Massey University, Auckland, New Zealand

Prediction of Students' Performance Through Data Mining.

Rahila Umer Baloch

2020.

A thesis presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science.

Abstract

Government funding to higher education providers are based upon graduate completions rather than on student enrollments. Therefore, unfinished degrees or delayed degree completions are major concerns for higher education providers since these problems impact their long-term financial security and overall cost effectiveness. Therefore, providers need to develop strategies for improving the quality of their education to ensure increased enrollment and retention rates.

This study uses predictive modeling techniques for assisting providers with real-time identification of struggling students in order to improve their course retention rates. Predictive models utilizing student demographic and other behavioral data gathered from an institutional learning platform have been developed to predict whether a student should be classed as at-risk of failing a course or not. Identification of at-risk students will help instructors take proactive measures, such as offering students extra help and other timely supports. The outcomes of this study will therefore provide a safety net for students as well as education providers in improving student engagement and retention rates.

The computational approaches adopted in this study include machine learning techniques in combination with educational process mining methods. Results show that multi-purpose predictive models that were designed to operate across a variety of different courses could not be generalized due to the complexity and diversity of the courses. Instead, a meta-learning approach for recommending the best classification algorithms for predicting students' performance is demonstrated.

The study reveals how process-unaware learning platforms that do not accurately reflect ongoing learner interactions can enable the discovery of student learning practices. It holds value in reconsidering predictive modeling techniques by supplementing the analysis with contextually-relevant process models that can be extracted from stand-alone activities of process-unaware learning platforms. This provides a prescriptive approach for conducting empirical research on predictive modeling with educational data sets. The study contributes to the fields of learning analytics and education process mining by providing a distinctive use of predictive modeling techniques that can be effectively applied to real world data sets.

Acknowledgments

Thanking those who have contributed to the completion of this journey. This thesis would not have been possible without your guidance, support, and encouragement.

I would like to express my sincere gratitude to my supervisory team; Dr. Teo Susnjak, Dr. Anuradha Mathrani and Dr. Suriadi Lim for their encouragement, direction, guidance, and support throughout the doctoral journey. They encouraged me to grow as an independent researcher and also helped me attain my full potential under their supervision. Their valuable supervision will enable me to undertake challenging research problems in the future. Special thanks to my committee members, Dr.Girija Chetty, Dr. Ian Bond, Dr. David Rozado for their time and interest in my research, and for their valuable feedback and insight.

I would also like to thank Higher Education Commission (HEC) Pakistan for funding my studies and support during my Ph.D. study. I would also like to thank School of Natural Computational Sciences (SNCS), Graduate Research School (GRS) for financial support for conferences travel grants.

I am fortunate to have so many friends who have invested in the attainment of my goals; Goomathy, Jun Ren, Azadeh, Nelofar, Hooman, Jumbo, JinJin, Zainal, Sibghat, Baryal, Ayesha, Shumaila, Amna, Arif shb and special thanks to Aroon for unconditional support and encouragement.

Last but not the least I would not have been standing at the finish line had it not been for the selfless love and prayers of my family. Their affection and encouragement helped me pass through the hard time. Special thanks to my elder sister Rashida Umer and her husband (Thank you Bobo and Alam Bhai, you were always there for me).

Dedication

To my family, for their everlasting, unconditional love and support. To the loving memories of my late father (Haji Muhammad Umer Sumalani). To the iron-will of my mother who taught us the power of imagination and strong will to achieve anything in life. To my brothers (Habib-ur-Rehman and Zia-ur-Rehman) for trusting in me. Despite the fact that we belong to a conservative society where education for girls is not normal, you all encouraged me and did not listen to those who consider girls' education to be less important than boys. I dedicate my success to the Men of my family, for their trust and support.

Publications

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. Journal of Research in Innovative Teaching Learning, 10(2), 160–176. https://doi.org/10.1108/JRIT-09-2017-0022
- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2017). Prediction of Students' Dropout in MOOC Environment, International Journal of Knowledge Engineering, Vol. 3, No. 2, December 2017 http://www.ijke.org/vol3/85KD015.pdf (ISSN: 2382-6185)
- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2020 in-press) Data Quality Challenges in Educational Process Mining: Building Process-Oriented Event Logs from Process-Unaware Online Learning Systems. Int. J. of Business Information Systems
- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. S(2018). A learning analytic approach: Using online weekly student engagement data to make predictions on student performance. 2018 International Conference on Computing, Electronic and Electrical Engineering (IEEE), 1–5.
- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2017). Predicting Student's Academic Performance in a MOOC Environment, 11th International Conference on Data Mining, Computers, Communication and Industrial Applications (DMCCIA-2017) http://dirpub.org/images/proceedingspdf/DIR1217002.pdf
- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2019) Mining Activity Log Data to Predict Student's Outcome in a Course ICBDE'19 Proceedings of the 2019 International Conference on Big Data and Education (ACM) Pages 52-58.
- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S.(2020) Current Stance on Predictive Analytics in Higher Education: Opportunities, Challenges and Future Directions. Interactive Learning Environments. (Accepted)
- R. Umer,Sohrab Khan., (2020) Prediction of Students' Failure using VLE and Demographic data: Case study Open University Data. In. J. Business Intelligence and Data Mining. (Accepted)
- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2020) Preparation for online education: Exploring students' engagement with LMS tools. Computers Education. (Ready for submission)

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2020) Application of Meta-learning in Selection of Classification algorithms for prediction of students' performance. Interactive learning Environment. (Ready for submission)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Student retention has been a critical issue in higher education institutions since their establishment. In early 1970s, Vincent Tinto [2], Spady [3] and Astin [4] are one of the earliest researchers who studied aspects related to student persistence and retention. They studied dropout characteristics and formulated a theoretical model that describes different interactions between individuals and institution that lead to dropouts. Prior to these studies, majority of the research referred to student persistence and retention as psychological studies [5][6] where the focus was on students' characteristics, their abilities and shortcomings that led to their dropout [7][8][9][10].

Student persistence was first explained by Alexander Astin, who developed the Input-Environment-Outcome model to study student persistence. This model revealed the importance of pre-existing characteristics of students (input variable) prior to entering college, environmental factors (environmental variables) of the institution and effects of the college outcome (output). The purpose of the model was to study the impact of various environmental variables on students behaviors, that is, see how students grow or adapt under varying environmental conditions. Later in 1975, Tinto [2] built upon Astin's effort to explain factors that influence student persistence. According to Tinto's theory, students have many inherent characteristics; moreover, on entering any college, their everyday interactions in the institutional settings may lead to increase or decrease in their commitment which in turn affects their integration into the intuition. Greater integration will lead to higher retention. Since that time, student retention included the concept of institutional responsibility that has an influence on student decisions regarding retention or dropping out [11]. As a result, retention rates have become one of the major indicators of overall effectiveness of many institutions

[5]. Therefore, higher retention rate has become imperative to institutional success, since as more students remain in institutes, they will pay tuition fees and overall, this will generate academic success.

Research shows that academic success is a key factor to improving the retention rates [12, 13]. One way to improve retention rates is to make early prediction of those students who are at-risk of failing the course [14][15][16] and to implement early intervention procedures. By having some analytical strategy which can enable predictions on students' performances, we can help these institutes make timely interventions and improve students' performance. The widespread use of tools like Student Management Systems (SMS) and Learning Management Systems (LMS) in higher education institutes have provided support in communication, the delivery of resources, the design of interactive learning activities and management of academic assessments; and at the same time have also provided them with large data sets related to students demographics and their academic records. Logs revealing details of students' interaction with LMS has facilitated new research that is aimed to improve students' academic performance [17] [18]. There are many success stories related to how data extracted from the tools that used students' data has helped improve the retention rate [19]. For example, Georgia State University using predictive analytics which improved graduation rates from 32% (2003) to 54% (2014) [20]. Purdue University, USA, predicted at-risk students as early as the second week which led to improving academic performance of those students who were identified as having low grades [21].

1.2 Problem Statement and Research Question

The main purpose of this study was to develop predictive models that utilize a combination of learning management system data, demographic data and cognitive data collected over the

course duration to accurately predict academic success across a range of students studying in a large tertiary institution located in New Zealand. By making models that can accurately predict student successes, this work can be used in the design of support systems that provide early interventions to help students who are likely to be unsuccessful. This will result in more successful students leading to improved retention rate and lowered dropout rate.

A student pursues many activities as they learn subject related concepts over their course duration. As the student interacts with the learning management system, they leave their digital traces. Eventually, each student has to take a final exam. Those students who reach a fixed threshold will be considered to have passed the course. Therefore, the problem is to predict the final outcome of the student based upon the activities they have undertaken during the course. These data traces collected by the e-learning platform like LMS along with other student related data extracted from the SMS can facilitate us in making such predictions.

The ultimate goal of prediction is to determine that a given student will be at-risk of failing the course or not in future. This task is performed using supervised learning (classification or regression) method where inputs are students' demographic data, assessment scores or interaction data from education tools (e.g., LMS) and output indicates the final outcome as pass or fail, at-risk or not at-risk, successful or not successful. The output variable could be a discrete or continuous variable. In case of a continuous variable, regression methods have been used to predict the final mark in the course and for a discrete variable, classification methods were used.

The problem statement for the current study is:

To what extent can a classifier predict the student's academic performance during the course, based on the students' interaction data with learning management system (LMS),

demographic and cognitive data?

The following research questions have provided the focus in this investigation.

- RQ1: What is the earliest possible time frame within a teaching semester for generating robust classifiers that are capable of reliably identifying students who are at-risk of failure?

- RQ2: Is it possibly to effectively combine process mining features with generic features in machine learning? If so, to what degree is classification accuracy can be improved with the newly augmented features?

- RQ3: Based on the LMS tools usage, is it possible to find distinct subgroups of courses that could be used in future for developing a generalized portable model for similar courses?

1.3 Significance of the Study

Higher education institutes are facing challenges due to low course enrollments and further with lower course completion rates. The issue of low dropout rate is fast becoming a priority, and universities are seeking out strategies for improving students' retention rate. According to the report of OECD [22] in Australia just 31% of students completed a 4-year degree programme, US had 49% completions while UK is on top with 71% completions. Lower retention rates are a serious threat to universities long term financial security. Universities, therefore, are focusing more on identifying strategies that ensure students successes and which can provide proactive actions to support students in their course work.

1.4 Scope

The scope of the study was to use students' data and develop information to support instructors in making decisions regarding interventions that can be made to help struggling students in the course. Data included in the study were limited to the students' records that were extracted from learning management systems (LMS) logs, demographic data and cognitive data collected during the course. This study involved 5 years of students' data from undergraduate, postgraduate levels of courses. However, due to large missing values, the study included only data from 2015 to 2017 data. Due to the size limitations of the data, results cannot be generalized for whole population.

1.5 Contributions

The main contributions of the thesis are presented as follows:

- This study demonstrates classification of blended courses (4,353) in heterogeneous clusters using behavioral data from LMS. The approach used for selection of maximum number of courses from pool of 10k courses for the study of overall online engagement pattern was based upon those which demonstrated relatively higher level of online activities. This study is a stepping-stone to move on-wards to the stage of development of more sophisticated predictive models and in enhancing our understanding of diverse patterns that occur in different courses. The findings from this study will help in building a generalized predictive model for predicting students' academic success for similar types of courses. Furthermore, quality of data plays a big role in development of such models. As such, data extractions from the LMS log without investigation of the course structure will affect the performance of the predictive model. Therefore, it is important

to investigate the instructional conditions of underlying courses, before utilizing LMS log for developing predictive model which can then be applied to unseen data set, and in making recommendations regarding the best classification algorithms. This could be beneficial for educators who lack knowledge of classification algorithms.

- A multi-label regression model is proposed for recommending classification algorithms to solve prediction problems in education domain. These regression models are trained using historical data of different courses. This study utilized data of 33 courses, and more than 20 classification algorithms were used to predict students' final outcome in the course. Meta-features were calculated for each data set which not only used statistical features, but domain knowledge was also included. Meta-features and performance of classification algorithms were integrated to make the training set for the regression model.

- This study contributes to the fields of learning analytics and process mining by providing lessons that have been learned in the extraction and conversion of process-unaware data to event logs for the purpose of analyzing online education data. Many quality issues generally faced with education data have been shared by using a specific running example (i.e., quiz-taking) that demonstrates these issues. The study further utilized a process mining approach to help in making early predictions to improve students' learning experience in Massive Open Online courses (MOOC). The impact of various machine learning techniques in combination with process mining features were investigated to measure effectiveness of these techniques. This study outlines a data driven approach to improve students' learning experience and decrease the dropout rate. Early predictions based on individual's participation can help educators in providing support

to those students who are struggling in the course.

1.5.1 Published Work

This thesis study has resulted in ten full-paper research articles comprising 5 journal publications, 3 papers in conference proceedings and 2 journal manuscripts are ready for submission. The thesis author is the primary researcher for all these articles.

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. Journal of Research in Innovative Teaching Learning, 10(2), 160–176. https://doi.org/10.1108/JRIT-09-2017-0022

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2017). Prediction of Students' Dropout in MOOC Environment, International Journal of Knowledge Engineering, Vol. 3, No. 2, December 2017 http://www.ijke.org/vol3/85KD015.pdf (ISSN: 2382-6185)

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2020 in-press) Data Quality Challenges in Educational Process Mining: Building Process-Oriented Event Logs from Process-Unaware Online Learning Systems. Int. J. of Business Information Systems

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. S(2018). A learning analytic approach: Using online weekly student engagement data to make predictions on student performance. 2018 International Conference on Computing, Electronic and Electrical Engineering (IEEE), 1–5.

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2017). Predicting Student's Academic Performance in a MOOC Environment, 11th International Conference on Data Mining, Computers, Communication and Industrial Applications (DMCCIA-2017) http://dirpub.org/images/proceedingspdf/DIR1217002.pdf

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2019) Mining Activity Log Data to Predict Student's Outcome in a Course ICBDE'19 Proceedings of the 2019 International Conference on Big Data and Education (ACM) Pages 52-58.

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S.(2020) Current Stance on Predictive Analytics in Higher Education: Opportunities, Challenges and Future Directions. Interactive Learning Environments. (Accepted)

- R. Umer,Sohrab Khan., (2020) Prediction of Students' Failure using VLE and Demographic data: Case study Open University Data. In. J. Business Intelligence and Data Mining. (Accepted)

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2020) Preparation for online education: Exploring students' engagement with LMS tools. Computers Education. (Ready for submission)

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2020) Application of Meta-learning in Selection of Classification algorithms for prediction of students' performance. Interactive learning Environment. (Ready for submission)

1.6 Thesis Outline

The first chapter has provided a brief overview of the problem domain, explained the need for this study, given the problem statement, stated the research questions, laid out the scope of this study and presented the study's contributions. The remaining thesis structure is outlined next.

- Chapter-2 provides a synthesis in the domain of education data mining, machine learning and education process mining, specifically regarding current methods used for making

predictions on students' academic performance. Additionally, this review identifies gaps and challenges in current research which consequently forge a path for future research. The systematic literature review presented below includes sections which covers the methodology for review, results of review provide an overview on types of data, and machine/data mining methods used for prediction of students' performance. The final section of the chapter reviews the current status of education process mining.

- Chapter-3 describes the general methods that were common across all experiments conducted in this research. It describes the data sets, feature types and their extraction process. Training procedures, evaluation methods and descriptions of the various environments and hardware that were used are also described. Specific details of methods used in experiments are described in relevant chapters.

- Chapter-4 presents clustering methods on more than 4000 heterogeneous courses, to divide them into small homogeneous groups. The aim of this study is to analyze the clusters of courses using students' LMS log data and study the patterns. These findings will help in future for developing a generalized predictive model for predicting students' academic success for similar courses.

- Chapter-5 presents an analysis of LMS data from different types of courses to see how accurately student academic performance can be forecasted when their weekly engagement data is integrated with assignment scores. This chapter highlights the importance of LMS data; which can give insights on student behaviors and can lead to development of accurate models that can be used for predicting the students' final outcome in enrolled courses. Finally, this chapter demonstrates the suitability of multi-variable regression algorithms to predict the performance of classification algorithms. This study proposes

an approach that utilizes historical data and machine learning experience using meta learning to recommend best subset of classification algorithms for predicting students' performance in a course.

- Chapter-6 is laid out in two parts. The first part focuses on preparation of event logs from educational data gathered from different courses delivered in a university setting. The second part of the chapter demonstrates use of a MOOC data set for preparing event log, extracting process mining features and conducting a study with generic features (extracted from Moodle event log and assessment scores) to make early predictions on students' performance. and to investigate how by incorporating process mining features, the performance of predictive models can be improved.

- Chapter-7 summarizes the study and provides a discussion on the overall results. Finally, the conclusions are drawn and recommendations for future research are proposed.

Chapter 2

State of the Art: Predictive Analytics in Higher Education

2.1 Introduction

In the recent past, use of predictive models for enabling real-time identification of struggling students have increasingly been recognized as a way forward to improve student retention rate. These predictive models are developed by utilizing student demographic data and other forms of behavioral data extracted from a range of resources within educational settings to predict performance of new students. Once those students who are struggling to engage with the course have been identified (i.e., they have been classed as at-risk of failing a course), then instructors can plan proactive measures to support these students in addressing their learning difficulties. In this manner, predictive models provide a safety net for students, who can then be provided with timely support, thereby positively impacting their performance.

This chapter provides a synthesis in the domain of education data mining, machine learning and education process mining on current methods used for prediction of students' academic performance. Additionally, this review identifies gaps and challenges in current research which consequently forge a path for future research. The systematic literature review presented below includes sections which covers the methodology for review, results of review provide an overview of type of data, and machine/data mining methods used for prediction of students performance. The final section of the chapter reviews the current status of education process mining.

2.2 Applications of LA and EDM

The emergence of new technology in the field of education is one of the major innovations that holds promise of enhancing the learning environment [23]. E-learning systems or web-based education systems no longer require students to be physically present in classrooms;

instead, they have replaced the traditional classroom with a virtual classroom. The advent of course management systems, combined with ubiquitous online communication and collaboration tools have enabled e-learning strategies that use virtual online environments across universities, community colleges, and schools [24].

Learning management systems have emerged from the field of e-learning and have become an integral part of higher education institutes. Figure 2.1 shows the evolution of LMS market. It is based on different data sources that includes surveys from US institutes, vendor reports, and press releases on LMS uses. They offer a variety of tools to enhance the teaching and learning experience, such as launching quizzes/tests, scheduling of various learning events (e.g., live lecture) or the enabling of seamless communication (e.g., discussion forums and chats). These functionalities provide students with a friendly, personalized and engaging learning environment. Further, by using other reporting strategies like dashboards, students can get real time feedback on their performance. Teachers too can benefit from these features and improve their pedagogical practices. According to the EDUCAUSE report entitled *The Current Ecosystem of Learning Management Systems in Higher Education: Student, Faculty, and IT Perspectives* [25], 74% percent of faculty think LMS is a tool to enhance teaching and 71% of faculty think LMS is a useful tool to enhance student learning (p.4).). Students have recognized LMS as an important part of their academic experience. According to the ESCAR report – *Study of undergraduate students and information technology, 2013* – the LMS has been rated as very/extremely important by students in relation to their academic success [25]. Teaching faculty and students have both recognized the importance of tools embedded in LMS, although the use of these tools among the faculty members is rather diverse. The majority of the faculty use LMS tools mainly for content delivery and only few utilize its full capacity [25][26].

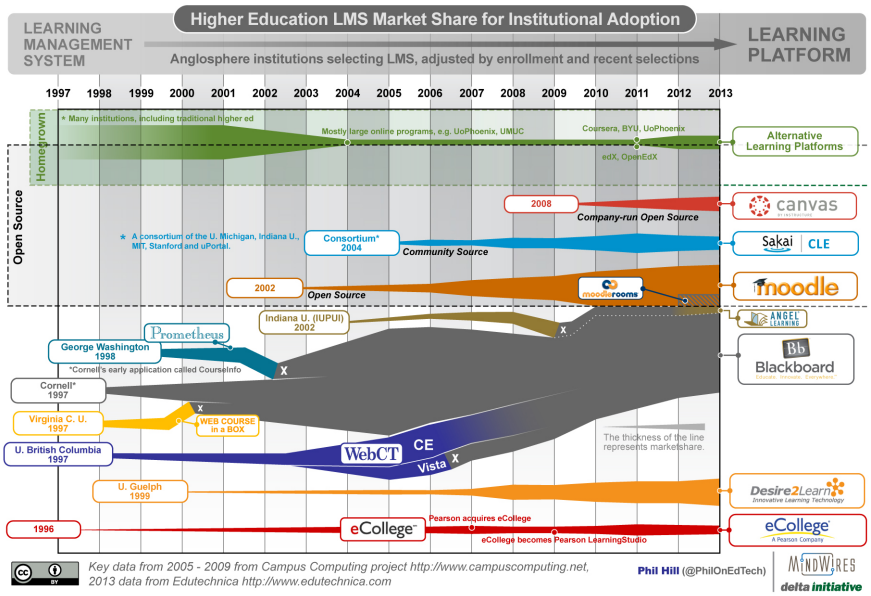


Figure 2.1: Higher education LMS market share 2013 [1]

Not only are these tools useful for supporting e-learning, they also collect a large amount of digital data that reflect ways in which students use the LMS. In other words, data collected by these systems serves as a proxy to gauge students' learning behavior. And, by analyzing this data, we can extract useful insights that can be used to improve students' performance and give better understanding on students' learning process to stakeholders which will ultimately lead to better managerial decision making.

Education Data Mining (EDM) and Learning Analytics (LA) are two different research communities that are exploring the extent to which historical data can be used to improve the quality of education delivery [27]. According to R. Baker and others [28], EDM is about "developing, researching and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist". Techniques from the data mining field can be used to extract useful knowledge from a large set of LMS data. Some of the examples of the application of education data mining include social network analysis, discourse ana-

lysis, predictive modeling, clustering, pattern mining, intelligent curriculum development and recommendation of contents amongst others.

Following are the other applications of EDM used in literature.

- **Student Modeling** Student modeling is used to characterize the learner and detect state, such as: cognition, emotion, domain knowledge, learning approach, achievements, learning preferences, skills. Student modeling is beneficial to adapt the teaching approach according to the learner's requirements.

- **Student behavior modeling** Student's behavior modeling is to detect the diverse behaviors of learners, such as: sleeping, motivation, guessing, inquiring, requesting help, willingness, access and response times, gambling etc. The goal is to adapt the system according to the behavior of learners.

- **Student performance modeling** Student performance modeling is one of the main targets of EDM approaches. The goal is to predict how well a student can accomplish a specific task in an optimal way. Task can be any learning goal or a response to a particular learning situation. Performance indicators used in literature are efficiency, evaluation, achievement, competence, elapsed time, correctness, deficiency, resource consumption etc.

- **Assessment** The target of this approach is to supervise and evaluate the learner's knowledge, skills and achieved outcomes. To analyze student's proficiency using static and dynamic testing and through offline and online assessments.

- **Student support and feedback** This approach is directly related to the improving the learning process. During the interaction of learners and system, student's support

is provided to improve the learning process, or to correct the misconceptions, bugs or faults, or to give suggestions during the decision process.

- **Curriculum, domain knowledge, sequencing, and teachers support** Curriculum design is the most laborious task for the teachers, where they have to author, seek, adapt and sequence the contents and design curriculum before giving instructions to students. Customized curriculum and teaching practices help learners in the acquisition of domain knowledge. Contents of the curriculum represents domain knowledge repositories and cognitive models of knowledge components to learned and skills to be trained. The curriculum content is sequentially delivered to students, and comprises regular evaluation options to enhance the teaching-learning experiences.

- **Tools development.** Another contribution of EDM is the design of tools that can perform much of the laborious tasks in the context of knowledge discovery. Because of the diverse nature of EDM, several diverse EDM tools are found in literature. Following are the main categories of tools developed in EDM.

  - **Extraction:** Supports the search, representation and storage of raw data from system to a suitable format that can be mined.
  - **Learning support:** Facilitates knowledge acquisition to solve course-related problems.
  - **Feature Engineering:** Analyzes and presents supporting features that can be mined.
  - **Visualization:** Supports the mining process and present the results of analysis for interpretation in an easy to understand format.

– **Analysis Support:** Deploys additional functionalities in evaluating student behaviors during their interaction with the system to help them develop cognitive skills for solving course-related problems.

Table 2.1 show the examples of tools developed in the domain of EDM.

Table 2.1: Description of Tools (EDM)

Ref.Id	Type	Name	Purpose
[29]	Extraction	ExtractAndMap	Depicts and deploys functionalities concerning data extraction from LMS
[30]	Extraction	Java desktop Moodle mining	Facilitate the extraction of log data and the execution of data mining process
[31]	Learning support	LQGen	Automatically generates proof problems that support and satisfy the conceptual requirements of the course instructor.
[32]	Feature engineering	Workbench	Suggests appropriate features related to the behavior of students
[33]	Visualization	EDM Vis	Facilitate the exploration, navigation and understanding of student's log through tree structure
[34]	Visualization	LiMS	Captures the data which describes the engagement of student with online learning environment
[35]	Visualization	Curriculum Customization Service	Facilitate online curriculum planning and observe the behavior of teachers
[36]	Visualization	Meerkat-ED	Visualizes snapshots of participants in the discussion forums, their interactions, and the tracking of the leader/peripheral students
[37]	Visualization	e-Learning Web Miner	Discover student's behavior and the way they navigate
[38]	Analysis support	Web-log based	Evaluates pedagogical processes occurring in LMS settings and student's attitudes
[39]	Analysis support	Continuous Improvement of e-Learning Courses	Uses association rule mining and collaborative filtering to make recommendations to improve e-learning course
[40]	Analysis support	Check My Activity	Support students to compare their own activity in Blackboard vs their peers(anatomized)
[41]	Analysis support	Brick	The idea is to develop pedagogical agents that monitors learner behavior through their actions to identify their behavior.
[42]	Analysis support	SIENA	The purpose to assess the abilities and knowledge of students and the other is to guide students self-study and self-evaluation to optimize focused learning.
[43]	Analysis support	eLAT	Enables teachers to explore and correlate content usage, user properties, user behavior, and assessment results through graphical indicators
[44]	Analysis support	DMOBE	It helps to extract learning patterns from student's performance it also helps tutor to mine their data to improve course optimization.
[45]	Analysis support	CurriM	Analyzes student and education responsible perspectives on curriculum mining and shows the achievements of a project interested in developing curriculum

2.3 Predicting students' academic performance

Predicting students' performance is one of the most researched topics of EDM. Improving retention rates has become a priority within the higher education domain. Despite numerous

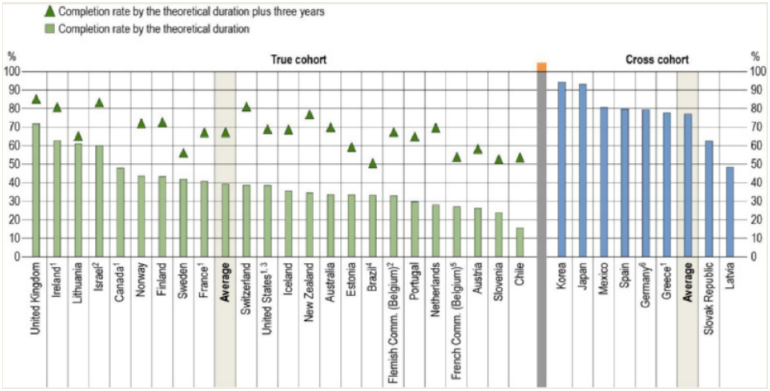


Figure 2.2: Completion rate of full-time students who entered a bachelor's or equivalent programme (2017). Source OECD (2019) Retrieved from https://bit.ly/35wpV9t Accessed 14th October 2019.

efforts, retention rates in higher education remains low [46, 47]. According to an OECD (Organization for Economic Cooperation and Development) report [48], "On average, 12% of students who enter a bachelor's programme full time, leave the tertiary system before the beginning of their second year of study. This share increases to 20% by the end of the programme's theoretical duration and to 24% three years later (Figure: 2.2). Another alarming fact is the percentage of students who no longer continue their enrollment after their first year of study. The OECD report notes that percentage range from 6% in the United States to 20% in Slovenia and the French community of Belgium (Figure: 2.3).

