.32	29.32	449.09	<2e-16
Nchapters	1	74.51	74.51	1141.38	<2e-16
Residuals	1295	84.54	0.07		
“Health Fall”					
nplay_video	1	19.91	19.91	156.74	<2e-16
Nchapters	1	60.55	60.55	476.81	<2e-16
Residuals	910	115.57	0.13		
"Electronics Spring"					
nplay_video	1	11.67	11.67	155.41	< 2e-16
Nchapters	1	41.42	41.42	551.37	< 2e-16
Residuals	528	39.66	41.42		
“Computer Science Spring”					
nplay_video	1	18.67	18.67	391.3	< 2e-16
Nchapters	1	50.99	50.99	1068.4	< 2e-16
Residuals	1354	64.61	64.61		
“Health Spring”					
nplay_video	1	8.26	8.26	73.780	<2e-16
Nchapters	1	52.98	52.98	473.017	<2e-16
Residuals	558	62.49			

5.3.2 Educational Level of Failing and Successful Learners Result

In this section, the distributions of educational level across success and failure learners is investigated. Table 5.8 illustrates the Chi-squared result. The DF stands for the degrees of freedom and can be defined as the number of independent values that vary in the final calculation. The result indicates a p-value of ($p < 0.05$) for all courses, except the “Computer Spring” and “Health Spring” courses, allowing for a rejection of the null hypothesis.

Figures 5.13 and 5.14 show the distribution of the success and failure of learners per each course in respect to their educational level. Overall, most completion learners are reported as being secondary, bachelors and masters qualified, with a smaller number of doctorate learners aiming to earn certification. An average of 20%-30% of learners who have bachelors or secondary degrees failed in “Electronic Fall”, “Electronic Spring” “Computer Fall”, and “Computer Spring” courses. The students with bachelors and masters qualifications dominated the highest rate of non-successful

learners for the “Health Fall” and “Health Spring” courses respectively. Around 46% of certificated learners in the “Electronic Fall” course have a secondary degree, whereas the percentage of such students drops to 36%-33percentage in the “Computer Spring” and “Computer Fall” courses. Most of the successful learners with a bachelor’s degree are shown in “Electronic Fall” course. An average of 44%-35percentage of certified students had a masters degree in the “Health Fall” and “Health spring” courses. Please refer to appendix 8 and 9.

Figures also show that learners with fewer secondary and doctorate qualifications reported the lowest percentage of participation across all courses. An average of 2% students with fewer secondary degrees failed in “Electronic Fall” and “Electronic Spring” courses, whereas conversely, the percentage of failure students in “Computer Fall” and “Computer Spring” courses is 2% higher, with doctorate qualifications applicable to approximately 2%-0.5percentage of the student participants. The course “Health Fall” has the highest number of doctorate qualifications learners. To summarise, the failure group includes double the number of students of the successful group per each educational level. As a result, educational levels cannot be considered the vital factor that influences student performance.

Table 5.8 Results of the Chi-squared Test

Course	χ^2 statistic	df	P-value
Electronics Fall	32.012	4	< 1.902e-06
Electronics Spring	3.4134	4	0.4912
Computer Science Fall	34.734	4	<5.268e-07
Computer Science Spring	64.434	4	<3.386e-13
Health Fall	23.936	4	<8.227e-05
Health Spring	1.8792	4	0.758

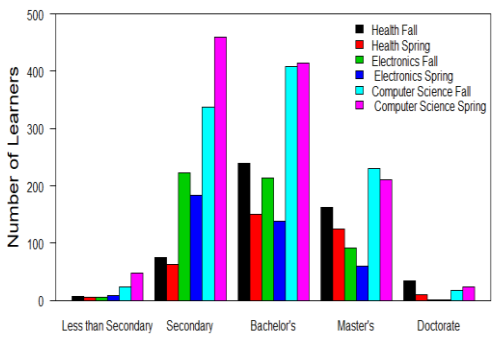


Figure 5.14 Failing Learners by Educational Level

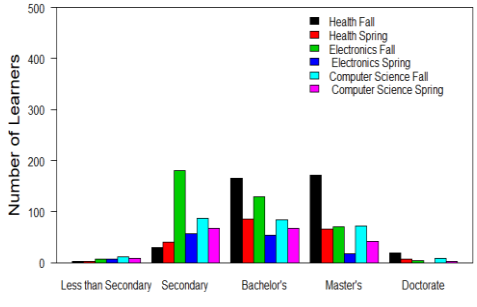


Figure 5.13 Successful Learners by Educational Level

5.3.3 Machine Learning Results

In the following section, the results of the second case study will be discussed. Machine learning has been used as the promising solution for detecting the student motivational category and identifying at-risk students early.

5.3.3.1 Motivational Prediction Model Results

The classification results of the motivational predictive model are presented in this section. Figure 5.15 compares an estimation of the performance of classifiers over the training set. The Figure shows that RF achieves the highest performance, whereas Nnet achieved the lowest performance. The simulation results are compared according to confusion matrix metrics and ROC curve, as listed in Tables 5.9 and 5.10 .Table 5.9 summarises the overall accuracy results, showing the best result of 0.802 yielded by the RF, whereas Nnet achieved the lowest result, with an average value of 0.734.

As can be seen in Table 5.10, the class “extrinsic” acquired the best performance for all classifiers, the ACC = 87%-90%. Conversely, the “amotivation” class yielded the lowest result for all classifiers. However, in the Nnet model the class “intrinsic” gives the poorest performance, with a value of 0.71.

In terms of sensitivity (true positive) analysis, the class “extrinsic” has gained the highest sensitivity for all the classifier models. As such, the RF produces the best performance, which achieves sensitivity=0.851, whereas Mlp gives the low sensitivity, with a value of 0.70. Over the three classes, the Nnet for class “intrinsic”, followed by Gbm, which achieve the values of 0.51 and 0.621 respectively, yields the poorest sensitivity.

Table 5.10 shows that specificity (true negative) is higher than sensitivity (true positive) across all classifiers. However, the class “amotivation” for Gbm provides better sensitivity. Again, the RF obtains the best specificity for the class “extrinsic”, which acquired the value of 0.965. All other models over the “extrinsic” label gain very good specificity with a range values of 0.91-0.96, whereas the Gbm acquired the lowest specificity = 0.764 for the class “amotivation.”

The ROC analysis was used to choose a decision threshold value for the true and false positive rate. Figure 5.16 shows the similarity of performance for all classifier models over the class “amotivation” and “extrinsic”, achieving a range of AUC values

within 83%-93% across all classes. The worst AUC gained by the Rpart model for class “intrinsic” gained a value of 0.79.

Table 5.10 also shows that the F1-Measure for RF has slightly better results than another model over the three classifiers, obtaining a value of 0.78, 0.88, 0.76 respectively. The lowest F1-Measure is reported for class “extrinsic” with a value of 0.62 in respect to the Nnet model.

Overall, the results show that there is no significant difference between the accuracy of an Mlp,Gbm and the Rpart tree. One possible explanation for the Mlp and Gbm slightly superior performance in comparison to the decision tree, is their ability to build internal abstractions to aid in the analysis of the complex relationship between the input features and the target. The hidden unit in the neural network creates a new feature space, which can be used to facilitate class discrimination. In addition, the Gbm reweight the week learners by adopt the ensemble methods. However, both Mlp and Gbm act as a black box, impeding the interpretation of feature contributions.

In contrast, decision trees provide an easily accessible representation, which may be used to understand which features have an impact on prediction. In our case, we find that “clickstream” followed by “ndays_act” features are the most important for the purposes of prediction.

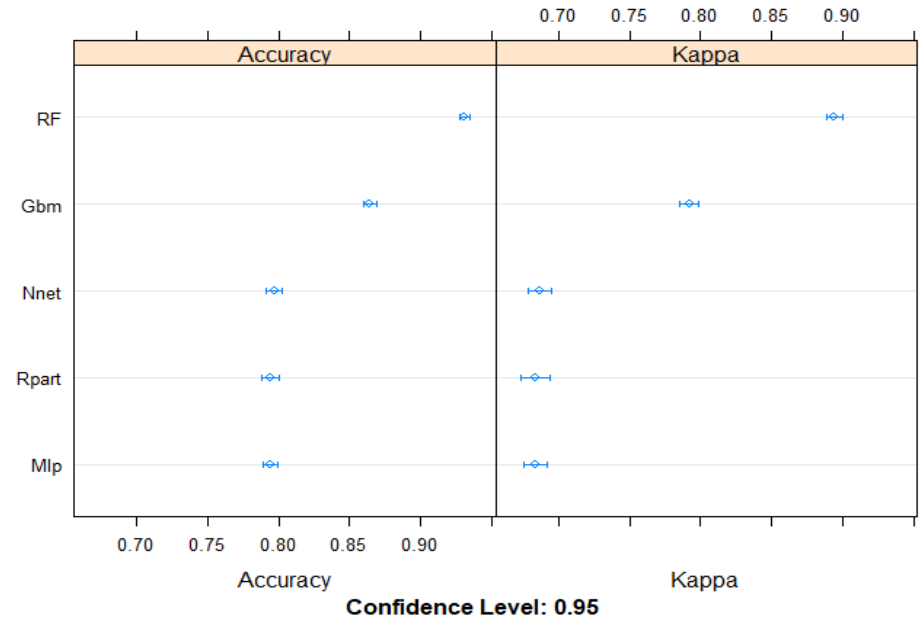


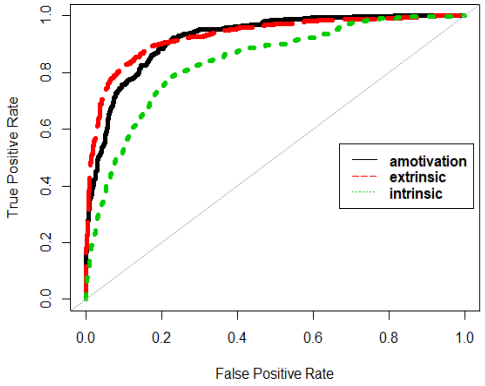
Figure 5.15 Estimation Accuracy for motivational predictive model

Table 5.9 Accuracy Result for Motivational Prediction Model

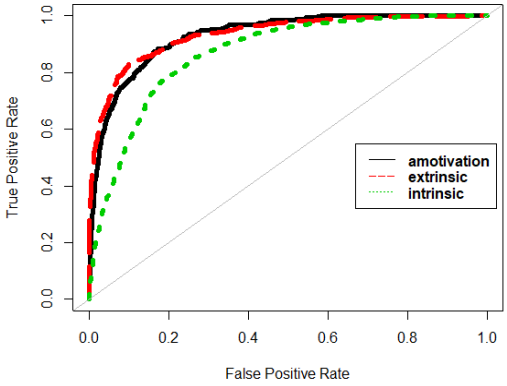
Classifier	Accuracy
Mlp	0.755
RF	**0.802**
Rpart	0.746
Gbm	0.774
Nnet	0.734

Table 5.10 Classification Performances Result for Motivational Prediction Modal

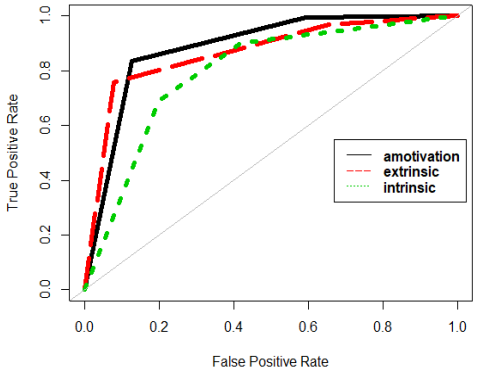
Classifier	Performance Metric				
Mlp	ACC.	F-Meas.	Sens.	Spec.	AUC
amotivation	0.756	0.705	0.702	0.809	0.920
extrinsic	0.878	0.847	0.819	0.938	0.929
intrinsic	0.814	0.714	0.750	0.879	0.837
RF	ACC.	F-Meas.	Sens.	Spec.	AUC
amotivation	0.804	0.761	0.754	0.855	0.929
extrinsic	**0.908**	**0.887**	0.851	**0.965**	**0.939**
intrinsic	0.845	0.757	0.809	0.882	0.866
Rpart	ACC.	F-Meas.	Sens.	Spec.	AUC
amotivation	0.741	0.6815	0.641	0.841	0.884
extrinsic	0.880	0.8426	0.844	0.915	0.865
intrinsic	0.817	0.7136	0.775	0.860	0.796
Gbm	ACC.	F-Meas.	Sens.	Spec.	AUC
amotivation	0.791	0.752	0.819	0.764	0.932
extrinsic	0.895	0.870	0.837	0.954	0.938
intrinsic	0.774	0.679	0.621	0.928	0.871
Nnet	ACC.	F-Meas.	Sens.	Spec.	AUC
amotivation	0.844	0.742	**0.867**	0.821	0.931
extrinsic	0.879	0.831	0.877	0.882	0.936
intrinsic	0.711	0.621	0.517	0.904	0.853



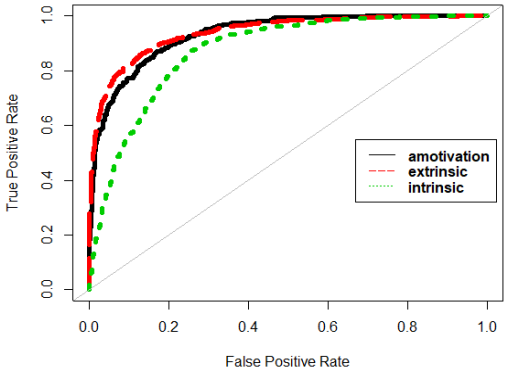
a) Mlp Classifier



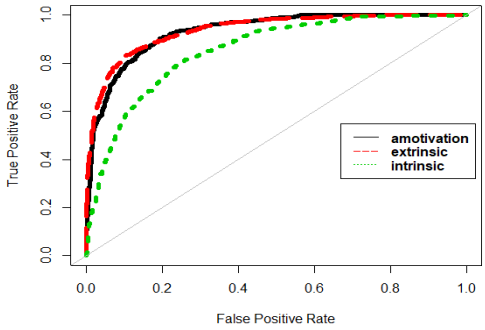
b) RF Classifier



c) Rpart Classifier



d) Gbm Classifier



e) Nnet Classifier

Figure 5.16 Roc Curve for motivational predictive model

5.3.3.2 Dropout Prediction Model Result Based on Student Performance

In this analysis the problem is followed by binary classification, wherein dropout students represented the positive class, whilst non-dropout students are assigned to the negative class. The empirical results compare the classifier models over two sets of

features, where students' demographic attributes and grades are considered in the first set of features, and student grade is eliminated in the second set of features.

Tables 5.11 and 5.12 show the simulation results obtained for each classifier respectively, over both sets of variables. As can be seen, accuracy is low for all classifiers over both sets of features. The Gbm acquired the highest accuracy in the first set of features, with a value of 0.53, while Glm offers the best accuracy in the second set of features, where a value is obtained of 0.57. The Rpart gives the lowest accuracy in both sets of features, achieving values of 0.332 and 0.189 respectively.

F-Measure is also used in this case study, and gives a better indication of the classifier model than the other performance metrics. The findings that result from the F-Measure are similar in accuracy. Gbm and Glm show the highest values for both sets of features, while Rpart achieves the poorest result.

Sensitivity (true positives) and specificity (true negatives) acquire the low values for all classifiers for both sets of features, with an approximate range of values from 0.40-.0.59. However, Rpart yields the highest specificity, with values of 0.72, 0.87 respectively in contrast with sensitivity where the lowest one is achieved by Rpart. The Receiver Operator Characteristic (ROC) and Area Under Curve (AUC) were also considered. Figures 5.17 and 5.18 show the ROC results for both sets of variables. The curves are shown to converge to roughly the same poor result on the plot, indicating a similarity in performance across models in both features sets, resulting in values of around 43% and 53%.

In general, not all the classifier models performed well. The poor results demonstrated that demographic features are incapable of distinguishing withdrawal and non-withdrawal students, due to such features having weak associations with the target class. The results reveal that behavioural features can sufficiently differentiate students who are at-risk of dropout.