This is a concern for higher education institutes because universities' long-term financial security is undermined if students leave the system without graduating. Higher education institutes are funded by govt based on completion outcome (i.e., by the graduates produced rather than students enrolled). Therefore, there is a constant pressure to increase the quality of education by improving the enrollment and retention rates. Higher education addresses the increasing emphasis on completion by developing strategies for supporting students and ensuring degree completion. Research shows that academic success is a key factor to improving the retention rates [12, 13]. One way to improve retention rates is to make early prediction

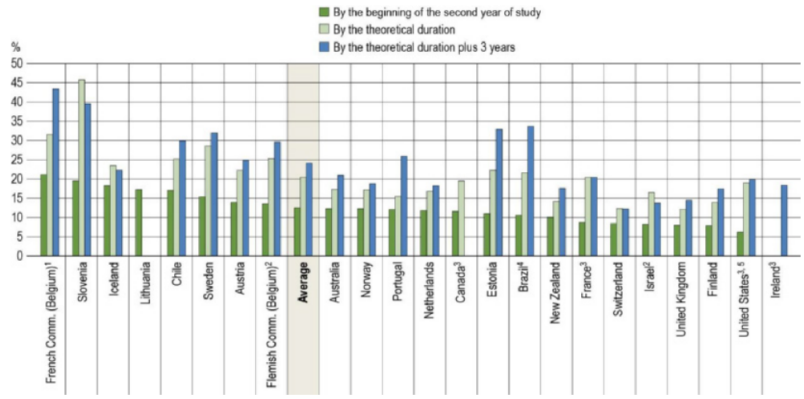


Figure 2.3: Share of full-time bachelor's students who are no longer enrolled in tertiary education (and have not graduated) at various time frames after entry (2017). Source (OECD 2019). Source OECD (2019) Retrieved from https://bit.ly/35wpV9t Accessed 14th October 2019.

of those students who are at-risk of failing the course [14][15][16] and to implement early intervention procedures.

The problem of predicting student performance could be stated as follows. A student pursues many activities to learn subject related concepts over the course duration. As the student interacts with the learning management system, they leave their digital traces. Eventually, each student has to take a final exam. Those students who reach a fixed threshold will be considered to have passed the course. Therefore, the problem is to predict the final outcome of the student based on the activities undertaken during the course. Data traces left collected by the e-learning platform like LMS, student management system (SMS) or any other related data to students have the ability to facilitate making this prediction.

The ultimate goal of prediction is to determine that a given student will be at-risk of failing the course or not in future. This task is performed using supervised learning (classification or regression) method where inputs are students' demographic data, assessment scores or interaction data from education tools (LMS) and output indicates the final outcome as pass or fail, at-risk or not at-risk, successful or not successful. The output variable could be a

Table 2.2: Search process.

Search Terms	(Prediction OR predictive model) AND (Success OR dropout OR at-risk) AND (Learning analytics OR analytic) AND (Machine learning OR education data mining).
Database	Science Direct ACM Digital Library Journal of Learning Analytics IEEE explorer
Fields	Title Abstract
Article type	Review Article, Conference Research

discrete or continuous variable. In case of a continuous variable, regression methods are used to predict the final mark in the course and for a discrete variable, classification methods are used.

This section provides details of systematic literature review which is guided by the following research questions. *Which type of data are commonly used for students performance predictions? What machine learning methods are considered useful for making accurate predictions that are related to student performance? And, What kind of evaluation measures are used for results analysis?*

The Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [49] methodology was followed for conducting this review. An extensive literature review on applications of machine learning in prediction of students' performance enabled in screening a list of works related to the chosen topic.

First, the scope of topic was defined, which is *prediction of students performance through data mining/machine learning*. Students could be across colleges or universities and predictions could be performed course-wise or degree-wise. Thus, terms were defined to search and

database to search for. Table 2.2 shows the terms and selected database. Selection of the database was dependent on the domain. Papers were browsed through related journals that are more relevant to the domain, which includes Science Direct, IEEE Xplore, Google Scholar and Journal of Learning Analytics. Keywords used for searching included but not limited to following: prediction, predictive model, machine learning, drop-out or at-risk or success, performance prediction, learning analytics, education data mining, learning management system and combination of these keywords. A breadth first search was conducted across searching libraries, after which each paper was physically screened to make decisions on keeping it or not. For each selected paper, their related work and papers that cited the selected paper were also reviewed for relevance.

This study investigated features that are co-related to the performance of the students. Search was limited to the papers where predefined terms were found in title, abstract or keywords. This allowed focus on the title and abstract of potential papers for identifying their relevance. Though conference papers were included as well, the focus was on research articles that were published in the related journals. Timeline restrictions were also imposed, whereby papers that were published before 2008 were not considered.

The last step involved a careful reading of identified papers and recording of features used for prediction, machine learning or statistical methods used, evaluation measures applied to test the predictive model and the sample size used by authors. After reading some papers, a pattern of features used in prediction were found, so categories of features were made as follows: academic features, demographic features, pre-academic features and virtual learning environment related features (details of each type are discussed in section 2.3.1). The same process was applied for noting the types of machine learning methods used and evaluation measures.

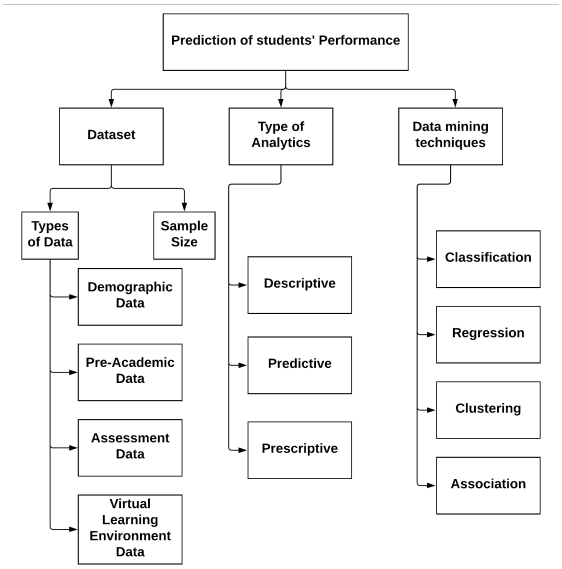


Figure 2.4: Documentation of the search results in selected databases.

In summary, each paper was reviewed in detail and examined for following characteristics: types of problem addressed, types of features used for prediction, types of machine learning method used for prediction, types of evaluation measure used for testing accuracy of predictive model, sample and attribute size used in the study. Documentation scheme for literature is shown in Figure 2.4.

Table 2.3: Inclusion and Exclusion criteria.

Inclusion criteria	**Exclusion criteria**
Those studies were focused that used machine learning or statistical method on students data to predict student performance either in course or overall degree program.	The papers that used MOOC datasets were excluded, as context was based on blended learning rather than completely online learning.
Peer-reviewed papers published during the period of 2008-2018 were included.	Workshops and posters were excluded, only full conference papers were included.
Studies with quantitative results.	Studies that do not present empirical data.
Full length articles published in impact factor journals.	Short conference papers or workshops

2.3.1 Types of Data

This section draws upon findings from literature reviews to answer the first research question which is *Which type of data are commonly used for prediction of students' performance?*

Student data was collected from a variety of sources. This study focuses on studies that utilized data collected from virtual learning environments like LMS and SMS, or data that are related to the previous academic scores. This form of data has been widely used and is considered to have a great impact on prediction of students' performance. Extensive research is done to find the correlation between the students' data with their performance. Students' data that are considered as academic include cumulative GPA, grades in pre-requisite course, grades in assignments/quizzes or age, ethnicity, gender and financial status amongst others. For the sake of simplicity, the student data were divided into four further categories: demographic data, pre-academic data, academic data and data extracted from virtual learning environment. Following sections reports on some key findings from the data categories:

Demographic Data

Individuals are shaped by their surroundings; therefore, features like family income, lifestyle, family environment, parents' education, community and race are important. These demographic factors influence their education in similar manner as it impacts other aspects of life. Demographic profiles comprising information like gender, age, marital status and ethnicity have been highly used in connection with students' performance [50][51][52]. There are many researchers that argue specifically on gender-based difference between the performance of students [53][54][55][56][57] [27]. These researchers argue that some demographic factors can affect academic performance of students at different study levels [58][59][60]. Other demographic characteristics that are used in literature are family income, socioeconomic status,

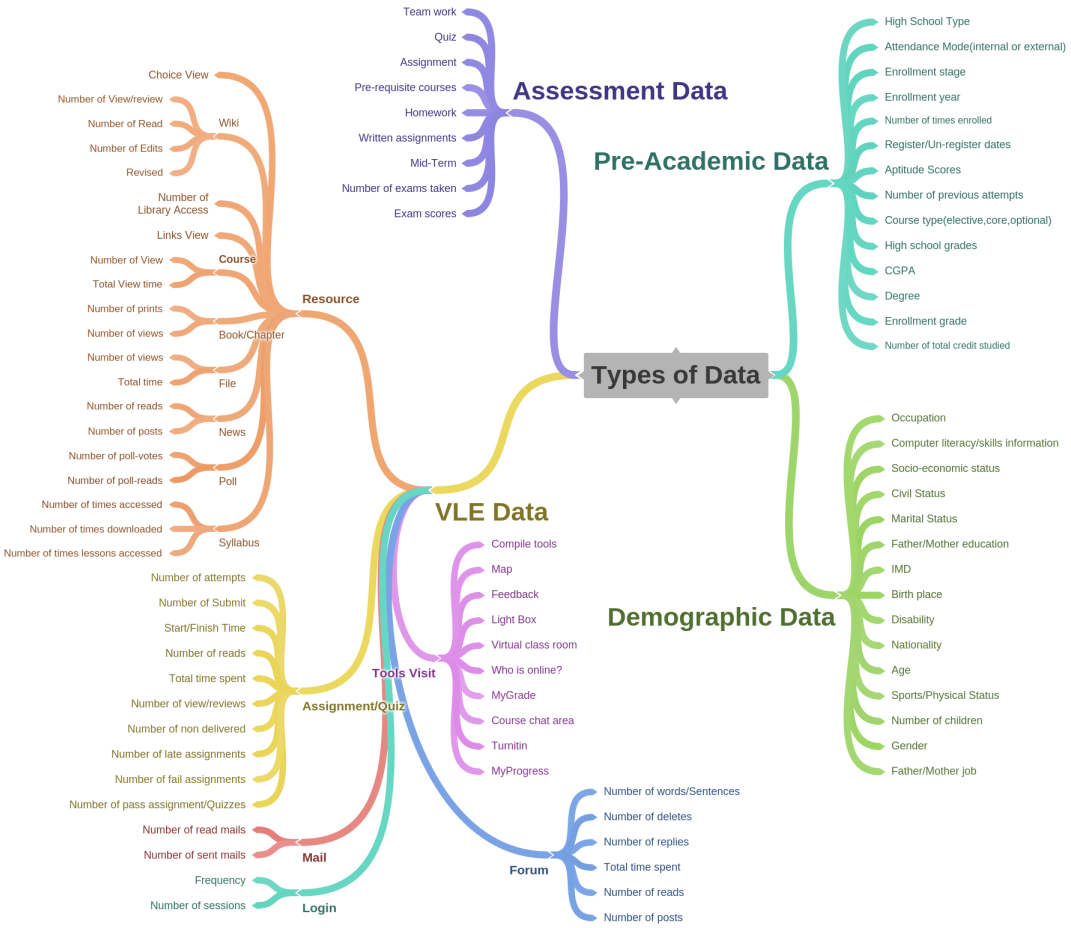


Figure 2.5: Snapshot of students' factors with relevant attributes.

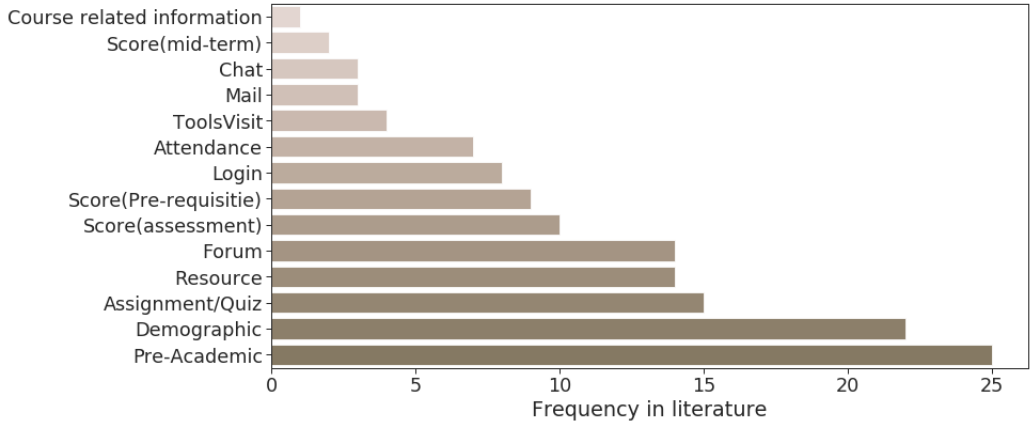


Figure 2.6: Most used features (Type of data) in literature.

race and ethnicity [61][62][63][64][65].

Pre-Academic Data

Few researchers have utilized background academic information related to students such as their CGPA, grades earned in high school, type of school (private, public), language scores, SAT scores, enrollment process etc. Many studies show that pre-academic information is helpful in predicting a student's future [64][66].

Virtual Learning Environment (VLE) Data

Students learning behaviors are monitored using VLE log of activities and have been used to assess their overall performance. Such activities include: visit to course contents, number of tools visited, number of messages read, post and reply, number of times and duration of sessions, number of login and logouts, number of mail messages sent, read etc. Most of studies have utilized the log of activities stored in VLE to make predictions of students performance [51][50]. Further division of VLE data is shown in Figure 2.5.

Assessment Data

Many researchers have utilized assessment scores that are available during the semester, since they are directly co-related to the performance of the students [67][66]. Examples of assessment data include: quiz scores, assignments, team projects, mid-term exams, written assignment etc. In most of the studies these are considered stronger predictors of students' performance.

Figure 2.6 presents the most used features that have served as inputs to predict students' performance. The top five input features identified as the most commonly used in prediction are: pre-academic data, demographic data, and in category of VLE data, features related to assignment and quizzes, features related to resource use and forum are abundant in literature.

2.3.2 Data mining techniques

This section draws upon findings from literature reviews to answer the second research question *What machine learning methods are considered useful for making accurate predictions that are related to student performance?*

The scope of our analysis is to determine the current methods used and discuss the limitation of the studies. Machine learning algorithms used in review are divided either as classification [68][69][70] or as regression methods [71][72][73]. Some of the research studies have used correlation analysis to measure the correlation of features with the final performance of students [74][51]. Few studies have used voting methods to combine decisions from different methods (e.g., ensemble method, stacking, etc) to further improve the performance of classifiers.

Rule-based classifiers

Rule-based methods are a type of machine learning algorithm that represent knowledge in a form of rules. Interpretability of classification results are favoured by these methods. Both numeric and categorical type of data are suitable for such analysis. Some of the methods used in literature are: DTNB (Decision Table Naive Bayes) [75], JRip [76], Nnge [77, 78] and Ridor [79]. Most used methods in the category of rule-based classifiers was JRip [67][80][81][82].

Tree-based Classifiers

Decision tree methods are widely used in literature. Knowledge extracted from data analysis results in a form of tree which is reasonably interpretable. Tree based methods are considered suitable for supervised learning with high accuracy and predictive power. These methods are used for both classification and regression and can map both linear and nonlinear relationships. Models are developed using features that are more informative and involves a

process of feature selection. Therefore, many studies use this method to identify important features. Following are some examples of tree-based algorithms used in literature; J48/C4.5 [83], LADTree [84], ADTree [85], CART [86] and Random Forest [87] etc. Random Forest classifier is the most commonly used method in literature.

Function-based Classifiers

Logistic Regression [88][89],[51][68], Multi-layer Perceptron [67] [66], RBFNetwork [90], SVM [91] [92] and SMO [66] [52] are examples of function-based algorithms, that extract knowledge in the form of a mathematical function. These algorithms are widely used in literature, with logistic regression being one of the most popular method found in literature.

Bayes-based algorithms

Bayes-based classifiers are based on the Bayes theorem of probability and referred to as probabilistic classifiers. One important assumption is that the features are mutually independent which is rarely satisfied. Examples of such classifiers includes; Naıve Bayes [93] [94] [70], Bayest net and Naïve Bayes simple [92].

Ensemble Classifiers

Some studies make use of ensemble classifiers [67][95][96] to enhance the performance of other classifiers. An ensemble model is created using more than one classifier and the final prediction is often made based on the majority voting scheme.

Instance-based Classifiers

Instance-based classification algorithms perform their main learning process at the instance level. They try to approximate a function that assigns class labels to instances. The instance classifier is combined with an underlying assumption, which links the class label of

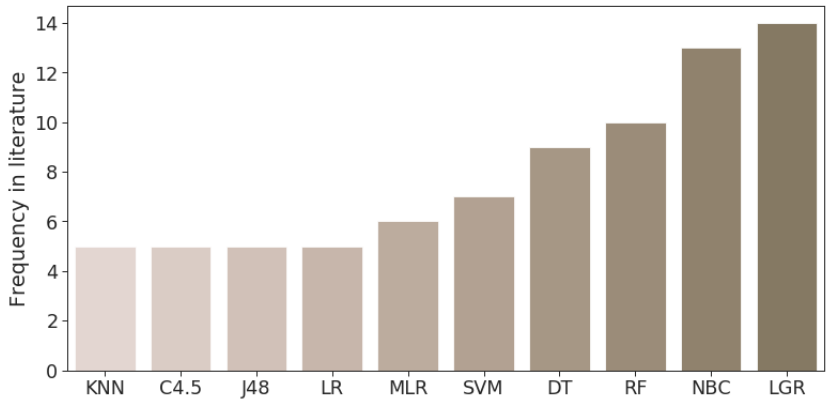


Figure 2.7: Top 10 most used methods in Literature.

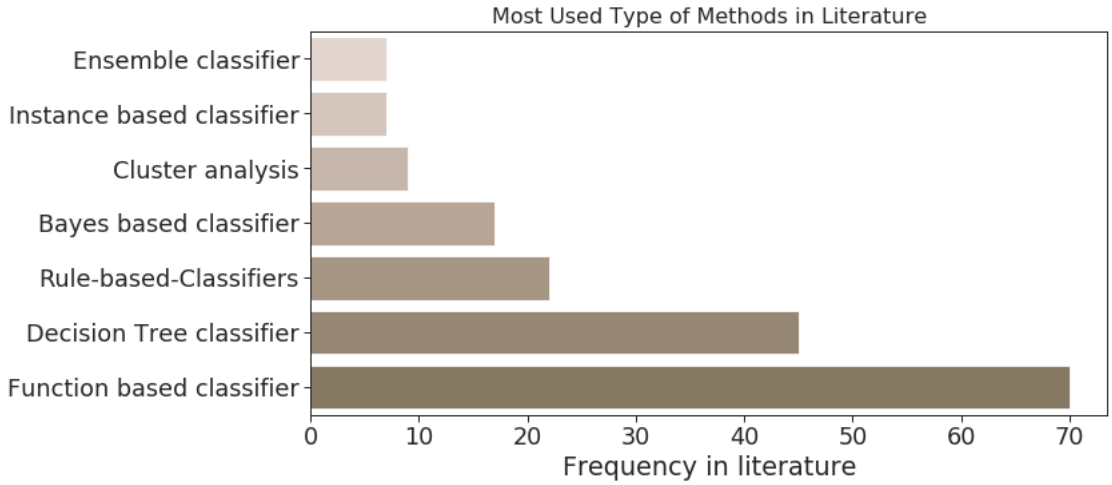


Figure 2.8: Most used methods in literature.

instances inside a bag with the bag class label. Examples of instance-based classifiers include: IBK [97] [82] and kNN [95][70] [98]. Researchers have used different machine learning algorithms for developing predictive models such as Naïve Bayes, J48, EM, Support Vector Machine. A summary of the methods used for predicting student performance is given in Table 2.4. Top ten methods used in literature are shown in Figure 2.7. Logistic regression (LGR), Naive Bayes classifier (NBC), Random Forest (RF) and Decision Tree (DT) are more frequently used with more than 30% studies using these (Figure-2.7).

Following methods have been used in literature for improving accuracy.

- **Ensemble Techniques:** Some studies have used ensemble techniques to enhance the performance of standalone classifiers. For example, Hu, Lo, and Shih [50] used AdaBoost in conjunction with CART, LGR and C4.5. AdaBoost [85] iteratively applies a classifier on each instance of the dataset. Misclassified instances are weighted higher for next iteration of learning, thereby the machine learning process focuses on learning more difficult samples. In the end AdaBoost makes use of a weighted majority voting for prediction. AdaBoost combined with CART classifier showed a higher performance.

  Another study [67] by Marbouti, Diefes-Dux, and Madhavan, used ensemble models by training more than one classifier (NBC, KNN and SVM) on dataset and predictions were done based on the majority vote of the classifiers. In order to improve overall accuracy, authors suggested combining models with low false negative error (NBC and SVM in example) and false positive errors (kNN) using an ensemble. Other ensemble approaches used are bagged trees, adaptive boosting trees and random forest [64].

- **Feature selection Method:** Few studies used feature selection methods to improve the performance of predictive models. Feature selection is the process of selecting subsets of features that have more predictive power and are strongly related to the predicted variable, and therefore are likely to improve the accuracy of models [99]. For example, Marbouti and others [67] used the Pearson correlation coefficient to extract the subset of features that are more related to the predicted variable(pass/fail) and only the correlation coefficients of more than 0.3 were considered. As a result of using feature selection method, accuracy of the models improved.

- **Data pre-processing:** Other approaches used in literature to improve the accuracy is data pre-processing [62]. Data pre-processing steps involve techniques such as treating

missing values and outlier detection. Data pre-processing has a significant effect on improving the accuracy by handling unreliable instances which could degrade the learning of algorithms.

- **Hyper-parameter optimization:** Hyper-parameters refer to the parameters of a machine learning algorithm used to build models (e.g., number of the decision tree, number of neighbors in kNN, etc.). The aim of hyper-parameter tuning is to find the set of parameters that give the lowest error rates on the test sets. In most of the studies in literature, either default values of parameters were used, or the studies did not discuss this aspect in detail. One study [62] analysed the effect of fine-tuning of EDM techniques when predicting students performance. According to their findings, fine-tuning of the techniques increased the final accuracy significantly.

2.3.3 Evaluation Metrics

This section draws upon findings from literature reviews to answer the third research question *What kind of evaluation measures are used for results analysis?* In order to evaluate the performance of a predictive model there are a variety of measures which can be used. A rigorous evaluation of prediction results is a critical step in the machine learning process. Classification accuracy is the most common metric used in literature for evaluating models which is defined as the percentage of correct classifications [67][69][81]. Predictive models with high accuracy are desired but due the simplicity of the measure, it may not fully convey cases where the accuracy between positive and negative classes is not the same. Therefore, simply accuracy measures are considered biased in situations where data is imbalanced between the number of samples in different classes. Many studies [100][74][82][101] have used other metrics that address the classification error of both majority and minority class. Other metrics used

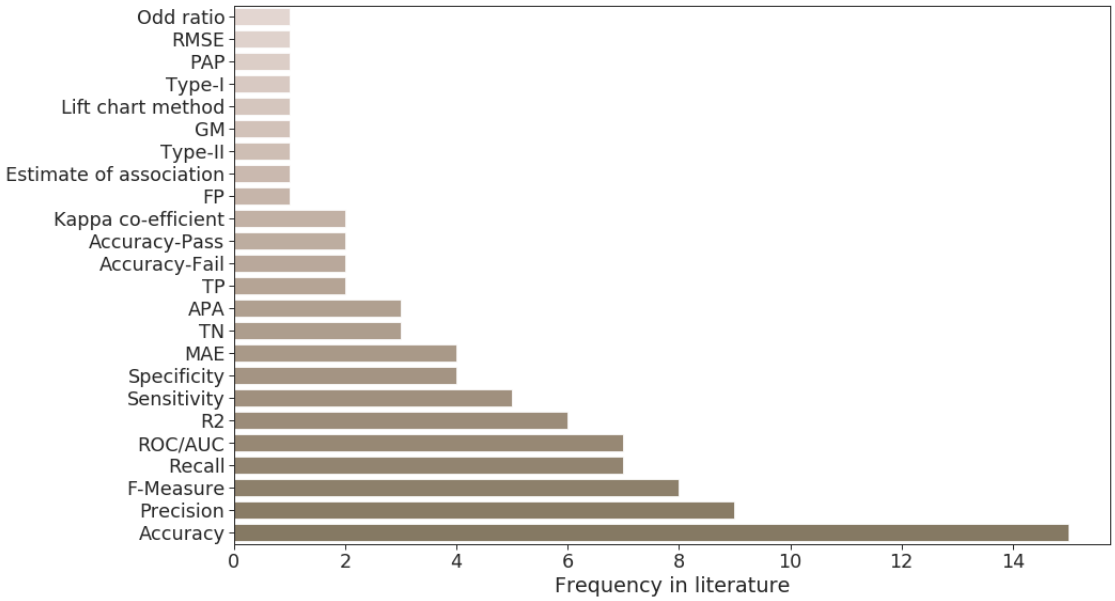


Figure 2.9: Evaluation measures used in Literature.

for classification are determined using the confusion matrix that includes true positive (TP), true negatives (TN), false positives (FP) and false negatives (FN).

Type 1 error is the probability that a student who is not at-risk of failing the course is misclassified as at-risk of failing the course. High Type I error means interventions are being made for students who are safe and do not need special attention and thus results in wastage of resources. Type II error is the probability that a student who is at-risk of failing the course is considered as a well performing student or not-at-risk of failing course. If the errors are high, then the system will not be capable of identifying at-risk students and therefore will be of little use as an early warning system. In majority of datasets, students who failed the course were the minority class. The goal of predictive models was to identify the students who are at-risk of failing the course with high accuracy. Therefore, some studies [68][70] considered accuracy of both positive (at-risk) and negative(not-at-risk) classes separately. Examples of such measures are recall, precision and false positive rate. An ideal classifier has a higher

value for recall (100% ideal i.e., 1) which measures the ability of the predictive model to identify students who are at-risk with 100% accuracy and zero false positive rate (i.e., not to predict any not-at-risk student as at-risk). Precision takes care of a second class or not-at-risk students failing the course.

Most common metric used for continuous variable prediction is R-square which represents the proportion of variance in the dependent variable which is explained by the independent variables [102]. Other metrics include Mean Absolute Error (MAE) and Root-mean-square error (RMSE) [103][104][73][91]. Average accuracy of a model for both classification and regression problem are measured using Average Prediction Accuracy (PAP) and Average Accurate Prediction (APA) [65][66]. Details of the evaluation measures used in literature are shown in Table 2.4.

2.4 Process Mining in Education

Process Mining (PM) is one of the techniques used in EDM which is process centric by nature [105][106]. PM uses event logs that are recorded in the information system and extracts process-related knowledge. The primary objective of PM is to discover, monitor and optimize the underlying process in a variety of application domains. Event logs used in process mining are considered as sequence of cases, where each case refers sequence of events and comprises of an event, activity and a time stamp. The result of process mining activities and analysis are process models or hierarchical flows that are visualized by powerful tools to better understand the processes [107].

In the context of LA and EDM, very few process-oriented approaches focus on the process as a whole [108]. Educational PM can be valuable for obtaining a better insight on the underlying educational processes. PM in education can be applied to construct educational

process models through model discovery of the observed behavior. Later, the obtained model can be used to project the information extracted from the logs and to check whether the observed behavior is reflected by the model, leading to different kinds of decision support [109].

There are various studies showing the potential of PM in the educational contexts. For example, Pechenizkiy et al. [110] used application of process mining tool like Prom in the context of educational data mining. The basic methods of process mining i.e. process discovery, process conformance checking, process performance analysis techniques were applied on on-line assessment data. The impact of immediate feedback with strict order of questions vs no feedback but flexible order of questions were observed in process mining context. In [109] a framework is introduced for integrating the domain knowledge in educational PM in order to facilitate interactive PM and help educators analyze educational processes based on resulting process models. In this work, the authors introduce and formalize patterns in an academic curriculum describing the possibilities and constraints of students. Having a process model can help students be aware of what they need to do, help the educators to check whether the curriculum was respected, and facilitate real-time detection of curriculum violations.

So far, the presented studies focused on a single process for the purpose of process model discovery from the learners' interaction data and conformance checking of the logs with the obtained model. Nevertheless, there are few efforts on comparing the process models and analyzing the differences from an educational point of view. For instance, Bannert et al. [111] have analyzed student processes of self-regulated learning and related the discovered process models with theories of self-regulated learning and meta cognition. In this study, the differences of learning behavior between successful students and less successful ones are highlighted via their temporal patterns. The authors explain that the successful students

perform regulatory activities with a higher frequency and in a different order than the less successful students. Also, the types of self-regulated learning activities of the two groups differ. In this study, the comparison of the two groups of students is mostly performed in a descriptive manner.

A recent study has proposed a numerical approach on comparative process mining using process cubes where events and process models are arranged in various dimensions [112]. In this study, the attributes are referred as the dimensions of a cube where the process models can be produced and compared.For instance, the authors compared the process of passed versus failed students of a course based upon data related to watching of video lectures. The results indicate that the average trace fitness for all students that passed was significantly higher than the one of the failed students. However, the results are not presented as a generalization but rather a starting point for a better analysis of learning processes. PM can be used in combination with other data mining methods to improve the discovery of the models with a better performance and comprehensibility.