Table 5.11 Classification Performances for dropout Model First Set of Features

Classifier	Acc.	F-Meas.	Sens.	Spec.	AUC
Mlp	0.502	0.637	0.497	0.538	0.458
RF	0.511	0.645	0.504	0.565	0.435
Rpart	0.332	0.424	0.279	**0.72**	0.488
Glm	0.481	0.645	0.509	0.489	**0.501**
Gbm	**0.533**	0.672	**0.544**	0.446	0.466
Nnet	0.497	0.633	0.493	0.522	0.456

Table 5.12 Classification Performances for dropout Model Second Set of Features

Classifier	Acc.	F-Meas.	Sens.	Spec.	AUC
Mlp	0.537	0.675	0.541	0.501	0.504
RF	0.498	0.636	0.492	0.547	0.464
Rpart	0.189	0.185	0.104	**0.878**	0.471
Glm	**0.571**	0.712	**0.59**	0.413	**0.535**
Gbm	0.447	0.58	0.43	0.581	0.489
Nnet	0.486	0.625	0.49	0.457	0.525

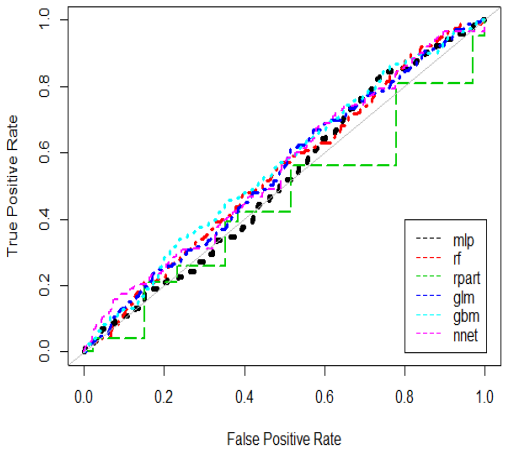


Figure 5.17 Roc First Set of Features

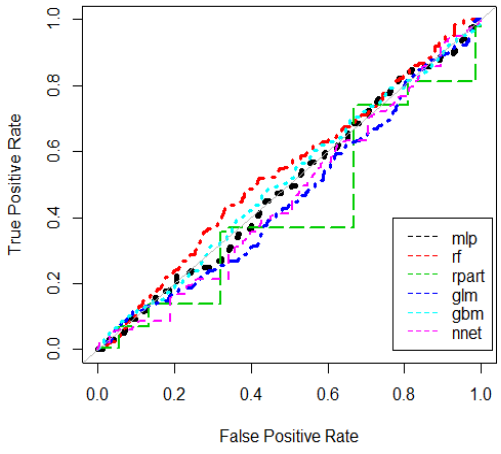


Figure 5.18 Roc Second Set of Features

5.3.3.3 Dropout prediction Model based on student motivational status Result

This section presents the results for the predictive model, according to the second definition of at-risk students. The table 5.13 shows the correlation analysis between the features and the target class. The results demonstrate a strong association between the behaviorur features at the second time interval with the target. The "Nchapters" acquired the coefficient value of 0.59, and conversely has a weak correlation with students' behavioral attributes at the first-time interval with the response variable.

The good relation between students' motivational status at the first time interval with the target at-risk students achieves a value of 0.34. The significant result indicates that student interventional motivation can be used as a robust predictor to estimate students who are at-risk from the dropping out in future courses.

To gain deeper insight, the statistical analysis has been applied. The result of statistical analysis is display in appendix 7. In all courses, the highest percentage of retention learners in the following course were unmotivated in the prior course. In contrast to students who were shown intrinsically and extrinsically motivated in the previous course, they were withdrawn from the course within a week. The appendix 7 shows that only 31% of amotivation students withdraw in the following course while the proportion of withdrawal students sharply increased for intrinsically and extrinsically motivated. It is noticeable that 84% -77% of intrinsically motivated and extrinsically motivated drop out in the next course.

The empirical results over each classifier have been compared in respect to their performance metrics, including accuracy, specificity and sensitivity, precision, recall, and AUC. Table 5.14 shows the results obtained for each classifier, respectively. As can be seen, the Nnet model shows the highest level of accuracy in comparison to other models, with the value of 0. 909, followed by RF. Both Mlp and Glm occupy the third highest accuracy models, having values of 0.870, 0.879 respectively. The Gbm and Rpart are considered the weaker classifiers, and achieve the lowest range of performance, with accuracies of 0.839 and 0.811, respectively.

Sensitivity is seen to be slightly higher than specificity. In particular, models Nnet, RF,Mlp and Glm obtained average values of 95%, 91%, and 88% respectively. Conversely, Rpart and Gbm achieve the lowest sensitivity. Rpart and Glm gained the

highest specificity with average values of 0.87. The worst specificity is yielded by RF. Figure 5.19 shows ROC for the predictive model. The results indicate a similarity of performance across all classifiers.

Table 5.13 Correlation Analysis
According to Second at- risk Student

Features	Correlation Coefficient
YOB	0.03
Gender	0.07
LOE_DI	0.01
final_cc_cname_DI	-0.04
ndays_act/t_1	0.10
Nevent/t_1	0.06
nplay_video /t_1	0.01
Nchapters/t_1	0.15
nforum_post/t_1	0.01
Explored/t_1	0.02
motivational statue/t_1	**0.34**
ndays_act/t_2	0.01
Nevent/t_2	-0.34
nplay_video /t_2	-0.32
Nchapters/t_2	**-0.59**
nforum_post/t_2	-0.10
Explored/t_2	-0.30

Table 5.14 Classification Performances for dropout Prediction Model

Classifier	Acc.	F-Meas.	Sens.	Spec.	AUC
Mlp	0.870	0.895	0.881	0.852	0.928
RF	0.888	0.911	0.919	0.837	0. 941
Rpart	0.811	0.836	0.770	**0.878**	0.906
Glm	0.879	0.901	0.883	0.873	0.937
Gbm	0.839	0.865	0.822	0.867	**0.944**
Nnet	**0.909**	0.9503	**0.9503**	0.8421	0.933

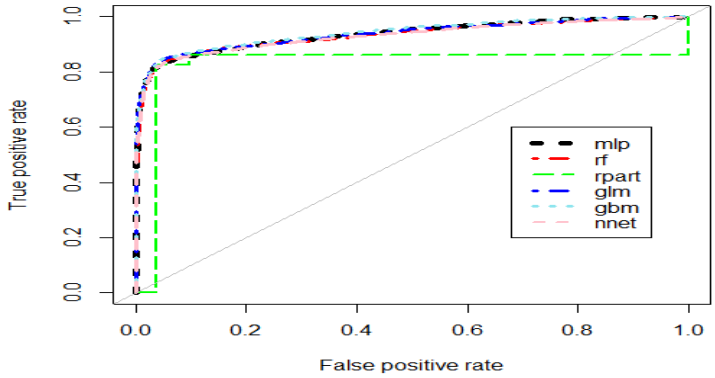


Figure 5.19. Roc Curve for Second Dropout Prediction Model

5.4 Experiment Three Results

The results relevant to the OULAD dataset is illustrated in the following section. The EDA results, features selection in addition to simulation results of prediction models have presented in next section.

5.4.1 Exploratory Data Analysis Results

The correlation analysis between the extracted behavioural features and the final student results is visualized in figure 5.20. The map shows a positive relationship between the behavioural features and the "final_result" class. The features "session.quiz" and "sum_click.quiz" have the highest degree of relationship, with the target, achieving a coefficient value 0.69 and 0.62 respectively. This is followed by "session.homepage" gaining a value of 0.51. The positive correlation with "final_result" is shown in other behavioural features. The approximate range of coefficient values are 0.20-0.36.

The weakest association of behavioural features with the target is acquired by "session.sharedsubpage" and "sum_click.sharedsubpage". It is notable that none of the feature has negative correlations with a given target. The positive correlation of behavioural features with target could demonstrate that our approach of feature extraction is sufficiently robust.The map in Figure 5.21 shows that there is a weak positive relationship between the demographic features and the "final_result". The attributes "gender" and "region" have a negative correlation with the target. In spite of the Harvard dataset and OULAD database having different structures, the behavioural features obtain the highest degree of correlation with learning outcome, while the demographic features yield low values for their relationship with student performance.

The OULAD dataset exhibits low variance. The number of Principal components was reported as 10 in this dataset. The Figure 5.23 describes the results of the Kaiser method. The Figure shows that nine components are chosen as the optimal components.

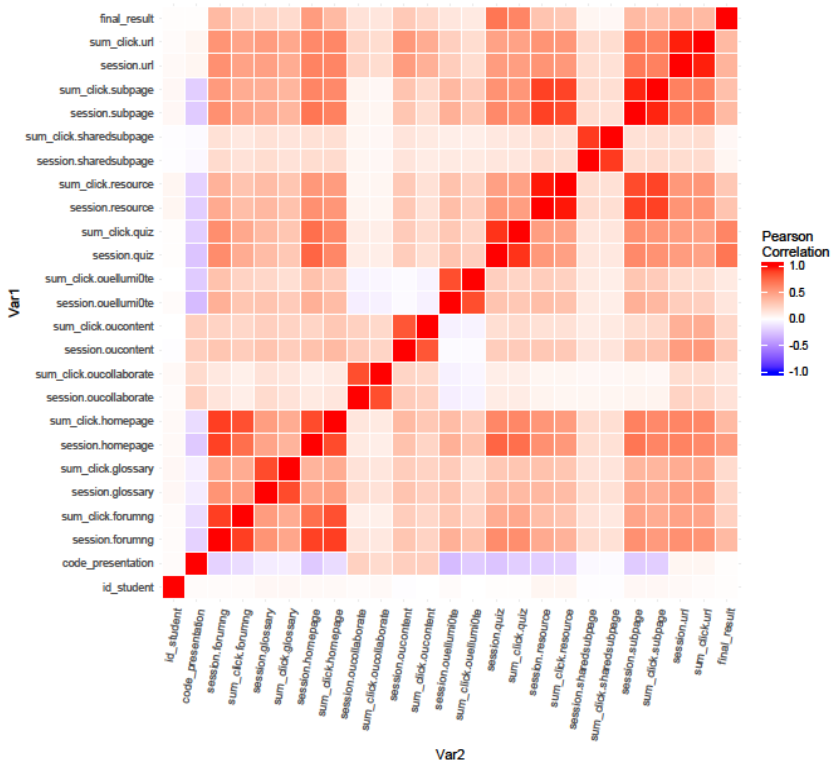


Figure 5.20 Heat Map for Behavioral Features OULAD Dataset

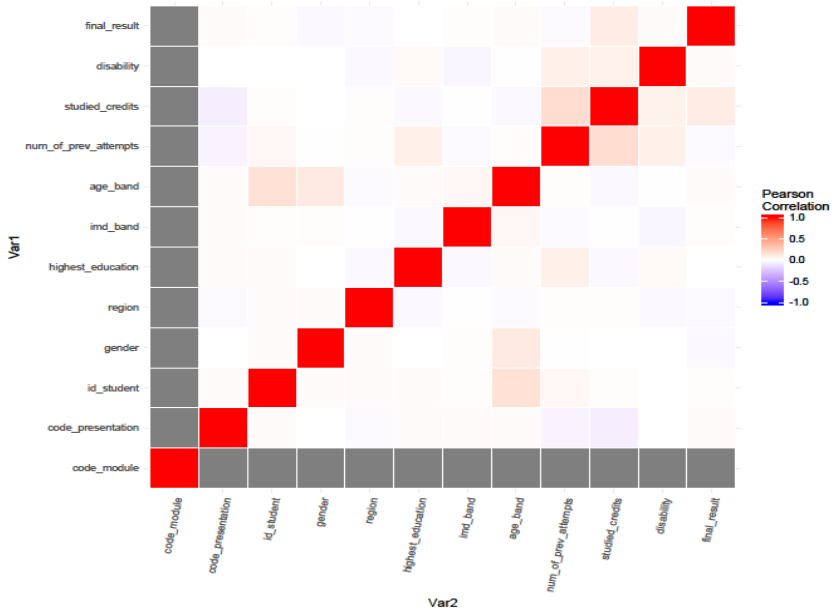


Figure 5.21 Heat Map for Demographic Features OULAD Dataset

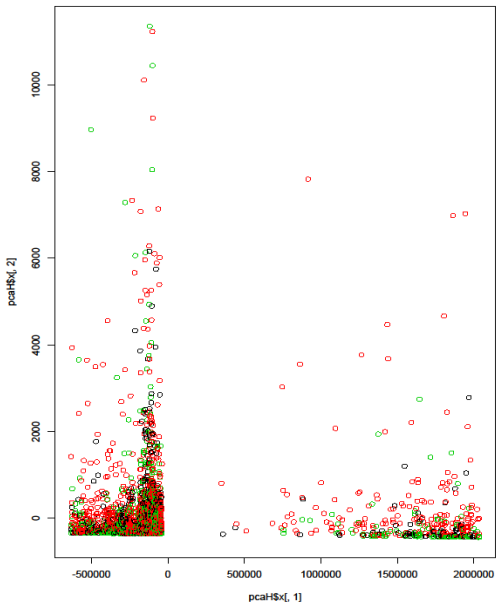


Figure 5.22 PCA for OULAD Dataset

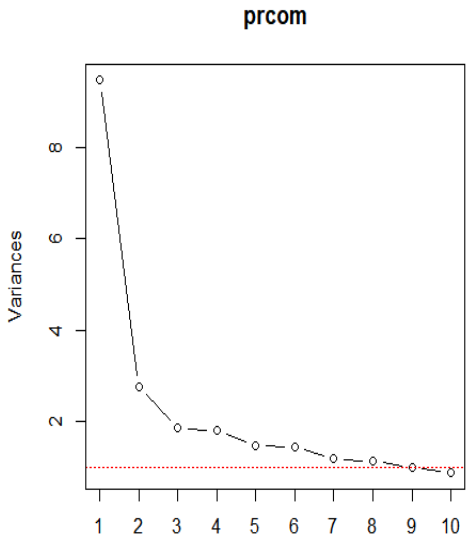


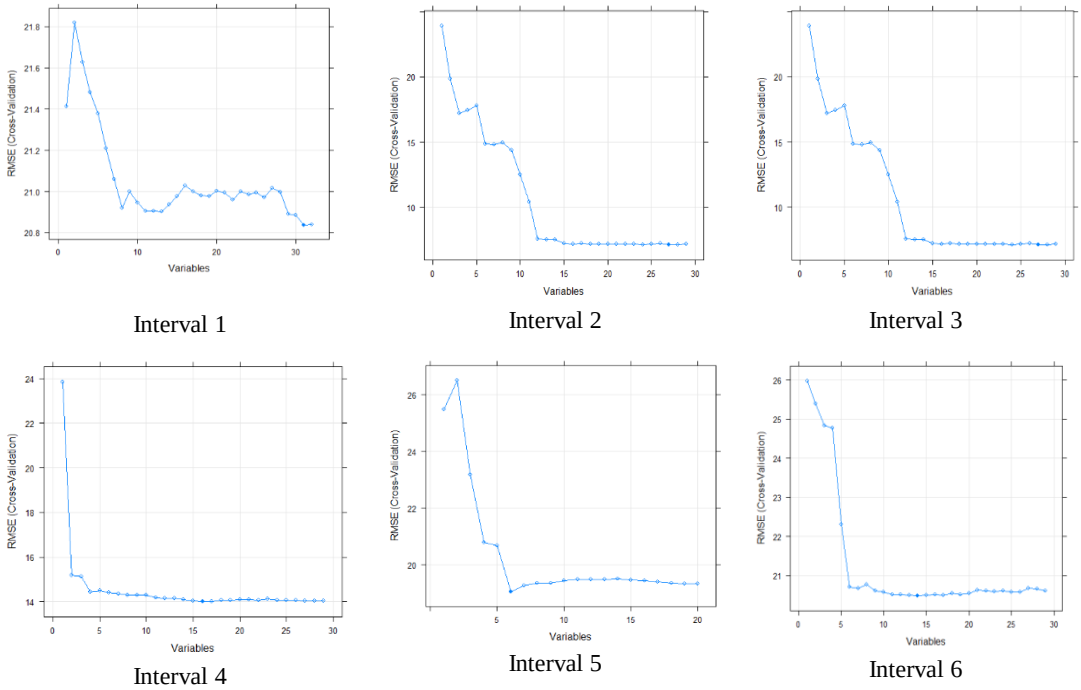
Figure 5.23 select PCA with Kaiser

5.4.2 Features Selection Results

RFE are considered only for regression analysis, the results of RFE across six intervals are listed in table 5.15 and Figure 5.24. The figure displays the results of the RFE, based on RMSE criteria. The results reveal that students' assessments grades during the previous slice occupy the position of highest top features across all intervals. "session.quiz/T6","sum_click.quiz/T5","session.homepage/T6","sum_click ,homepage/T6" are the top five features for intervals 5 and 6, respectively. The "session.homepage" and "sum_click.homepage" of the previous and current time slice are the most important features across all intervals. It also appears that "sum_click.forumng" and "session.forumng" have been selected as top features, only for interval 1 and 2. In general, the activities ("homepage","quiz"," subpage") robustly predicted the students' grades over all time intervals.