In [113], PM is applied in combination with clustering to improve the obtained educational process models in a study where the students followed an online course using Moodle (a learning management system). In this work, first the cluster analysis is performed to group the students based on their Moodle usage data and their grades, then PM is applied on the clusters separately to obtain the process models, resulting in a better performance / fitness and comprehensibility / size.

2.4.1 Data Quality for Process mining

This section presents previous work done in the domain of data quality for process mining analysis. Process mining manifesto [114] created by the IEEE task force on process mining

provides a mechanism to measure the quality of the event data and to rank them (from 1 star to 5 stars). Best quality event logs (5 stars) are characterized as complete and trustworthy as they are recorded automatically, whereas poor quality (1 and 2 stars) event logs are incomplete and considered not as reliable since they are mostly recorded manually. The manifesto suggests that only event logs rated from 3 to 5 stars are suitable for process mining analysis. Four broad categories, namely missing data, incorrect data, imprecise data and irrelevant data are common issues that affect event log quality when used for process mining [115]. Such data inconsistencies and data anomalies pose huge challenges for the analyst as process-mining is data-driven. Processes provide a workflow perspective where various tasks require time-ordered operational data that can be queued to demonstrate dynamic behaviors. Calvanese et al. [116] used an ontology-based approach to extract event logs from a relational database. The data stored in databases are flattened as XES file using both domain ontology and event ontology. Although, this technique can provide access to the data in databases using query unfolding and by applying ontology-based data access (ODBA) methods [117], however performance issues can occur when dealing with large databases.

Many studies have highlighted different challenges that are faced during the extraction process of data from a process-unaware system. For instance, Van Der Aalst [118] presented a different approach to transform relational database into event logs. They used classification to create multiple event logs for making comparisons; however, there was no consideration to business-related decisions, rather their study mainly discussed theoretical challenges of extracting event logs. Real world business operations are driven by external environments, organizational policies and managerial decisions as businesses strive to have a competitive advantage. Another study by Pérez-Castillo et al. [119] created event logs by extracting logs from a non-process-oriented system based on correlation of events and with similarity

between attributes. Jans and Soffer [120] also discuss issues in extraction of event data from a relational database using an end-user's perspective. Using an example of procure-to-pay process, they described a structured procedure to extract data from a relational database and convert to an event log. Even so, they stress on the multiplicity of decision-making factors that influence the selection of process instance and associated activities, which can impact the quality of the event log.

Further, Selig [121] demonstrated how results of process analysis of data produced by process-unaware systems are different than the one produced by process mining system. Using an example of purchasing process from a SAP module of an enterprise system platform, Selig shows the emergence of ambiguous cases such as divergence and convergence. The author proposes continuous data extractions from the enterprise system and transforming it into an event log by considering proper case attributes, the notion of a case and the granularity of events to address data quality issues. Kim et al. [122] presented a detailed taxonomy of 'dirty' data as a framework for understanding the origin of such data. They proposed a metric for measuring the quality of data and have referred to data as 'dirty' if the results of data analysis are not up to expectations due to the low quality. Moreover, their study explored the impact of such data on data mining and provided techniques for dealing with unclean data. A different approach is to consider data quality problems in relation to its source or origin (i.e., single-source versus multi-source) and its granularity level (i.e., at instance-level or at schema level). Rahm and Do [123] emphasis the use of transformation techniques for cleaning of data to cover both instance and schema perspectives in an integrated manner and have presented commercial tools with their limitations for data cleaning. While these data quality taxonomies cover some of the problems that have an impact on process mining analysis, they are not completely related to the process mining perspective. Suriadi et al. [124] used pattern-

based approaches to identify common data quality problems which were distilled from their experience and labeled them with pattern terminology. Their patterns are validated with use of event logs from practice and have been evaluated by research experts in the domain to serve as knowledge repository for event log preparation and provide recommendations on improving data quality.

Different tools are available to extract event data from database and convert it to XES file. For example, XESame [125] provides a platform where data is selected and matched with XES elements. However, there is no direct access to the database and database is only used for storage purpose. Other similar commercial tools examples are Minit and Celonis. Best part of these tools is high efficiency in data extraction, but the downside is that they cannot handle huge amounts of data, especially if the computer memory is exceeded because the transformation takes place in memory. Similar to previous tools, data cannot be accessed directly from database. To address the issues of memory and have direct access to the database, another technique was proposed by van Dongen and Shabani [126] which used comparatively less memory; however, there is no empirical evidence yet to ascertain its time-performance.

2.5 Summary

Data mining has been widely used in diverse fields including business, finance, manufacturing, fraud detection and marketing [127]. It has also gained much popularity in the field of education and has created many research opportunities for education institutions to provide enhanced capabilities that can meet the 21st century educational needs [128]. First section of the chapter has identified the application of education data mining and listed some of the tools developed to facilitate the education process. In the next section, an extensive

literature review of prior studies that have utilized machine learning techniques to predict student performances by using historical data is presented. By predicting how a student will perform, educators can provide timely interventions to students who may be at-risk of failing the course and consequently the quality of education delivery can be improved. List of all studies that were included in review, are summarized in Table: 2.4. The last section provides a review of the works in the education process mining domain and highlights some of the recent works using education data. Additionally, this review has identified challenges and importance of the quality of event logs in the domain of process mining.

NOTE: Chapter 2 is a partial re-print of following article: The thesis author was the primary investigator of this article.

R. Umer, Susnjak, T., Mathrani, A., Suriadi, S.(2020) Current Stance on Predictive Analytics in Higher Education: Opportunities, Challenges and Future Directions. Interactive Learning Environments. (Accepted)

Table 2.4: Details of literature included in study.

Ref.	Technique	Feature Types	Total features	Sample	Models	Evaluation Measure	Best classifier
[50]	Classification	Assign, Forum, Resource, Login	14	330	LGR, CART, C4.5, AdaBoost	Accuracy, Type-I, Type-II	AdaBoost +CART
[51]	Network analysis, Co-relation, Classification	Assign, Mail, Forum, Resource, Login, ToolsVisit	22	118	LGR, MLR	Accuracy R2	Not stated
[66]	Classification	Pre-Academic, Score (pre-requisite), Score(mid-term),	8	323	MLR, RBFN SVM	APA, AP, R2	SVM
[52]	Classification	Assign, Resource, Forum, Mail, Chat, Demo, ToolsVisit, Pre-Academic	11	32593	RTV-SVM	Sensitivity, Specificity	RTV-SVM
[67]	Classification	Score(Assessments)	06-14	120	LGR, SVM, DT, MLP, NBC, KNN, ensemble (SVM, KNN, NBC)	ACC	Not stated
[64]	Classification	Pre-Academic, Demo,	13-19	2459	SVM, DT, NBC, RF, BagTree, BoostTree	Accuracy, Sensitivity, Precision	RF
[80]	Classification	Pre-Academic, Assign, Forum, Resource, Demo	27	4010	NBC, SMO, J48, Jrip	Precision, Recall, F-measure, Kappa AUC.	Jrip
[69]	Regression Classification	Score(Assessments)	13 or 6	104	LGR, FFNN SVM, PESFAM DTNB, JRip, Nnge, Ridor, ADTree, J48, LADTree, RF, LGR, MLP, RBFNetwork,	Accuracy, precision, recall, specificity	LGR
[81]	Classification Clustering	Forum	10	114	SMO, BayesNet, NB-Simple, Clustering (EM, HierarchicalCluster, sIB, SimpleKMeans, Xmeans, FarthestFirst)	Accuracy, F-measure	Not stated
[91]	Clustering	Pre-Academic, Assing, Resource, Demo	20	136	BART, RF, PCR, KNN, NN, SVM Multivariate Adaptive Regression Splines	MAE	BART
[129]	Correlation	Score(mid-term), Pre-Academic, Forum, Login, Assign	11 or 8	530	Hierarchical regression model. Descriptive statistics and correlation analysis	R2	Not stated
[130]	Classification Regression	Pre-Academic, Assign, Forum, Resource, Demo, ToolsVisit	22	4139	MLR, LGR	AUC	Not stated
[65]	Regression Classification	Pre-Academic, Score(pre-requisite), Assign, Chat, Forum, Demo	36	119,366	Linear regression, Robust linear regression, RF	R2, MAE, RMSE, APA, PAP	RF
[92]	Classification	Pre-Academic, Score(Assessment), Forum, Login, Demo	14 or 19	15150	LGR, J48, SVM/SMO, NBC	Accuracy, Precision, Recall, FP rate	Not stated
[62]	Classification	Pre-Academic, Score(assessment), Demo	14	161-262	SVM, J48, NN, NBC	F-Measure Precision, Recall	SVM
[74]	Co-relation Classification Regression	Assign, Forum, Resource, Login	23	4989	Linear regression, LGR	R2, ACC, TN	Not stated
[61]	Classification	Pre-Academic, Assign, Login, Demo	18	429	J48, LGR, Helinger distance, DT, RF	ACC(Fail) ACC(Pass) Recall Precision AUC	NBC

Table 2.4 continued from previous page

Ref.	Technique	Feature Types	Total features	Sample	Models	Evaluation Measure	Best classifier
[95]	Classification	Pre-Academic, Score(assessment), Assign, Resource, Demo	11	32,593	NBC, KNN, CART, Voting (NBC, KNN, CART),	F-Measure Precision Recall	Not stated
[70]	Classification	Pre-Academic, Scores(pre-requisite)	8	210	DT, Rule Induction, KNN, NBC, NN, RF	Accuracy, Kappa, Precision	Not stated
[131]	Classification	Pre-Academic, Demo	14	3599	LGR	Odd ratio	Not stated
[102]	Regression	Attendance	10	356	Multiple linear regression models	R2	Not stated
[98]	Regression	Assign, Forum,	9	438	ADLinear, PolQuadraticLMS Kernel, KNN,J48/C45, CART, AprioriC, CN2, Corcoran, XCS, GGP, SIA,MaxLogitBoost, SAP, GAP, GP, Chi, NNEP, RBFN,GANN, MLP	ACC	Not stated
[88]	Classification	Pre-Academic, Score(pre-requisite), Demo	11	1789	LGR, NN, DecisionList, BayesianNetwork, DiscriminantAnalysis, J48/C45, CART, Quest, CHAID, Ensemble (Bayesian Network, CHAID, LGR)	ACC	Not stated
[104]	Regression	Attendance, Score(assessment), Demo	17	-	M5, BP, LWR, SMOreg, LR, M5rules	MAE	Not stated
[132]	Clustering	Forum, Assign, Resource	6	140	EM (cluster), K mean(cluster)		Not stated
[82]	Classification	Pre-Academic, Demo, Attendance, Score(pre-requisiite), Score(assessment)	61	419	ICRM, NBC, SMO, IBK, JRip, J48/C45	ACC, TP, TN, GM	
[133]	Classification	Demo, Score(Pre-requisitie), Pre-Academic	8	11496	LGR, ANN, DT, RF, LGR + ANN, DT+RF	ACC	Not stated
[89]	Classification	Score(Pre-requisite), Attendance, Pre-Academic	3+	293	LGR	ACC, Sensitivity, Specificity	
[72]	Regression	Pre-Academic, Demo, Attendance, Resource, Score(assessment)	18	197	J48/C45, AODE, KNN, NBC	ACC	KNN
[134]	Classification	Score(pre-requisite), Attendance, Score(assessment)	8	50	Decision Trees		
[135]	Classification Clustering	Demo, Pre-Academic	18	3360	Rule Induction, Naïve bayes, K-Mean Clustering	ACC	Not stated
[68]	Classification	Demo, Pre-Academic, Resource, Toolsvisit, Chat, Mail, Forum, Login, Score(assessment)	106-560	365-2498	LR, NBC, SVM-R,XGB, LR-W, RF, SWM-W-R	AUC, F-measure, Precision, Recall	XGBoost
[136]	Classification	Resource, Forum, Assign	20	754-9679	LR (Multinomial), LR	Estimate of association	
[73]	Regression	Scores(assessments)	4	60	FFNN, MLR	MAE	NN
[101]	Classification	Pre-Academic, Demo, Score(pre-requisite)	33-120	629-1532	DT, NBC	TPR, TNR	NBC
[137]	Classification	Resource, Login,	5	1200	SVM	ROC AUC, F1	Not stated
[138]	Classification	Resource, Score(assessment), Assign,	Not stated	Not stated	RF, NBC, KNN, LDA	F-measure	Not stated
[96]	Classification	Pre-Academic, Demo,	29	16,066	ANN, DT, SVM, LGR, RF, boostTree, BaggTree , Ensembles /Information Fusion	Sensitivity	Ensemble
[127]	Classification	Demo, Pre-Academic	12	Not stated	Decision Trees	Lift chart method	Not stated
[139]	Classification	Pre-Academic, Demo, Attendance, Course related information	88	15,833	Maximum likelihood probit analysis	Sensitivity Specificity AUC	Not stated

Chapter 3

Research Methodology

3.1 Introduction

This study has used data analytics techniques in an educational domain to provide deeper insights on student progression through a course, that is, draw out predictions as to which students are at risk of not satisfying course requirements, or are likely to withdraw. Data sources like learning management system (LMS), enrollment management system (EMS) and student management system (SMS) have been used. This chapter provides an overview of the methodologies that have been utilized in the study. The three objectives of the research are:

- To assess the available data to identify which variables/factors are the most meaningful for achieving high predictive accuracy.
- To investigate various machine learning, data mining and process mining methods in the field of education and examine their feasibility so as to identify which algorithm(s), or ensembles of algorithms have higher accuracy and are therefore, most suitable for generating predictive models.
- To collect and assess student logs (e.g., time spent on different course activities) for analyzing their relationships with course achievement. And, to further identify differences between student outcomes based on their background information, online engagement data, prior achievements/grades or current progress in their courses.

This chapter describes the general methods common for all experiments conducted in this research. It describes the datasets, feature types and their extraction process. Training procedures, evaluation methods and the description of the environment and hardware used is also described. Further, specific details of methods used in particular experiments have been

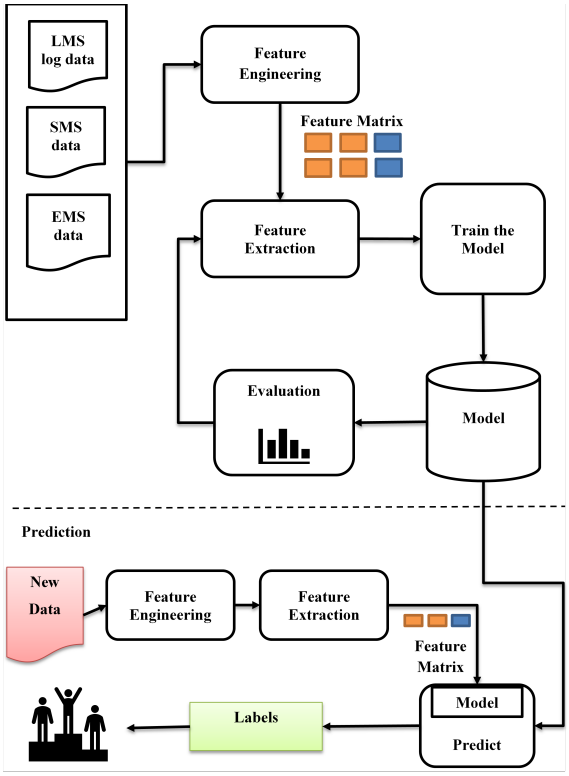


Figure 3.1: Supervised learning: Prediction for classifying students course outcome using LMS, SMS and demographic data.

described in the relevant chapters.

Figure 3.1 shows chain of steps that starts from the raw data and ends in prediction of the student's outcome for each course. As demonstrated in Figure 3.1, first stage involves preparation of the input data. At this stage data is collected from different data sources and converted to the required format before the feature matrix is set for further training stages. In the following section, each step of the process which starts with preparation of data from raw data and feature extraction are described. Three main datasets extracted from different sources were used. The details of each dataset and general methods for developing predictive models are described next.

3.2 Dataset 1: Main Dataset

Data were collected from an Australasian tertiary education provider. The following data sources: Learning Management System (LMS), Enrollment Management System (EMS) and Student Management System (SMS) provided the starting point. The first step involved the task of data extraction from the data source and then acquiring it in the desired format, since the structure and format of LMS data was not directly suited for data mining; Therefore, transformation process entailed data pre-processing including data cleaning, treatment of missing values, transformation into the desired format and integrating data from diverse sources. All individual course-related data were collected from two databases: SMS and LMS. While the SMS data was formatted and included course final scores, number of enrolled students and scores for assessments, the same cannot be said for LMS data. Most of the selected courses were blended courses having options to meet instructors face-to-face to discuss course related problems; although the tertiary provider delivered courses across three different modes, namely, internal mode (where students are taught on a weekly basis on-campus), distance mode (where students are off-campus and mainly use electronic means although some face-to-face opportunity may be provided via one or more contact workshops) and block mode (which refers to internal courses where the class contact time is compressed to four or five full-day teaching). Both undergraduate and postgraduate courses with at least one student registration that have run since 2015 were included. The undergraduate courses were divided into four levels based on their complexity (as UG-1, UG-2, UG-3 and UG-4). Moreover, all courses were taken from different subject areas spread across five different academic disciplines, namely, Business, Arts, Health Sciences, Humanities and Social Sciences and STEM (see Table: 3.1).

Table 3.1: Course distribution according to the demographics.

	Course Level					
Course Demographic	PG	UG-1	UG-1	UG-3	UG-4	Total
Business School	347 (7.971%)	132 (3.032%)	227 (5.214%)	227 (5.214%)	NaN	933
Arts	1 (0.022%)	162 (3.721%)	148 (3.399%)	112 (2.572%)	124 (2.848%)	547
Health Sciences	111 (2.549%)	94 (2.159%)	128 (2.940%)	123 (2.825%)	20 (0.459%)	476
Humanities and Social sciences	471 (10.82%)	281 (6.455%)	307 (7.052%)	275 (6.317%)	40 (0.918%)	1,374
STEM	200 (4.594%)	195 (4.479%)	304 (6.983%)	311 (7.144%)	13 (0.298%)	1,023
Total						4,353

3.2.1 Independent Variables

The learning management system stores logs that detail information of each individual user's activities originating from different tools within the system. User actions such as mouse clicks, content views and keyboard strokes provide information on time spent on a given resource (e.g., quiz, assignment, forum etc), content view counts or frequency of other related activities that may be performed within the system. Log data stores capture real time interactions within the system, that is each module's activity level measure along with temporal information is stored in the log files, which can be grouped by course, by participants and by time.

Each interaction that occurs in the module is stored in a csv (comma separated values) file. Additional information on how many interactions of each type of activity have occurred for each user can be further retrieved by writing relevant queries. This study has utilized tracking variables from LMS as independent variables. These tracking variables collected from LMS are described in Table 3.2. Further, several other variables have been calculated using original dataset. Table 3.3 gives the list of calculated variables. Table 3.3 gives the list of calculated variables. Table 3.4 shows the coding used for independent variables.

In addition to the learning management system variables, the study has utilized information extracted from enrollment management system (EMS). EMS comprises detailed student

Table 3.2: Definitions of independent variables collected from LMS log data

Variable	Description	Type
Forum post created	The total number of forum posts created by the student within the course.	Continuous
Discussion post read	The total number of discussion posts viewed by the student. If a student views the same discussion posts multiple times, the system logs each view as a separate entry.	Continuous
Assignment upload	The number of uploaded assignments	Continuous
Quizzes completed	The number of quizzes completed online	Continuous
Quiz view summary	The number of views of the quiz summary results.	Continuous
Resource view	The total views of source documents uploaded to the system	Continuous
Assignment viewed	The total number of assignments viewed by the student. If a student views the same assignment multiple times, the system logs each view as a separate entry.	Continuous
Assignment submit	The number of assignments submitted by the student.	Continuous
Course content viewed	The total number of times students viewed the course content. If a student views the same content multiple times, the system logs each view as a separate entry.	Continuous
URL s viewed	The total number of URL s viewed by the student. If a student views the same URL multiple times, the system logs each view as a separate entry.	Continuous
Book tool print	The total number of times students used book tool to print.	Continuous
Book viewed	The total number of times students viewed a book	Continuous
First/last access	The day number when a student first and last accessed course materials. This variable is calculated with reference to the course start date.	Continuous
Weekdays engagement	The number of activities students perform in LMS on weekdays.	Continuous
Weekend engagement	The number of activities students perform in LMS on weekends	Continuous
Day time activities	The number of activities students perform with respect to day time. This variable is calculated by dividing the event timestamp into day time (morning, evening, afternoon and night time).	Continuous

Table 3.3: List of calculated variables.

Variable	Description	Type
Success	Academic success was defined as the student complete the course within time and scores greater than 55.	Discrete, interval
Class Size	Class size is calculated by counting the number of students enrolled.	Discrete, interval
Number of Distinct Activities	For each course total number of unique activities(LMS tools) are calculated.	Discrete, interval
Number of Distinct Actions	For each course the total number of unique actions (i.e, view, create, delete, start, submit etc.) are calculated.	Discrete, interval
Total Unique Events	For each course, the total number of unique events (i.e, a combination of unique activities and actions) are calculated.	Continuous, ordinal
Total Events	Count of total events in a course.	Continuous, ordinal

profile information (i.e., their demographic data, high school standing, entrance requirements etc.). The independent variables collected from student management system are shown in Table 3.5. Students grades for assignments, quizzes, project and final exam were collected from student management system. All assessments and their respective scale with respect to final score were also provided. For some courses all assessments were aggregated and stored in one variable.

3.2.2 Dependent Variables

To develop a predictive model, combination of variables collected from LMS, EMS and SMS have been utilized to help in accurately predicting students' academic success within a course. Academic success is measured by using the final grade in the course, which has been defined as a dependent variable. Students final grades were obtained from SMS database. The grades were provided in letter format (e.g. A, B, A+, A-, WD, etc.). Also, the final score for courses were extracted from the SMS database. Due to the differences in grading schemes across different offerings in semesters, two other variables were created from the final score of the course. Students who scored more than 55 marks in the course were grouped as not-at-risk, while all others were considered to be at-risk of failing the course. Students who

Table 3.4: Coding for calculated variables.

Variable	**Coding**
Success	0 = Grade less than 55 1= Grade equal or more than 55
English Equivalency Test	0= English test is not required 1= English test is required
Citizenship	0= Not from New Zealand 1= From New Zealand
Course Size	0= Small (1-21) 1= Medium class (enrolled students= 22-36) 2=Medium large(enrolled students=37-65) 3= Large class (enrolled students= 66-300) 2= Very Large class (enrolled students more than 300)
Distinct Actions	0= Not diverse (n is less than 2) 1= Little diverse(n is equal to 3) 2=Somewhat diverse (n is equal to 4) 3= Very diverse (n is greater than 4)
Distinct Activities	0= Not diverse (n is less than 2) 1= Little diverse(n is equal to 3) 2=Somewhat diverse (n is equal to 4) 3= Very diverse (n is greater than 4)
Course Level	0= Under Graduate Level 1 1= Under Graduate Level 2 2=Under Graduate Level 3 3=Under Graduate Level 4 4=Post Graduate Level

had withdrawn from the course were also included in the at-risk category.

Table 3.5: Definition of independent variables collected from EMS.

Variable	**Description**	**Type**
English Equivalency Test	English test is required or not.	Discrete, nominal
Enrollment Programme Title	Program name for enrolment	Discrete, nominal
Basis for Admission	Recognized entrance qualification for admission.	Discrete, nominal
Citizenship	The nationality of the student (self-reported)	Discrete, nominal
Age	Age of students at the time of admission	Continuous, ordinal
Ethnicity	The ethnicity of the student (self-reported)	Discrete, nominal
Highest School Qualification	Highest school qualification at the time of admission	Discrete, nominal
Gender	The gender of the student (self reported)	Discrete, nominal

Table 3.6: Grading Schema

Not-at-risk								**At-risk**				
A+	**A**	**A-**	**B+**	**B**	**B-**	**C+**	**C**	**C-**	**D**	**E**	DNC*	WD*
90-100	85-89.99	80-84.99	75-79.99	70-74.99	65-69.99	60-64.99	55-59.99	50-54.99	40-49.99	0-39.99	NA	NA

3.3 Dataset 2: MOOC data

The Center for Advanced Research through Online Learning (CAROL) [140] facilitates researchers by providing them with data from online MOOC courses. In Stanford University, data are collected when students interacted with each other or with online materials. For instance, data was related to how learners manipulated controls on video players as they viewed portions of a class, submitted solutions to problem sets and made posts on course message boards or other peer grading activities. Some demographic data was also collected. CAROL has made some of this data available for use to researchers and instructors, to help improve both instructional delivery and provide a basis for gathering more general insights into teaching and learning in digital environments. The protocol of accessing data required an online application that was filled and submitted by the primary supervisor. The researcher team were required to demonstrate awareness on ethical know-how when conducting research on human subjects prior to receiving the CAROL learner data. The researcher team have completed training on human ethics at their university which was submitted in the online application. Research data have next been derived from courses offered over three platforms: NovoEd, Coursera, and Lagunita and a Stanford instance of the OpenEdX platform. These data include raw events taken directly from the course tracking logs and instructional data. Application for data request was approved for following courses: Science Writing Fall (2013, 2014 and 2015) and Economics Summer(2014 and 2015). Table 3.8 shows the list of tables that were included in the CAROL database. Table 3.7 shows the list of tables that were included in the CAROL database.

Table 3.7: List of the requested Table from CAROL database.

Table Name	Description
Activity Grade	Assignment grades; includes right/wrong for each problem part, the learners' solution choice for each answer pard, and the first and last solution submission times.
Video Interaction	An excerpt from tracking logs, focusing on participants' interactions with video
Performance	Daily cumulative assignment performance per learner. This is the average grade over all assignments up to the day before the query.
Demographics	Combines learner self-reported demographic information across multiple tables.
Final Grade	Final grade is an anonymized. FinalGrade contains the learners' grades as computed by the platform at the end of the course.
CourseInfo	Facts about courses, such as start/end dates, academic year and quarter in which a course was offered.
EdxProblem	Metadata on problems offered to learners in courses.
EdxVideo	Metadata on videos in courses.
Event Xtract	A much slimmed view of the OpenEdX tracking log events. The view only includes fields that are currently in use by the platform.
Time-on-task	Estimate for how long participants worked online on their courses. Computed from tracking log events, and partitioned into 30 minute sessions.
User Grade	Raw grades as recorded by the platform.

3.4 Dataset 3: Open University Data

The Open University dataset [141]. was also used in this study. The Open University learning analytics (OULA) aims to predict students' performance at early stages to improve the retention rate. The dataset obtained from Open University comprised seven courses over the duration of 2013-2014 academic year for more than 30,000 students. It included demographic information and engagement data from the virtual learning environment along with assessment scores throughout the semester. Students result in each course has been divided into four categories: pass, fail, distinction and withdrawn. Students are required to get more than 45 marks to pass the course, otherwise they are considered as fail. The distinction students are those who achieved more than 85 scores in their result, while those student who left the course during the semester or unregistered have been categorized as withdrawn. The Open University dataset is publicly available online for use by researchers. List of attributes used are shown in Table 3.8.

Table 3.8: List of attributes in Open University Dataset.

S. No.	Attribute	Description	Type
1	Gender	Gender of the student	Binary
2	Imd_and	Index of multiple deprivations band	Nominal
3	Highest qualification	Highest qualification at the time to entry for course module.	Binary
4	Age_and	Range of students age (0-35, 35-50)	Binary
5	Num-of-prev-attempt	Number of previous attempted to pass the given course	Numeric
6	Studied credits	Number of total credits student is currently enrolled.	Numeric
7	Disability	Student is whether is disable or not	Binary
8	Region	Geographic region where students belong to. Total unique regions were XYZ	Binary
9	date_egistration	The day student registered for the course	Numeric
10	Activity_ype	Number of clicks in the activity performed during the semester in different roles (forum, resource etc.)	Numeric
11	Final_esult	Students result in the course	Nominal

3.5 Data Preparation for Analysis

3.5.1 Removal of extraneous records

The records that had no corresponding final grade in a course was removed from all existing datasets. Following are the situations when the final grade was found to be missing.

- Guest accounts were often created by faculty members within the system. Faculty have full privileges like any other users; therefore, their activities are also tracked. But they have no official standing and so no grades are awarded. It is applicable in situations, when students do not enroll officially but may want to audit the course (i.e., access the course for their own learning).

- Activities of those students who do not continue or withdraw from the course are still stored in log table but with no final grades have been allocated.

- Students who do not continue or have withdrawn were removed from the dataset.

Other records that were deleted included

Other records that were deleted

- Duplicate values for demographic variable as we found that same student was given multiple values of ethnicity.
- The scope of the courses was limited to offerings in years from 2015 to 2017; hence, rest of the courses were deleted from the dataset.