Table 5.15 High Ranking Features OULAD Dataset across Six Time Intervals

Interval	Five Top Features
Interval 1	"session.homepage/T2" "sum_click.forumng/T2" "session.resource/T2" "session.forumng/T2" "sum_click.subpage.T2"
Interval 2	"Score.Assessment 1" "session.homepage/T3" "session.quiz/T3" "session.homepage /T2" "sum_click.forumng/T3
Interval 3	"Score. Assessment 2" "sum_click.homepage/T3" "sum_click.subpage/T3" "session.homepage/T3" "session. subpage/T3"
Interval 4	"score. Assessment 3" "sum_click.quiz/T5", "session.quiz/T5", "sum_click.homepage/T5", "session.homepage/T5"
Interval 5	"score. Assessment 4", "session.quiz/T6", "sum_click.quiz/T5", "session.quiz/T5" " sum_click.homepage/T6"
Interval 6	"Score. Assessment 5", "session.homepage/T6", "sum_click.homepage/T6", "sum_click.homepage/T7" " session .homepage/T7"



Interval 1 Interval 2 Interval 3

Interval 4 Interval 5 Interval 6

Figure 5.24 High Ranking Features OULAD Dataset across Six Time Intervals

5.4.3 Students Assessments Grades Predication Model Results

The regression analysis has been applied to predict students' assessment grades over six time intervals. Two sets of features are considered. In the first set of features, dynamic behavioural features and student performance are used as predictors, while only the top five features are employed in the second set of features.

The empirical results from the first set of features have acquired slightly better performance than the selected features. However, the Mlp and Rpart in the second set of features offer better results. The RMSE metric measures the difference between the predicted values and the actual observations. The lowest RSME value demonstrates better performance of the predictive model. In terms of the first set of features, RF obtains the best RMSE achieving values of 8.131 over interval 3. For the second set of features, the Gbm model gives the best RMSE, with a value of 11.230 over the interval 1.

As can be seen, for both set of features the Gbm models over interval 1, 2, 4, 5, and 6 give the best RMSE, with averages values of 11-22. RF occupies the second best model. The RMSE increases by 3% in Glm across all intervals and acquired an average value of 11-23. Mlp acquired the worst RMSE, with a value of 44.904 and 39.215 at interval 6 for both sets of features. In general, Mlp achieves the poorest RSME across all intervals for both sets of features.

The $R^2$is the percentage of variation in the predicted variables used to evaluate the goodness of fit predictive model into the regression line. This is called the coefficient of determination. The best R^2 was obtained by RF, which gives the values of 0.937 and 0.853 at interval 3 and interval 4 for the first and second set of features respectively. Again, for both sets of features the RF acquired the best R^2across all other intervals, with approximate range values of 0.85-0.64. All models over interval 1 for the second set of features gain poor R^2 results, with an average value of 0.170-0.274.

Table 5.16 Regression Result for Predication Students' Assessments Grades Model

All Features Set

Interval	Model	RMSE	R^2
Interval 1	Mlp	18.162	0.483
	RF	11.395	0.790
	Rpart	11.647	0.780
	Glm	11.382	0.790
	Gbm	11.230	0.204
Interval 2	Mlp	29.517	0.427
	RF	18.221	0.640
	Rpart	22.400	0.459
	Glm	19.116	0.604
	Gbm	18.145	0.644
Interval 3	Mlp	20.564	0.641
	RF	**8.131**	**0.937**
	Rpart	13.760	0.18
	Glm	15.376	0.773
	Gbm	11.506	0.873
Interval 4	Mlp	18.205	0.768
	RF	13.904	0.853
	Rpart	16.264	0.799
	Glm	15.354	0.821
	Gbm	13.892	0.853
Interval 5	Mlp	29.665	0.428
	RF	19.900	0.716
	Rpart	25.001	0.553
	Glm	21.708	0.663
	Gbm	19.871	0.718
Interval 6	Mlp	44.904	0.147
	RF	21.529	0.688
	Rpart	25.259	0.569
	Glm	22.058	0.673
	Gbm	21.425	0.691

Selected Features

Interval	Model	RMSE	R^2
Interval 1	Mlp	24.324	0.170
	RF	21.366	0.272
	Rpart	22.457	0.198
	Glm	22.344	0.204
	Gbm	**11.230**	0.274
Interval 2	Mlp	26.374	0.558
	RF	18.062	0.643
	Rpart	22.628	0.444
	Glm	18.831	0.612
	Gbm	18.321	0.634
Interval 3	Mlp	27.574	0.405
	RF	16.225	0.764
	Rpart	24.304	0.434
	Glm	23.106	0.489
	Gbm	22.600	0.512
Interval 4	Mlp	16.733	0.790
	RF	13.905	**0.853**
	Rpart	16.264	0.799
	Glm	15.787	0.811
	Gbm	14.074	0.849
Interval 5	Mlp	26.271	0.54
	RF	19.259	0.737
	Rpart	25.001	0.553
	Glm	21.831	0.659
	Gbm	19.478	0.728
Interval 6	MLP	39.215	0.067
	RF	22.794	0.650
	Rpart	25.919	0.546
	Glm	23.332	0.633
	Gbm	22.761	0.650

5.4.4 Final Students Performance Prediction Model Results

The classification analysis results from the first case study are presented as follows. As can be seen in a table 5.17 all classifiers obtain similar ideal results, the highest performance achieved by Gbm with the value of 0.868 while RF, Nnet producing the value of 0.854, achieved the lowest accuracy.

As can be seen in table 5.18, the class "Withdrawn" acquired the best accuracy of all Classifiers reaching an average value of 0.99 whereas the class "Fail" gives the lowest performance, with approximate range of accuracy between 0.76-0.80.

The sensitivities are high over all classifiers for class “Withdraw” and “Pass”. The best sensitivity achieved by Rpart reported the values of 0.99 and 0.92. The class “Fail” gained very low sensitivities across all classifiers. This is expected since the number of records with target class “Fail” are limited hence, the algorithm cannot learn well.

With regards, to true negative instance, the Gbm and Nnet produce the best result, specificity =0.998 for class “Withdrawn”. The poorest specificity gained by Rpart for class “Pass” obtained the values of 0.81. As can be seen, the best F1-Measure gained by Gbm yielded value of 0.993, 0.864, 0.772, for the class “Withdrawn”,” Pass” and “Fail” respectively. The lowest F1-Measure is shown for Rpart with the value of 0.67 over class “Fail”.

ROC is used in this study to choose a decision threshold value for the true and false positive rate across each class. Figure 5.26 lists ROC curves. Overall, a range of AUC values between 0.99-0.82 for all classes was obtained.

As previously mentioned, the demographic behavioural and temporal features in classification analysis were combined. In this model, the total numbers of variables are 35. As a result, the predictive model may suffer from the overfitting issue. In this case, we compare classifiers results in terms of train and test error, which could give an indication of the overfitting problem. Figure 5.25 displays the result of overfitting evaluation. It can be observed that training and test error are low for all classifiers. The lowest test and train error was obtained by Gbm .The RF ,Nnet obtained a similar test error with an approximate percentage of 14% .The training errors are slightly higher in these classifiers. The largest error was acquired by the Mlp model. Although, all models fit well for most classifers , Mlp suffers from overfitting.

Table 5.17 Accuracy Result for Final Students Performance Model

Classifier	Accuracy
Mlp	0.858
RF	0.854
Rpart	0.862
Gbm	**0.868**
Nnet	0.854

Table 5.18 Results for Final Students Performance Prediction Model

Classifier	Performance Metrics				
MLP	ACC.	F-Meas.	Sens.	Spec.	AUC
Pass	0.858	0.850	0.892	0.824	0.916
Fail	0.782	0.690	0.631	0.932	0.886
Withdrawn	**0.993**	0.992	0.989	0.996	0.996
RF	ACC.	F-Meas.	Sens.	Spec.	AUC
Pass	0.855	0.843	0.844	0.866	0.924
Fail	0.808	0.713	0.712	0.904	0.892
Withdrawn	**0.995**	0.993	0.991	0.990	0.995
Rpart	ACC.	F-Meas.	Sens.	Spec.	AUC
Pass	0.866	0.865	0.923	0.810	0.867
Fail	0.767	0.671	0.582	0.952	0.821
Withdrawn	**0.997**	0.991	**0.996**	0.992	**0.997**
Gbm	ACC.	F-Meas.	Sens.	Spec.	AUC
Pass	0.872	0.864	0.903	0.841	0.925
Fail	0.802	0.722	0.665	0.939	0.900
Withdrawn	**0.994**	0.993	0.991	**0.998**	**0.997**
Nnet	ACC.	F-Meas.	Sens.	Spec.	AUC
Pass	0.856	0.8471	0.870	0.843	0.925
Fail	0 .795	0.7045	0.670	0.920	0.900
Withdrawn	**0.994**	0.9934	0.991	**0.998**	0.998

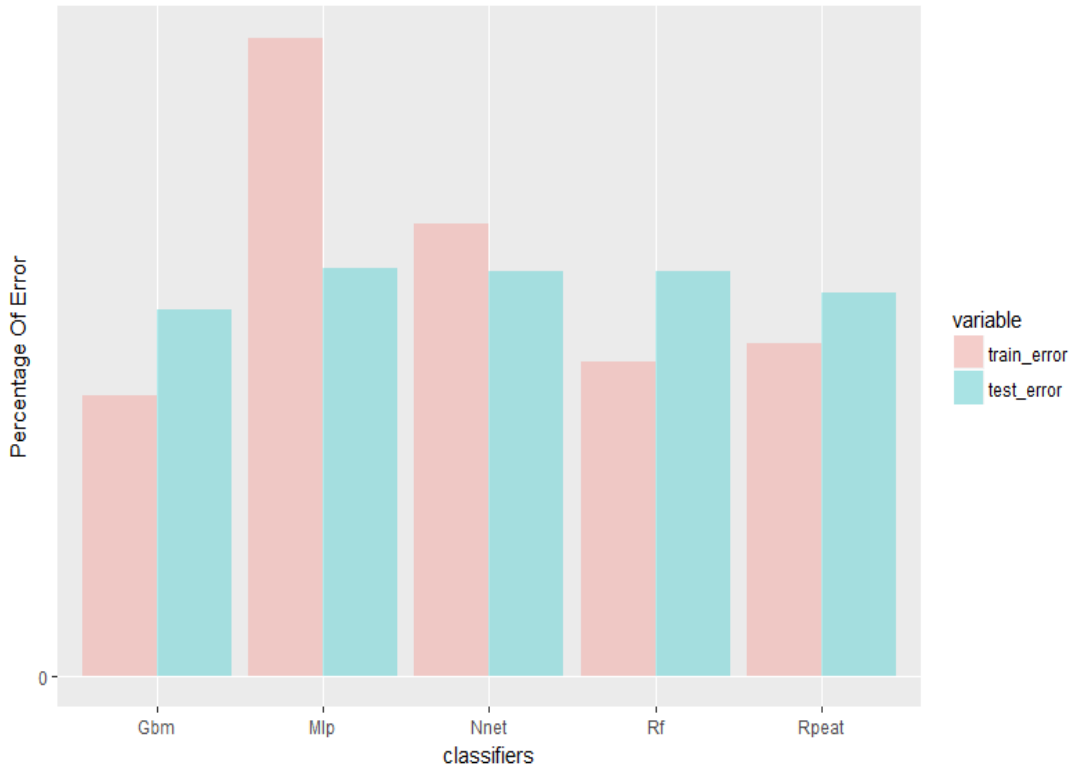
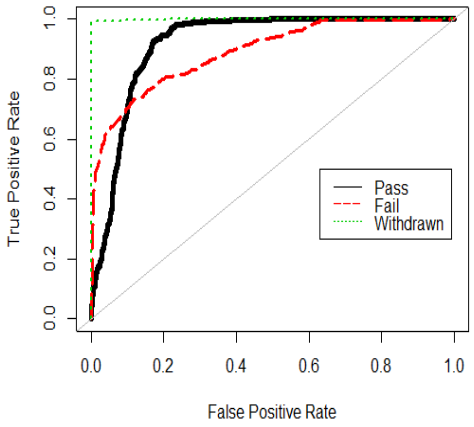
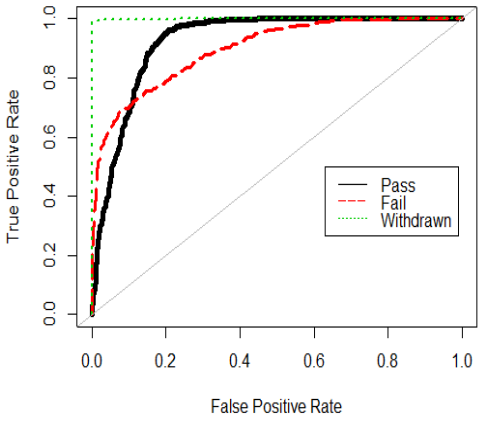


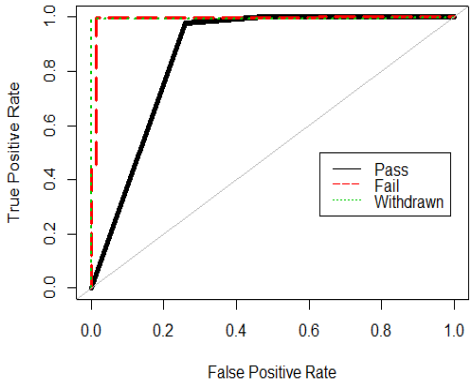
Figure 5.25 Comparing Computational Training and Test Time



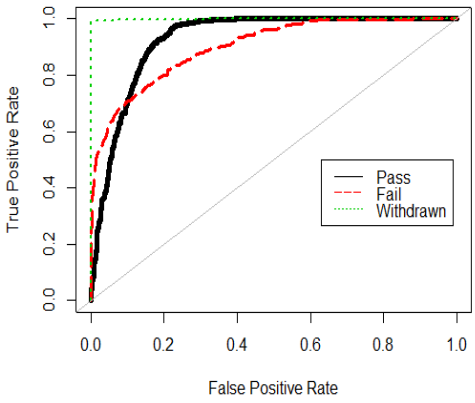
a) Mlp Classifier



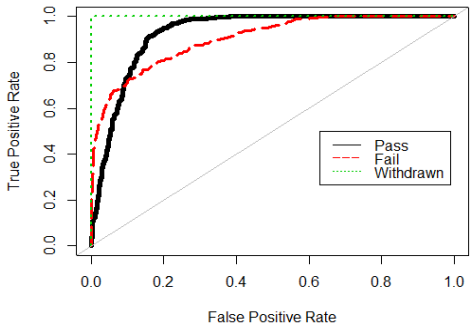
b)RF Classifier



c)Rpart Classifier



d)Gbm Classifier



e)Nnet Classifier

Figure 5.26 Roc Curve for Final Students Performance Prediction Model

5.4.5 Dropout Prediction Model based on Latent Engagement Result

Model responses from the experimental procedure were obtained for the classifiers under study, namely the EDDA, Mlp, RF, Rpart, Gbm ,Glm and Nnet classifiers, designed within the context of a discrete binary outcome. Ground truth labels were defined as "Withdrawal" and "No withdrawal" respectively. Different model types are used to represent the clusters. The best model is selected for each interval based on cross-validation error, and test error. As can be seen in table 5.21, some models do not fit the data in interval 1, 3, 5 and 6.

The table 5.19 list the result of EDDA model over six intervals. The results indicate that the number of estimated parameters varies significantly over time. However, a similar number of latent variables were found to be present among intervals 3 and 5. A possible interpretation is that the student's hidden status of engagement could vary at the beginning and end of the course while in the middle and the end of the course it was stable. Although, the estimated numbers of hidden variables are approximately similar for intervals 1 and 2, the clusters per each EDDA model are different in shapes and sizes for both intervals.