3.5.2 Scaling of variables

Due to the difference in the use of learning management tools by faculty and students, all learning management system variables were transformed to Z scores to get normal distribution. This method is used to scale the variables in order to fall within a range of 0 and 1. All the learning management systems variables will have a mean of zero and standard deviation of one. Therefore, the Z-scores help to standardize data. It takes the value of the observation and converts it to the number of standard deviations from the mean. Following is the formula for calculating Z score. z-Score Formula

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

Here, X is the value of independent variable and μ is the class mean for the independent variable and σ is the class standard deviation for the independent variable. Z score for an item indicates how far and in which direction, it has deviated from the class distribution mean.

3.5.3 Discretization

Considering the skewed distribution of numerical features, they were transformed into discrete ordinal values using 4 levels of quantile score for each variable (i.e,. 25%, 50%, 75%,

and 100%). For the other features with more missing values, were transformed them into binary variables. Additional to this, many categorical variables had several unique values which were divided to form more manageable groups (See Table 3.10 and 3.9).

Table 3.9: Cross categorization of Ethnicity.

Category	**Ethnicity**
European	NZ European/Pakeha, Other European, British/Irish, Italian, Australian, German, Polish, Dutch, Greek, South Slav
Pacific People	Samoan, Other Pacific Groups, Fijian, Tongan, Niuean, Tokelauan
Middle Eastern /Latin American /African	Latin American, African, Middle Eastern
Maori	New Zealand Maori, Cook Island Maori
Asian	Indian, Chinese, Korean, Other South East Asian,Filipino, Sri Lankan, Other Asian, Japanese, Vietnamese, Cambodian
Other	Other, Not Stated, Unknown, Null values if any

Table 3.10: Basis for Admission.

Variable	Categories
New Zealand University Entrance	Graduate Status, Admission with Equivalent Status (Entrance Level), NCEA University Entrance,˜Admission with Equivalent Status (Credits), PhD Approval in Principle
Discretionary Entrance	Discretionary Entrance, Summer Session
Personal Interest	Sub-degree Not yet Matriculated, Adult Admission, Work Experience, Under 20 Special Admission
High School Certificate	Higher School Cert, NZ UE prior to 1986, NCEA Level 2 or 6th Form Cert, NCEA Level 1 or School Cert, NZ University Entrance Sixth Form Certificate (without UE) NZ UE Bursary and Scholarship Exam (1994 - 2004), Junior Scholarship, Overseas qual (IB Cambridge), School Certificate, Overseas school qualification, NZ Bursaries Examination (pre-1993), NCEA Level 1, NCEA Level 2, NCEA Level 3, NCEA Level Two Attained, NZ Scholarship, International Baccalaureate, Cambridge International Examination, 14 or more NCEA credits at any level, Combined Burs/NQF Credits Atta, Other school qualification, 14+ CR any level A, NCEA Level 3 or Bursary/Schol
No formal school qualification	No formal qualification, Not Specified.

3.6 Ethics and Human Subjects' Consideration

As part of the Human Research Ethics approval, a written plain language statement was provided to the data owners to gain their written consent for using data for the proposed research. As part of the agreement, the primary researcher signed a confidentiality agreement. According to the signed agreement, the primary researcher could only use the anonymized data for the purpose of their research. The primary researcher is allowed to publish the research findings within academic publications as long as the identity of the students are anonymized. At no time students identity has been made available to the researcher.

3.7 Predictive Model Development

To build predictive model for early prediction of students' performance several machine learning methods have been applied. Classification process in predictive modeling is used to predict the students' final result in a course which is unknown; therefore could be one of the following labels: at-risk or at-not-risk. Selection of machine learning methods has been based on the most popular methods that are currently used in this domain [142]. Classification process comprises of two steps: (1) use the historical data to train the model which is called training set, and (2) classification of the unseen data (which is called the testing set) by the trained model developed in the first step [143]. In machine learning, the training and testing datasets are considered as subsets of an assumed universal dataset which contains all the possible data pairs from the real world. Models are developed by using the training set to learn the properties of universal dataset. Testing dataset is then used to evaluate the performance of the learned model. Learned properties by the model should be applicable for both training and testing dataset, otherwise the model is not considered accurate [144]. Table 5.8 shows the list of classification algorithms used in this study.

3.7.1 Evaluation of predictive models

A number of evaluation criteria has been used to measure the predictive ability of a model in this study. The following metrics commonly used in literature have been utilized to measure the predictive power of developed model:

- **Accuracy** is the total number of correct predictions divided by the total number of predictions made for a dataset. However, it is considered an inappropriate measure in case of imbalanced dataset.

Table 3.11: Classification algorithms used to prepare training set for meta-model.

Classifiers	**Description**
J48	Non-commercial decision tree C4.5 [83]
RandomForest	Forest of random trees [87]
PART	A version of C4.5 using decision list [145]
DecisionTable	Simple decision table [146]
JRip	Propositional rule learner [76]
OneR	Uses minimum-error attribute [147]
ZeroR	0-R classifier
IBk	K-NN classifier [148]
KStar	Entropy-based [149]
LWL	Locally Weighed Learning [150, 151]
NaiveBayes	Naive Bayes classifier [93]
AdaBoost M1	Boosting a nominal class classifier [152]
Bagging	Reduces variance [153]
Stacking	Combines the output from others [154]
LogitBoost	Additive logistic regression classifier [155]
RandomCommittee	Ensemble of randomizable base classifiers
Vote	Uses majority vote to label new instance [156, 157]
Logistic	Logistic regression with a ridge estimator [158]
MultilayerPerceptron	Neural network with back propagation
SimpleLogistic	Linear logistic regression [159, 160]

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3.2}$$

True positives (TP) represent the instances that were correctly classified as positives. False Positives (FP) and False Negatives (FN) are number of instances that were incorrectly classified as positive and negative respectively. Accuracy of 1 means that the model identifies all the positive and negative cases correctly.

- **F-Measure** is the harmonic mean of precision and recall. Where precision measure the number of positive class predictions that were actually positive class and recall measure the number of positive class predictions made out of all positive class in dataset. F-

measure is a single score that takes both precision and recall in consideration.

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.5}$$

- **AUC-ROC Curve** is a performance measurement criterion for classification problem which utilizes various threshold settings. ROC is a probability curve and area under curve (AUC) represents measure of separability which tells the capability of model to distinguish between two classes. Higher the value of AUC, better the model will be in separating the two classes. The ROC curve is plotted with true positive rate (TPR) also called recall against the false positive rate (FPR), TPR is on y-axis and FPR is on x-axis. Following equations are involved for calculating AUC-ROC.

$$TruePositiveRate = \frac{TP}{TP + FN} \tag{3.6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.7}$$

$$FalsePositiveRate = 1 - Specificity = \frac{FP}{TN + FP} \tag{3.8}$$

3.8 Software

Following subsections have described the software that was used in the study.

3.8.1 Machine learning toolkit

All steps such as data pre-processing, making subsets, scaling variables, clustering, classification was done using the programming language Python's, module scikit-learn. Scikitlearn is an increasingly popular machine learning library written in Python. It is designed to be simple and efficient to make it useful for diverse kind of users in terms of their expertise. The primary aim of the project is to provide an efficient implementation of classic, well-established machine learning algorithms. It includes supervised and unsupervised learning algorithms, tools for model evaluation and selection, as well as tools for data pre-processing and feature engineering. It provides efficient implementations using state-of-the-art algorithms that are accessible to non-machine learning experts and their applications are reusable across scientific disciplines. Both tools integrate wide range of classification, regression and clustering algorithms for different kinds of problems. In addition, for plotting of graphs, the python module matplotlib [161], Seaborn and Power BI software [162] were used.

3.8.2 Process mining toolkit

For process mining techniques th ProM tool[1] was utilized. ProM is a generic open-source framework which combines most of the existing process mining techniques as plug-ins into a single tool. It is a state-of-the-art tool that provides a graphical user-friendly front-end environment for ease in implementing process mining algorithms.

3.8.3 The Jupyter Notebook

The IPython Notebook is now known as the Jupyter Notebook. It is an open source web application that provides an interactive computational environment, in which one can combine live code, execution, mathematics, plots, equations, visualizations, explanatory text

[1]http://www.promtools.org/doku.php

and rich media. This application was used for purpose of data cleaning, data transformation, statistical modeling and machine learning.

3.9 Summary

The overarching goal of this study is to use data analytics techniques for the maximization of student learning outcomes and enhance the overall quality of their educational experience. In doing this, various data mining/machine learning algorithms have been applied to student data (extracted from various sources) as these students progressed through a course, in order to predict which students are considered to be at risk of not satisfying their course requirements or are likely to withdraw. The purpose of this chapter was to describe the methodology used in the conduct of this study. The chapter provides details of data sources and describes the three datasets that were extracted from these sources. Additionally, it describes the general methods that were common for all experiments in the research. Training procedures, evaluation methods and the description of the environment and hardware used are also described.

Chapter 4

Student Engagement Patterns Using Hierarchical Clustering Methods.

4.1 Introduction

This chapter reports on data mining techniques that have been applied to student data that has been collected from student management system (SMS) and learning management systems (LMS). In learning analytics studies, data captured in log files as a result of interactions occurring in LMS, or student data stored in the SMS, or learner data that has been extracted from social media platforms, have been used for gaining understanding on learning processes. For instance, log files recorded in LMS comprises timestamped data of events that occur when student views of course content, or when they attempt a quiz or participate in a discussion forum [28]. Data mining techniques are then applied to find learning patterns from this log data to describe student behaviors. Next, interpretation of these patterns helps further understanding of student learning process so that they can be supported in improving their learning. This process of interpretation and reporting of results is described as learning analytics [163, 164]. Many studies have focused on addressing problems such as improving student retention rates in higher education by development of predictive models that enable early identification of students who may be at risk of failing the course or of attrition.

This study has used two data sources for the development of predictive models. First data source is the student management system, which includes student grades, demographic information, skills, school grades, financial information, language scores etc. The second data source is the log data generated by the LMS, that showcases timestamped records of students interactions with the online platform. The prediction of student academic standing has been performed as a regression problem by predicting the final scores or by classification problem where the dependent variable belongs to one of two categories, that is, it is either *fail* or *pass*, (or *at-risk* or *not-at-risk*).

The success of the predictive model depends on the quality of data that is used to develop the predictive model. Therefore, before developing a predictive model for a course (or a larger generalized model for all courses), it is important to investigate how LMS tools have been utilized in that course. The difference in the use of technology or tools has an effect in the generalization power of the predictive models for predicting academic success, and insufficient knowledge on LMS usage can result in over or underestimates in student success. However, it is not easy to investigate all courses individually; and a systematic approach is needed to group courses based on LMS usage by students. Research in learning analytics and related fields has achieved great attention for understanding and optimizing learning process. Much of the researchers have focused on addressing the problems improving retention rates of higher education. The focus has been on specifically developing predictive models that supports in early identification of students at-risk of failing the course or of attrition. Two types of data or their combinations, have been used in developing such models. First, data stored in student management system which include their grades, demographic information, skills, school grades, financial information, language scores etc. Second, from log data generated by LMS, when students interact with online environments, each activity along with timestamp is recorded. Course-related data from two databases: SMS and LMS provided the raw data inputs for this study. Data from the SMS included the students' final exam scores in that particular course, the number of student enrollments in the course, assessments scores across various student assignments, etc. Further, the study had access to large-scale LMS data comprising more than 10 thousand courses from one higher education institution. To find how LMS tools have been used in different courses, a hierarchical clustering approach was applied. That is, based on the LMS tools usage, distinct subgroups were defined that provided the grounding for futuristic development of a general model and also a portable predictive model

for similar courses. Clustering methods were applied on more than 4000 heterogeneous courses to divide them into smaller homogeneous groups. The study aimed to analyse the clusters of courses using LMS log data (pertaining to students) and study emergent patterns to develop a generalized predictive model (applicable to courses within the clusters) for predicting aspects related to student academic performance.

4.2 Motivation

The motivation of this chapter is as follows.

1. Present an approach used for acquiring data and selecting maximum number of courses that demonstrate relatively high level of online activities from the pool of 10k courses.

2. Describe the clustering approach employed for extracting common features from the 4,353 courses.

3. Study the overall pattern of online engagement in these selected courses.

4.3 Data Acquisition

In this section we describe the data processing steps.

4.3.1 Data Sources

Data mining processing steps entails following steps: pre-processing data, apply data mining algorithm for analysis and reporting the results [165]. The first step is the task of data extraction from the data source and then acquiring it in the desired format, since the structure and format of LMS data is not directly suited for data mining and therefore it requires a transformation process. This task of data pre-processing includes data cleaning, treatment of missing values, transformation into the desired format and integrating data from

diverse sources. In this study all individual courses were collected from two databases: SMS and LMS. While the SMS data was formatted and included course final scores, number of enrolled students and scores for assessments, the same cannot be said for LMS data.

The LMS data included information on total students enrolled and log data that reported all activities performed by individual users in the LMS. That is, every user action or every click is captured and stored in log files. Users provided they have enough system privileges, can query the user actions using a functionality called report logs [166]. The query output retrieves each interaction that occurs in the module which is stored in a csv (comma separated values) file; further additional information on how many interactions of each type of activity have occurred for each user are also stored. All individual student identifiers were removed for this study. The output data from the Interactions module was imported to perform more detailed data analysis. Additionally, supplementary descriptive graphs were created using python library *seaborn*.

Most of the courses were blended courses and had options to meet instructors face-to-face to discuss course related problems. Generally, courses are taught in three different modes: internal mode (where students are taught on a weekly basis on-campus), distance mode (where students are off-campus and mainly use electronic means although some face-to-face opportunity may be provided via one or more contact workshops) and block mode (which refers to internal courses where the class contact time is compressed to four or five full-day teaching). This study included both undergraduate and postgraduate courses that have run since 2015 with at least one student registration. The undergraduate courses were divided into four levels based on their complexity. Moreover, all courses were taken from different subject areas spread across different academic disciplines.

4.4 Data Preparation and Processing

This data processing steps are described next. Dataset preparation comprised two phases: feature engineering for cluster analysis and selection of courses.

4.4.1 Phase 1: Feature Engineering for cluster analysis

In order to measure online engagement, following active variables from each course were calculated by dividing courses into different clusters.

1. **Class Size**

   Class size is calculated by counting the number of students enrolled for a particular course. Each course was divided into five categories based on the number of enrolled students as small (student numbers are in range 1 to 21), medium (students numbers are in range 22 to 36), medium-large (student numbers are in range 37 to 65), large (student numbers are in range 66 to 300) and very large (when student numbers exceed 300).

2. **Number of Distinct Activities**

   In LMS, an activity is a general name for group of features that are used for a specific purpose in a course. It is related to how students interact with their peers or the instructor. A list of activities in standard LMS are shown Table 4.1. For each course, the number of distinct activities have been calculated and divided into following subcategories: not diverse (when number of distinct activities are less than 2), little diverse (when number of distinct activities equals 3), somewhat diverse (when number of distinct activities equals 4) and very diverse (when number of distinct activities are more than 4). Additional to this, all relevant activities have been grouped together

Table 4.1: Definitions and categories of standard LMS activities.

S. No	Category	Activity	Definition
1	Assignment Activities	Assignments	Instructor are enabled to grade and comment on uploaded assignments by students.
		(LTI) External Tool	Allow students to participate with learning resources as an external tool on other websites.
2	Tools for Student Management	Attendance	The Attendance activity is designed for instructors to be able to take attendance during class, and for students to be able to view their own attendance record.
3	Communication and Collaboration	Data	Enables particepants to create, maintin and search a database.
		Wiki	A collection of pages that are editable.
		Workshop	Peer assessment.
		Glossary	Enables participants to maintain a list of definitions.
		Forum	Allow participants to have asynchronous discussion.
		Dialogue	Dialogues allow students or instructor to start a private conversation with another user in the same course.
4	Interactive Delivery of Content	Lesson	Content delivery in lesson form.
		Url	Instructor can send the student to any place they can reach on their web browser, for example Wikipedia
		Scorm	Enable scorm packages in course.
		Resource	A resource is an item that a teacher can use to support learning, such as a file or link.
		Folder	For helping organize files and one folder may contain other folders
		Page	The student sees a single, scrollable screen that an instructor creates with the robust HTML editor
		Book	Multi-page resources with a book-like format.
		Journal	A journal entry is one in which students type directly into a text field in Moodle.
5	Assessments and Surveys	Quiz	Allow instructors to design quizzes that could be automatically graded.
		Feedback	For collecting different responses e.g. conducting survey
		Survey	Collecting students' response about course, contents and teaching style.
		Choice	An instructor asks multiple choice questions
		Questionnaire	It allows instructor to create a wide range of questions to get student feedback e.g. on a course or activities.
		Chat	Enables participants to have a synchronous disucssion.

into five major groups/categories; *assignment activities*, *tools for student management*, *communication and collaboration*, *interactive delivery of content*, and *assessments and surveys*. These grouping of activities will help in analyzing difference between clusters at a higher level.

3. **Number of Distinct Actions**

Score of actions can be performed within each activity. For example, in activities *forum* actions can be performed as *view message, review message, delete message* etc. As shown in Figure 4.1 *View* is the action which could be performed for almost all activities. Therefore, it was further necessary to consider measurements that reflect diversity in terms of actions for the courses. Diverse actions for each course was divided into following subcategories: not diverse (when number of distinct actions less than 2), little diverse (when number of distinct actions equal to 3), somewhat diverse (when number of distinct actions equal to 4) and very diverse (when number of distinct actions

are more than 4).

Figure 4.1: List of possible actions in each category of activities.

4. **Total Unique Events**

Total unique events are calculated by counting unique combination of both activities and actions. Figure 4.2 shows an example of a course with 7 number of distinct activities and 5 number of distinct actions. Number of edges in the directed graph shows total distinct combination of activities and actions. Events were further divided into two categories; major and minor. Events that are used in less than 65% of the courses are counted as least used activities and are marked *minor activities* (See Table: 4.3) and events that are used more than 65% of the courses are marked as *major events* (see Table: 4.2). Considering the skewed distributions of minor and major activities, all activities were next converted into ordinal values. Major events were converted with 4 levels using quantile score for each variable (i.e., 25%, 50%, 75% and 100%) and minor events were redefined as binary class (i.e., (used, not-used).

Table 4.2: Descriptive statistics for most occurred(major) events in courses.

Activities	Action	Mean	Std	Min	25%	50%	75%	Max	Missing Percent(%)	Frequency (%)	Frequency
Core	Viewed	10706.10	23251.01	4.00	1111.00	3840.00	10397.00	379740.00	0.02	99.98	4352.00
Forum	Viewed	4886.24	14855.48	4.00	168.00	812.00	3296.00	291764.00	7.99	92.01	4005.00
Resource	Viewed	4444.03	9796.12	4.00	328.00	1338.00	4387.00	157416.00	9.35	90.65	3946.00
Assign	Viewed	3344.90	5966.29	4.00	300.00	1368.00	3912.00	74800.00	20.63	79.37	3455.00
Gradereport_user	Viewed	1006.03	2261.97	4.00	48.00	268.00	984.00	31092.00	21.43	78.57	3420.00
Gradereport_overview	Viewed	153.20	329.34	4.00	16.00	52.00	152.00	4952.00	31.70	68.30	2973.00
Url	Viewed	930.03	3114.28	4.00	48.00	176.00	664.00	61468.00	39.12	60.88	2650.00
Assignsubmission_file	Uploaded	378.98	571.70	4.00	84.00	204.00	424.00	7260.00	41.17	58.83	2561.00
Assignsubmission$_{file}$	Created	305.36	460.98	4.00	72.00	164.00	340.00	5804.00	41.19	58.81	2560.00
Assign	Submitted	317.58	503.82	4.00	72.00	168.00	348.00	7012.00	41.90	58.10	2529.00
Assignsubmission_file	Updated	83.02	142.23	4.00	12.00	36.00	88.00	1700.00	47.74	52.26	2275.00
Folder	Viewed	2673.65	6741.85	4.00	124.00	544.00	2170.00	82156.00	48.50	51.50	2242.00

Table 4.3: Descriptive statistics for least occurred (minor) events.

Activities	Action	Mean	Std	Min	25%	50%	75%	Max	Missing Percent(%)	Frequency (%)	Frequency
Forum	Created	397.87	1011.50	4.00	24.00	92.00	316.00	19688.00	50.84	49.16	2140.00
Forum	Uploaded	311.13	829.84	4.00	16.00	72.00	257.00	13220.00	52.22	47.78	2080.00
Book	Viewed	9504.16	23166.07	4.00	508.00	2248.00	8310.00	359264.00	59.13	40.87	1779.00
Forum	Searched	48.02	171.46	4.00	4.00	12.00	28.00	3864.00	67.40	32.60	1419.00
Assignsubmission_comments	Created	19.25	72.60	4.00	4.00	8.00	20.00	2452.00	68.30	31.70	1380.00
Forum	Deleted	48.02	149.82	4.00	4.00	12.00	32.00	1996.00	72.00	28.00	1219.00
Page	Viewed	1875.09	10740.96	4.00	68.00	248.00	970.00	238116.00	72.34	27.66	1204.00
Forum	Updated	56.94	149.22	4.00	8.00	16.00	44.00	2164.00	72.62	27.38	1192.00
Core	Graded	3808.89	10554.61	4.00	136.00	512.00	2182.00	143212.00	78.80	21.20	923.00
Booktool_print	Printed	54.81	118.68	4.00	8.00	16.00	48.00	1804.00	80.04	19.96	869.00
Quiz	Viewed	29052.58	65578.26	4.00	874.00	4428.00	25474.00	685344.00	80.45	19.55	851.00
Quiz	Started	2522.49	6258.31	4.00	124.00	440.00	1768.00	60356.00	81.42	18.58	809.00
Quiz	Submitted	2442.92	6102.67	4.00	112.00	412.00	1740.00	58340.00	81.71	18.29	796.00
Quiz	Reviewed	3260.39	8693.16	4.00	136.00	512.00	1974.00	87568.00	81.92	18.08	787.00
Choice	Viewed	1594.71	2573.97	4.00	243.00	744.00	1773.00	18708.00	88.61	11.39	496.00
Choice	Submitted	327.00	472.22	4.00	95.00	172.00	356.00	3372.00	90.08	9.92	432.00
Choice	Updated	185.52	319.51	4.00	24.00	82.00	200.00	2316.00	91.32	8.68	378.00
Assign	Accepted	320.92	490.29	4.00	76.00	160.00	338.00	4328.00	91.48	8.52	371.00
Core	Updated	3211.09	13871.96	4.00	68.00	370.00	1508.00	184240.00	91.78	8.22	358.00
Dialogue	Viewed	1242.89	2668.08	4.00	80.00	464.00	1356.00	26652.00	92.49	7.51	327.00
Forum	Disabled	7.84	7.50	4.00	4.00	4.00	8.00	68.00	92.56	7.44	324.00
Assignsubmission_comments	Deleted	8.01	12.03	4.00	4.00	4.00	8.00	144.00	92.90	7.10	309.00
Dialogue	Created	224.47	464.03	4.00	32.00	92.00	236.00	4260.00	94.12	5.88	256.00
Glossary	Viewed	2420.92	9894.31	4.00	68.00	292.00	1256.00	112932.00	95.01	4.99	217.00
Forum	Enabled	6.59	4.70	4.00	4.00	4.00	8.00	32.00	95.50	4.50	196.00
Questionnaire	Viewed	426.12	880.67	4.00	36.00	120.00	352.00	4848.00	95.93	4.07	177.00
Questionnaire	Submitted	141.83	242.36	4.00	18.00	52.00	126.00	1160.00	96.99	3.01	131.00
Assignment_upload	Uploaded	164.34	271.82	4.00	14.00	72.00	180.00	1760.00	97.82	2.18	95.00
Assignment_upload	Submitted	113.05	160.84	4.00	12.00	52.00	148.00	828.00	97.91	2.09	91.00
Feedback	Viewed	355.21	467.20	4.00	52.00	164.00	440.00	2284.00	97.91	2.09	91.00
Assignsubmission_onlinetext	Created	265.02	552.38	4.00	56.00	128.00	292.00	4812.00	98.02	1.98	86.00
Assignsubmission_onlinetext	Uploaded	325.72	668.75	4.00	65.00	144.00	356.00	5816.00	98.02	1.98	86.00
Glossary	Created	210.78	379.86	4.00	20.00	66.00	166.00	1888.00	98.12	1.88	82.00
Chat	Viewed	280.10	326.59	4.00	60.00	180.00	392.00	1460.00	98.14	1.86	81.00
Core	Created	88.71	231.64	4.00	4.00	20.00	88.00	1540.00	98.32	1.68	73.00
Assignsubmission_onlinetext	Updated	74.57	138.37	4.00	13.00	28.00	64.00	1004.00	98.39	1.61	70.00
Feedback	Submitted	161.70	207.13	4.00	26.00	80.00	201.00	880.00	98.48	1.52	66.00
Core	Assigned	55.81	241.30	4.00	4.00	8.00	17.00	1536.00	98.53	1.47	64.00
Quiz	Abandoned	14.00	21.69	4.00	4.00	4.00	12.00	120.00	98.67	1.33	58.00
Lesson	Started	2233.26	3671.61	4.00	108.00	436.00	1668.00	14004.00	98.69	1.31	57.00
Lesson	Viewed	17974.04	31877.73	4.00	984.00	2784.00	13204.00	153700.00	98.69	1.31	57.00
Glossary	Updated	55.56	107.75	4.00	8.00	24.00	52.00	684.00	98.74	1.26	55.00
Lesson	Ended	802.16	1337.26	8.00	44.00	152.00	641.00	4856.00	98.85	1.15	50.00
Gradereport_grader	Viewed	42.61	48.11	4.00	8.00	24.00	64.00	232.00	98.87	1.13	49.00
Core	Deleted	308.67	831.58	4.00	8.00	24.00	133.00	4424.00	98.90	1.10	48.00

5. **Total Events**

Total events are calculated by summing all events that were performed in a course by students.

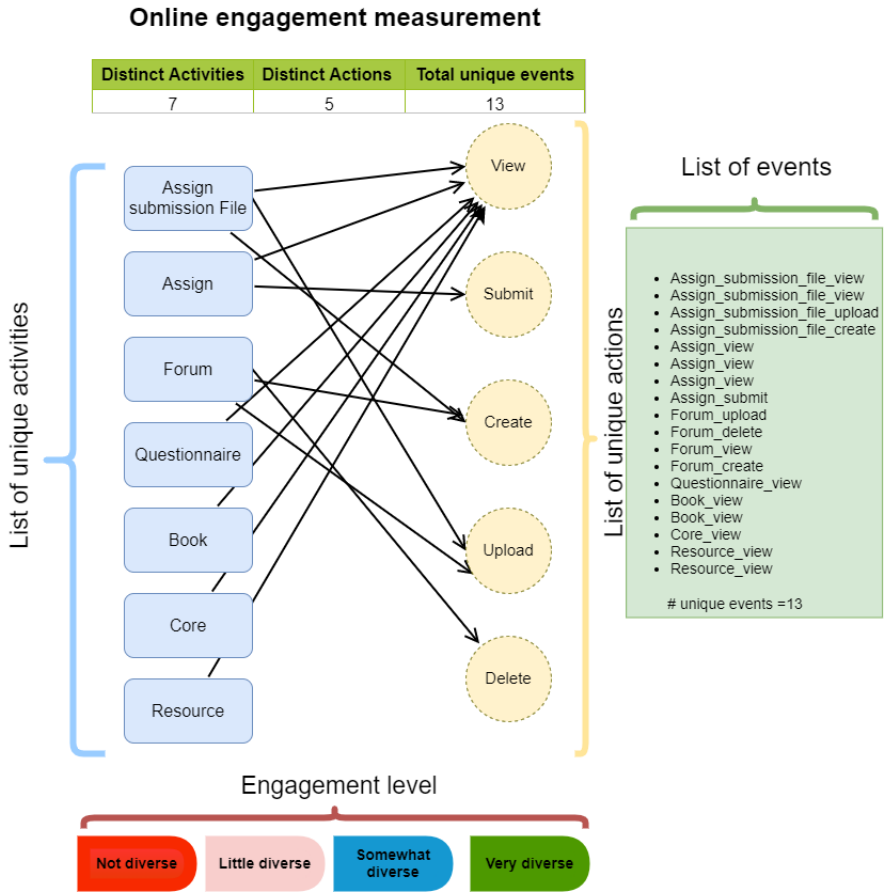


Figure 4.2: An example of a course with 7 number of distinct activities and 5 number of distinct actions. Number of edges in the directed graph shows total distinct combination of activities and actions.

4.4.2 Phase 2: Selection of Courses for Cluster Analysis

Raw data produced after integration of the two data sources (i.e., LMS and SMS) contained data spread over more than 10k courses. However, most of the courses were not usable in terms of usage of LMS tools. Therefore, a strategy was configured for selecting courses which met a minimum criterion in terms of LMS tool usage. Selection of courses was followed in a systematic way by observing following inclusion and exclusion rules (Figure: 4.3).