The highest BIC value was achieved at interval 2 with a value of -406158.9. As a result, the strongest and best model was obtained at interval 2 while weakness EDDA was acquired by interval 1. The latent engagement affects 10% of students at interval 4 while only 1% of students are affected by such engagement at interval 1. Table 5.20 give more detail in respect to volume, shape and orientation of clusters.it is clear that model type was EEI across intervals 3 and 5. It means clusters in both models are the same sizes and shapes but have a different orientation. BIC value of EDDA models are reported -978205.4 and -957081.5 respectively.

F1-Measures are suitable summary computations in the presence of imbalanced class data, as is the case in the current study. In general, unbalanced data can result in misclassification through biased selection of the majority class.The empirical results in table 5.22 shows that the highest F1-Measure is acquired by RF followed by Mlp, Rpart , and Gbm classifiers, with values of 0.991 **,**0.987 across interval 3. In intervals 1 and 4, the Rpart model achieved the highest F1-Measure, obtaining average values of 0.898 and 0.965. The EDDA occupied the second best classifier gaining range values of 0.840 -0.958. The EDDA classifier obtained a moderate to the similar range

of accuracy with F1-Measure values of 0.899-0.923 across interval 5 and interval 6. The lowest F1-Measure is seen at interval 3 resulting from the EDDA classifier, yielding a value of 0.778.

Considering the sensitivities outcomes, the RF classifier achieves the highest value at interval 3 with a value of 0.987. Although the good sensitivity is produced by the EDDA model compared with the other classifiers over intervals 2,3,4,5 and 6, with average values of 0.864 -0.941, the EDDA model acquired low sensitivity over interval 1. The RF model achieved the lowest sensitivity at interval 1 yielding values of 0.715.

All classifier models obtained a good specificity values over all intervals ranging over between 0.98 -0.81. However, Mlp, Nnet and EDDA models present slightly lower specificity with values of 0.786, 0.710 respectively over interval 1. Rpart obtains the lowest range of specificity at interval 1 with a true negative percentage of 66%. ROC is used in this study to select a decision threshold value for the trade-off between true and false positive rates across each time interval. Figure 5.27 lists the ROC curves respectively. Overall, the range of AUC values falls within 0. 99-0. 0.87 for intervals 2, to 6. Conversely, intervals 1 acquired the lowest AUC values, yielding 0.88-0.77 respectively.

Table 5.19 EDDA Model Result

Interval	No. Training Example	No. Estimate Parmeters	Log. likelihood	BIC	Model Type
Interval 1	1225	1130	-412321.23	-1080271.2	VII
Interval 2	2303	1023	-199119.4	-406158.9	EEV
Interval 3	4683	87	-488735.1	-978205.4	EEI
Interval 4	4473	462	-372407.1	-748697.7	EEE
Interval 5	4788	84	-478184.8	-957081.5	EEI
Interval 6	938	53	-317146	-634654.6	EII

Table 5.20 EDDA Model Type

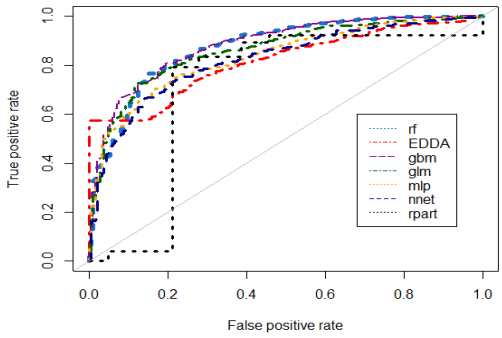
Model Type	Volume	Shape	Orientation
VII	Variable	Equal	NA
EEV	Equal	Equal	Variable
EEI	Equal	Equal	Coordinate axes
EEE	Equal	Equal	Equal
EII	Equal	Equal	NA

Table 5.21 EDDA Model Type According to Test Error

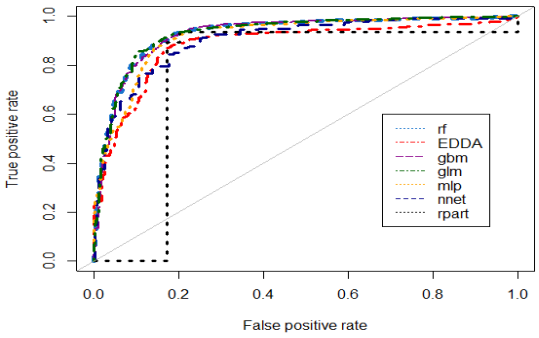
Interval	Model _Type	BIC	10-fold CV	Test error
Interval 1	VVV	-98735.7	0.5654	0.10732
	VII	-1080271.2	0.5314	0.10774
	EEI	-250721.9	0.2604	0.34805
	VEI	-232015.3	0.3069	0.44051
	EVI	-247679.1	0.2604	0.36631
	VVI	-228184.0	0.2832	0.39517
	EEE	-126804.7	0.2375	0.35041
Interval 2	EII	-1837857.0	0.4355	0.2067
	VII	-1832638.7	0.4355	0.2067
	EEI	-490586.8	0.1841	0.2144
	VEI	-457386.4	0.2353	0.2987
	EVI	-481823.7	0.2075	0.2589
	VVI	-443390.9	0.2179	0.2712
	EVE	-408269.0	0.1489	0.2038
	VEE	-380239.9	0.2205	0.2729
	VVE	-379471.8	0.2092	0.2530
	EEV	-406158.9	0.1406	0.1915
	VEV	-370181.3	0.1706	0.2015
	EVV	-403886.3	0.1636	0.2255
	VVV	-364961.3	0.1793	0.2044
Interval 3	EII	-3482765.8	0.5016	0.2392
	VII	-3423472.8	0.5443	0.2392
	EEI	-978205.4	0.1129	0.1391
	VEI	-907554.8	0.125	0.16950
Interval 4	EII	-3246627.7	0.4386	0.3688
	VII	-3241475.4	0.4379	0.3688
	EEI	-893190.1	0.0990	0.1276
	VEI	-819512.1	0.1041	0.1165
	EVI	-882401.3	0.0836	0.0772
	VVI	-788149.7	0.0561	0.0573
	EEE	-748697.7	0.0507	0.0544
	EVE	-738797.7	0.0672	0.0603
	VEE	-672752.1	0.1019	0.1329
	VVE	-663385.8	0.0666	0.0644
	EEV	-737424.3	0.0684	0.0761
	VEV	-640366.4	0.0771	0.0930
	EVV	-732692.3	0.0753	0.0889
	VVV	-625234.9	0.0751	0.0913
Interval 5	VII	-3389710.4	0.5524	0.4121
	EEI	-957081.5	0.1856	0.1902
	VEI	-760913.8	0.2280	0.2330
Interval 6	EII	-634654.6	0.3752	0.1247
	VII	-632052.7	0.3827	0.1247
	EEI	-144698.8	0.1567	0.1588

Table 5.22 Classification Performances for Dropout Prediction Model

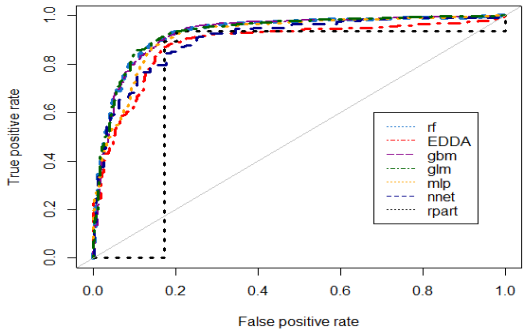
Intervals		Performance Metrics				
	Classifiers	ACC.	F-Meas.	Sens	Spec	AUC
Interval 1	EDDA	0.7462	0.840	0.750	0.710	0.775
	MLP	0.7473	0.838	0.742	0.786	0.821
	RF	0.7326	0.826	0.715	**0.874**	0.875
	Rpart	0.8281	**0.898**	**0.846**	0.664	0.810
	Gbm	**0.8814**	0.8397	0.736	0.841	0.881
	Glm	0.7485	0.8394	0.736	0.847	0.861
	Nnet	0.7473	0.8399	0.742	0.786	0.822
Interval 2	EDDA	0.8576	0.905	0.864	0.832	0.881
	Mlp	0.8576	0.905	0.856	0.861	0.891
	RF	0.8518	0.900	0.843	**0.883**	0.902
	Rpart	0.8793	0.881	**0.892**	0.827	0.871
	Gbm	0.8442	0.894	0.833	**0.883**	**0.926**
	Glm	**0.9155**	0.584	0.909	0.915	0.925
	Nnet	0.8412	0.894	0.849	0.810	0.907
Interval 3	EDDA	0.9354	0.778	0.894	0.941	0.961
	Mlp	0.9806	0.987	0.981	0.977	0.987
	RF	**0.9871**	**0.991**	**0.987**	**0.984**	**0.994**
	Rpart	0.9806	0.987	0.981	0.979	0.980
	Gbm	0.9803	0.987	0.979	0.983	0.993
	Glm	0.9776	0.985	0.976	0.980	0.987
	Nnet	0.9786	0.985	0.978	0.977	0.990
Interval 4	EDDA	0.9479	0.958	0.941	0.958	0.983
	Mlp	0.9451	0.955	0.936	0.958	0.980
	RF	0.9485	0.958	0.931	0.977	0.988
	Rpart	**0.9573**	**0.965**	**0.954**	0.961	**0.990**
	Gbm	0.9567	0.965	0.953	0.961	0.990
	Glm	0.9561	0.964	0.949	**0.966**	0.986
	Nnet	0.9438	0.955	0.949	0.934	0.968
Interval 5	EDDA	0.8788	0.899	0.921	0.819	0.906
	Mlp	0.8841	0.907	**0.967**	0.765	0.925
	RF	0.9093	0.923	0.934	0.874	0.929
	Rpart	0.8811	0.901	0.927	0.816	0.873
	Gbm	**0.9204**	**0.932**	0.931	**0.905**	**0.947**
	Glm	0.9093	0.922	0.922	0.891	0.942
	Nnet	0.911	0.923	0.915	0.905	0.945
Interval 6	EDDA	0.8743	0.923	0.869	0.906	0.913
	Mlp	0.9123	0.948	0.920	0.851	0.934
	RF	**0.9396**	**0.964**	**0.938**	**0.945**	**0.973**
	Rpart	0.9123	0.948	0.922	0.843	0.924
	Gbm	0.9045	0.943	0.909	0.867	0.937
	Glm	0.8772	0.916	0.871	0.914	0.916
	Nnet	0.8626	0.911	0.857	0.898	0.907



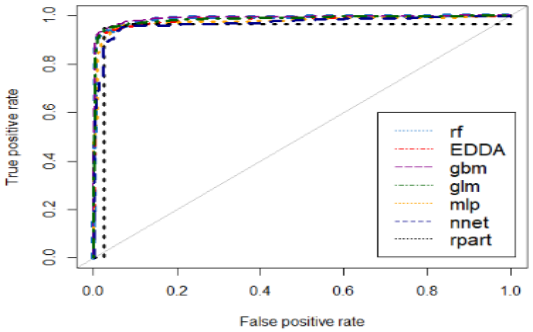
a) Interval 1



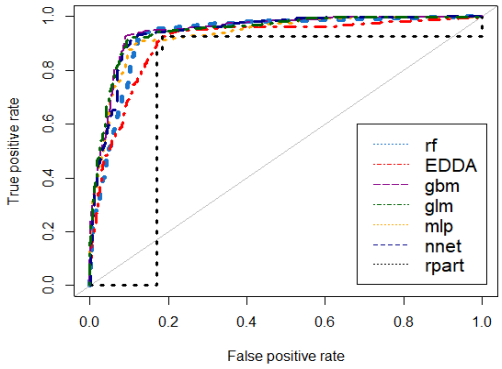
b)Interval 2



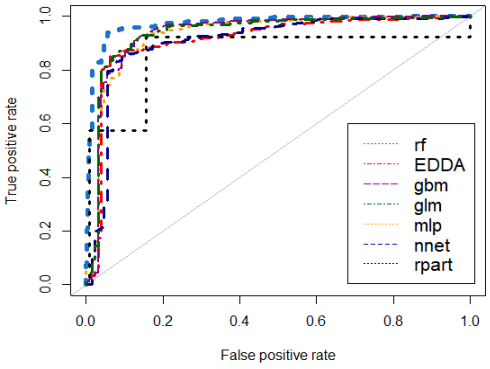
c)Intervel3



d)Interval 4



e)Interval 5



f)Interval 6

Figure 5.27 Roc Curve for Dropout Prediction Model

5.5 Results Discussion

The supervised machine learning utilized to predict learner outcome in Harvard courses. In order to evaluate which feature has a greater influence on the student outcome in an online setting, the FRE is used to ascertain the highest rank features. The results of the FRE illustrate that the behavioural features:”nchapters","nplay_video","ndays_act","nevents" and "explored" are the most important features. The machine learning results for all features and selected features are compared, whilst observing a number of similarities between them in terms of performance metrics. The results of the F1–Measure show that Gbm and RF obtain the highest performance for the first and second set of features respectively, whereas Rpart acquires the lowest performance over both sets of features.

The main reason that Gbm and RF achieved the highest performance is that both classifiers are ensemble machine learning models that are capable of reducing the variance and decreasing the bias of the dataset. In particular, the gradient boosting advantages a multiplicity of weak base classifiers to form a strong classifier with adaptive re-weighting of the data during the iterative learning process. The RF model creates many classification trees during the learning process; each node of the tree randomly selects the most important features, and will produce its own classification result. The RF classifier uses the majority of votes to predict the final result (Bharathidason and Venkataeswaran, 2014).

Nevertheless, the Gbm acquired good performance in the second set of features. The RF achieved the best performance, since RF is capable of ascertaining the best behavioral features that discern the positive and negative class through adopting a voting mechanism (Ganjisaffar, Caruana and Lopes, 2011).

Although both sets of features approximately acquire similar accuracy, there is a considerable difference in their computational complexity. The fastest models were LG and the slowest models were RF. The RF was the lowest algorithm - the time required to train RF in the second set of features is less than that of the second fastest model in the first set of features.

The unsupervised machine learning is used to examine the whether students who are different in their engagement types, they are similar in their achievement. The results

show that students have the similar characteristic although, they engage in different activities

The descriptive statistical analysis is utilized to compare the successful learning with the failing in respect to geographical location. The result shows that Europe and Asia rank the highest in terms of learner success rates, whereas the highest ratio of failure groups is distributed across various regions. Due to the availability of advanced technological integration of universities around the world, students in any region may interact with courses through the internet. However, language is considered a barrier that could affect student performance. The course is delivered in English; hence, second language learners might be less motivated to exchange knowledge with other participants. Geographical location may not be considered a crucial factor that could significantly affect student performance in online courses. Student performance could be more heavily associated with the engagement levels of students. Specifically, it is more relevant to the number of chapters read by students.

A Chi-squared test was applied to investigate the presence of a significant difference in the education level of learners between the success and failure groups. The results reveal that the distribution of students according to their educational levels is significantly different. The lower qualified students show the largest percentage of successful completion learners in some courses, due to how these courses are designed to serve students who had less knowledge in a particular topic. An analysis of such descriptive statistics could be conducted by enhancing learning resources through the early identification of failure students

As discussed earlier, the target is to identify the at-risk students and understands the reason behind student dropout from various angles. A trajectories analysis is capable of investigating how the level of student engagement and motivation could influence at-risk students. With a trajectories analysis, the failure factor could be inspected by temporally tracking the students' engagement levels across courses. In addition, deterioration of student motivation could be detected over time by tracing learner intervention motivation prior to students deciding to quit the course.

The traditional statistical analysis has a limited ability to perform a trajectories analysis due to the fact that the statistical analysis is not designed to capture arbitrary non-linear patterns. As a result, such procedures require an expert assumption about

the form of the data before analysis, relying on the notion of a super-population in the form, which must be chosen on a prior basis (Chu *et al.*, 2007). Moreover, in the present context, hypothesis tests and inference procedures are not conducive to an analysis of the behavioral intervention over a period since the data is not guaranteed to satisfy classical statistical constraints. To investigate whether or not factors such as learner motivation and level of engagement could influence at-risk students in danger of withdrawal in the next course, a further analysis was therefore performed using machine learning techniques that do not rely on classical assumptions.

The Harvard dataset did not explicitly define the students' motivational label. Therefore, LA is used to derive learners' motivations, based on IM theory. To examine how motivation trajectories influences at-risk students, the LA is utilized to categorise the taxonomy of learners, based on the IM concept. Students' motivation can be broadly divided into three statuses, "amotivation", "extrinsic" and "intrinsic".