- **Exclusion Rule 1:** The scope of the courses were limited; hence, the focus was to include those courses which were offered in years from 2015 to 2017, while rest of the

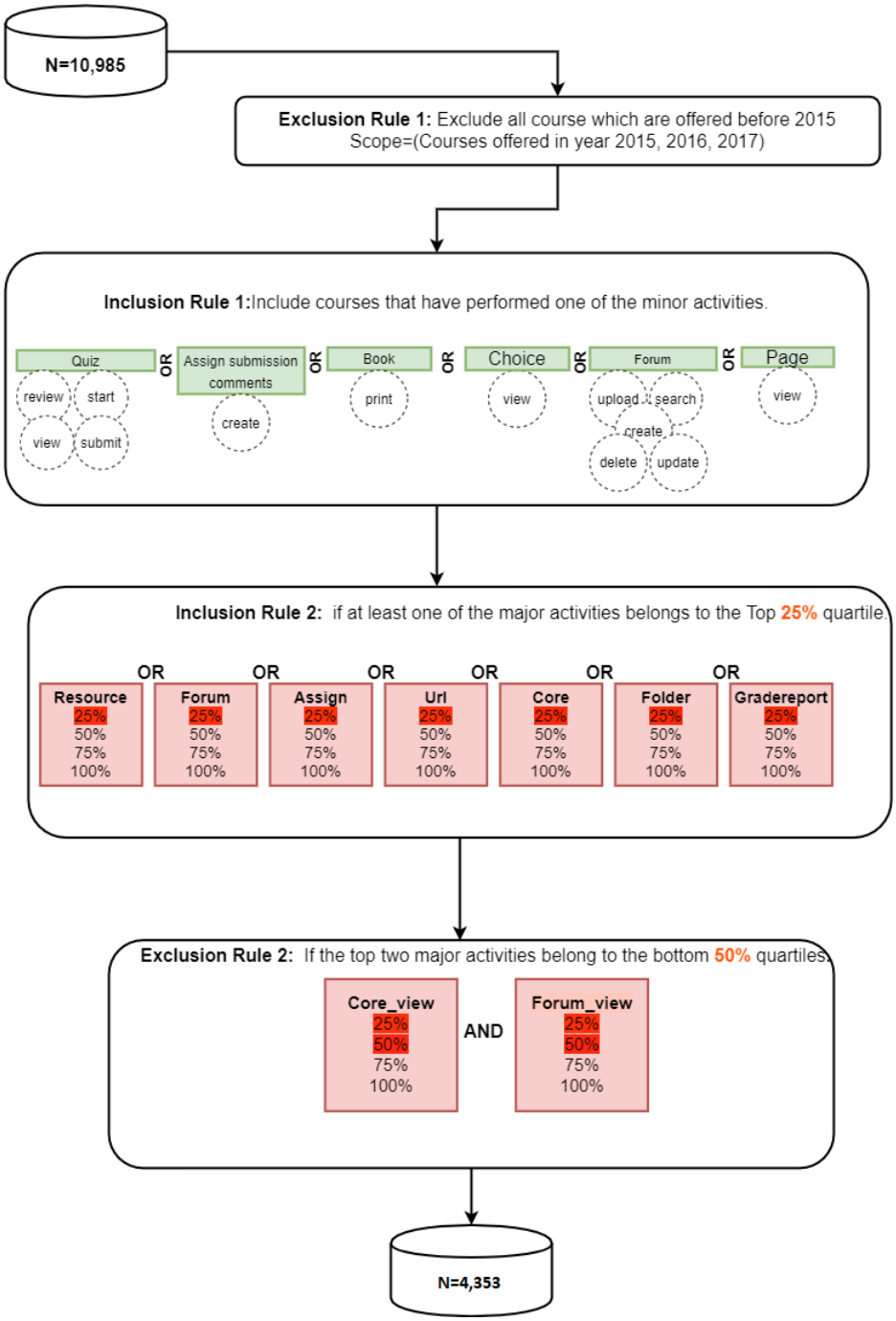


Figure 4.3: Systematic selection of the courses for cluster analysis.

courses were deleted from the dataset.

- **Inclusion Rule 1:** In the second phase, descriptive statistics of unique events were examined to enable selection of the most frequent events in the course by calculating

the frequency of events in total courses. All courses that have performed at least one of the minor activities were included as are listed in Table 4.3.

- **Inclusion Rule 2**: Include courses, if one of major activities belong to the top 25% quartile.

- **Exclusion Rule 2:** Exclude courses where top two major activities belong to the bottom 50% quartiles.

4.5 Hierarchical Agglomerative Clustering

Cluster analysis is a type of exploratory data analysis which gives some intuition about the structure of the data. Clustering is defined as a process of dividing the data into subgroups also called clusters, such that each data in the subgroup share similar properties. These homogeneous clusters within the data are calculated using similarity measures (e.g., Euclidean-based distance or correlation-based distance). Clustering methods are broadly divided into two categories: hierarchical and partitional. Hierarchical clustering assumes cluster's structure in nested structural levels, unlike partitional clustering that considers a single common level for all clusters.

In this study, agglomerative clustering method is used which is one of the most common hierarchical clustering techniques. Hierarchical method creates a hierarchical decomposition of the given set of data objects. The agglomerative approach that is based on the hierarchical decomposition is also called bottom-up approach. The clustering technique assumes that when each data point is similar enough to other data points, the data at the starting can be assumed to be clustered in one cluster. Suppose there are 4 data points in the beginning. First, each of these points will be assigned to a cluster to result in 4 clusters. Then, at each

iteration, closest pair of clusters are merged and this step is repeated until only a single cluster is left. At each step, clusters are merged or added; therefore, this type of clustering is also known as additive hierarchical clustering. Similarity is calculated by the distance between the two centroids of clusters. The points having the least distance are referred to as similar points and therefore are merged as one point. Evaluation of cluster methods are not simple like classification methods, where onc can measure the performance of methods by comparing the values with available class labels of data. In case of clustering most of the time class labels are not available, therefore it is called unsupervised learning. Since agglomerative clustering algorithms discussed above, require ($wherek = numberofclusters$)) as an input and do not learn from the data itself, there is no right answer for the number of clusters for a given problem. Expert knowledge may help to evaluate the clustering results. There are methods to evaluate the performance of clustering method for a given cluster number. In this study silhouette analysis is used to get an intuition about k. Silhouette analysis can be used to measure the degree of separation between clusters. For each data point i, the silhouette s_i is calculated as follows:

$$s(i) = \frac{b^i - a^i}{max(a^i, b^i)}$$

Silhouette score can take values in range $[-1, 1]$. Observation with value s_i close to 1 is considered well clustered. A small value (close to 0) means that data points lie between two clusters and observation with a negative value are probably placed in wrong cluster.

4.6 Results

The online engagement patterns that emerged and how classification of courses were done are presented next.

4.6.1 Overall patterns of online engagement

The findings showed that the class size of courses varied from 1 to 1327 students (Mean=74.57, SD=125.74) as shown in Table 4.4. The majority of the courses (30.7%) selected for cluster analysis fall in the category of Small where number of enrolled students are in the range of $1 - 21$. Second largest category is of Large which is (26.3%) of selected courses. Largest category in course type is *Postgraduate* which is almost 26% of the total selected courses (Figure 4.4).

Table 4.4: Descriptive statistics of active variables for 4353 courses.

	Label.	Mean	Std	Min	25%	50%	75%	Max	Mode
Class Size	V1	74.57	125.74	1.00	16.00	35.00	81.00	1,327.00	14
Distinct Action	V2	5.45	3.25	1.00	3.00	6.00	8.00	20.00	1
Distinct Acitvities	V3	8.64	2.95	1.00	7.00	9.00	11.00	21.00	9
Total Unique events	V4	22.55	11.78	1.00	13.00	22.00	30.00	73.00	23
Total Events	V5	39,319.12	102,536.90	4.00	2,740.00	10,532.00	32,892.00	1,421,580.00	212

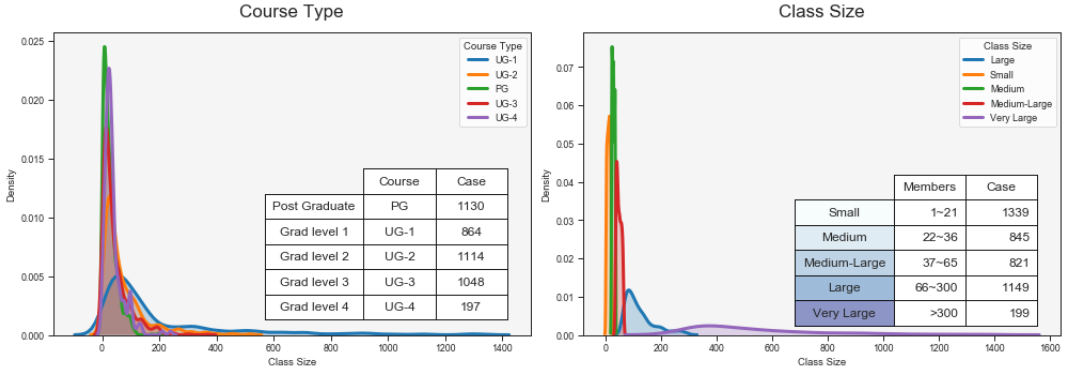


Figure 4.4: Course type and Class Size categories in 4353 courses (selected for cluster analysis).

Among the 4,353 courses, majority of the courses' study mode falls in category *internal*

Table 4.5: Division of select courses for cluster analysis according to study mode.

Study Mode	**UG-1**	**UG-2**	**UG-3**	**UG-4**	**PG**	**Total**
External	257	305	333	39	281	1215
Internal	536	786	683	146	523	2674
Block	69	23	32	12	326	462
Blended	2	0	0	0		2
Total						4353

Table 4.6: Selected Course distribution according to the demographics.

				Course Level		
Course Demographic	PG	UG-1	UG-1	UG-3	UG-4	Total
Business School	347 (7.971%)	132 (3.032%)	227 (5.214%)	227 (5.214%)	NaN	933
Arts	1 (0.022%)	162 (3.721%)	148 (3.399%)	112 (2.572%)	124 (2.848%)	547
Health Sciences	111 (2.549%)	94 (2.159%)	128 (2.940%)	123 (2.825%)	20 (0.459%)	476
Humanities and Social sciences	471 (10.82%)	281 (6.455%)	307 (7.052%)	275 (6.317%)	40 (0.918%)	1,374
STEM	200 (4.594%)	195 (4.479%)	304 (6.983%)	311 (7.144%)	13 (0.298%)	1,023
Total						4,353

(61.4%) while second largest category is of external (27.9%) (see Table 4.5). majority of the courses (i.e., about 31.5%) were found to belong to majoring subjects within the 'Humanities and Social Sciences' discipline (e.g., psychology, sociology, education, etc.), while second largest class (i.e., 23.5%) belonged to subject majors within the STEM disciplines (i.e., Science, Technology, Engineering Mathematics) refer Table 4.6.

Distinct activities and distinct actions are two log variables that are used to measure the engagement level in these courses. Distinct activities can be calculated by adding all distinct activities that were performed in the course by students. Some digital content like web pages, text messages, HTML files are directly created via Moodle system, while external content (e.g., research material, videos, audio etc) known as resources are uploaded. Student and content interaction activities occur as and when students interact with the content resources such as when they browse and access different resources. List of all such activities that occurred in courses

List of all such activities that occurred in courses are listed in Table 4.2 and Table 4.3.

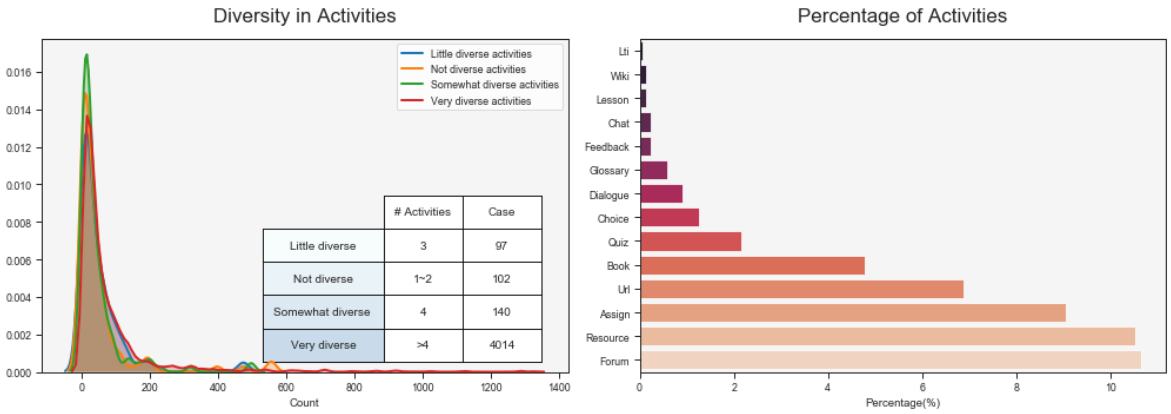


Figure 4.5: Diversity of activities in selected courses ($n = 4,353$).

Activities related to forum, resource, assignments and url were mostly performed by students (Table 4.7). Activities are further explained in terms of actions and most popular among them is *Viewed* action. Other performed actions were *created*, uploaded, *accepted*, *started* etc. Table 4.8 shows the frequency of actions performed in selected courses. For example *created* is second most performed action in 70% of the courses.

Table 4.7: Descriptive statistics of Moodle activities (mean) in 4353 courses.

	Assign	Book	Choice	Dialogue	Forum	Quiz	Resource	Url
Frequencies	3,455.00	1,779.00	496.00	327.00	4,013.00	851.00	3,946.00	2,650.00
Mean	12.52	18.40	4.75	2.37	11.31	25.90	15.01	2.63
Std	8.97	15.30	6.37	4.35	11.84	18.67	12.46	3.76
Max	56.34	71.66	37.97	59.45	72.35	80.39	72.05	50.00
Percentage(%)	79.37	40.87	11.39	7.51	92.19	19.55	90.65	60.88

Table 4.8: Descriptive statistics of Moodle action (% of total activities) in 4353 courses.

	Viewed	Created	Updated	Uploaded	Accepted	Deleted	Printed	Reviewed	Started	Submitted
Frequencies	4,353.00	3,047.00	2,747.00	3,029.00	371.00	1,361.00	869.00	787.00	849.00	2,950.00
Mean	91.69	1.58	0.70	1.58	0.58	0.12	0.12	2.37	1.99	1.50
Std	12.69	1.30	1.88	1.19	0.46	0.41	0.29	2.15	1.63	1.17
Max	100.00	20.00	31.17	11.92	3.83	8.93	5.77	13.37	14.29	9.86
Percentage(%)	100.00	70.00	63.11	69.58	8.52	31.27	19.96	18.08	19.50	67.77

Figure 4.5 and 4.6 show few most performed activities and actions in selected courses. Majority of the selected courses are in category of very diverse in terms of both unique

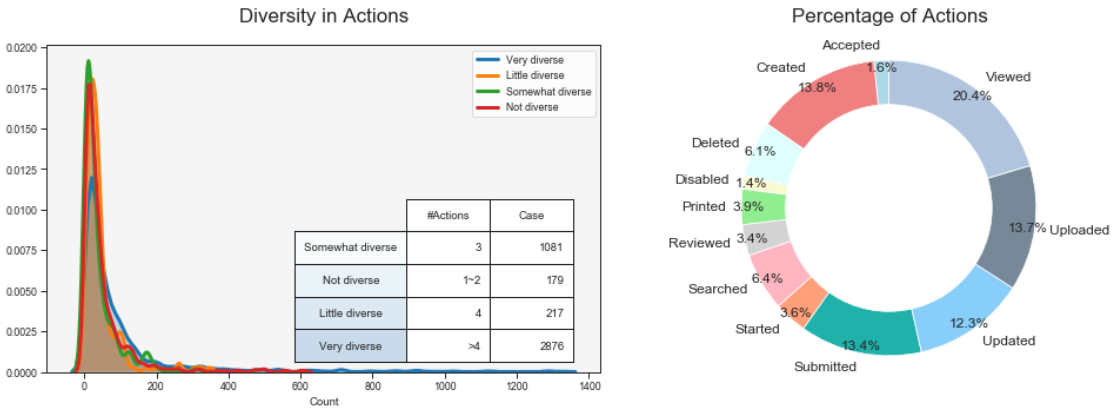


Figure 4.6: Diversity of action in selected courses ($n = 4,353$).

activities and 'unique actions'. More than 9% of the unique activities belong to Moodle activities of forum, resource and assignments and most of them were in the category of *viewed* (20.4%). This shows that majority of the courses performed were passive activities like *view*. Also the majority of events that occurred belong to category *view*.

4.6.2 Classification of courses

In this study, cluster analysis was performed on aggregated data from LMS log data from 4,353 courses. Determining the number of clusters is an essential part for partitioning data into groups. In literature there are scores of methods used. Generally these methods can be divided into two major groups; direct methods (examples of direct methods are silhouette and elbow methods) and statistical methods (that use methods to compare evidence against null hypothesis for example gap statistic). Apart from these methods there are more than 30 indices that are used to identify optimal number of clusters. In this study all 30 indices were used to decide the optimal number of clusters using majority voting. As shown in Figure 4.7 according to the majority voting best number of clusters is 4.

After the application of 30 indices for finding the optimal number of cluster values (Figure:4.7), among all indices 12 proposed four as the best number of clusters. According to

Table 4.9: Clustering validity indices results. NOC is an abbreviation for Number of Clusters.

Index	**KL**[167]	CH[168]	**Hartigan** [169]	**CCC**[170]	**Scott** [171]	**Marriot** [172]	**TrCovW** [173]	**TraceW**[173]
NOC	4	4	4	4	4	4	4	4
Value	5.30	205.12	350.27	12.63	3906.58	4.74E+46	51134.587	544.27
Index	**Friedman** [174]	**Rubin** [174]	**Cindex**[175]	**DB** [176]	**Silhouette** [177]	**Duda** [178]	**PseudoT2** [179]	**Beale** [180]
NOC	4	4	5	3	3	NA	NA	4
Value	2.74	-0.22	0.48	2.02	0.21	NA	NA	0.97
Index	**Ratkowsky** [181]	**Ball** [182]	**PtBiserial** [183]	**Frey** [184]	**McClain** [185]	**Dunn** [186]	**SDbw** [187]	**SDindex** [188]
NOC	5	4	5	3	3	5	5	3
Value	0.10	813.81	0.30	3.69	0.04	0.30	0.99	4.43

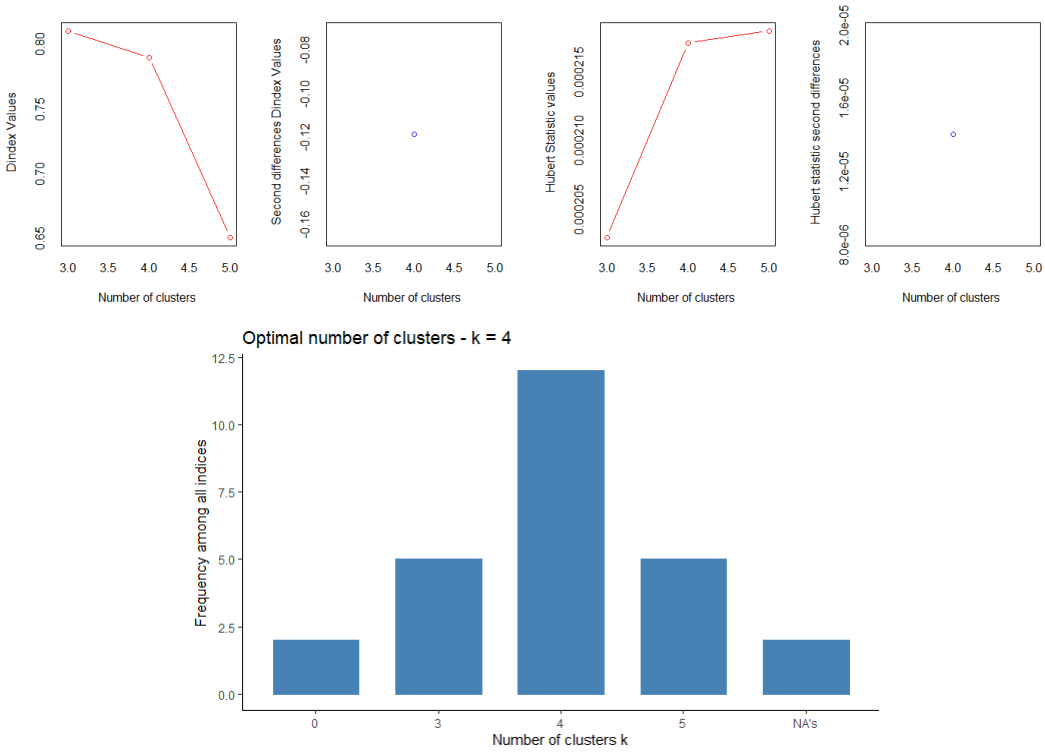


Figure 4.7: The above plot presents the frequency among all 30 indices used for the determination of the optimal number of clusters. Statistic points out $K = 4$ as the best number of partitions to the clustering

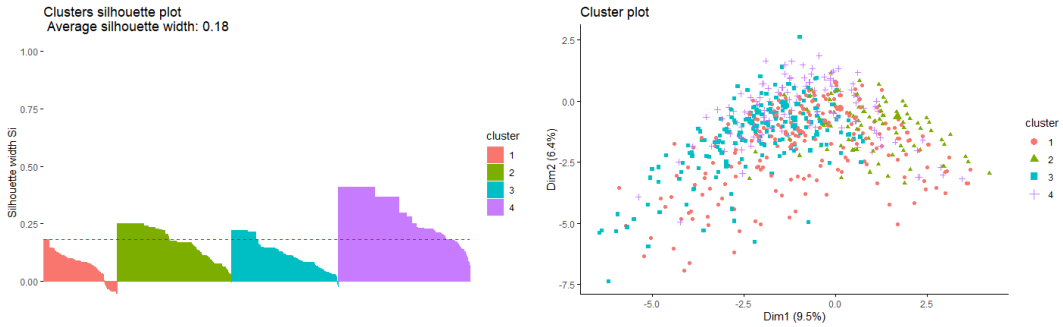


Figure 4.8: A silhouette plot (a) used hierarchical clustering (agglomerative) on data with number of clusters=4 with visualization of the data (b).

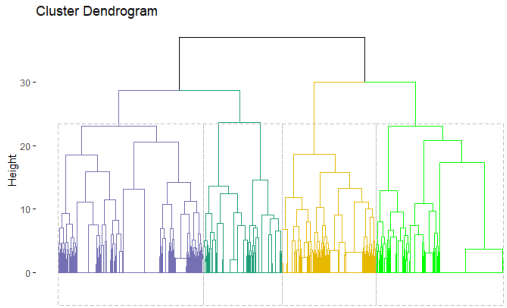


Figure 4.9: Cluster dendrogram for the agglomerative hierarchical clustering on data with number of clusters=4

the majority vote four was selected as best number of clusters. Within 4-class clusters, we compared each cluster by using the course profiles and online activities as shown in Figure 4.10.

Cluster 1 ($n = 1418$ or 32.58%), demonstrates higher usage of online activities for forum (94.36%), then assignment (93.79%) and resource (89.35%). Majority of the courses are of small ($n = 449$ or 31.66%) where the enrolled students are in range of 1 to 21. The majority of the courses in cluster 1, were from *Humanities and Social Sciences* ($n = 581$ or 40.97%) discipline, then *Business* ($n = 369$ or 26.02%) finally *STEM* ($n = 249$ or17.55%) (Table 4.11). Further analysis was performed to identify general features regarding online activities (Table 4.10).

Cluster 1 showed higher level use of distinct activities and 76.87% of courses used more

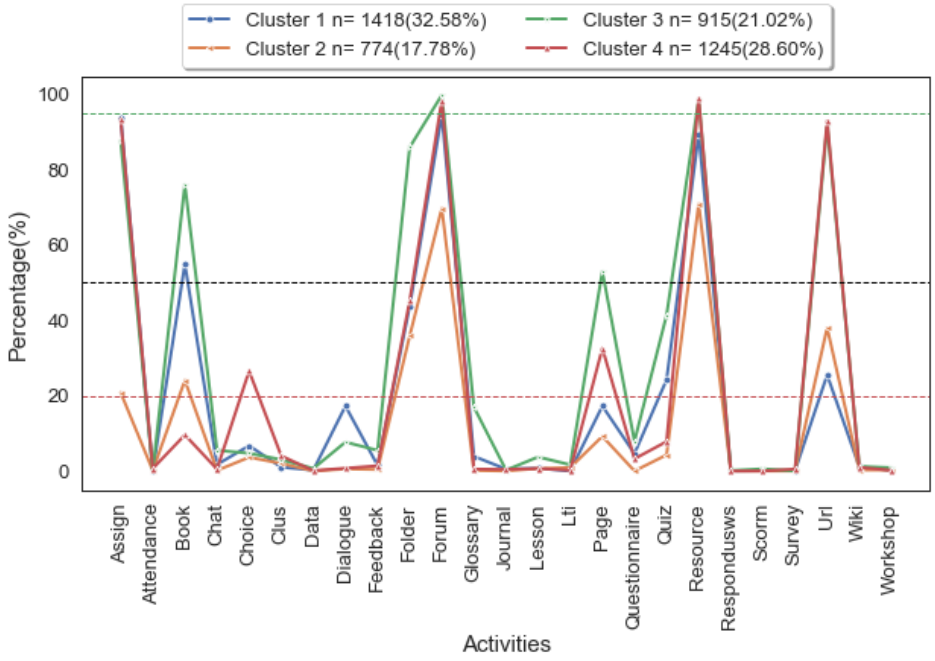


Figure 4.10: Percentage of use of online activity items in four clusters of selected courses.

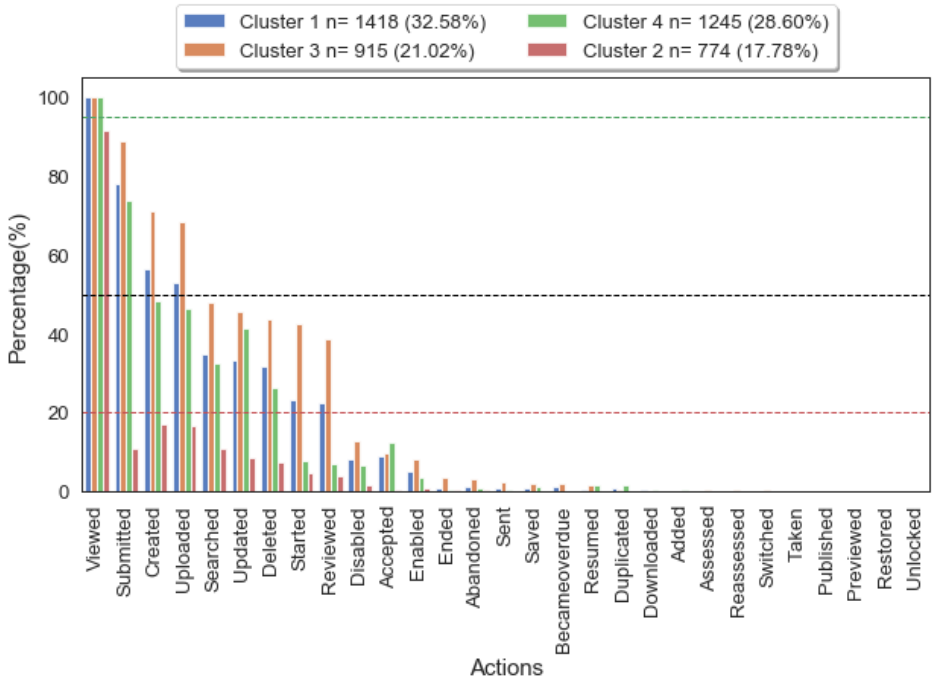


Figure 4.11: Percentage of use of online action in four clusters of selected courses.

Table 4.10: Active and descriptive features of 4 clusters ($n = 4,353$).

	Variables	Cluster 1 $n = 1418(32.58\%)$	Cluster 2 $n = 774(17.78\%)$	Cluster 3 $n = 915(21.02\%)$	Cluster 4 $n = 1245(28.60\%)$
Class Size	Small (1-21)	449 (31.66 %)	338 43.67%)	185 (20.22%)	366 (29.40 %)
	Medium (22-36)	279 (9.68 %)	147 (18.99%)	182 (19.89%)	237 (19.04%)
	Medium-Large (37-65)	258 (18.19%)	124 (16.02%)	169 (18.47%)	270 (21.69%)
	Large (66-300)	364 (25.67%)	144 (18.60%)	332 (36.28%)	309 (24.82%)
	Very Large (more than 300)	68 (4.80%)	21 2.71%)	47 (5.14%)	63 (5.06%)
Study Mode	Block mode	182 (12.83%)	79 (10.21%)	65 (7.10%)	136 (10.92%)
	External mode	452 (31.88%)	134 (17.31%)	358 (39.13%)	270 (21.69%)
	Internal mode	784 (55.29%)	561 (72.48%)	490 (53.55%)	839 (67.39%)
	Internal and External mode			2 (0.22%)	(0.00%)
Distinct Actions	Not diverse (n<2)	210 (14.81%)	564 (72.87%)	49 (5.36%)	257 (20.64%)
	Little diverse ($n = 3$)	38 (2.68%)	58 (7.49%)	12 (1.31%)	71(5.70%)
	Somewhat diverse ($n = 4$)	80 (5.64%)	42 (5.43%)	27 (2.95%)	68 (5.46%)
	Very diverse (n> 4)	1090 (76.87%)	110 (14.21%)	827 (90.38%)	849 (68.19%)
Distinct Activities	Not diverse (n<2)	3 (0.21%)	97 (12.53%)	0 (0%)	2 (0.16%)
	Little diverse ($n = 3$)	13 (0.92%)	81 (10.47%)	0 (0%)	3 (0.24%)
	Somewhat diverse ($n = 4$)	40 (2.82%)	92 (11.89%)	0 (0%)	7 (0.56%)
	Very diverse (n> 4)	1362(96.05%)	504 (65.12%)	915 (100%)	1233 (99.04%)
Course Level	PG	507 (35.75%)	182 (23.51%)	228 (24.92%)	212 (17.03%)
	UG-1	225 (15.87%)	100 (12.92%)	225 (24.59%)	314 (25.22%)
	UG-2	317 (22.36%)	208 (26.87%)	243 (26.56%)	346 (27.79%)
	UG-3	321 (22.64%)	237 (30.62%)	201 (21.97%)	289 (23.21%)
	UG-4	48 (3.39%)	47 (6.07%)	18(1.97%)	84(6.75%)

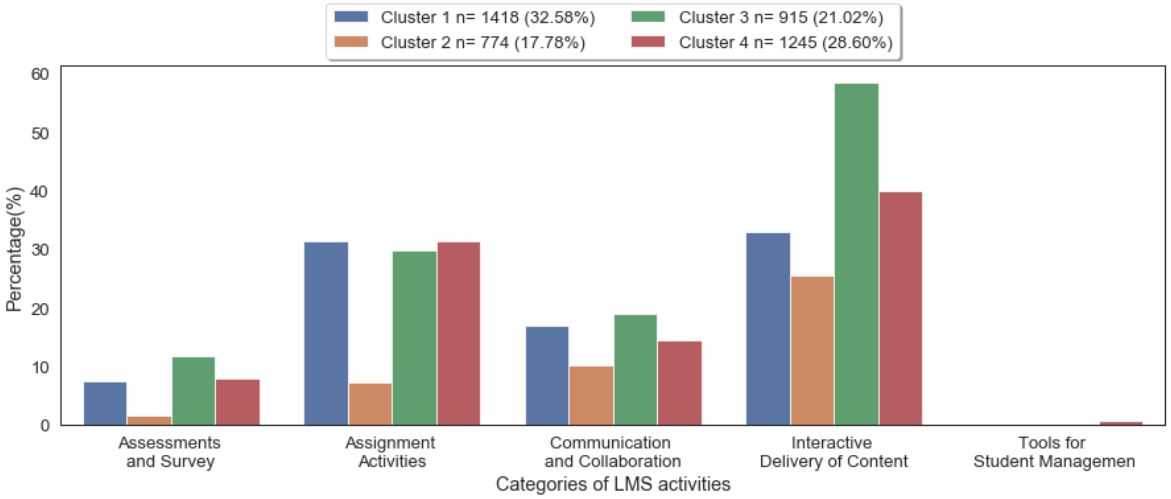


Figure 4.12: Average percentage of each clusters in the use of LMS online activities.

Table 4.11: Demographic characterstics of selected courses ($n = 4,353$).

Domain	**Total (%)**	**PG**	**UG-1**	**UG-2**	**UG-3**	**UG-4**
Cluster 1 n= 1418 (32.58%)						
Business School	369 (26.02%)	145 (10.22%)	43 (3.032%)	81 (5.712%)	100 (7.052%)	
Arts	41 (2.891%)		10 (0.705%)	7 (0.493%)	13 (0.916%)	11 (0.775%)
Health Sciences	178 (12.55%)	58 (4.090%)	23 (1.622%)	41 (2.891%)	45 (3.173%)	11 (0.775%)
Humanities and Social sciences	**581 (40.97%)**	225 (15.86%)	122 (8.603%)	111 (7.827%)	97 (6.840%)	26 (1.833%)
STEM	249 (17.55%)	79 (5.571%)	27 (1.904%)	77 (5.430%)	66 (4.654%)	
Cluster 2 n= 774 (17.78%)						
Business School	137 (17.70%)	57 (7.364%)	14 (1.808%)	36 (4.651%)	30 (3.875%)	
Arts	105 (13.56%)	1 (0.129%)	12 (1.550%)	32 (4.134%)	23 (2.971%)	37 (4.780%)
Health Sciences	73 (9.431%)	9 (1.162%)	11 (1.421%)	21 (2.713%)	29 (3.746%)	3 (0.387%)
Humanities and Social sciences	183 (23.64%)	61 (7.881%)	34 (4.392%)	32 (4.134%)	49 (6.330%)	7 (0.904%)
STEM	**276 (35.65%)**	54 (6.976%)	29 (3.746%)	87 (11.24%)	106 (13.69%)	
Cluster 3 n= 915(21.02%)						
Business School	203 (22.18%)	54 (5.901%)	44 (4.808%)	62 (6.775%)	43 (4.699%)	
Arts	6 (0.655%)			3 (0.327%)	1 (0.109%)	2 (0.218%)
Health Sciences	140 (15.30%)	18 (1.967%)	42 (4.590%)	42 (4.590%)	34 (3.715%)	4 (0.437%)
Humanities and Social sciences	**294 (32.13%)**	131 (14.31%)	54 (5.901%)	57 (6.229%)	51 (5.573%)	1 (0.109%)
STEM	272 (29.72%)	25 (2.732%)	85 (9.289%)	79 (8.633%)	72 (7.868%)	11 (1.202%)
Cluster 4 n= 1245 (28.60%)						
Business School	224 (17.99%)	91 (7.309%)	31 (2.489%)	48 (3.855%)	54 (4.337%)	
Arts	395 (31.72%)		140 (11.24%)	106 (8.514%)	75 (6.024%)	74 (5.943%)
Health Sciences	85 (6.827%)	26 (2.088%)	18 (1.445%)	24 (1.927%)	15 (1.204%)	2 (0.160%)
Humanities and Social sciences	**315 (25.30%)**	53 (4.257%)	71 (5.702%)	107 (8.594%)	78 (6.265%)	6 (0.481%)
STEM	226 (18.15%)	42 (3.373%)	54 (4.337%)	61 (4.899%)	67 (5.381%)	2 (0.160%)

than 4 distinct actions. More than 50% of the courses in cluster 1 used 4 distinct activities (forum, assignment, resource and book) and more than 20% of the courses used 7 distinct activities (Figure:4.6). With regard to actions, more than 50% of the courses used three distinct non-passive (other than view) actions that include; submitted (77.8%), created(56.28%) and uploaded (52.89%). In order to see overall impact of engagement we divided the LMS activities into five major categories; assignment activities (i.e, assignment, lti, journal), communication and collaboration (chat, forum, wiki, glossary and workshop), assessment and survey (quiz, questionnaire and choice), tools for student management (attendance), interactive delivery content (lesson, page, folder, url, SCORM and book). For each activity percentage was calculated in each cluster (For example assignment is used in more than 90% of the courses in cluster 1). Later, activities and their percentage values are grouped in their relevant category and average of percentage value has been calculated and assigned to LMS category. For example cluster 1 scored more than 30% in the category of *assignment activities* (Figure: 4.12). Cluster 1 scored highest among other clusters in the category of *assignment based activities*.

Cluster 2 with smallest portion ($n = 774$ or 17.78%) showing lowest usage of online activities compared to other clusters. Majority of the courses are small ($n = 338$ or 43.67%) in class size. Also, majority of the courses in cluster 2 belong to the STEM ($n = 276$ or 35.65%). Cluster 2 shows no significant diversity action, majority of the courses are not diverse ($n = 564$ or 72.87%), where number of distinct actions are less than 2. More than 80% of the courses in the cluster 2, only used passive action (view only) and which is used only to view resource, assignments. Cluster 2 showed higher usage of activities like forum, resource and Url and overall. Cluster 2 ranked lowest in all LMS categories which shows that most of the courses are inactive and only used LMS for viewing resources.

Courses in Cluster 3 (n = 915 or 21.02%) stand 3rd with regard to usage of online activities. All courses in the cluster fall in the category of very diverse in activities (n = 915 or 100%) and more than 50% of the courses used 7 distinct activities. Most used activity items are; forum, assignment, book and resource. Majority (n = 332 or 36.28%) of the courses belong to large class size where number of students are in the range of 66 to 300. Majority of the courses belong to Humanities and Social Sciences (n = 294 or 32.13%). Cluster 3 scored highest percentage in the category of *interactive delivery of content*, *communication and collaboration* and *assessment and surveys*. This makes cluster 3 reveal more diverse courses which used range of different activities such as glossary, feedback, lti, lesson and book which other clusters rarely used.

Cluster 4 (n = 1245 or28.60%) has also demonstrated higher usage of online activities after cluster 3 and cluster 1. Majority of the courses in this cluster are small (n = 366 or 29.40%) and are from *Arts* (n = 395 or 31.72%). Majority of the activities performed are in the activity item: forum, resource and url. In LMS activities, after cluster 3, cluster 4 courses scored highest percentage in the category of *interactive delivery of content* and is the only cluster which used *tools for student management (attendance)*. Cluster 4 also showed high level of diversity in online learning activity (n = 1233 or 99.04%) and diversity in actions (n = 849 or 68.19%) as well. More than 40% of the courses used more than 5 non-passive actions such as submitted, created, uploaded etc (Figure:4.10).

4.7 Summary

Based on research findings from literature review, this study hypothesized that courses could be divided into distinct subgroups based on the characteristics of course as well as based upon associated online behaviors of the enrolled students. In this study, classification

of courses was attempted by mining the Learning Management System log data. The aim was to investigate the use of large-scale LMS data in one institution for more than 10 thousand courses and find out how various disciplines have made use of LMS tools. The investigation was based on the LMS tools usage, to see if distinct subgroups could be found that could provide futuristic ground for general and portable predictive model for similar courses. Clustering methods were applied on more than 4000 heterogeneous courses to divide them into smaller homogeneous groups. The aim of this study is to analyze these clusters by using students LMS log data and study emergent patterns. Overall, the results reveal that it is possible to see patterns in blended courses using online behavioral data recorded in LMS logs. This study clustered 4,353 courses that were systematically sampled from a large pool of 10k courses, that were representative of relatively high level of online activity. Using hierarchical clustering approach, 4 clusters of courses were identified. Cluster 1, comprising 32.5% of the courses is the largest cluster and demonstrated high level of online activities in the category of assignment, forum and resource. More than 50% of the course used non-passive actions in activity item assignment specifically. Cluster1 can be titled as assignment-based activities since almost 22% courses have utilized quiz module of Moodle. Cluster 2 which is smallest in number (n = 774), is the group of most inactive courses which used Moodle; it was used only for accessing course materials and most of the time was used only for passive activities like view. Cluster 2 showed higher usage of activities like forum, resource and URL. Cluster 3 (n = 915) was most diverse in terms of online activity items utilized. More than 50% of the courses utilized seven distinct activity items such as glossary, feedback, lti, lesson and book which other clusters rarely used. Cluster 4 also showed high levels of online activities as cluster 1, especially in the category of 'interactive delivery of contents' and few of courses in the cluster used 'attendance' tool which no other cluster used. It is not possible

to distinguish between the clusters in terms of academic disciplines. This study found that majority of the courses were from 'Humanities and Social Sciences' discipline. All clusters had majority courses from 'Humanities and social sciences', except cluster 2(inactive courses) where majority of the courses were from STEM areas. It is important to note the imitations in the conduct of this study. For example, types of data used in this study included binary, numerical and categorical variables, and data was not normally distributed which made it challenging. In converting the data into binary and ordinal categories, there is loss of information. Another way is to make meaningful categories. Different input parameters within the same dataset can produce different clusters. But applying cluster on all combination of parameters is computationally expensive. The selection of the active variables is crucial and requires expert knowledge. Moreover, missing values have been omitted, whereas imputations techniques could have been used to overcome this limitation. The choice of distance measure is Euclidean which is the default value, while other distance measures could be investigated.

Chapter 5

Predictions of Students' Performance Using Online Engagement Data

5.1 Introduction

Higher education institutes are facing challenges due to low course enrollments and further lower course completion rates. The issue of low dropout rate is fast becoming a priority, and universities are seeking for strategies to improve students retention rate. According to the report of OECD [189] in Australia just 31% of students' completed a 4-year degree programme, US had 49% completions while UK is on top with 71% completions. Lower retention rates are a serious threat to universities long term financial security. Universities, therefore, are focusing more on identifying strategies that ensure students successes and which can provide proactive actions to support students in their course work.

Having some analytical strategy which can enable predictions on students performances can help these institutes to make timely interventions for improving students' performance. The widespread use of tools (e.g., SMS, LMS) have supported higher education institutes in providing seamless online communication, in delivery of learning and teaching resources, designing interactive learning activities and managing academic assessments. Besides they also provide them with large datasets that are related to students demographics student academic records and log files. These logs are based upon students interactions with the LMS and have offered us with new research directions that can help in improving students' academic performance [17] [18]. There are many success stories related to how data extracted from the tools that used student data has helped to improve the overall retention rate [19].For example, Georgia State University used predictive analytics which led to improvement in graduation rates from 32% (2003) to 54% (2014) [20]. Purdue University, USA, predicted the at-risk students as early as the second week, which led to increased support for these students resulting in improvements in their academic performance [21].

In this chapter, LMS data from different types of courses has been analyzed to see how accurately student academic performance can be forecasted when their weekly engagement data is integrated with assignment scores. This study highlights the importance of LMS data; which can give insights on student behaviors and can lead to development of accurate models that can be used for predicting the students final outcome in enrolled courses. Finally, this chapter demonstrates the suitability of multi-variable regression algorithms to predict the performance of classification algorithms. This study proposes an approach that utilizes historical data and machine learning experience using meta-learning to recommend best subset of classification algorithms for predicting students' performance in a course.

5.2 Motivation

The motivation of the chapter is as follows.

- Investigate the effectiveness of LMS log data to identify students who are likely to drop-out from one month MOOC courses.

- Analyze the impact of LMS log data and assessment scores across different types of courses (distance vs internal) to enable early predictions related to students final outcomes.

- Propose a multi-label regression model that utilizes meta-knowledge of education datasets and predict best classification algorithms' performance and recommend the best algorithm that has higher expected accuracy.

Table 5.1: List of events included in dataset.

S.No	Activity Item	Description
1	Problem	Working on course assignment
2	Video	watching course video
3	Access	Accessing other source than video and problem
4	Wiki	Accessing course wiki
5	Discussion	Accessing the course forum
6	Navigate	Navigating other parts of course
6	Page-close	Closing the web page

5.3 Experiment 1: Prediction of Dropout Using MOOC dataset

This experiment's focus is on predicting students' dropout by using event logs and to investigate the importance of student engagement activities and further evaluate their correlation with student retention. In this case study, MOOC datasets have been used for making predictions on student dropouts based on a count of their online activities.

5.3.1 MOOC Dataset

This study has utilized MOOC datasets. Five courses were chosen and used the event log of enrolled students. Events logs contain timestamps for a list of event types (as shown in 5.1). The sum of total events performed in a day were counted. The main dataset comprised thirty variables for representing the sum of total events performed each day by a student. The last variable is either 0 or 1 which represents the dropout status of the student (i.e., not dropout or dropout respectively.) There were total of 40 courses in the dataset; however, for final dataset, those courses with larger number of students have been considered. A detailed description of the courses is given in Table 5.2. A student enrolled in the course will be considered as a dropout if they leave no record till 10 days after the last day of that course.

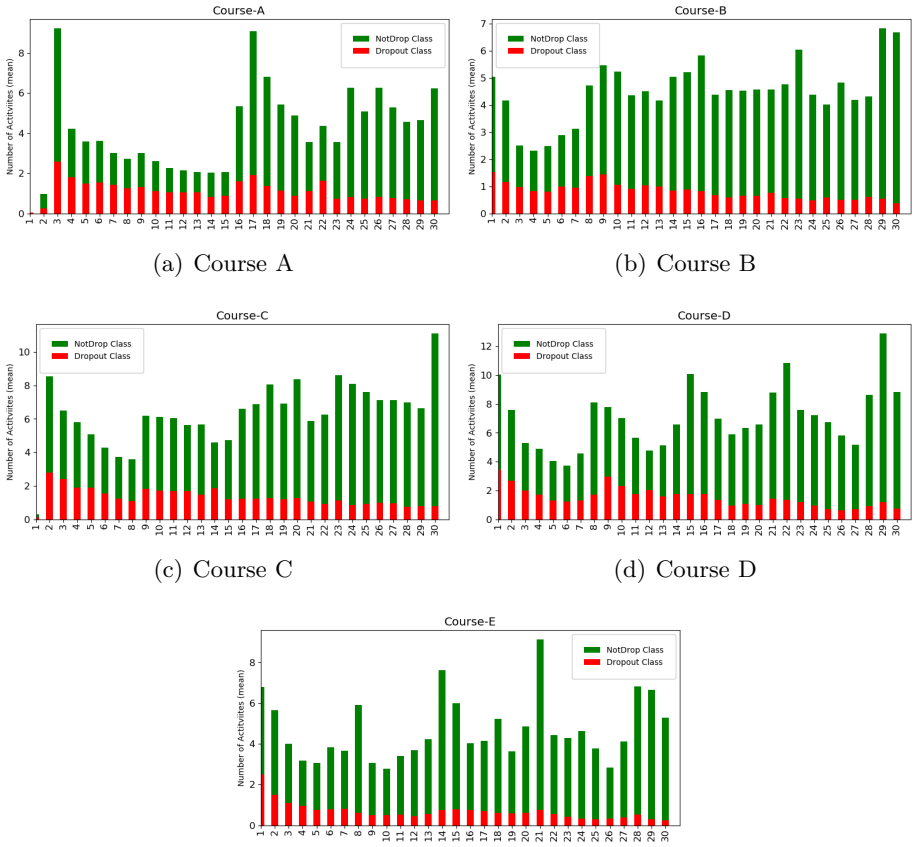
Table 5.2: Number of students enrolled in the selected courses.

Course	Enrolled students	Drop-out	Not-Dropout
A	120004	7186	3136
B	7775	6479	1296
C	10322	7186	3136
D	9382	6501	2881
E	3005	2597	408

Figure 5.1 shows the mean number of activities students performed daily and compared two groups (Dropout vs Non-Dropout) in a course. It is quite obvious that there is significant difference between two groups in terms of online engagement. Student who dropped out are seen to be less active throughout the duration of the course compared to the not-drop-out class. A common observation across all courses was that after mid-way into the course, the activity levels increased.

5.3.2 Experimental Design

The objective of this case study was to identify those students who are most likely to drop out by using their log activities. The question was: Can counts of log activities should be used to predict the likelihood of a student dropping out of the course? For classification we used following machine learning algorithms: Naıve Bayes (NB), Random Forest (RF), Logistic Regression (LR) and K Nearest Neighbor (KNN). These methods are widely using in Education Data Mining (EDM) and are considered well suited for such a domain. Experiments were performed for five datasets (i.e., one dataset for each course). To evaluate the performance of each machine learning techniques on the test set, three performance criteria were used: F-Measure, AUC, and accuracy. To estimate the generalization capability of the model for future dataset, 10-fold cross validation technique was used. Performance of the classification methods were then evaluated using overall accuracy, F1-score and using ROC curve. These



(a) Course A

(b) Course B

(c) Course C

(d) Course D

(e) Course E

Figure 5.1: Average number of activities performed by two groups of students during the 30-days course

classification methods have been used for the prediction of students' final status in one of two classes: Dropout or Non-dropout. Prediction was based on the count of activities that students perform daily during the course. Figures 5.1 shows the average number of activities performed by two groups of students (Dropout and Non-Dropout) for different courses. The duration of the courses shown was 30 days. The figure shows difference in activity levels between two extreme groups for five courses.

5.3.3 Results and Discussion

This section answers the following research question in light of the analysis *Which machine algorithm predicts the likelihood of students dropping out with high accuracy?*

Table 5.3: Performance of different classification algorithms for MOOC dataset.

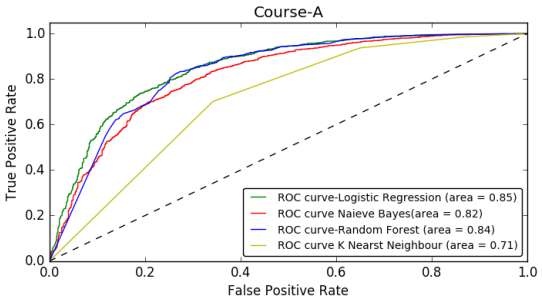
	LGR		KNN		NB		RF	
Dataset	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Course A	0.839	0.829	0.803	0.780	0.808	0.807	0.806	0.753
Course B	0.88	0.875	0.858	0.833	0.858	0.861	0.858	0.813
Course C	0.840	0.832	0.798	0.784	0.818	0.814	0.797	0.776
Course D	0.833	0.824	0.785	0.768	0.816	0.812	0.793	0.771
Course E	0.893	0.876	0.874	0.848	0.867	0.866	0.881	0.839

Table 5.3 shows the performance comparison for different classifiers used (i.e., NB, RF, LR and KNN). In dataset since counts of positive and negative class are not same, this makes it an imbalanced dataset. So, a baseline classifier would give more weight to the majority class resulting in a biased result. Therefore, in this case study, the results were compared and based on the F1 score and area under ROC curve, the best classifier was chosen. Results show that Logistic Regression outperformed other algorithms both on basis of overall accuracy and F1-score. For all datasets, the overall accuracy showed 1 to 2% more than F1-score; however, due to the imbalanced nature of the dataset, F1-score is considered as a final metric for comparison. Maximum accuracy obtained is 0.89 and F1-score is 0.876 for course E, which is a small dataset compared to the other datasets. It contains almost 3000 records and the difference in activity level between two groups is quite huge which further made classification easy and two groups were separated with high accuracy. Other classifiers' performance is similar, usually 3% less accurate than Logistic Regression. For all courses that were observed, similar results and maximum F1 score were obtained by Logistic Regression. Overall maximum F1-score obtained is 0.875 for course B and course E. One similarity between courses B and E is that,

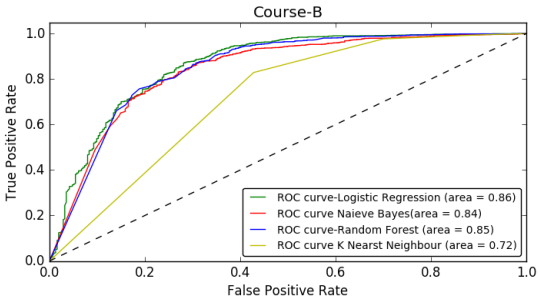
in both courses the activity level for dropout students was very low and remained almost constant after first week. However, the non-dropout students were active and their activity level was not constant throughout the duration of the course. Further comparative analysis between the classifiers considered ROC. ROC curves for each of the classifiers are shown in Figure 5.2. The x axis is the false positive rate, or in our case it is the percent of students that continued the course and whom were identified as likely to drop out. The Y axis is the true positive rate or the percent of all dropouts were correctly identified as likely to drop out.

Performance across the classifiers is comparable. Minimum AUC (area under curve) is 0.71 by KNN and maximum AUC (0.85) is gained 0.85 by Logistic Regression in course A. For all courses we get similar kind of results. Maximum AUC achieved is 0.89 for course C and D by logistic regression. Course C and D are among the largest datasets with almost 10,000 students enrolled and activity level for not dropout students is more than the other course. Above results show that event log data can be a strong signal for predicting students' dropout. However, these results can be regarded as baseline which can be further improved by integrating more features. Nevertheless, it is useful in those cases when we need to make early predictions during the first or second week of the courses, especially since we may not have assessments, rather just have the event logs that reveal student interactions with the learning management system.

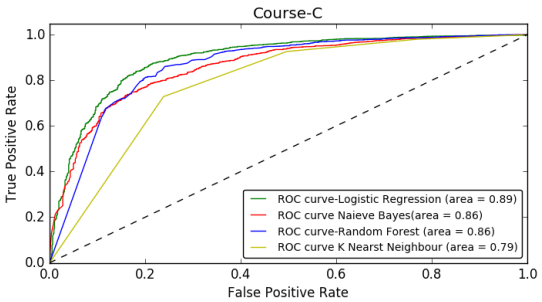
Identification of probable dropout students is only helpful when accurate prediction is made as early as possible, ideally before the mid of the course so that timely interventions can be made. Given this context, we performed predictions after every six days. A new dataset was divided into five sets: Day-6, Day-12, Day-18, Day-24, and Day-30. Here, Day12 means the event logs comprise activities until the 12th day of the course. Figure 5.3 shows



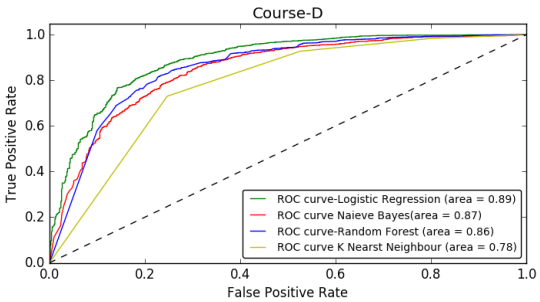
(a) Course A



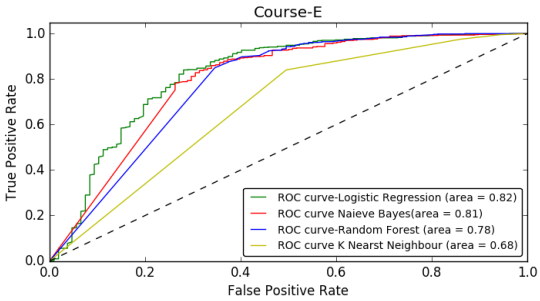
(b) Course B



(c) Course C



(d) Course D

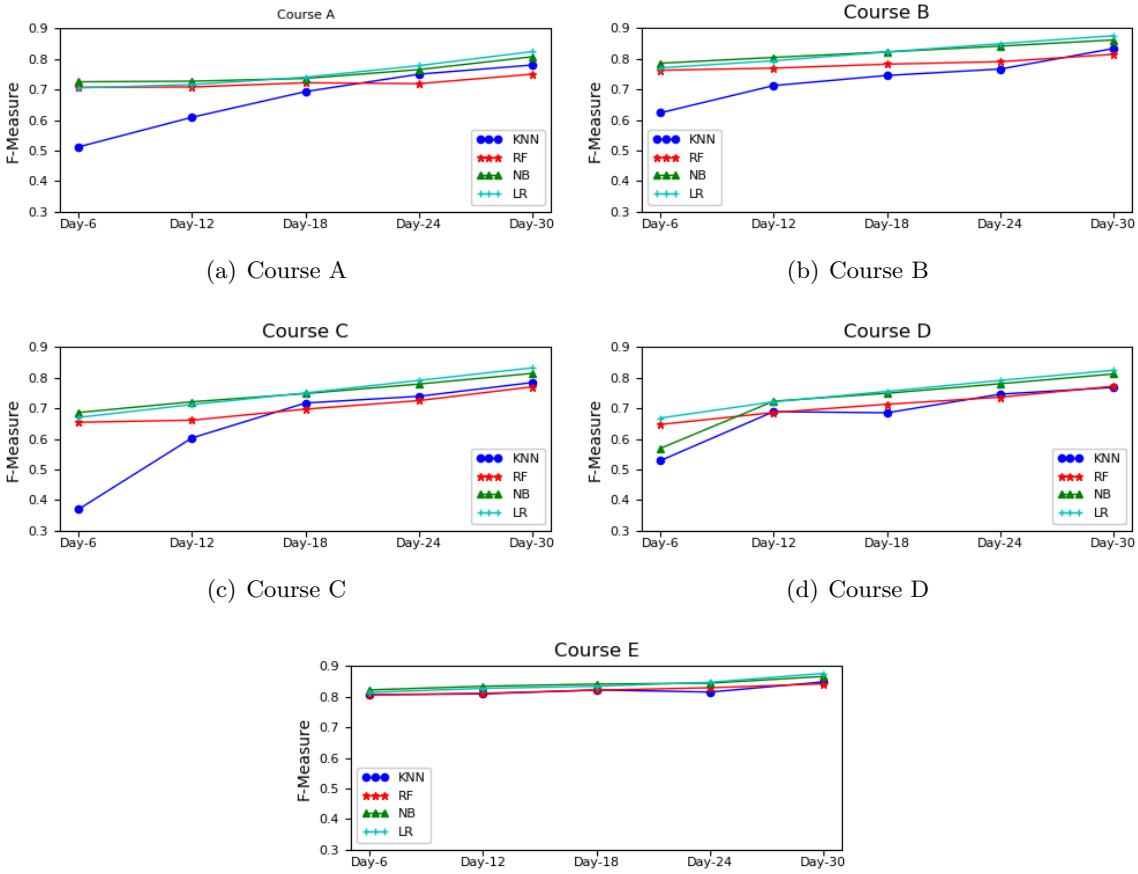


(e) Course E

Figure 5.2: Comparative results of the machine learning algorithms for prediction of dropout

the performance of machine learning techniques for making predictions of dropout during the course. Y axis show the F1-score and X-axis shows the dataset used for prediction. Overall results show that prediction accuracy improves with time as more engagement data becomes available. The best scores that we achieved are for course E, that is, minimum F1-score achieved is 0.81 after 6 days which increased to 0.87 by the end of the course. Logistic regression and Naive Bayes performed better than Random Forest and K Nearest Neighbor.