The result of ML shows all classifiers acquire good performance, the RF achieve the best F1-Measure over three classes obtaining a value of 0.88, 0.76, 0.75 respectively. The lowest F1-Measure is reported for class "extrinsic" with a value of 0.62 in respect to the Nnet mode.

At-risk students are defined in terms of two factors, namely performance and motivational status. Two temporal predictive models are designed with the objective of helping educators provide timely intervention and support for at-risk students.

Only demographic features are considered in the first temporal predictive model. These features include students' date of birth, educational level, sex, geographical location and GPA (grade), to evaluate how students' performance during previous courses could affect at-risk students in the following course. All the demographic features are accounted in the first set of the features, whereas the students grade of the previous courses are excluded in the second set of features.

The result of ML over both set of features are very low. The F1-Measure shown for all classifiers less than 72%. The results reveal that sensitivities and specificities are relatively low for all classifiers. This is expected due the fact that the demographic features are inadequate to differentiate withdrawal students from non-withdrawal students. The empirical result of machine learning demonstrate that failure is unlikely to be the key reason prevents students from completing the following courses.

The second temporal predictive model uses the students' behavioural features in the following course in addition to students' motivational status during the previous course. A correlation analysis is applied to measure the association link between students' motivational categories and at-risk students. The results show that students' motivational statuses are relevant to at-risk students. The result shows significant improvements in the second temporal predictive model. This is because student behavioural features and motivational categories are better reflected in the at-risk student than the demographic features.

The best F1-Measure obtained by Nnet with a value of 0.95 follow by RF and Glm. The lowest F1-Measure was gain by Gbm and Rpart with values of 0.865and 0.836 receptively. Sensitivities (withdrawal students) are slightly higher than specificities (non-withdrawal students). The finding reveals that all classifiers acquire good performance. The simulation result demonstrates that Feedforward Neural Network with a single hidden layer achieves the highest performance. This was attributed to the learning algorithm. The Broyden-Fletcher-Goldfarh-Shanno (BFGS) optimisation algorithm was utilized to train the neural network with a single layer while the backpropagation (BP) is used to train a neural network with two hidden layers. The BFGS is optimisation algorithms is used for solve complex nonlinear problem. The BFGS has an advantage over BP as its capacity to find the second derivative of the activation function. More specifically, the BFGS change the gain of activation function adaptively for each neuron. The gradient descent compute error considers the weight and gain values as a consequence, the search direction of the learning algorithm is improved and achieve better convergence rate resulting to enhance the accuracy performance of neural network and better convergence rates.

Again, the database is balanced regarding the second at-risk students' model. However, the number of withdrawal students' records is slightly higher than that of non-withdrawal students' records. This could have an influence on learning of classifier, since it is easier for the classifier to predict the positive class (withdrawal student). In this study, sensitivities are more of a priority than the specificities, since it is considered worthwhile to predict withdrawal students, rather than non-withdrawal.

The machine learning results reveal that motivation trajectories is a valuable factor for estimating dropout learners in the online course. It can be used as significant indicator of at-risk students. The proposed temporal predictive models could help decision makers identify at-risk students in the early stages of their studies. Moreover, educators may gain a richer understanding by considering students' motivational status as the main reason behind learners' withdrawal within the online course setting.

With regards to OULAD dataset two student performance predictive models are designed consider regression and classification analysis. The results of predicting students' assessments grades model show that the best RMSE and R^2were acquired over interval 3 while the worst RMSE and R^2were given by interval 5 and 6. This could be attributed to a number of students record over interval 3 highest than another interval as a result; the algorithm can learn more and the model will fit well. In contrast, the lowest number of training example shown was over interval 6.

The final student performance predictive model revealed ideal sensitivities and specificities for all classifiers Although, the sensitivities and specificities are balanced for all classifiers over class "Withdrawn" and "Pass", the specificities are higher than corresponding sensitivities for class "Fail". This is because the database is skewed in favour of choosing the majority classes "Withdrawn" and "Pass". In this case, predicting withdrawal students is more of a priority than predicting success and fail students as it is worthwhile to predict students who withdraw from the entire course rather than students who stay engaged with the course.

The primary reason the machine learning models obtain higher performance in classification than regression is related to the type of features sets. As such, in classification analysis, the static behavioral features in conjunction with the temporal features and demographic features are used as the input variables in the prediction of the final student performance model while only dynamic behavioral features are employed to estimate students' assessments grades. As the correlation analysis demonstrates there is a weak association between the students' performance with demographic features accordingly, the demographic features cannot be accountable for the sufficient features. The temporal features that include the date of student registration and deregistration in an online course are robust predictor features that adequately affect student achievement. It could be impossible to combine the temporal

features with behavioral features with respect to regression analysis as the database includes student temporal information for the entire period.

To examine how the levels of student engagement and latent engagement influence at-risk students, the temporal prediction models have been introduced in OULAD. The model could help educator detect the disengaged students since the beginning of the course. The simulation results of machine learning demonstrated that engagement has a significant impact on students' withdrawal within the MOOCS environment. The estimation of the student engagement status could help to identify at-risk students early.

The results reveal that the EDDA model is capable of inferring latent characteristics of the students. The number of latent variables decreased about 90% at the end of the course therefore; the student emotional status could be more stable at the end of the course. The number of the hidden variables are few compared with the number of examples across all intervals. As a result, latent engagement cannot be considered the vital factor that distinguishes the at-risk student and not at-risk students. The simulation results have shown that all classifiers achieve the high performance over all intervals. In contrast, the RF and EDDA show low performance at intervals 1 and 3 respectively.

The empirical results reveal that the EDDA models in the middle and the end acquired best sensitivity of course while EDDA achieves the lowest sensitivity in interval 1. The number of unobserved variables was approximate 50% higher at interval 1 than interval 4. This implication of increasing the number of hidden variable could affect the true positive rate within a particular interval. As a result, latent variables could lead to misclassification problems. Although, the proportion of latent variables at interval 2 is slightly less than interval.

5.6 Chapter summary

This chapter has prsensted the simulation results of three experiments.The results are demonstrated that supervised machine learning is capable to traceing the students' activities over time. The results of students performance prediction models have achieved a high value of accuracies. The AUC acquired approximate values of 0.88-0.99 across two datasets. The finding of students' assessments grades predication

model gives the lower result than the final student performance model, in such model the temporal features has been excluded.

The results of descriptive statistics show the vast difference between failing and successful learners in term of engagement level. As such, The successful student read material three times more than failing students, however, in some courses the failing learning watched videos twice time than their pairs. The result of Chi-squared shows failing and successful students different with respect to their educational levels. More specific, the successful completion learners reported as the low qualified learners.

The chapter also explains in details result of dropout Prediction model, the finding indicated that student performance trajectories could not be considered as a vital factor that impacts student performance. The AUC value was shown 0.50 for dropout model, however, the finding shows that students motivational statues is an important factor that significantly impacts the at-risk student. The finding state that around 70 %of demotivated student in the previous course re-engage in the following course. The finding has recommended considering the motivation trajectories as one of the predictors for at-risk students.

Nevertheless, the results status that students' failure in the prior course cannot influence the student decision to participate in the future courses. However, a student's failure in previous assessment can be the main reason for preventing the students from engaging in the following assessments within a single course.

The result of dropout Prediction Model based also posit that engagement levels can be used to detect at-risk students. The engagement levels of disengaged students are shown lowly. Due the simulation results of traditional machine learning are quite similar to conventional machine learning models therefore, latent engagement cannot be consider the direct reason of student withdrawal.

Chapter 6: Conclusion and and Future Work

6.1 Conclusion

A comprehensive understanding of the reasons behind student withdrawal in the online environment has been provided in this thesis. More specifically, this project focuses on those students who engage in the course through participation in digital learning materials and then drop out for a particular reason. This project takes an initial step towards delivering effective intervention support for the at-risk student. In this work, lack of motivation, low engagement level and latent engagement has been suggested as significant factors that hinder students from completing the courses.

In this thesis, the proposed motivation predictive model successfully achieves the first objective of the prediction of students' motivational status. LA is utilised to categories the learners' motivation status according to IM theory. The students' motivation is classified into three categories; ML is used to estimate the students' motivational category. The resulting findings of six classifiers reveal that adoption of IM theory to predict student motivation in future is highly successful and promising. The work in this project is different from that carried out in the literature. In their work, the students have been categorized as either motivated or unmotivated. In our work, the interpretation of motivation is based on psychological theories.

As the second objective, two dropout prediction models are designed taking into consideration the Harvard dataset. The models can help the educator to deliver early intervention support for at-risk students. The findings of ML show that low student performance in the previous course cannot be considered a crucial factor that could influence the participation of students in a subsequent course. However, the students' motivational status is a valuable indicator in identifying dropout learners in the online course. In addition, the result demonstrates that most intrinsically and extrinsically motivated students in the previous course were motivated in the following course. On the other hand, unmotivated students in a prior course were engaged in following course. The findings could assist the educator to monitor changes in students' motivational status. Hence, the educator can easily identify those students who need support and provide additional learning materials.

The third objective is also achieved by investigating the dynamic link between engagement in a sequence of assessments within a single course and withdrawal rate. Latent engagement has semantic meaning but cannot be directly observed in the dataset. It can be measured from students' activities. Temporal prediction models, which have been introduced, consider the OULAD dataset. The Mixture model is utilised to infer Latent variables. The findings of ML indicate that few numbers of students' withdrawal from the course are due to latent engagement. Consequently, the latent variable cannot significantly discriminate between the dropout and non-dropout students. In addition, the results reveal that the engagement level is a crucial indicator that is directly related to dropout.

As the fourth objective, the effect of behavioural engagement on students' performance is investigated through tracking students' activities. Student performance predictive models have been proposed in our work over two datasets. The input predictors consist of behaviour features followed by the date of students' registration and deregistration from the courses in addition to demographic features. The machine learning results show that the students' performance in a particular assignment relies on students' mark in the previous assignment within single Courses.

Although machine learning results demonstrate that students' engagement level is a significant predictor of student performance in online settings over both datasets, the engagement style might be more relevant. As such, the results show the students' performance prediction model in the Harvard dataset achieves better accuracy than the prediction model in OULAD dataset due to, the exclusion of videos lectures in OULAD dataset. This is due to video lectures providing interactive means of learning that could help the student to process information more easily than conventional teaching materials. Additionally, the videos are more suitable for auditory and visual learners. The students have the option to listen, watch or read the learning material. The findings suggest using an educational YouTube channel to deliver the online courses. The findings also demonstrate that the date of student deregistration from the course is a valuable predictor that is significantly correlated to student performance across two datasets.

With the students' assessment grades prediction model, the data does not provide the last date of students' activity prior to undertaken assessments. The findings' are

recommended taking into account the temporal features in predicting subsequent assessments grades.

In conclusion, various factors influencing at-risk students have been evaluated over Harvard and OULAD datasets. Both dataset results demonstrate that clickstream features can be reliable predictors, which are remarkably relevant to the prediction of students' outcomes and subsequent assessment grades and for the estimation of students' dropout. The temporal features are also important attributes. For instance, the number of days that students interact with the course can be strongly used to identify the students who are in danger of withdrawal.

This project provides the significant contribution towards the early identification of at-risk student in MOOCs. The project tackles the withdrawal problem in the online setting and provides some valuable insights and recommendations that might lead to the fully automated intervention of the decision maker. The educators can handle the dropout issue by providing more teaching assistants or deliver additional teaching material in a personalised way to at-risk students.

6.2 Future Work

In the current study, many factors have been evaluated to flag and detect at-risk students in the online context. It is possible to suggest other approaches that could improve proposed predictive models. In the following paragraphs, the future research work on enhancing this research project is introduced.

Emotion is a critical factor that has a significant impact on modulating the learning process and offers insight into monitoring the students' attitudes toward MOOCs. Many psychologists' and neuroscientists' findings have observed the influence of both positive and negative aspects of student cognitive engagement on withdrawal rate (Wen, Yang and Rosé, 2014). The sentiment analysis provides more information, which is valuable for tracking students' behavior and monitoring students' opinions. Although, many studies pay more attention to inferring the student negative emotional effects such as boredom, frustration and fatigue that would lead a student to withdraw from the course. These studies rely on post-course surveys to obtain students' feedback on the quality of the course; few students actually respond and fill in surveys.

Sentiment analysis can be utilized to discover the students' opinions toward e MOOCs. As such, the post-forum can be used to capture students' attitudes and flag those who tend to drop out from the course. Different emotional status can be inferred from discussion forums such as frustration, fatigue, and boredom. This emotional status provides the students with motivational encouragement and stimulation to facilitate an interactive learning environment; including feedback modalities such as visually oriented hints. Additionally, the instructors would be able to recognize the reasons underpinning student withdrawal from the course in a more precise manner.

Harvard dataset and OULAD dataset did not include the text of discussion forums. The datasets only contain information that was relevant to the number of clicks that the student performs on the forum. Therefore, a web-based system could be designed to capture the student sentiments in an online course through the post forum. The system can follow an emotional coping strategy for the tracking of the users' emotional status. The emotional coping strategy is a psychological strategy that aims to reduce the stress, minimize emotional arousal and promote the effect of progress in a specific task. The system will allow students to write a short sentence to express their emotional statuses. The LA can be employed to infer the students' emotional statuses from the written text for example; LA could be tracking events that could cause students' depression or frustration. Furthermore, the instructor would encourage students to release their negative emotions in such things as interactive video games and deliver immediate support.

Machine learning is an effective approach that has seen widespread use in the online context for the purpose of facilitating automated detection of at-risk students. Although, much more work has adopted the machine learning using students' activity data in predicting of at-risk students, few tasks incorporate the student sentiments in the detecting of at-risk students. To this end, the various techniques of the machine-learning model can be used to automatically predict students who are in danger of dropout from the course considering sentiments. For instance, the RNN can be used to measure the students' sentiments and discover the impact on at-risk students through inferring the sequences of temporal events.

In the future more advanced machine learning models can be used, deep machine learning can be considered given its capability to represent the complex representation of students behaviour features without the need to engineer features and could enhance

the accuracy of the dropout prediction model (Gan *et al.*, 2015). Various models of deep learning can be used such as, Deep belief neural network, Deep relational learning, Convolutional neural network and Stacked Autoencoders (Gan *et al.*, 2015)(Wang, Shi and Yeung, 2017).

Because a large volume of data can be captured from MOOCs platform, deep learning might be more suitable for high dimensional data. Features could be extracted from large data without the need for human intervention. Also, the tuning and selecting of the optimal parameters without any human help, as a result, a more robust predictive model can be obtained by adopting deep learning(Gan *et al.*, 2015).

References

Adams, C. *et al.* (2014) 'A phenomenology of learning large : the tutorial sphere of xMOOC video lectures', *Distance Education*. Routledge, 35(2), pp. 203–217. doi: 10.1080/01587919.2014.917701.

Agresti, A. (2007) *An Introduction to Categorical Data Analysis, Statistics*. New York (USA): John Wiley & Sons Inc. doi: 10.1002/0471249688.

Ahmadi, H., Mottaghitalaband, M. and Nariman-Zadeh, N. (2007) 'Group method of data handling-type neural network prediction of broiler performance based on dietary metabolizable energy, methionine, and lysine', *Journal of Applied Poultry Research*, 16(4), pp. 494–501. doi: 10.3382/japr.2006-00074.

Aleksander, M. de G. *et al.* (2009) 'A brief introduction to Weightless Neural Systems', (April), pp. 2099–305.

Alias, U. F., Ahmad, N. B. and Hasan, S. (2015) 'Student behavior analysis using self-organizing map clustering technique', *ARPN Journal of Engineering and Applied Sciences*, 10(23), pp. 17987–17995. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84953426340&partnerID=40&md5=1e2f0dca13a9ebfbb6787263e3e5796b.

Altrabsheh, N., Cocea, M. and Fallahkhair, S. (2016) 'Predicting students ' emotions using machine learning techniques', in. Portsmouth: International Conference on Artificial Intelligence in Education, pp. 537–540.

Asif, R. *et al.* (2017) 'Analyzing undergraduate students' performance using educational data mining', *Computers & Education*. Elsevier Ltd, 113, pp. 177–194. doi: 10.1016/j.compedu.2017.05.007.

Atakulreka, A. and Sutivong, D. (2007) 'Avoiding Local Minima in Feedforward Neural Networks by Simultaneous Learning', *Lecture Notes in Artificial Intelligence*, 4830, pp. 100–109. doi: 10.1007/978-3-540-76928-6_12.

Baek, J. and Shore, J. (2016) 'Promoting Student Engagement in MOOCs', in *Proceedings of the Third ACM Conference on Learning @ Scale*. Edinburgh, Scotland, UK, pp. 293–296. doi: 10.1145/2876034.2893437.

Baker, R. S. J. D. and Siemens, G. (2014) 'Educational Data Mining and Learning Analytics', in Rupp, André A., Leighton, J. P. (ed.) *Cambridge Handbook of the Learning Sciences*. doi: 10.1007/978-1-4614-3305-7.

Balakrishnan, G. and 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.*, pp. 1–15. Available at: http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.pdf.

Banfield, R. E. *et al.* (2007) 'A comparison of decision tree ensemble creation techniques', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), pp. 173–180. doi: 10.1109/TPAMI.2007.250609.

Barak, M., Watted, A. and Haick, H. (2016) 'Motivation to learn in massive open online courses: Examining aspects of language and social engagement', *Computers*

& Education. Elsevier Ltd, 94, pp. 49–60. doi: 10.1016/j.compedu.2015.11.010.

de Barba, P. G., Kennedy, G. E. and Ainley, M. D. (2016) 'The role of students' motivation and participation in predicting performance in a MOOC', *Journal of Computer Assisted Learning*, 32(3), pp. 218–231. doi: 10.1111/jcal.12130.

Baxt, G. (1995) 'Application of artificial neural networks to clinical medicine', pp. 1135–1138.

Beal, C., Mitra, S. and Cohen, P. R. (2007) 'Modeling learning patterns of students with a tutoring system using Hidden Markov Models', in *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, pp. 238–245.

Bekkar, M., Djemaa, H. K. and Alitouche, T. A. (2013) 'Evaluation measures for models assessment over imbalanced data sets', *Journal of Information Engineering and Applications*, 3(10), pp. 27–38. Available at: http://www.iiste.org/Journals/index.php/JIEA/article/view/7633.

Bensmail, H. *et al.* (1996) 'Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition', *Journal of the American statistical Association*, 91(436), pp. 1743–1748.

Bharathidason, S. and Venkataeswaran, J. C. (2014) 'Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees', *International Journal of Computer Applications*, 101(13), pp. 26–30.

Biau, G. and Scornet, E. (2015) 'A Random Forest Guided Tour', 25(2), pp. 197–227. doi: 10.1007/s11749-016-0481-7.

Birkett, A. N. and Goubran, R. A. (1995) 'ACOUSTIC ECHO CANCELLATION USING NLMS-NEURAL NETWORK STRUCTURES', in *In Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., International Conference on IEEE*, pp. 2035–3038.

del Blanco, A. *et al.* (2013) 'E-Learning Standards and Learning Analytics', *2013 Ieee Global Engineering Education Conference*, pp. 1255–1261. doi: 10.1109/EduCon.2013.6530268.

Botički, I., Budišćak, I. and Hoić-Božić, N. (2008) 'Module for Online Assessment in Ahyco LEARNING MANAGEMENT SYSTEM', 38(2), pp. 115–131.

Bouguila, N. and Ziou, D. (2007) '[2007][10] High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length.pdf', 29(10), pp. 1716–1731.

Bouveyron, C. (2014) 'Adaptive Mixture Discriminant Analysis for Supervised Learning with Unobserved Classes', *Journal of Classification*, 31(1), pp. 49–84. doi: 10.1007/s00357-014-9147-x.

Brinton, C. G. *et al.* (2014) 'Learning about Social Learning in MOOCs : From Statistical Analysis to Generative Model', 7(4), pp. 346–359.

Buabeng-Andoh, C. (2012) 'Factors influencing teachers ' adoption and integration of information and communication technology into teaching: A review of the literature', *International Journal of Education and Development using Information and Communication Technology*, 8(1), pp. 136–155.

Bullinaria, J. A. (2015) ‘Learning in Multi-Layer Perceptrons, Back-Propagation’, *Neural Computation : Lecture 7*, pp. 1–16. Available at: https://www.cs.bham.ac.uk/~jxb/inc.html.

Chandrashekar, G. and Sahin, F. (2014) ‘A survey on feature selection methods’, *Computers and Electrical Engineering*. Elsevier Ltd, 40(1), pp. 16–28. doi: 10.1016/j.compeleceng.2013.11.024.

Chaplot, D. S., Rhim, E. and Kim, J. (2015) ‘Predicting student attrition in MOOCs using sentiment analysis and neural networks’, *Workshops at the 17th International Conference on Artificial Intelligence in Education, AIED-WS 2015*, 1432, pp. 7–12. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84944342705&partnerID=40&md5=76eb43de1aa0507d73d876d711602306.

Cho, M.-H. and Heron, M. L. (2015) ‘Self-regulated learning: the role of motivation, emotion, and use of learning strategies in students’ learning experiences in a self-paced online mathematics course’, *Distance Education*. Routledge, 36(1), pp. 80–99. doi: 10.1080/01587919.2015.1019963.

Chu, C.-T. *et al.* (2007) ‘Map-Reduce for Machine Learning on Multicore’, in *Advances in Neural Information Processing Systems 19*. Canada: NIPS’06 Proceedings of the 19th International Conference on Neural Information Processing Systems, pp. 281–288. doi: 10.1234/12345678.

Clow, D. (2012) ‘The learning analytics cycle’, in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK ’12*. Uk, p. 134. doi: 10.1145/2330601.2330636.

Clow, D. (2013) ‘MOOCs and the funnel of participation’, *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK ’13*, p. 185. doi: 10.1145/2460296.2460332.

Cobos, R., Wilde, A. and Zaluska, E. (2017) ‘Predicting attrition from massive open online courses in FutureLearn and edX’, *CEUR Workshop Proceedings*, 1967(March 2017), pp. 74–93.

Coffrin, C. *et al.* (2014) ‘Visualizing patterns of student engagement and performance in MOOCs’, *Proceedins of the Fourth International Conference on Learning Analytics And Knowledge - LAK ’14*, pp. 83–92. doi: 10.1145/2567574.2567586.

Czado, C. and Tu, M. (2004) ‘Presentation : Introduction to GLM’s. Introduction to GLM’s’, *Presentation*, pp. 1–30.

Czepiel, S. A. (2012) ‘Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation’, *Class Notes*, pp. 1–23. Available at: papers3://publication/uuid/4E1E1B7E-9CAC-4570-8949-E96B51D9C91D.

D Seaton, J Reich, S Nesterko, T Mullaney, J. W. (2014) *6.00x Introduction to Computer Science and Programming MITx on edX – 2012 Fall*. New York (USA).

Dalipi, F., Imran, A. S. and Kastrati, Z. (2018) ‘MOOC Dropout Prediction Using Machine Learning Techniques : Review and Research Challenges’, in. Sweden: 8 IEEE Global Engineering Education Conference (EDUCON), pp. 1013–1020.

Daniels, A. and Mascini, J. (1943) ‘A LOGICAL CALCULUS OF THE IDEAS

IMMANENT IN NERVOUS ACTIVITY', *the bulletin of mathematical biophysics*, 5(4), pp. 15–13. doi: 10.1179/1743275812Y.0000000008.

De'ath, G. (2007) 'Boosted regression trees for ecological modeling and prediction', *Ecology*, 88(1), pp. 243–251. doi: 10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2.

Dejong, G. (1986) 'Explanation-Based Learning : An Alternative View', 1(2), pp. 145–176.

Demšar, J. (2006) 'Statistical Comparisons of Classifiers over Multiple Data Sets', *Journal of Machine Learning Research*, 7, pp. 1–30. doi: 10.1016/j.jecp.2010.03.005.

Dimitrios, B. *et al.* (2013) 'Traditional Teaching Methods Vs. Teaching Through the Application of Information and Communication Technologies in the Accounting Field: Quo Vadis?', *European Scientific Journal*, 99(2828), pp. 1857–7881.

Domingos, P. (2012) 'A few useful things to know about machine learning', *Communications of the ACM*, 55(10), pp. 78–87. doi: 10.1145/2347736.2347755.

Doria, N. S. F., Freire, E. O. and Basilio, J. C. (2013) 'An algorithm inspired by the deterministic annealing approach to avoid local minima in artificial potential fields', in *16th International Conference on Advanced Robotics, ICAR .*, pp. 0–5. doi: 10.1109/ICAR.2013.6766480.

Dutton, H., Cheong, P., & Park, N. (2004) 'The Social Shaping of a Virtual Learning Environment: The Case of a University-Wide Course Management System.', *Electronic Journal of e-learning*, 2(2), pp. 1–12. Available at: http://www.inf.ufes.br/~cvnascimento/artigos/issue1-art3-dutton-cheong-park.pdf.

Fei, M. and Yeung, D.-Y. (2015) 'Temporal Models for Predicting Student Dropout in Massive Open Online Courses', *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 256–263. doi: 10.1109/ICDMW.2015.174.

Ferguson, R. and Clow, D. (2015) 'Examining engagement', *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*, pp. 51–58. doi: 10.1145/2723576.2723606.

Fernández, A. *et al.* (2018) 'SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary', *Journal of Artificial Intelligence Research*, 61, pp. 863–905.

Ferré, L. (1995) 'Selection of components in principal component analysis: A comparison of methods', *Computational Statistics & Data Analysis*, 19(6), pp. 669–682. doi: 10.1016/0167-9473(94)00020-J.

Figueiredo, M. A. T. and Jain, A. K. (2002) 'Unsupervised learning of finite mixture models', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3), pp. 381–396. doi: 10.1109/34.990138.

Fop, M., Murphy, T. B. and Raftery, A. E. (2016) 'mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models', 8(1), pp. 1–29.

Fraley, C. and Raftery, a E. (2002) 'Model-based clustering, discriminant analysis,

and density estimation', *Journal of the American Statistical Association*, 97(458), pp. 611–631. doi: 10.1198/016214502760047131.

Fraley, C. and Raftery, A. E. (2007) 'Bayesian regularization for normal mixture estimation and model-based clustering', *Journal of Classification*, 24(2), pp. 155–181. doi: 10.1007/s00357-007-0004-5.

Friedman, J. H. (2002) 'Stochastic gradient boosting', *Computational Statistics and Data Analysis*, 38(4), pp. 367–378. doi: 10.1016/S0167-9473(01)00065-2.

Friedman, T. H. R. T. J. (2001) *The Elements of Statistical Learning Data*. New York, NY, USA: Springer series in statistic.

Gallego Arrufat, M. J., Gamiz Sanchez, V. and Gutierrez Santiuste, E. (2015) 'Trends in Assessment in Massive Open Online Courses', *Educacion Xx1*, 18(2), pp. 77–96. doi: 10.5944/educXX1.12935.

Gan, Z. *et al.* (2015) 'Learning Deep Sigmoid Belief Networks with Data Augmentation', *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 38, pp. 268–276. Available at: http://jmlr.org/proceedings/papers/v38/gan15.html.

Ganjisaffar, Y., Caruana, R. and Lopes, C. V. (2011) 'Bagging gradient-boosted trees for high precision, low variance ranking models', *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, (c), p. 85. doi: 10.1145/2009916.2009932.

Geigle, C. (2017) 'Modeling MOOC Student Behavior With Two-Layer Hidden Markov Models', in *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pp. 205–208.

Gharehchopogh, F. S. and Khalifelu, Z. A. (2011) 'Using Intelligent Tutoring Systems in Instruction and Education', *International Journal of Innovation Management and Technology*, 2(5), pp. 250–254. Available at: http://search.proquest.com.ezproxy.lancs.ac.uk/docview/1441451187?rfr_id=info%3Axri%2Fsid%3Aprimo.

Ghaznavi, M. R., Keikha, A. and Yaghoubi, N.-M. (2011) 'The Impact of Information and Communication Technology (ICT) on Educational Improvement', in *International Education Studies*, pp. 116–125. doi: 10.5539/ies.v4n2p116.

Glynn, S. M. *et al.* (2011) 'Science motivation questionnaire II: Validation with science majors and nonscience majors', *Journal of Research in Science Teaching*, 48(10), pp. 1159–1176. doi: 10.1002/tea.20442.

Goldenberg, A., Kubica, J. and Komarek, P. (2003) 'A Comparison of Statistical and Machine Learning Algorithms on the Task of Link Completion', in Analysis, I. K. W. on L. and Behavior., for D. C. (eds), p. 8.

Granitto, P. M. *et al.* (2006) 'Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products', 83, pp. 83–90. doi: 10.1016/j.chemolab.2006.01.007.

Gribble, P. (2014) 'Analysis of Covariance (ANCOVA) Introduction to Statistics Using R (Psychology 9041B)', pp. 1–14.

Guttag, J. (2014) *6.00x Introduction to Computer Science and Programming MITx on edX – 2013 Spring*. New York(USA).

Guyon, I. and Elisseeff, A. (2003) 'An Introduction to Variable and Feature Selection', *Journal of Machine Learning Research (JMLR)*, 3(3), pp. 1157–1182. doi: 10.1016/j.aca.2011.07.027.

Ham, J. *et al.* (2005) 'Investigation of the random forest framework for classification of hyperspectral data', *IEEE Transactions on Geoscience and Remote Sensing*, 43(3), pp. 492–501. doi: 10.1109/TGRS.2004.842481.

Hanley, J. A. and McNeil, B. J. (1982) 'The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve', *Radiology*, 143(1), pp. 29–36. doi: 10.1148/radiology.143.1.7063747.

Hao Cen, Kenneth Koedinger, B. J. and Carnegie (2006) 'Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement', in *In International Conference on Intelligent Tutoring Systems*. Springer, Berlin, Heidelberg, pp. 164–175. doi: 10.3389/fimmu.2017.00318.

Hargreaves, D. H. (1994) 'The new professionalism: The synthesis of professional and institutional development', *Teaching and Teacher Education*, 10(4), pp. 423–438. doi: 10.1016/0742-051X(94)90023-X.

Hartnett, M., George, A. S. and Zealand, N. (2011) 'Examining motivation in online distance learning environments: Complex, multifaceted and situation-dependent', *The International Review of Research in Open and Distributed Learning*, 12(6)(2005), pp. 20–38. Available at: http://0-web.b.ebscohost.com.catalog.lib.cmich.edu/ehost/pdfviewer/pdfviewer?sid=17e46517-fc6f-4292-81b2-8094b5523a21%40sessionmgr112&vid=1&hid=105.

He, J., Bailey, J. and Rubinstein, B. I. P. (2015) 'Identifying At-Risk Students in Massive Open Online Courses', *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 1749–1755.

Hew, K. F. (2015) 'Towards a Model of Engaging Online Students: Lessons from MOOCs and Four Policy Documents', *International Journal of Information and Education Technology*, 5(6), pp. 425–431. doi: 10.7763/IJIET.2015.V5.543.

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(2), pp. 320–341. doi: 10.1111/bjet.12235.

Hew, K. F. and Cheung, W. S. (2014) 'Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges', *Educational Research Review*, 12, pp. 45–58. doi: 10.1016/j.edurev.2014.05.001.

Ho, A. D. *et al.* (2014) 'HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013', *SSRN Electronic Journal*, (1), pp. 1–33. doi: 10.2139/ssrn.2381263.

Ho, A. D. *et al.* (2015) 'HarvardX and MITx : Two Years of Open Online Courses Fall 2012-Summer 2014', *SSRN Electronic Journal*, (10), pp. 1–37. doi: 10.2139/ssrn.2586847.

Hofer, J. and 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(9), pp. 1147–1158. doi: 10.1177/0146167211408329.

Hone, K. S. and El Said, G. R. (2016) 'Exploring the factors affecting MOOC retention: A survey study', *Computers and Education*. Elsevier Ltd, 98, pp. 157–168. doi: 10.1016/j.compedu.2016.03.016.

Hosseini, H. G., Luo, D. and Reynolds, K. J. (2006) 'The comparison of different feed forward neural network architectures for ECG signal diagnosis', *Medical Engineering and Physics*, 28(4), pp. 372–378. doi: 10.1016/j.medengphy.2005.06.006.

Huang, B. and Hew, K. F. (2016) 'Measuring Learners' Motivation Level in Massive Open Online Courses', *International Journal of Information and Education Technology*, 6(10), pp. 759–764. doi: 10.7763/IJIET.2016.V6.788.