As time increased, the prediction accuracy increased faster with KNN in comparison to other classifiers.



(a) Course A

(b) Course B

(c) Course C

(d) Course D

(e) Course E

Figure 5.3: Comparative results of the machine learning algorithms for prediction of dropout

5.4 Experiment 2: Weekly predictions of students' performance

In this experiment, the LMS data from courses have been analyzed to see how accurately we can forecast student performances when their weekly engagement data is integrated with assignment scores. This study investigates the effects of LMS trace data and the grades of assessment scores on the prediction of students' outcome in the course. Prediction of students'

outcome is classified in to two classes; 'at-risk' (students who are at risk of failing the course) and 'not-at-risk' (students who are not at risk of failing the course).

5.4.1 Dataset

Data were collected from an Australasian tertiary education (See section 3.2 for details) provider. The data extracted from two different sources: one from official LMS (Moodle) which produced detailed activity logs for each student and resulted in a total of 6,83672 raw records and the second source was the provider's official student management system (SMS) which contained 400 students' profile information, assessment scores and other information that are required during the registration process. Trace data for detecting student interactions with LMS (Moodle) was extracted. Following are some examples of Moodle features/components that were accessed during the course: assignments, resources, forum, book, quizzes and chats. This trace data represents the number of times student used particular component of Moodle, hence the data type is numerical. Next, data pre-processing steps were undertaken so as to generate a dataset that is in a format on which algorithms can be applied. This step is not straight forward, because there are various sources of data. The main challenge is to identify the relevant data, capture it, extract it and integrate all data to get useful information. One of the challenges faced during the extraction process was that both sources have different primary key for course related information, so script was written to extract information and integrate from both sources. The final dataset thus generated included academic scores and activity data that was captured from log files and which is mostly referred as LMS engagement data. Data extracted from both sources were encrypted, in order to take care of student's privacy. Other pre-processing steps further involved handling missing values, looking for outliers and changing some data formats to align it with the overall database

structure.

Datasets have been extracted from two types of courses

1. **Distance courses-** In distance courses, students were required to access digital study resources, contribute over online discussion forums and complete other online tasks and activities.

2. **Internal courses-** In internal courses, students were provided course materials through LMS. However, it was not compulsory to be active in online forums, so there are less online activities compared to distance courses.

Weekly Datasets

The datasets were divided into two classes; *at-risk* and *not-at-risk*. These classes are extracted based on the values of final grade code of courses. Because the data was not balanced, there was a need to divide the grade-code in two classes. Division of grade codes in two classes are given in Table 5.4. Those students whose final grade is less than 55 and those who had withdrawn (and were not continuing their study) have been considered as at-risk of failing the course.

Table 5.4: Grading Schema

Not-at-risk								**At-risk**				
A+	**A**	**A-**	**B+**	**B**	**B-**	**C+**	**C**	**C-**	**D**	**E**	DNC*	WD*
90-100	85-89.99	80-84.99	75-79.99	70-74.99	65-69.99	60-64.99	55-59.99	50-54.99	40-49.99	0-39.99	NA	NA

The course has been structured in a weekly format, that is, since these students were provided with weekly learning materials, hence they engaged with the new material and also performed expected tasks on a weekly basis. Logs were maintained on a daily basis; however, for matching the logs to the course structure, the events too were aggregated into weekly

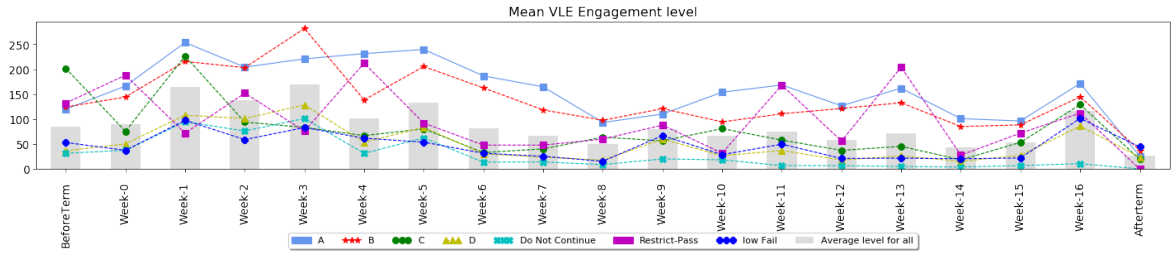


Figure 5.4: Average assignment scores and LMS engagement level for different groups of students.

Table 5.5: Summary statistics of datasets.

S	Course title	Code	Type	Course Size (Number of students)	Number of assessments	Number of exams	At-Risk: not at-risk	V1*	V2*
1	Introductory mathematics	C1	Internal	93	5	1	49:44	29	49396
2	Introductory mathematics	C2	Distance	78	5	1	28:50	56	130304
3	Software development	C3	Internal	94	5	1	47:47	39	203696
4	Introduction to finance	C4	Distance	125	2	1	72:53	49	300276

V1* Total number of unique events performed in the course.
V2* Total number of records in the event log of a course.

format. As the results of assignments scores become available, they are included in the LMS engagement data to train the model. Once dataset was extracted from a course, it was further divided into 16 sub-datasets that represented weekly times.

Following equation shows how we get the dataset for week number n.

$$w(n) = \sum_{i=0}^{n} W^i$$

For example, week-4 dataset can be defined as $w4 = w0 + w1 + w2 + w3 + w4$

Predictor variable for a given week is count of total online activities done that week and second variable could be assessment scores (if available at that time).

The data captured over the duration of the course and for the duration before the course started is referred as *before-term*, and the data for the time after the course ended is referred as *after-term*. All logged activities were considered, though some of these activities might not be related to learning directly, for example like downloading material. But such activity

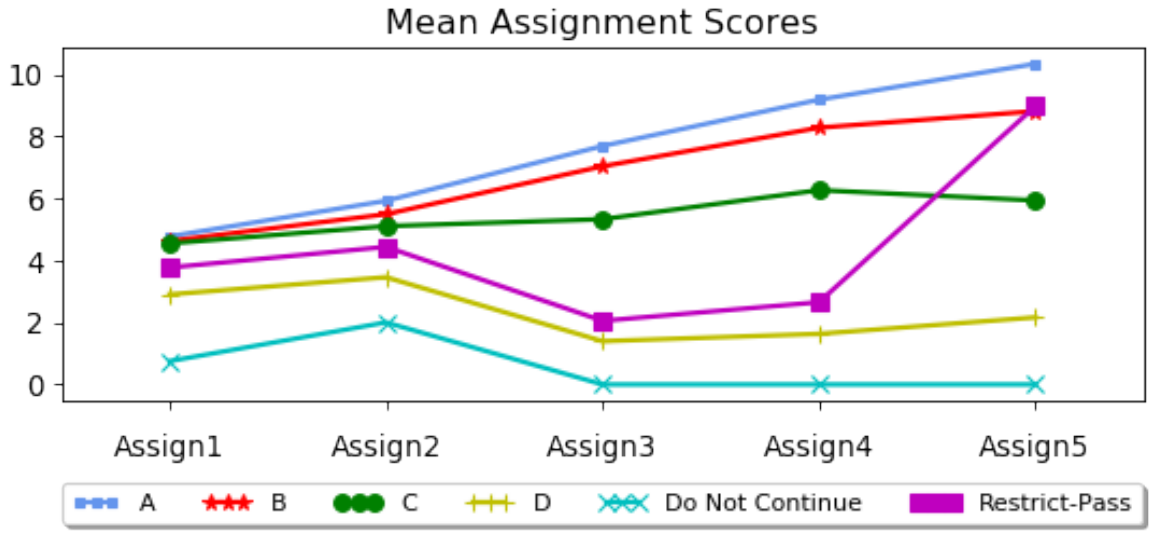


Figure 5.5: Average assignment scores for different groups of students

indicates that the student was engaged with the course. Assessment scores extracted from SMS complemented the analysis. Students have been divided in different groups (i.e., A, B, C, D, do not continue etc.) based on their final grade in course. Figure 5.4 shows an example of a course with average engagement level of students during the course for different groups of students and Figure 5.5 shows average scores in assignments of all groups for a course having a total of five assignments, and where the first assignment submission was due in the second week of the semester.

5.4.2 Experimental Design

To classify students in two group as either *at-risk* or *not-at-risk* of failing the course, following machine learning algorithms were used: Random Forest classifier (RF), Naïve Bayes (NB), Logistic regression (LR), Linear Discriminating analysis (LDA) and ensemble classifier which used weighted majority vote of single classifiers(RF, LDA, LR, NB). In this experiment, four datasets extracted from four different courses were analyzed. Brief description of the courses are as follows (Table 5.5).

1. **C1:** First dataset contains information about 93 undergraduate students extracted from course *Introductory University Mathematics*. This course is designed to increase the mathematical concepts and skills and is partially taught online. There are a total five assessments and one final exam over 16 weeks of course teaching.

2. **C2:** Second dataset is extracted from same course as C1 but now instruction mode was completely online. All resources were provided using digitally media and students completed online activities and were submitted their assignments online. The structure of the course in terms of assessment was same as C1. Total number of students enrolled in this course was 78.

3. **C3:** Third dataset is extracted from course *Application Software Development* that was taught internally on campus. Total number of students in the course was 94 out of which half of the students belonged to *at-risk* class.

4. **C4:** Fourth dataset is extracted from course *Introductory Financial Accounting*. This course is fully taught online and the number of students enrolled in the course were 125. There were two assessments for 16 weeks and one exam. After week 4 there is an online quiz and after week-10, the assignment scores were made available. 53 students out of 125 were considered as at risk of failing the course where 72 were considered safe.

All the classifiers were trained using 10-fold cross validation in order to counter the possibility of over fitting or biased results due to the small sample sizes. Cross-validation is preferred method that evaluates trained models using *unseen data*. In 10-fold cross validation, data-set(training) is divided into equal folds(subsets). The model is repeatedly trained on nine folds and tested on the remaining one-fold. The final accuracy results are the mean of the 10 independent training models.

5.4.3 Evaluation Metrics

The performance of machine learning classifiers was compared, in classifying the student's outcome in the course in two classes i.e., *at-risk* and *not-at-risk* of failing the course. Due to imbalanced datasets, we evaluated the performance of the classifiers with more than one measure: accuracy and F measure. F-measure is considered better evaluation method than accuracy in case of imbalanced datasets. F-measure is defined as harmonic mean of precision and recall.

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.1}$$

Where precision (also called as positive predictive value) is the proportion of positives in the accepted set $\frac{TP}{FP+TP}$ and recall (also called positive rate) is the proportion of all positives that were included in the accepted set $\frac{TP}{TP+FN}$.

5.4.4 Tools

For conducting experiments, this study used Python module Scikit-learn. Scikit-learn is increasingly popular machine learning library that integrates wide range of algorithms for supervised, unsupervised and semi-supervised problems. It provides a user-friendly environment for non-experts in machine learning and is reusable in different scientific platforms. The library is distributed under BSD license proving opportunities of its application and use in both academic and commercial setups. For visualization and plotting of results another Python module Matplotlib has been used.

5.4.5 Results and Discussion

As shown in Table 5.5, the size of the class varied from 78 to 125 students in class (mean= 97.7, SD= 19.7). The courses were taught in duration of sixteen weeks. Number

of assessments in each course varied from 2 to a maximum of 5 assessments. Two variables; V1 and V2 represent the online activity level in the courses. V1 represents the total unique activities performed in the course, whereas V2 shows total number of records in the log of the course. Activities presented in Moodle are recorded as the interaction between student and instructor or of the student with available resources. Some of the examples of activities are: view-course, create chat, submit-quiz etc. In the present study, all possible activities that could occur in Moodle were included. This has been restricted up to students and log activities of instructor or other users of Moodle have been excluded. Courses are classified in two types; distance and internal. In distance courses, teaching is done completely online while internal courses are partial online, that is while students still used the online platform (LMS) for communication and access of online resources, it was not compulsory. This study used following classification methods to predict student's outcome (at-risk or not-at-risk) in the course: Random Forest, LDA, Naıve Bayes, Logistic Regression and Ensemble classifier. The evaluation results are presented in Figure 5.6. Experiments have been repeated for each course separately. Sixteen datasets were created for each course, resulting in sixteen results for each course and total of 64 results across all courses. Following subsections present the results of the analysis for each course.

Dataset C1

F1-score achieved after every weekly prediction is shown in figure 5.6. Results show the prediction accuracy of each classifier. Minimum F1- score achieved in week-1 is 0.33 which improved to 0.90 at end of week-16. Here, the important thing is prediction accuracy at week-8 which is near the middle of the course, since the aim is to identify students who are at risk of failing the course as early as possible. Performance of each classifier varied throughout

the experiments. Maximum F1-score achieved after week-1 is 0.47 by Random Forest and Decision Tree. After week-1 there was 10% improvement in accuracy which resulted in 0.57 maximum F1 score by Random Forest and lasted to 0.59 for week-3. After week-3, when the first assignment scores became available the accuracy boosted by 34%. Maximum F1-score achieved after week-3 is 0.85 by Random Forest and LDA. From week-4 to week-10, maximum F1 varies between 0.85 to 0.87 showing 1 to 2% improvement. After week-10 accuracy reached to 0.9 and this stayed almost same until week-16. Figure 5.6 shows that there is always improvement in accuracy once assignment scores are available and these scores are then included in the predictor variables. Vertical lines in the figure shows the week number when assignment scores become available. Overall Random Forest and Ensemble Classifiers provided maximum accuracy throughout the experiments. Accuracy improved significantly after week-3 when assignment-1 scores become available, which adds to the discriminatory power of all predictive models. This also shows that assessments scores have more discriminatory power than LMS data. As this course was taught internally, LMS activities might not be enough to differentiate between two groups of students. Number of unique event or activities are fewer and total log records also show that internal courses may not be rich in terms of LMS data when compared with other courses(see Table: 5.6). Results also show that LMS data in courses where online activities are not compulsory or where LMS is used just for merely accessing resources, then the data are not enough to enable predictions regarding student outcomes in such courses.

Dataset C2

Figure 5.6 shows results of predictive models for dataset C2. This dataset is extracted from a course which is totally taught online and it can be seen that this dataset is rich in LMS

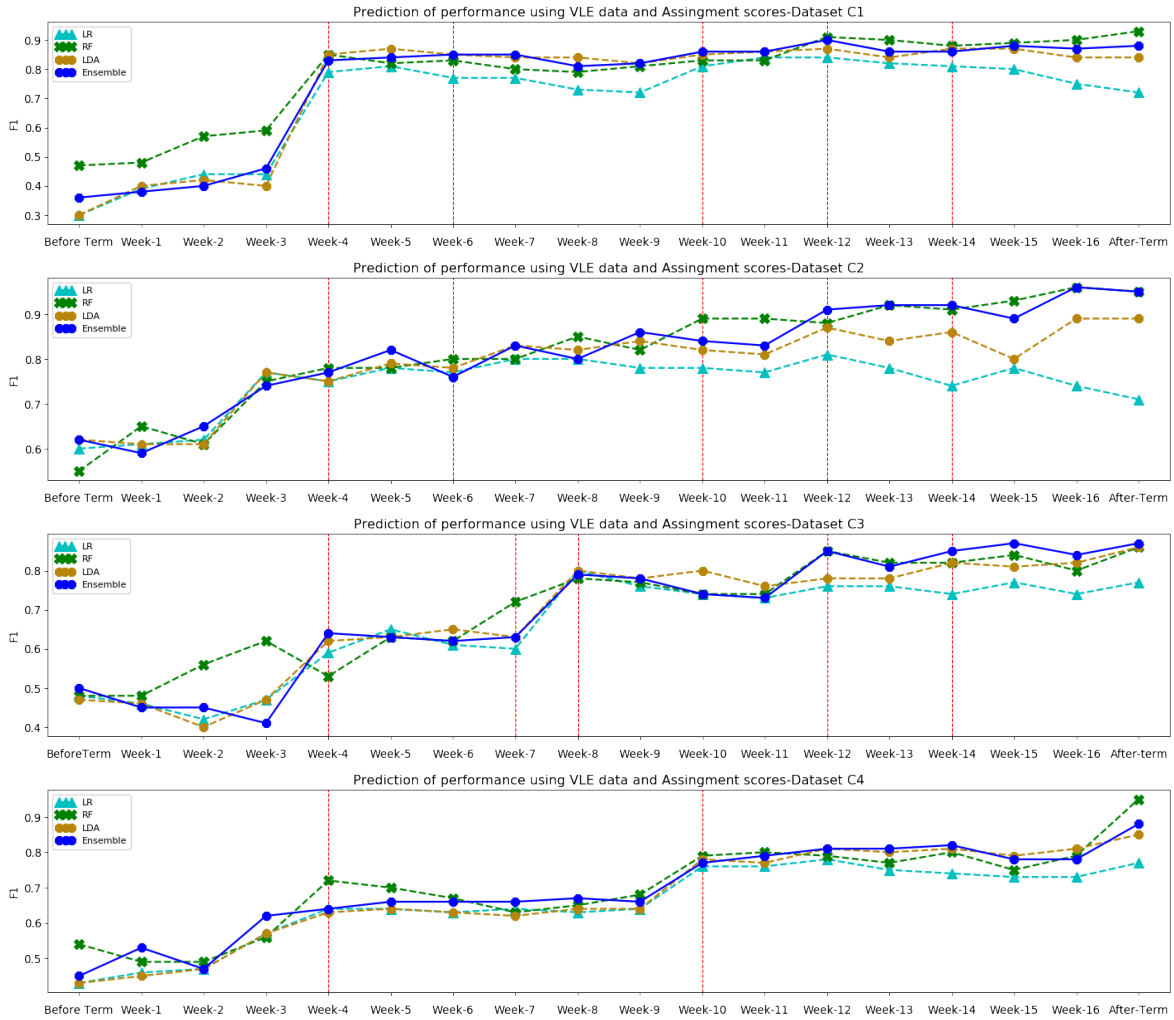


Figure 5.6: Prediction results of classifiers; predicting students' outcome in a course as *at-risk* or *not-at-risk* of failing the course. Evaluation measured used in the experiment is F1.

Table 5.6: Summary statistics of dataset their mean activities per week and maximum F1-score achieved by classifier for each weekly dataset. Here V1* is mean of activities count per week and V2* is maximum F1-Score achieved per week.

Dataset	Variable	Week-1	Week-2	Week-3	Week-4	Week-5	Week-6	Week-7	Week-8
C1	V1*	58.06	44.59	54.23	30.91	39.37	17.31	5.22	11.37
	V2*	48%	57%	59%	85%	87%	85%	85%	84%
C2	V1*	163.24	135.39	166.48	99.70	130.94	81.11	66.48	49.82
	V2*	65%	66%	77%	80%	86%	85%	84%	81%
C3	V1*	121.32	123.70	165.06	29.57	45.06	149.02	110.26	104.26
	V2*	52%	56%	62%	64%	65%	65%	72%	80%
C4	V1*	210.06	162.38	124	129.61	145.77	169.07	90.44	116.76
	V2*	53%	56%	62%	73%	71%	67%	66%	66%

data compared to the previous dataset. Number of assessments and their due dates are like dataset C1. However, results of predictive models are different from C1. Maximum F1-score achieved in week-1 is 0.65, almost all classifiers' performance is in similar range $(0.60 - 0.65)$. After week-2 there is 11% improvement in the accuracy which has resulted in maximum F1-score of 0.78 by Random Forest. After week-4 to week-12, F1 score remained in the range of 0.81 to 0.87. Accuracy improved when assessment scores become available, but range of improvement is from 2 to 5%. Maximum boost to accuracy achieved after week-3 which was like the previous dataset C1. Both C1 and C2 datasets are extracted from similar course, the only difference is about the pattern of teaching (internal vs distance). Results show that in this case, LMS data in is more useful than the data extracted from internal course. Accuracy is more stable in distance course and assignment-1 score added more value to the predication accuracy in both internal and distance course. Although One can not deny the importance of LMS data in internal courses as it can still achieve accuracy of 55% without assessment scores after week-1. However, assessment scores have got more discriminating power in both courses (C1 and C2).

Dataset C3

Figure 5.6 shows the prediction results of classifiers for dataset C3. This dataset is extracted from a course which is taught partially online. Which means that LMS was used to communicate with students and for access of digital resources. However, there were no compulsory activities to be performed in LMS. There were five assessments during the sixteen weeks of the course. Assignment-1 scores were available after week-3. This dataset is balanced, that is the ratio of students at-risk to the students not-at-risk is 1:1. Maximum F1-score achieved after week-1 is 0.52 by Random Forest. After week-2 there is 6% improvement

in accuracy and maximum F1-score achieve is 0.62 by random forest classifier. From week-1 to week-6 accuracy ranges from 52 to 65%. After week-4 when scores of first assignments were made available, this provided only 1% improvement in the accuracy (F1 increases from 0.64 to 0.65). After week-6, accuracy improved by 12% and reached to a maximum of 72%. After week-8 accuracy remained in the range of 80 to 85%. Assessment scores did not improve the accuracy to the level that was observed in other datasets. However, the accuracy improved after every week by 1 to 4% on an average. This dataset shows that both LMS data and assessment scores are equally important.

Dataset C4

Figure 5.6 shows the prediction accuracy of predictive models for dataset C4. This dataset was extracted from a course which is taught completely online. Number of assessments in the course are two. Maximum F1- score achieved after week-1 is 0.53 by ensemble method. After week-2 there is 6% improvement in accuracy. First assignment score is available after week-3 which increased accuracy from 62 to 73%. Unlike other datasets, accuracy decreases by 1 to 9% after week-4 until week-9. After week-9 scores for assignment two become available and which improves the accuracy. Maximum F1-score achieved after week-9 is 0.8 in week-12. This dataset is rich in LMS data compared to other datasets. However, number of assessments are less, which effected the overall accuracy of predictive models. Moreover, there is no consistency in the scores of predictive models. Unlike other datasets, accuracy is not constantly improving after each week which confirms the value of assessment scores. It is shown in Table 5.6 that average online activities of students during the course is more than other datasets, which means students have used LMS more frequently, so it may not be as discriminating when compared to the assessment scores.

Identifying the best machine learning algorithms for a given task is a challenging task, especially in the field of education. In this method we will exploit meta-learning information to identify best algorithms in terms of accuracy.

5.5 Application of meta-learning in predictive analytics

There are scores of algorithms used in different domains of machine learning, statistics, artificial intelligence to extract knowledge from large volumes of data. Identifying the best machine learning algorithms for a given task is a challenging task, especially in the field of education. It takes considerable effort in terms of time and knowledge to select the most suitable algorithm with different combination of parameter settings to solve a task. According to the no-free-lunch theorem there is not a single universal learning algorithm which performs uniformly better than the other. Therefore, no single recommendations can be made for an arbitrary data, rather the performance of classification algorithms depends on the characteristics of underlying data. This problem was originally defined by Rice [190] which defines a framework to select and apply algorithm(s) for a given task (Figure: 5.7) .

Meta-learning covers any type of learning that are based on prior experience with related tasks. Meta-learning is sub-domain of machine learning [191] which is used to recommend best learning algorithms for a given task by learning from the past experiences in related learning tasks. It has been used in past for optimization as well [192]. Following are the steps involved in meta-learning for predictive models

- **Dataset characterization** Dataset characterization is the process of extracting high-quality characteristics or meta-features of dataset that can provide distinguishing information regarding the performance of algorithms. It can be performed using techniques such as statistical, information theoretic or model-based approaches.

- **Algorithm selection** Algorithm selection is the process that maps an input space composed of datasets to an output space which is composed of predictive models. The goal of algorithm selection process is to learn from data and its characteristics and automatically assign a best performing algorithm. Performance criteria for predictive models could be accuracy, running time, memory etc.

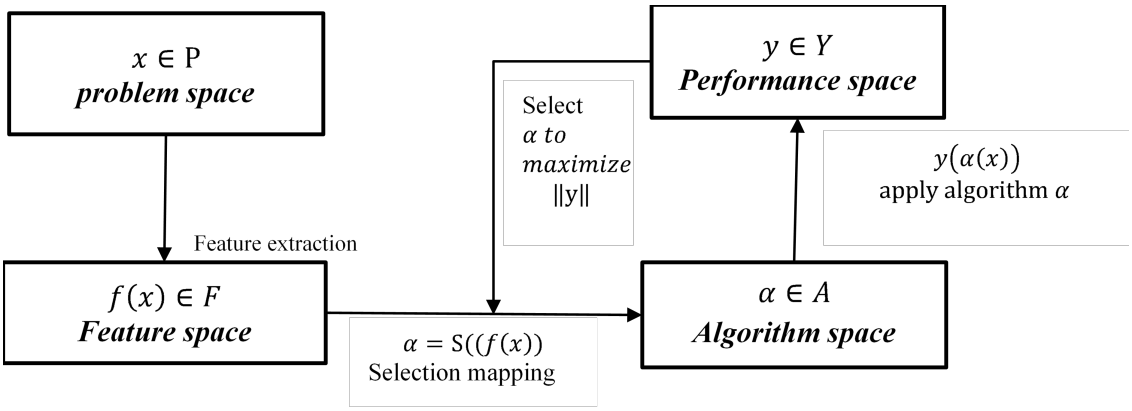


Figure 5.7: Rice's framework for algorithm selection

5.5.1 Meta-Learning Steps

Meta-learning definitions varies according to the domain in general; however, in simplest terms meta-learning refers learning to learn. The goal of meta-learning is to learn the relationship between data characteristics and the performance of the algorithm. Research in the field of meta-learning has focused on identifying most optimal meta-features that characterize the datasets well. Based on these meta-features, different classification, regression and ranking based methods are proposed.

Definition: For a given problem x where $x \in \mathcal{P}$ with set of features $f(x) \in \mathcal{F}$, find a mapping $M((f(x))$ in algorithm space A, such that selected algorithm α maximize the performance mapping $Y(\delta(\alpha \in A)) \in y$.

Above definition can be described as follows. The problem of selecting best machine

learning algorithm can be considered as a search problem with set of individual algorithms in search space as the aim is to identify the algorithm or set of algorithms exhibiting best performance. A general framework for recommending best machine learning algorithm is illustrated in 5.8. According to the framework, data repository containing a set of training examples from where meta-learning knowledge is acquired. These training examples relate to a subset of a problem.

Characterization measure of each dataset stored in the data repository are extracted, and which is labelled as meta-features in figure 5.8. Extraction of meta-features could fall to one of the following categories: simple, statistical, based on information theory and complexity. When performance measures of the algorithm are extracted as and when the algorithm is applied on the problem, then such measures are called target features. Set of input-meta features along with target-meta features are stored in meta-knowledge, which serves as training set for meta-model. Meta-model trains using meta-knowledge automatically maps the problems' characteristics with the input meta-features to enhance the algorithm's performance for test set. Meta-learning differs from base-learning since it makes use of meta-knowledge from previous experience. Therefore, one important step in meta-learning is the extraction of meta-knowledge from previous experiences that could benefit the search process; therefore, overall performance solely depends on the quality of the meta-knowledge.

5.5.2 Empirically Evaluation

The proposed framework was applied to recommend the best performing algorithms for classification of students' performance. Data repository consisting of 465 datasets, were used for prediction of student's performance in each course at the end of the semester.