Huang, S. and Fang, N. (2010) 'Ac 2010-190: Regression Models for Predicting Student Academic Performance in an Engineering Dynamics Course', in *Age*. In American Society for Engineering Education. American Society for Engineering Education, p. 1.

Hung, J.-L. *et al.* (2017) 'Identifying At-Risk Students for Early Interventions—A Time-Series Clustering Approach', *IEEE Transactions on Emerging Topics in Computing*, 5(1), pp. 45–55. doi: 10.1109/TETC.2015.2504239.

Irani, J., Pise, N. and Phatak, M. (2016) 'Clustering Techniques and the Similarity Measures used in Clustering: A Survey', *International Journal of Computer Applications*, 134(7), pp. 975–8887. Available at: http://www.ijcaonline.org/research/volume134/number7/irani-2016-ijca-907841.pdf.

Jadhav, Urmi, and A. S. (2016) 'Effect of varying neurons in the hidden layer of neural network for simple character recognition', *International Journal on Recent and Innovation Trends in Computing and Communication*, 4(6)(June), pp. 266–269.

Jakub Kuzilek, Martin Hlosta, Drahomira Herrmannova, Zdenek Zdrahal, J. V. and A. W. (2015) 'OU Analyse: Analysing at-risk students at The Open University', in Conference, 5th International Learning Analytics and Knowledge (LAK) (ed.) *Learning Analytics Review*. New York (USA): LAK, pp. 1–6. Available at: http://libeprints.open.ac.uk/42529/1/__userdata_documents4_ctb44_Desktop_analysing-at-risk-students-at-open-university.pdf.

Jeong, H. *et al.* (2008) 'Using hidden markov models to characterize student behaviors in learning-by-teaching environments', in *In International conference on intelligent tutoring systems*, pp. 614–625. doi: 10.1007/978-3-540-69132-7-64.

Jiang, S. *et al.* (2014) 'Predicting MOOC Performance with Week 1 Behavior', in *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, pp. 273–275.

Jones, M. S. and K. M. L. (2014) 'MOOCs as LIS Professional Development Platforms: Evaluating and Refining SJSU's First Not-for-Credit MOOC', *Journal of Education for Library and Information Science*, 55(4).

Jyoti Bora, D. and Kumar Gupta, A. (2014) 'A Comparative study Between Fuzzy

Clustering Algorithm and Hard Clustering Algorithm', *International Journal of Computer Trends and Technology*, 10(2), pp. 108–113. doi: 10.14445/22312803/IJCTT-V10P119.

Kabakchieva, D. (2013) 'Predicting student performance by using data mining methods for classification', *Cybernetics and Information Technologies*, 13(1), pp. 61–72. doi: 10.2478/cait-2013-0006.

Kai, W. *et al.* (2008) 'An expanded training set based validation method to avoid overfitting for neural network classifier', in *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, pp. 83–87. doi: 10.1109/ICNC.2008.571.

Karamizadeh, S. *et al.* (2014) 'Advantage and drawback of support vector machine functionality', *I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings*, (I4ct), pp. 63–65. doi: 10.1109/I4CT.2014.6914146.

Kesim, M. and Alt, H. (2015) 'A Theoretical Analysis of Moocs Types From A Perspective of Learning Theories', *Procedia - Social and Behavioral Sciences*. Elsevier B.V., 186(Wclta 2014), pp. 15–19. doi: 10.1016/j.sbspro.2015.04.056.

Khalil, H. and Ebner, M. (2014) 'MOOCs Completion Rates and Possible Methods to Improve Retention - A Literature Review', *EdMedia: World Conference on Educational Media and Technology*, 2014(1), pp. 1305–1313. Available at: /p/147656/.

Kizilcec, R. F., Piech, C. and Schneider, E. (2013) 'Deconstructing Disengagement : Analyzing Learner Subpopulations in Massive Open Online Courses', in. In Proceedings of the third international conference on learning analytics and knowledge ., pp. 170–197. doi: 10.1145/2460296.2460330.

Kloft, M. *et al.* (2014) 'Predicting MOOC Dropout over Weeks Using Machine Learning Methods', in *Knowledge Management and E-Learning*, pp. 60–65.

Koch, P., Konen, W. and Hein, K. (2010) 'Gesture recognition on few training data using Slow Feature Analysis and parametric bootstrap', *Neural Networks (IJCNN), The 2010 International Joint Conference on*, 0(3), pp. 1–8. doi: 10.1109/IJCNN.2010.5596842.

Koedinger, K. R. *et al.* (2013) 'New Potentials for Data-Driven Intelligent Tutoring System Development and Optimization', *AI Magazine*, 1, pp. 27–41. doi: http://dx.doi.org/10.1609/aimag.v34i3.2484.

Koedinger, K. R., Mclaughlin, E. A. and Stamper, J. (2012) 'Automated Student Model Improvement', in *Educational Data Mining, proceedings of the 5th International Conference on*, pp. 17–24. doi: 10.978.17421/02764.

Kotsiantis, S. B., Kanellopoulos, D. and Pintelas, P. E. (2006) 'Data Preprocessing for Supervised Leaning', *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE*, 1(1), pp. 111–117.

Kraska, T., Talwalkar, A., Duchi, J.C., Griffith, R., Franklin, M.J. and Jordan, M. . (2013) 'MLbase : A Distributed Machine-learning System'. In Cidr, 1, pp. 2–1.

Kumar, A., Naughton, J. and Patel, J. M. (2015) 'Learning Generalized Linear

Models Over Normalized Data', *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, pp. 1969–1984. doi: 10.1145/2723372.2723713.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2016) 'Open University Learning Analytics Dataset', *Proceedings of the Data literacy for Learning Analytics Workshop, LAK 2016 Conference*. doi: 10.1002/asi.

Kwak, Hanock, and B.-T. Z. (2018) 'UNDERSTANDING LOCAL MINIMA IN NEURAL NET- WORKSBY LOSS SURFACE DECOMPOSITION', in, pp. 1–12.

Laboratories, T. B., Avenue, M. and Hill, M. (1995) 'Random Decision Forests Tin Kam Ho Perceptron training', *In Document analysis and recognition, proceedings of the third international conferen*, 1, pp. 278–282.

Lakkaraju, H. *et al.* (2015) 'A machine learning framework to identify students at-risk of adverse academic outcomes', in *In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1909–1918.

Lambert, J. B., Johnson, S. C. and Xue, L. (1998) 'On the piecewise analysis of networks of linear threshold neurons', *Journal of the American Chemical Society*, 11(4), pp. 691–697. doi: 10.1021/ja00093a015.

Lan, A. S. *et al.* (2016) 'Behavior-Based Latent Variable Model for Learner Engagement', pp. 64–71. Available at: http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.pdf.

Langley, P. (2011) 'The changing science of machine learning', *Machine Learning*, 82(3), pp. 275–279. doi: 10.1007/s10994-011-5242-y.

Lawrence, S. and Giles, C. L. (2000) 'machine', *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, pp. 114–119 vol.1. doi: 10.1109/IJCNN.2000.857823.

Leal, E. A., Miranda, G. J. and Carmo, C. R. S. (2012) . 'Self-determination theory: An analysis of student motivation in an accounting degree program', *Academic Journal -R. Cont. Fin. - USP, Sao Paulo*, 24(62), pp. 162–173.

Leban, G. (2006) 'VizRank : Data Visualization Guided by Machine', *Data Mining and Knowledge Discovery*, pp. 119–136. doi: 10.1007/s10618-005-0031-5.

Li, S., Tang, Q. and Zhang, Y. (2016) 'A Case Study on Learning Difficulties and Corresponding Supports for Learning in cMOOCs| Une étude de cas sur les difficultés d'apprentissage et le soutien correspondant pour l'apprentissage dans les cMOOC', *Canadian Journal of Learning and Technology/La revue canadienne de l'apprentissage et de la technologie*, 42(2).

Li, W. *et al.* (2014) 'Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning', in. In Proceedings of the EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs, pp. 3130–3137. doi: 10.1109/IJCNN.2016.7727598.

Liang, K.-Y. and Zeger, S. I. (1986) 'Longitudinal Data Analysis Using Generalized Linear Models', *Biometrika*, 73(June), pp. 13–22.

Liaw, a and Wiener, M. (2002) 'Classification and Regression by randomForest', *R news*, 2(December), pp. 18–22. doi: 10.1177/154405910408300516.

Liyanagunawardena, T. R., Parslow, P. and Williams, S. A. (2014) 'Dropout: MOOC Participants' Perspective', *Proceedings of the European MOOC Stakeholder Summit 2014*, pp. 95–100. Available at: http://centaur.reading.ac.uk/36002/.

Lloyd, S., Mohseni, M. and Rebentrost, P. (2013) 'Quantum algorithms for supervised and unsupervised machine learning', pp. 1–11. doi: 10.1038/nature23474.

Louppe, G. *et al.* (2013) 'Understanding variable importances in forests of randomized trees', *Advances in Neural Information Processing Systems 26*, pp. 431–439. doi: NIPS2013_4928.

Lykourentzou, I. *et al.* (2009) 'Dropout prediction in e-learning courses through the combination of machine learning techniques', *Computers and Education*. Elsevier Ltd, 53(3), pp. 950–965. doi: 10.1016/j.compedu.2009.05.010.

Marcialis, G. L. and Roli, F. (2005) 'Fusion of multiple fingerprint matchers by single-layer perceptron with class-separation loss function', *Pattern Recognition Letters*, 26(12), pp. 1830–1839. doi: 10.1016/j.patrec.2005.03.004.

Martimort, D. (1996) 'Exclusive Dealing , Common Agency , and Multiprincipals Incentive Theory', *The RAND journal of economics*, 27(1), pp. 1–31.

Mattingly, K. D., Rice, M. C. and Berge, Z. L. (2012) 'Learning analytics as a tool for closing the assessment loop in higher education', *Knowledge Management and E-Learning*, 4(3), pp. 236–247.

Meier, Y. *et al.* (2016) 'Association for Library and Information Science Education', *IEEE Transactions on Signal Processing*, 64(4), pp. 959–972. doi: 10.1109/TSP.2015.2496278.

Minaei-bidgoli, B. *et al.* (2003) 'Predicting Student Performance : an Application of Data Mining Methods With the Educational Web-Based System Lon-Capa', in *Frontiers in Education, 2003. FIE 2003 33rd Annual*. Westminster, United States, p. T2A–13. doi: 10.1109/FIE.2003.1263284.

Mitros, P. F. *et al.* (2013) 'Teaching electronic circuits online: Lessons from MITx's 6.002x on edX', *Proceedings - IEEE International Symposium on Circuits and Systems*, pp. 2763–2766. doi: 10.1109/ISCAS.2013.6572451.

Moe, W. W. and Fader, P. S. (2004) 'Capturing evolving visit behavior in clickstream data', *Journal of Interactive Marketing*, 18(1), pp. 5–19. doi: 10.1002/dir.10074.

Mtebe, J. S. and Raisamo, R. (2008) 'Evaluating Usability in Learning Management System Moodle', in, p. 30th. doi: 10.1002/j.1681-4835.2014.tb00436.x.

Mullan, J. (2016) *Learning Analytics in Higher Education*. London.

Natekin, A. and Knoll, A. (2013) 'Gradient boosting machines, a tutorial', *Frontiers in Neurorobotics*, 7(DEC). doi: 10.3389/fnbot.2013.00021.

Nawar, S. and Mouazen, A. M. (2017) 'Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon', *Sensors*

(Switzerland), 17(10), pp. 1–22. doi: 10.3390/s17102428.

Nawi, N. M., Ransing, M. R. and Ransing, R. S. (2006) 'An improved learning algorithm based on the Broyden-Fletcher-GoldfarbShanno (BFGS) method for back propagation neural networks', *Proceedings - ISDA 2006: Sixth International Conference on Intelligent Systems Design and Applications*, 1, pp. 152–157. doi: 10.1109/ISDA.2006.95.

Nelder, J.A. and Baker, R. . (2014) 'Generalized Linear Models', pp. 1–4.

Nilsson, N. J. (2005) 'Introduction to machine learning: An early draft of a proposed textbook'.

Nkuyubwatsi, B. (2016) 'Positioning Extension Massive Open Online Courses (xMOOCs) Within the Open Access and the Lifelong Learning Agendas in a Developing Setting', 3(1), pp. 14–36.

Noordam, J. C. *et al.* (2000) 'Geometrically Guided Fuzzy C-means Clustering for Multivariate Image Segmentation', *In Pattern Recognition, 2000. Proceedings. 15th International Conference*, 1, pp. 462–465.

Nunes, C. M. *et al.* (2004) 'Feature Subset Selection Using an Optimized Hill Climbing Algorithm for Handwritten Character Recognition.', in *Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop*, pp. 1018–1025. doi: http://springerlink.metapress.com/openurl.asp?genre=article&issn=0302-9743&volume=3138&spage=1018.

Olinsky, A., Kristin, K. and Brayton, K. B. (2012) 'Assessing Gradient Boosting in the Reduction of Misclassification Error in the Prediction of Success for Actuarial Majors', *Case Studies In Business, Industry & Government Statistics*, 5(1), pp. 12–16. Available at: http://revues-sfds.math.cnrs.fr/ojs/index.php/csbigs/article/view/226.

Oruç, Ö. E. and Kanca, A. (2011) 'Evaluation and Comparison of Diagnostic Test Performance Based on Information Theory', *International Journal of Statistics and Applications*, 1(1), pp. 10–13. doi: 10.5923/j.statistics.20110101.03.

Osborne, J. W. (2010) 'Improving your data transformations : Applying the Box-Cox transformation', *Practical Assessment, Research & Evaluation*, 15(12), pp. 1–9.

Osborne, J. W. and Jones, B. D. (2011) 'Identification with Academics and Motivation to Achieve in School: How the Structure of the Self Influences Academic Outcomes', *Educational Psychology Review*, 23(1), pp. 131–158. doi: 10.1007/s10648-011-9151-1.

Pahor, M. (2009) 'Principal Component Analysis', *Statistics*, 2(1), pp. 37–52.

Pal, M. and Mather, P. M. (2003) 'An assessment of the effectiveness of decision tree methods for land cover classification', *Remote Sensing of Environment*, 86(4), pp. 554–565. doi: 10.1016/S0034-4257(03)00132-9.

Pal, N. R. *et al.* (2005) 'A Possibilistic Fuzzy c-Means Clustering Algorithm', *IEEE Transactions on Fuzzy Systems*, 13(4), pp. 517–530. doi: 10.1109/TFUZZ.2004.840099.

Pal, N. R. and Bezdek, J. C. (1995) 'On Cluster Validity for the Fuzzy c-Means Model', *IEEE Transactions on Fuzzy Systems*, 3(3), pp. 370–379. doi: 10.1109/91.413225.

Pamuk, S. (2012) 'The Need for Pedagogical Change in Online Adult Learning : A Distance Education Case in a Traditional University.', 11(2), pp. 389–405.

Pereira, F., Mitchell, T. and Botvinick, M. (2009) 'Machine learning classifiers and fMRI: a tutorial overview.', *NeuroImage*. Elsevier Inc., 45(1 Suppl), pp. S199–S209. doi: 10.1016/j.neuroimage.2008.11.007.

Perner, P., Zscherpel, U. and Jacobsen, C. (2001) 'A comparison between neural networks and decision trees based on data from industrial radiographic testing', *Pattern Recognition Letters*, 22(1), pp. 47–54. doi: 10.1016/S0167-8655(00)00098-2.

Phyu, T. N. (2009) 'Survey of Classification Techniques in Data Mining', *IN Proceedings of the International MultiConference of Engineers and Computer Scientists*, I, pp. 18–20. doi: 10.1007/s10916-014-0018-0.

Piech, C. *et al.* (2015) 'Deep Knowledge Tracing', in *In Advances in Neural Information Processing Systems*, pp. 505–513. Available at: http://arxiv.org/abs/1506.05908.

Piotrowski, A. P. and Napiorkowski, J. J. (2013) 'A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling', *Journal of Hydrology*. Elsevier B.V., 476, pp. 97–111. doi: 10.1016/j.jhydrol.2012.10.019.

Podgorelec, V., Kokol, P. and Rozman, I. (2002) 'Decision Trees: An overview and Their Use in Medicine', *Journal of Medical Systems*, 26(5), pp. 445–463. doi: 10.1023/A.

Qiu, J. *et al.* (2016) 'Modeling and Predicting Learning Behavior in MOOCs', in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. California, USA, pp. 93–102. doi: 10.1145/2835776.2835842.

Ramesh, A. *et al.* (2013) 'Modeling Learner Engagement in MOOCs using Probabilistic Soft Logic', *NIPS Workshop on Data Driven Education*, 16, pp. 1–7.

Ramesh, A. *et al.* (2014) 'Learning Latent Engagement Patterns of Students in Online Courses', in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence Learning*, pp. 1272–1278.

Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., Chuang, I. and Ho, A. (2014) 'PH207x: Health in Numbers and PH278x: Human Health and Global Environmental Change: 2012-2013 Course Report', in, pp. 2012–2013.

Renz, J., Schwerer, F. and Meinel, C. (2016) 'openSAP: Evaluating xMOOC Usage and Challenges for Scalable and Open Enterprise Education.', *International Journal of Advanced Corporate Learning*, 9(2), pp. 34–39. Available at: http://10.0.15.151/ijac.v9i2.6008%5Cnhttps://ezp.lib.unimelb.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=117786557&site=eds-live&scope=site.

Ridgeway, G. (2007) 'Generalized Boosted Models: A guide to the gbm package', *Compute*, 1(4), pp. 1–12. doi: 10.1111/j.1467-9752.1996.tb00390.x.

Ritter, S., Anderson, J.R., Koedinger, K.R. and Corbett, A. (2007) 'Cognitive Tutor: Applied research in mathematics education.', in, pp. 249–255.

Rodriguez, G. (2013) 'Generalized Linear Model Theory', *Encyclopedia of statistical sciences*, 4.

Rounds, E. M. (1980) 'A combined nonparametric approach to feature selection and binary decision tree design', *Pattern Recognition*, 12(5), pp. 313–317. doi: 10.1016/0031-3203(80)90029-1.

Rovai, A. P. (2002) 'Building sense of community at a dstance', in *International Review of Research in Open & Distance Learning*, pp. 1–16. doi: 10.19173/IRRODL.V3I1.79.

Russell, N., Cribbin, L. and Murphy, T. B. (2012) 'upclass : An R Package for Updating Model-Based Classification Rules'. Available at: https://cran.r-project.org/web/packages/upclass/vignettes/upclass.pdf.

Ryan, R. M. and Deci, E. L. (2000) 'Intrinsic and Extrinsic Motivations : Classic Definitions and New Directions', *Contemporary Educational Psychology*, 67, pp. 54–67. doi: 10.1006/ceps.1999.1020.

Sachin, R. Barahate, and M. S. V. (2012) 'A Survey and Future Vision of Data mining in Educational Field', in *In Advanced Computing & Communication Technologies (ACCT)*, pp. 96–100. doi: 10.1109/ACCT.2012.14.

Sadiku, M. N. O., Tembely, M. and Musa, S. M. (2017) 'Deep Learning', *International Research Journal of Advanced Engineering and Science*, 2(1), pp. 77–78.

Samuel, A. L. (1959) 'Some studies in machine learning using the game of checkers', *IBM Journal of research and development*, 3(3), pp. 210–229. doi: 10.1016/0066-4138(69)90004-4.

Sarkar, S. (2012) 'The Role of Information and Communication Technology (ICT) in Higher Education for the 21st Century', in *Science Probe*, pp. 30–41.

Schmidhuber, J. (2015) 'Deep Learning in neural networks: An overview', *Neural Networks*. Elsevier Ltd, 61, pp. 85–117. doi: 10.1016/j.neunet.2014.09.003.

Sclater, N. (2008) 'Web 2.0, personal learning environments, and the future of learning management systems', *EDUCAUSE Center for Applied Research Research Bulletin*.

Shah, A. D. *et al.* (2014) 'Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study', *American Journal of Epidemiology*, 179(6), pp. 764–774. doi: 10.1093/aje/kwt312.

Shaliz, C. (2012) 'Logistic Regression', *Advanced Data Analysis from an Elementary Point of View*, pp. 223–237. Available at: http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf.

Shapiro, H. B. *et al.* (2017) 'Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers', *Computers and Education*. Elsevier Ltd, 110, pp. 35–50. doi: 10.1016/j.compedu.2017.03.003.

Siemens, G. (2013) ‘Learning Analytics : The Emergence of a Discipline’, *American Behavioral Scientist*, 57(10), pp. 1380–1400. doi: 10.1177/0002764213498851.

Sinclair, J. and Kalvala, S. (2016) ‘Student engagement in massive open online courses’, *International Journal of Learning Technology*, 11(3), p. 218. doi: 10.1504/IJLT.2016.079035.

Sing, T. *et al.* (2009) ‘ROCR: Visualizing the performance of scoring classifiers’, *R package version*, 1, p. 4.

Sørebø, Ø. *et al.* (2009) ‘The role of self-determination theory in explaining teachers’ motivation to continue to use e-learning technology’, *Computers and Education*. Elsevier Ltd, 53(4), pp. 1177–1187. doi: 10.1016/j.compedu.2009.06.001.

Spearman, C. (1904) ‘General Intelligence , Objectively Determined and Measured’, *The American Journal of Psychology*, 15(2), pp. 201–292.

Statistics, M. (1986) ‘Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models’, *The Annals of Statistics*, 13(1), pp. 342–368.

Staubitz, T. *et al.* (2014) ‘TOWARDS SOCIAL GAMIFICATION - IMPLEMENTING A SOCIAL GRAPH IN AN XMOOC PLATFORM’, in *InProceedings of the 7th International Conference of Education, Research and Innovation*, pp. 17–19.

Suganya, R. and Shanthi, R. (2012) ‘Fuzzy C-Means Algorithm-A Review’, *International Journal of Scientific and Research Publications*, 2(11), pp. 2250–3153. Available at: www.ijsrp.org.

Tahar, N. F. *et al.* (2010) ‘Students’ attitude toward mathematics: The use of factor analysis in determining the criteria’, *Procedia - Social and Behavioral Sciences*, 8(5), pp. 476–481. doi: 10.1016/j.sbspro.2010.12.065.

Tan, A. C. and Gilbert, D. (2003) ‘An empirical comparison of supervised machine learning techniques in bioinformatics’, *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, 19, pp. 219–222. Available at: http://dl.acm.org/citation.cfm?id=820189.820218.

Tatiana, S. (2016) ‘Participation in Massive Open Online Courses : the Effect of Learner Motivation and Engagement on Achievement’, in *Achievement (No. WP BRP 37/EDU/2016)*. National Research University Higher School of Economics, pp. 1–16.

Taylor, C., Veeramachaneni, K. and O’Reilly, U.-M. (2014) ‘Likely to stop? Predicting Stopout in Massive Open Online Courses’. Available at: http://arxiv.org/abs/1408.3382.

Trowler, V. (2010) ‘Student engagement literature review’, *The higher education academy*. UK: Department of Educational Research, pp. 1–15. doi: 10.1037/0022-0663.85.4.571.

Tu, J. V. (1996) ‘Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes’, *Journal of Clinical Epidemiology*, 49(11), pp. 1225–1231. doi: 10.1016/S0895-4356(96)00002-9.

Turner, J. C. and Patrick, H. (2008) 'How does motivation develop and why does it change? Reframing motivation research', *Educational Psychologist*, 43(3), pp. 119–131. doi: 10.1080/00461520802178441.

Tvarozek, J. and Brza, T. (2014) 'Engaging Students in Online Courses through Interactive Badges', *Proceedings of the International Conference on e-Learning '14*, pp. 89–95.

VanLehn, K. (2006) 'The Behavior of Tutoring Systems', *International Journal of Artificial Intelligence and Education*, 16(3), pp. 227–265. Available at: http://dl.acm.org/citation.cfm?id=1435351.1435353.

Vecci, L., Piazza, F. and Uncini, A. (1998) 'Learning and approximation capabilities of adaptive spline activation function neural networks', *Neural Networks*, 11(2), pp. 259–270. doi: 10.1016/S0893-6080(97)00118-4.

Velazquez-Iturbide, J. A., Hernan-Losada, I. and Paredes-Velasco, M. (2017) 'Evaluating the Effect of Program Visualization on Student Motivation', *IEEE Transactions on Education*, 60(3), pp. 238–245. doi: 10.1109/TE.2017.2648781.

Vuk, M. (2006) 'ROC Curve , Lift Chart and Calibration Plot', *Metodolos˙ki zvezki*, 3(1), pp. 89–108. doi: 10.1.1.126.7382.

Wang, F. and Chen, L. (2016) 'A Nonlinear State Space Model for Identifying At-Risk Students in Open Online Courses', *Proceedings of the 9th International Conference on Educational Data Mining*, pp. 527–532.

Wang, H., Shi, X. and Yeung, D.-Y. (2017) 'Relational Deep Learning: A Deep Latent Variable Model for Link Prediction.', *AAAI Conference on Artificial Intelligence*, pp. 2688–2694.

Wang, W., Yu, H. and Miao, C. (2017) 'Deep Model for Dropout Prediction in MOOCs', in *In Proceedings of the 2nd International Conference on Crowd Science and Engineering.*, pp. 26–32. doi: 10.1145/3126973.3126990.

Wang, Z. *et al.* (2017) 'Interaction pattern analysis in cMOOCs based on the connectivist interaction and engagement framework', *British Journal of Educational Technology*, 48(2), pp. 683–699. doi: 10.1111/bjet.12433.

Weaver, D., Spratt, C. and Nair, C. S. (2008) 'Academic and student use of a learning management system: Implications for quality', *Australasian Journal of Educational Technology*, 24(1), pp. 30–41. doi: 10.14742/ajet.1228.

Weisberg, S. (2001) 'Yeo-Johnson Power Transformations', in Department of Applied Statistics, U. of M. (ed.), pp. 1–4.

Wen, M. *et al.* (2014) 'Linguistic Reflections of Student Engagement in Massive Open Online Courses', *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pp. 525–534. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84909951147&partnerID=40&md5=80f121bcfc587505feaee3a3d6675c59.

Wen, M., Yang, D. and Rosé, C. P. (2014) 'Sentiment Analysis in MOOC Discussion Forums : What does it tell us ?', in. Proceedings of Educational Data Mining.

West, D. M. D. M. (2012) ‘Big Data for Education: Data Mining, Data Analytics, and Web Dashboards’, *Governance Studies. Brookings, US: Reuters*, 4(1).

Wikramanayake, G. N. (2003) ‘e-Learning: Changes in Teaching and Learning Styles’, in *22nd National Information Technology Conference*, pp. 118–124.

Wilson, G. (2004) ‘Online Interaction impacts on learning: Teaching the teachers to teach online’, *Australasian Journal of Educational Technology*, 20(1), p. 33.

Wolff, A. *et al.* (2014) ‘Developing predictive models for early detection of at-risk students on distance learning modules’, *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge - LAK ’14*, 1, p. T2A13-T2A18. doi: http://www.informatik.uni-trier.de/~ley/db/conf/lak/index.html.

Yang, D. and Rose, C. P. (2013) ‘“ Turn on , Tune in , Drop out ”: Anticipating student dropouts in Massive Open Online Courses’, in. In Proceedings of the NIPS Data-Driven Education Workshop., pp. 13–20.

Ye, C. and Biswas, G. (2014) ‘Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information’, *Journal of Learning Analytics*, 1(3), pp. 169–172. Available at: http://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/4212.

Yousef, A. M. F. *et al.* (2014) ‘What drives a successful MOOC? An empirical examination of criteria to assure design quality of MOOCs’, *Proceedings - IEEE 14th International Conference on Advanced Learning Technologies, ICALT 2014*, pp. 44–48. doi: 10.1109/ICALT.2014.23.

Yun, C. *et al.* (2007) ‘An Experimental Study on Feature Subset Selection Methods’, *7th IEEE International Conference on Computer and Information Technology (CIT 2007)*, pp. 77–82. doi: 10.1109/CIT.2007.81.

Zhao, J., Mathieu, M. and LeCun, Y. (2016) ‘Energy-based Generative Adversarial Network’, in *arXiv preprint arXiv:1609.03126.*, pp. 1–17. doi: 10.1016/j.neunet.2014.10.001.

Zhou, M. (2016) ‘Chinese university students’ acceptance of MOOCs: A self-determination perspective’, *Computers and Education*. Elsevier Ltd, 92–93, pp. 194–203. doi: 10.1016/j.compedu.2015.10.012.

Zimmerman, R. K. *et al.* (2016) ‘Classification and Regression Tree (CART) analysis to predict influenza in primary care patients’, *BMC Infectious Diseases*. BMC Infectious Diseases, 16(1), p. 503. doi: 10.1186/s12879-016-1839-x.

Zutshi, S., O’Hare, S. and Rodafinos, A. (2013) ‘Experiences in MOOCs: The Perspective of Students’, *American Journal of Distance Education*, 27(4), pp. 218–227. doi: 10.1080/08923647.2013.838067.

Appendix

Appendix 1. Number of videos and chapters view by failing learners

Course	Number of videos watch by failing group	Number of chapters read by failing group
Health Fall	3144383	3283
Health Spring	22039	1332
Electronics Fall	183983	3673
Electronics Spring	58641	2439
Computer Science Fall	230913	6137
Computer Science Spring	154775	6465

Appendix 2. Number of videos and chapters view by successful learners

Course	Number of videos watch by Successful group	Number of chapter read by Successful group
Health Fall	355583	5543
Health Spring	41710	1602
Electronics Fall	307040	6074
Electronics Spring	85876	2156
Computer Science Fall	171451	4339
Computer Science Spring	78659	3040

Appendix 3. Number of chapter read by successful students per continent

Learners Group	Africa	Asia	Australia	America	Europe
Successful Health Fall	915	1980	148	1121	1379
Successful Health Spring	200	445	50	931	1707
Successful Electronics Fall	113	3278	45	931	1707
Successful Electronics Spring	68	1289	64	271	542
Successful Computer Science Fall	70	874	30	1273	2058
Successful Computer Science Spring	17	944	18	940	1121

Appendix 4. Number of chapter read by failing students per continent

Learners Group	Africa	Asia	Australia	America	Europe
Failing Health Fall	651	815	74	1243	532
Failing Health Spring	172	359	20	478	314
Failing Electronics Fall	249	1723	63	729	954
Failing Electronics Spring	155	1194	51	522	534
Failing Computer Science Fall	320	1194	77	2307	1683
Failing Computer Science Spring	239	1810	56	2561	1847

Appendix 5. Number of videos watch by successful students per continent

Learners Group	Africa	Asia	Australia	America	Europe
Successful Health Fall	67016	99218	10582	70383	108384
Successful Health Spring	4330	4886	943	21972	9579
Successful Electronics Fall	3352	121223	188	53836	128441
Successful Electronics Spring	5550	26613	232	14986	38735
Successful Computer Science Fall	1882	23346	1543	44642	100038
Successful Computer Science Spring	92	16656	101	44642	35809

Appendix 6. Number of videos watch by failing students per continent

Learners Group	Africa	Asia	Australia	America	Europe
Failing Health Fall	34009	29936	1790	57754	23373
Failing Health Spring	2735	6663	645	7612	4536
Failing Electronics Fall	34323	59871	1241	31129	58498
Failing Electronics Spring	3142	23155	989	14184	17412
Failing Computer Science Fall	20184	54388	3213	84698	70652
Failing Computer Science Spring	4978	42261	1322	60655	46247

Appendix 7. Percentage of motivational status in trajectories courses

	Number of motivation student in pervious courses	Number of extrinsically student in pervious courses	Number of intrinsically student in pervious courses
Withdrawal	750	986	1261
Non Withdrawal	1957	294	234

Appendix 8. Number of Successful Learners per educational level

Learners Group	Less than Secondary	Secondary	Bachelor's	Master's	Doctorate
Successful Health Fall	2	29	166	172	19
Successful Health Spring	3	40	86	66	7
Successful Electronics Fall	7	180	129	70	4
Successful Electronics Spring	7	57	53	18	0
Successful Computer Science Fall	12	87	84	72	24
Successful Computer Science Spring	8	68	68	41	8

Appendix 9. Number of Failing Learners per educational levels

Learners Group	Less than Secondary	Secondary	Bachelor's	Master's	Doctorate
Failing Health Fall	7	75	239	162	34
Failing Health Spring	6	63	151	124	10
Failing Electronics Fall	6	223	214	92	1
Failing Electronics Spring	8	184	138	60	1
Failing Computer Science	23	337	408	230	17
Failing Computer Science Spring	48	460	414	210	24