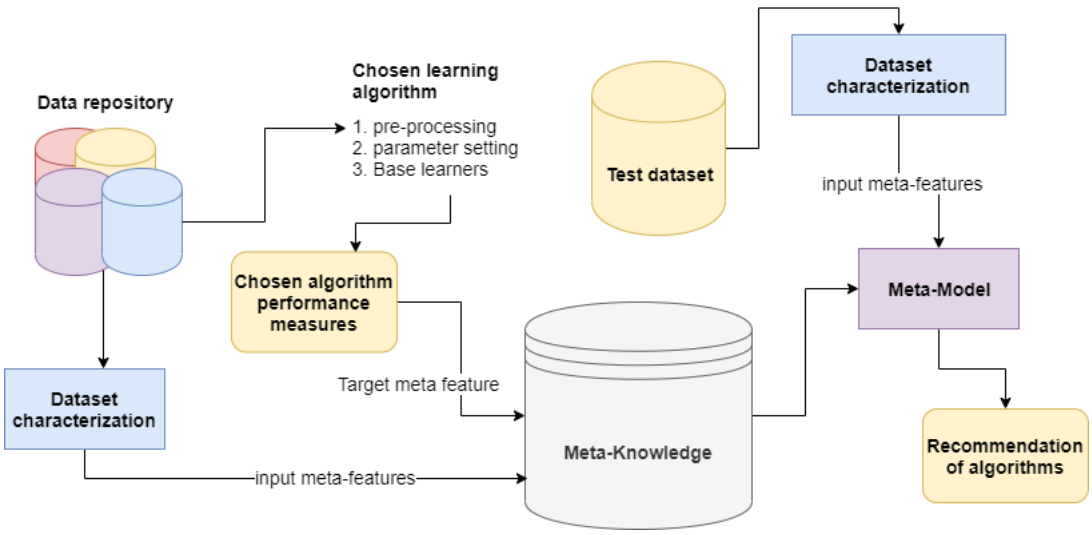


Figure 5.8: Meta-learning approach for algorithm selection (modified from)

Preparing datasets

Dataset is created from two different sources: LMS and SMS. The selected courses were taught between 2015 to 2017 and were from different domains of undergraduate and post-graduate levels. The frequency of log activities during the semester were considered as independent variables along with demographic information of students. The predicted variable is the final outcome of the student in the course as a binary class (at-risk or not-at-risk). Dataset for $n = 33$ courses is generated using above discussed approach in the attribute-value format and stored into the data repository. Data repository contained 465 number of datasets with different number of features in each. Due to the diverse usage of LMS, there were scores of missing values for some LMS log features for some courses.

Classification algorithms evaluation

In previous section, dataset preparation steps were described where dataset for 33 courses were prepared and subdivided into weekly datasets. Each dataset comprised independent variables that is made of log activities and demographic information of students and pre-

Table 5.7: List of courses used in the study.

Id.	Course Title	Semester	Enrolment	Not-At-Risk	At-Risk
26896_158100_1601	Computer Applications and the Information Age	1601	100	42	58
29685_158100_1603	Computer Applications and the Information Age	1603	33	29	4
25418_158225_1502	Systems Analysis and Modelling	1502	81	61	20
25960_158247_1601	Database Design	1601	48	22	26
31308_158247_1701	Database Design	1701	56	20	36
25409_158258_1502	Web Development	1502	24	19	5
26052_158258_1503	Web Development	1503	127	103	24
29366_158258_1602	Web Development	1602	61	52	9
27702_158337_1601	Database Development	1601	13	11	2
31128_158337_1701	Database Development	1701	85	64	23
25793_110109_1503	Introductory Financial Accounting	1503	80	46	34
26215_110109_1601	Introductory Financial Accounting	1601	105	70	35
27794_110109_1601	Introductory Financial Accounting	1601	50	30	20
29455_110109_1603	Introductory Financial Accounting	1603	115	74	41
30723_110109_1701	Introductory Financial Accounting	1701	116	74	45
26204_110309_1601	Advanced Financial Accounting	1601	215	131	84
30472_110329_1701	Advanced Financial Accounting	1701	135	74	61
26832_110329_1601	Advanced Management Accounting	1601	228	128	100
30472_110309_1701	Advanced Financial Accounting	1701	274	167	107
25047_110303_1502	Integrative Accounting	1502	31	23	8
25048_110303_1502	Integrative Accounting	1502	72	70	2
25049_110303_1502	Integrative Accounting	1502	15	12	3
28816_110303_1602	Integrative Accounting	1602	39	27	12
28999_110303_1602	Integrative Accounting	1602	55	46	9
29152_110303_1602	Integrative Accounting	1602	24	23	1
24838_115107_1502	Management Information Systems	1502	186	112	74
25531_115107_1502	Management Information Systems	1502	155	112	43
27363_115107_1601	Management Information Systems	1601	174	123	51
28564_115107_1602	Management Information Systems	1602	615	349	266
29748_115107_1603	Management Information Systems	1603	65	37	28
25861_115101_1503	Statistics for Business	1503	222	126	96
27367_115101_1601	Statistics for Business	1601	322	166	156
29610_115101_1603	Statistics for Business	1603	40	26	14

dicted variable (or binary class) having two values that helped in determining the student's final outcome in the course (i.e., at-risk or not-at-risk of failing the course). In this section classification algorithms have been used to predict students outcome in final course. Choice of selection algorithms are based on the popular methods used in the literature for this domain and also by considering the expertise of the target audience (instructors). End-users of this system might not be aware of the domain knowledge; therefore, white-box methods were chosen so that their results can be easily interpreted and are highly comprehensible. Table 5.8 shows the list of algorithms used in the study. Each dataset makes an instance in the training

Table 5.8: Classification algorithms used to prepare training set for meta-model.

Classifiers	**Description**
J48	Non-commercial decision tree C4.5 [83]
RandomForest	Forest of random trees [87]
PART	A version of C4.5 using decision list [145]
DecisionTable	Simple decision table [146]
JRip	Propositional rule learner [76]
OneR	Uses minimum-error attribute [147]
ZeroR	0-R classifier
IBk	K-NN classifier [148]
KStar	Entropy-based [149]
LWL	Locally Weighed Learning [150, 151]
NaiveBayes	Naive Bayes classifier [93]
AdaBoost M1	Boosting a nominal class classifier [152]
Bagging	Reduces variance [153]
Stacking	Combines the output from others [154]
LogitBoost	Additive logistic regression classifier [155]
RandomCommittee	Ensemble of randomizable base classifiers
Vote	Uses majority vote to label new instance [156, 157]
Logistic	Logistic regression with a ridge estimator [158]
MultilayerPerceptron	Neural network with back propagation
SimpleLogistic	Linear logistic regression [159, 160]

set. Number of features in each dataset is different. Therefore, it is necessary to make fixed number of features from each dataset. We have used Principal Components Analysis (PCA) to generate fixed number of features. PCA is a powerful statistical technique, that has been

Table 5.9: Classification algorithms' performance for an example dataset with n=161.

Algorithm	Accuracy	RMSE	Fscore	Kappa	PRC	AUC
J48	65.26	0.48	0.53	0	0.61	0.5
RandomForest	52.83	0.55	0.49	-0.06	0.62	0.45
PART	65.26	0.48	0.53	0	0.61	0.5
DecisionTable	65.26	0.48	0.53	0	0.61	0.5
JRip	59.63	0.49	0.5	-0.04	0.6	0.47
OneR	55.26	0.66	0.54	-0.01	0.61	0.49
ZeroR	65.26	0.48	0.53	0	0.61	0.5
IBk	56.58	0.64	0.57	0.07	0.63	0.53
KStar	55.37	0.57	0.55	-0.02	0.67	0.54
LWL	57.76	0.53	0.5	-0.02	0.62	0.4
NaiveBayes	60.37	0.52	0.55	-0.04	0.66	0.49
AdaBoostM1	60.37	0.5	0.55	-0.04	0.64	0.5
Bagging	53.42	0.55	0.51	-0.03	0.63	0.46
Stacking	65.26	0.48	0.53	0	0.61	0.5
LogitBoost	49.63	0.54	0.47	-0.14	0.62	0.41
RandomCommittee	52.72	0.63	0.51	-0.06	0.59	0.37
Vote	65.26	0.48	0.53	0	0.61	0.5
Logistic	61.51	0.49	0.52	-0.04	0.66	0.45
SimpleLogistic	61.51	0.49	0.51	-0.04	0.59	0.44

used to find patterns in high dimensional data and express the data in a way that highlights similarities and differences. Once patterns are found in data, then it can be used to compress the data without loss of much information. Using PCA we will get same fixed number of features for each dataset regardless of the original dataset structure. Each algorithm was run for each dataset (n=465) stored in dataset repository. All the classifiers were trained using 10-fold cross validation in order to counter the possibility of over fitting or biased results due to the small sample sizes. Each dataset was split randomly into ten folds which were mutually exclusive and had approximately same number of classes. Each algorithm was run for ten times and evaluation measures, namely, accuracy, F-measure and AUC-ROC curve criterion were used to compare algorithm performance.

Table 5.10: Features used for training set.

Feature	Description
Algorithm	Classification algorithm name.
ClassInst	Ratio of no of classes to instances
AttrClass	Ration of number of features to number of classes.
Median(feature)	Middle value in the dataset
Mean (feature)	Average value in the dataset
Std (feature)	Standard deviation
Skewness (feature)	Measure of asymmetry of the probability distribution.
Entro-Class	Class entropy of target attribute.
TotalCoor	Amount of information shared among variables.
Performance	Measure of performance fore each algorithm.
Quiz	Number of activities (mean) performed using quiz module.
Assignment	Number of activities (mean) performed using assignment module.
Forum	Number of activities (mean) performed using forum module.
Total unique LMS events	Total unique events performed in course by students.

Meta features from LMS log data

Next step was to extract meta features from dataset. List of meta features extracted from each dataset are given in Table 5.10. Apart from the statistical features for each dataset, some variables specific to the LMS log activities used in that particular course were used. As already mentioned. that courses were of different types, that is, were either distance based or taught internally. So, there was much diversity in the LMS usage by students or instructors. Because majority of the independent variables are from LMS log, there was further need to add variables within each dataset which could help differentiate the diversity of the courses in terms of LMS activities usage. We used same variables which were engineered in previous chapter for cluster analysis. Details of the features are given in Table 5.10.

Once the meta-features were extracted for each dataset then this information was integrated with the performance of each algorithms using only one evaluation matrix. In our case, F-measure was chosen as the final metric based on which algorithms were to be compared. Final training set for meta model is the combination of name of classification algorithms,

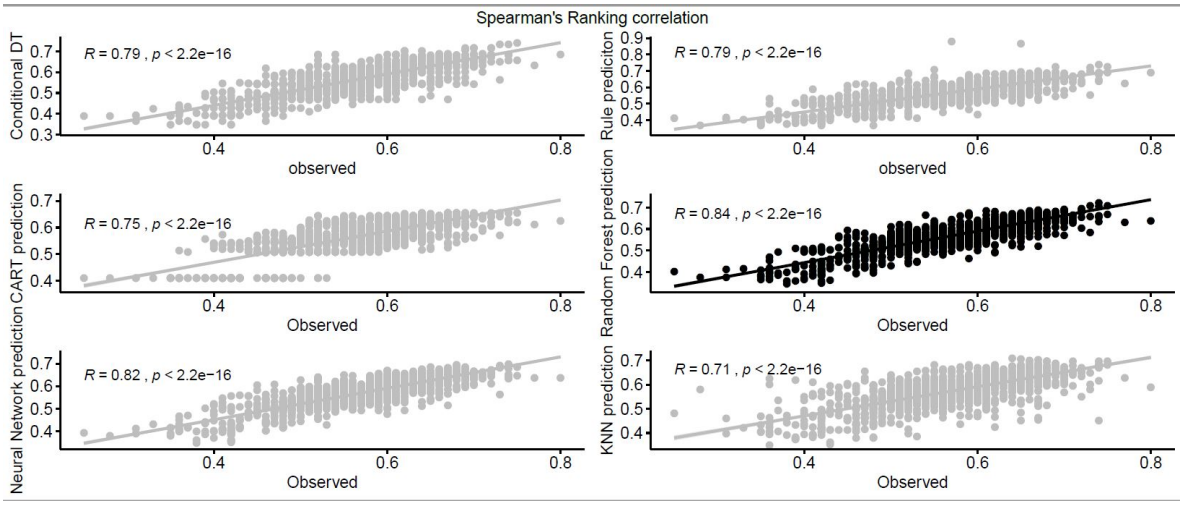


Figure 5.9: Comparison of regression models using Spear man's ranking co-relation.

statistical meta-features, meta-features related to LMS log data, and in last the dependent variable is the continuous value which is the F-measure of corresponding algorithm (for which value is 1 for rest of the algorithm names this value will be zero).

Training regression model

Next step is to train the regression model to make predictions on the performance of classification algorithms for a given unseen dataset and recommend the best classification algorithm. This study utilized following regression tree models; CART (classification and regression tree), Conditional Tree, Random Forest, Neural network, KNN and rule based. Spear-man's rank correlation test was used to assess the performance of regression models and select best regression model with high rank. Figure 5.9 shows the result of regression models and Random Forest stands in highest rank when compared to rest of the regression models. The RMSE(root mean square error) for all models are less than 1 which shows that all regression models are competitive in predicting the performance of classification algorithms (Figure 5.9).

5.6 Summary

Learning management system(LMS) is widely used in higher education institutions and is used mostly for providing course materials. However, LMS is much more capable than just providing learning materials. LMS supports teaching practices by providing an environment that can support online interactions with the learning material and enhance communication among peers and instructors. Activities performed during the course can help in identifying student behaviors that could lead to prediction of likely future grades. However, prediction depends on the dataset; the more the range of data, more will be the accuracy of model. Although LMS provides huge learning tasks; use of LMS tools are diverse, moreso since some institutes use LMS tasks extensively while some just use it for communication purpose. Extensive use of LMS tasks increases the digital traces of students which are very useful in prediction. In this chapter, the LMS data of a different types of courses have been analyzed and have shown how it helped improve the prediction accuracy when the LMS data was integrated with assignment scores.

In first section, event logs of five MOOCs courses were used to predict students that are most likely to have dropped out. Machine learning algorithms used for the classification are Random Forest, Logistic Regression, K Nearest Neighbor and Naive Bayes. Results show that techniques used in this study are able to make predictions on dropouts. However, it can be further improved by integrating more features that are directly linked to the learning process like assessments, quizzes grades etc. Nevertheless, it is useful in cases when we need to make early predictions, such as during the first or second week of the courses when assessments do not yet exist, but instead, the event logs of students' interaction with learning management system is available. This study investigated the fact that students who engage more in the

course are less likely to dropout. Results also show that prediction accuracy is better in courses where there is a significant difference between engagement levels of two groups of students (at-risk and not-at-risk). In second experiment, different types of courses were selected (main dataset) in terms of their instruction format (distance vs internal) and compared several machine learning methods that includes, Random Forest, Logistic Regression, Naıve Bayes, LDA and Ensemble method to predict the outcome of students in the course using LMS trace data and assessment scores. There are four datasets extracted from four different courses. Each dataset has been further divided into sub datasets (sixteen) to predict outcomes after every week. Total number of experiments conducted were 64. There are two kinds of attributes in each dataset, one is the count of total online activities and second is assessment scores (if available at that time of the week). The aim was to predict the outcome after every week into two classes; at-risk and not-at-risk of failing the course. Results confirm that LMS data have got discriminating power, but not more than assessment scores. Courses in which students used LMS more frequently and number of assignments are more than 3, the accuracy of predictive model for such courses were high. Our findings show that combination of LMS data and assessment scores can improve the accuracy of predictive models. In addition to this, it is also confirmed that more LMS data doesn't directly improve the accuracy, which means just count of activities is not enough; it needs to be investigated further that what kind of activities differentiate between groups of students. This study confirms the importance of LMS data with combination of assessment scores in prediction of student's academic performance. However, it is not enough to generalize the conclusion as the data used in the study is limited to one institution and there are more data that can be used for classification. In the last section, a multi-label regression model was proposed for recommending classification algorithms to solve a prediction problem in education domain. These regression models are trained using

historical data of different courses. This study utilized data of 33 courses, and more than 20 classification algorithms were used to predict students final outcome in the course. Meta-features were calculated for each dataset which not only used statistical features but domain knowledge was also included. Meta-features and performance of classification algorithms were integrated to make training set for regression model. Several regression algorithms were used to predict the performance of classification algorithm for a given dataset using the meta-features. Regression models compared significantly and the best regression model was selected using Spearman's co-relation rank, which can then be used to rank classification algorithms for a given data set. These kind of applications are useful for recommending set of algorithms with promising performance to the end-user who may lack domain knowledge.

NOTE: Chapter 5 is a partial re-print of following three articles:
The thesis author was the primary investigator of this article.

- R. Umer, Susnjak, T., Mathrani, A., Suriadi, S. (2017).Prediction of Students' Dropout in MOOC Environment, International Journal of Knowledge Engineering, Vol. 3, No. 2, December 2017 http://www.ijke.org/vol3/85KD015.pdf (ISSN: 2382-6185)

- R. Umer, T. Susnjak, A. Mathrani and S. Suriadi, "A learning analytics approach: Using online weekly student engagement data to make predictions on student performance," 2018 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, 2018, pp. 1-5, doi: 10.1109/ICECUBE.2018.8610959. ©2018 IEEE

- Umer, R., Mathrani, A., Susnjak, T., Lim, S. (2019, March). Mining Activity Log Data to Predict Student's Outcome in a Course. In Proceedings of the 2019 International Conference on Big Data and Education (pp. 52-58).

Chapter 6

Application of Process Mining to Improve Predictions

6.1 Introduction

Currently most studies use LMS logs and assessment scores as independent variables for predicting students' performance, or, they may rely on self-reported student data. LMS logs captured during the course can serve as proxy variables for analyzing students behavior; however, merely counting the frequency of activities or timestamps may not give an overall picture of their behavioral aspects. An emerging field of research, educational process mining (EPM), has led the way for advanced learning analytics, where learner produced activity data can be contextualized to understand students' learning styles/habits that influence their academic performance. Moreover, these habits can then be represented visually through process models [81]. Few researchers have used process models for analyzing student behavioral patterns. However, process models have limited use if they are not used timely. In this study, process mining features were incorporated along with standard features and used for building early prediction capabilities on student performance. The objective of this chapter is to describe how process mining features have been used to capture different behavioral aspects of different student groups and thereby enable development of a predictive model that can be used for early prediction in student performance.

6.2 Motivation

The motivation of this chapter is as follows.

1. Documentation of the process of preparation of event logs using educational data that has been gathered from different course deliveries in a university setting.

2. Description of challenges in data usability and data integration from process-unaware databases and demonstration of the construction of a process-oriented event log. Further

recommendations are proposed.

3. Application of data mining/machine learning algorithms to student datasets (i.e., LMS log and assessment scores), as students are progressing through a course to enable prediction of at-risk students (i.e., those students who are not satisfying course requirements, or are likely to fail).

4. Application of data mining/machine learning algorithms to student datasets (i.e., LMS log and assessment scores only) and integrate process mining features to study the degree to which accuracy of predictions can be improved.

6.3 Chapter Layout

This chapter is laid out in two parts. The first part focuses on preparation of event logs from educational data gathered from different courses delivered in a university setting. Data have been gathered from two platforms: (1) the student management system (SMS) comprising student grades and profile information and (2) the learning management system (Moodle) comprising course-related instructional tasks and which are used for communicating with students over the course duration. Moodle is a popular open-source learning platform and considered an exemplar of LMS by many higher education institutions [193]. The overall intent is to first generate event logs from incompatible underlying process-unaware platforms for then conducting educational process mining analysis. This chapter describes data extraction, case identification and data quality improvements in using process-unaware data from online learning platforms. More than 50 courses were analyzed for event log formations; however, in demonstration of a running example, data related to student participation in an online quiz of one course has been described for constructing a reasonable event log for EPM purposes.

This study contributes to the fields of learning analytics and process mining. Practice lessons learned in the extraction and conversion of process-unaware data to event logs for the purpose of analysing online education data are conveyed. Quality issues that generally face education data are explained with a specific running example (i.e., quiz-taking). Limited studies (e.g., [120]), have shared such practice experiences; therefore, this work contributes significantly to the process mining body of knowledge. Moreover, analytics is an ever-evolving field with expansions in data capture, data modeling, data analysis, data predictive and data-driven decision-making approaches [194].

The two main parts of the chapter are as follows:

1. **Building a Process-Oriented Event Log: Challenges in Process Mining:** In order to incorporate process mining features, the first step is to build an event log from students' activity log data from Moodle. However, after a few attempts on using the main dataset provided from Moodle, a different approach was used. Moodle is process-unaware; therefore, a workaround was required to extract contextually event logs from such process-unaware systems. This section gives a detailed account of the extraction and conversion of process-unaware data to event logs for the purpose of analyzing online education data.

2. **Prediction of Performance by Incorporating Process Mining Feature:** This part of the chapter demonstrates use of a MOOC dataset for preparing event log, extracting process mining features and conducting a study with generic features (extracted from Moodle event log and assessment scores) to make early predictions on students' performance.

Part 1: Building a Process-Oriented Event Log: Challenges in Process Mining

6.4 Process Mining

The objective of process mining is to conduct data analyses to project a process-oriented perspective and answer questions related to business process flows. Figure 6.1 shows the workflow for process mining. First, a plan is articulated for understanding the available data and the business domain which is aligned with the expected outcomes. Next, the raw data which might have originated from multiple sources (e.g., flat file, excel sheet, database table etc.) is extracted. This phase of data collection is quite challenging; more so since data are typically not structured or sometimes the metadata is missing. Data management issues such as the integration and sharing of data are common challenges faced by any analytics teams in organizations. Moreover, there is abundance of data, and all data are not meant to be used; instead, datasets are extracted based on the questions or scope of the problem. Scoping the problem requires a combination of quantitative and software skills alongside required

business domain knowledge [195]. Event log is created which is further filtered depending on the granularity level of the problem under analysis. The filtered event log is the input for process mining applications: process discovery, conformance and enhancement.

1. **Discovery:**

   By analyzing event-based data (i.e., event log), analysts can reveal process flows to show how different processes are being performed for a given task. In education domain, a discovery example would relate to detection of patterns of various activities that students perform while taking a quiz.

2. **Conformance:**

   Here the event log is analyzed to verify whether the existing business model was being followed or not. Event log is again an input for conformance checking.

3. **Enhancement:**

   Enhancement is used to analyze other process perspectives by exploiting the discovered model, such as resource analysis, performance analysis, as well as decision mining.

Event logs for process mining comprises following elements.

- **Case:** Each entry in the log maps to a single case. A case is also called process instance which reflects the single execution of a process.

- **Events:** Events are single entries in the log. Events are mapped to cases which are ordered based on time-stamp. Events relate to specific activities and associated attributes.

- **Variant:** A specific sequence of activities referred as variants is used to compare different cases. An event log comprises a set of events (or time-stamped order of performed

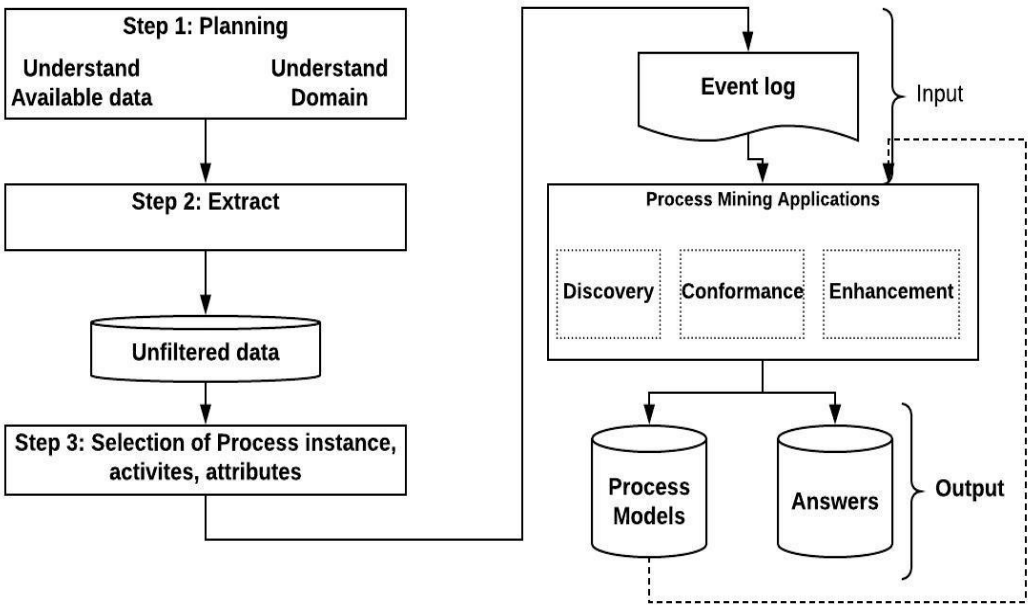


Figure 6.1: Life cycle model describing major steps for process mining project consists of planning, data extraction and selection of case, activities and attributes.

activities). Every event is mapped to a case which has specific orders of events. Each unique sequence of activities makes a variant.

- **Attributes** Certain specification or attributes can give more information at either the event or the case level.

In 2010, the IEEE Task Force on Process Mining adopted XES (Extendable Event Stream) as a standard for process mining event logs [114]. XES is XML based, in which each event log entry comprises an event type, a time stamp and associated attributes. Attributes are assigned at different levels, i.e., log level, case level, or event level. An XES log can contain a number of cases (traces in XES language), where each case describes a sequence of events pertaining to this case. This requires that every event is already assigned to a case and all events related to a case are known.

6.5 Event extraction

Next step was to understand the underlying Moodle database from where relevant data can be collected to meet the primary goal, that is, to get a process-oriented perspective from the static datasets. The required datasets are often scattered in multiple tables. Therefore, this step requires an understanding of the underlying database for making proper selection of tables. A process perspective requires time-stamped information of the activities undertaken to link the actions pertaining to a particular *'case'* sequentially. Extracting this information may include merging additional tables by using appropriately defined unique identifiers. Therefore, a schematic wide view of all tables and their relationships with each other must be known.

1. **Understanding available data**

   Moodle has a built-in logging system that tracks and stores navigational activity performed by the user [196]. In this manner all student activities are recorded in a standard logging system. These logs are stored in relational database (MySQL or PostgreSQL). In total there are more than 346 tables in Moodle database. However, data extraction may not involve all of them, since all modules (quiz, assignment, lesson, surveys etc.) are not necessarily used in a course delivery. The table 'logstore-standard-log' keeps tracks of all activities performed in Moodle. We can filter this table to extract more specific logs, such as by course, by participant, by day or by session (or any combination of these). These logs can give detailed views of activities performed by students during the course duration. For example, in a quiz activity, we can determine how much time students took to complete the quiz, their scores and the number of attempts among other similar bits of information. Figure 6.2 shows the high-level activities performed

during the quiz, which starts with 'Start Quiz' step and ends in 'Submit Quiz'. A quiz can be viewed or reviewed multiple times, so there could be a self-loop in these steps. Basic activities that are related to the quiz-taking process are stored in log table

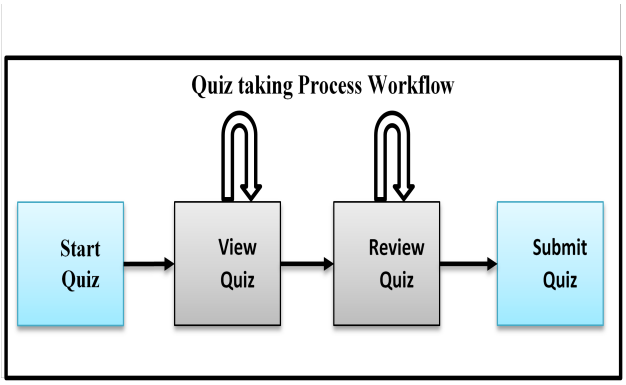


Figure 6.2: Basic activities in process of quiz-taking

with time stamps. Log table provides only high-level picture of the process. In order to get more details of activities undertaken, additional tables that are associated with quiz-module need to be merged. The quiz module enables teachers to design quizzes as a part of students learning activity. A quiz consists of multiple questions of different types (e.g., calculated, multi-choice, description, essay etc.). Table 6.1 describes the tables related to quiz module.

Figure 6.3 represents a part of the database and displays six tables relevant to our running example. Table quiz contains information about quiz such as course name, total marks, number of attempts allowed, etc. Table quiz-grades stores overall grades for each student on the quiz, based on their various attempts. Records in quiz-grades table refer to the quiz in the quiz table. One quiz can be attempted multiple times; so, the quiz-attempts table stores information about each quiz attempts. Quiz's attempts are made of multiple question attempts. The question-usages is a bridging table that connects each quiz-attempt to a set of question attempts. Question-attempts table

Table 6.1: Details of all related tables in quiz-module.

Table Name	**Description**
mdl_quiz	Has quiz information like name, grading methods, number of attempts allowed, quiz time open, quiz time close, maximum grade, etc.
mdl_course-modules	Stores information about courses and its modules.
mdl_quiz-attempts	Stores information about quiz attempt, attempt start time, attempt finish time, grade in attempt, attempt state (e.g. in progress, complete, to do etc.)
mdl_quiz-grades	Stores final grades of students in quizzes.
mdl_question-usages	A unique id is assigned to each attempt made on a set of questions. A question usage is made up of a number of question_attempts.
mdl_question-attempts	Each row here corresponds to an attempt at one question, as part of a question-usage. A question-attempt will have some question-attempt-steps.
mdl_question	Stores questions that are in the quiz
mdl_question-attempt-steps	Stores one step in in a question attempt
mdl_logstore-standard-log	Stores information about activates performed in Moodle with their timestamp

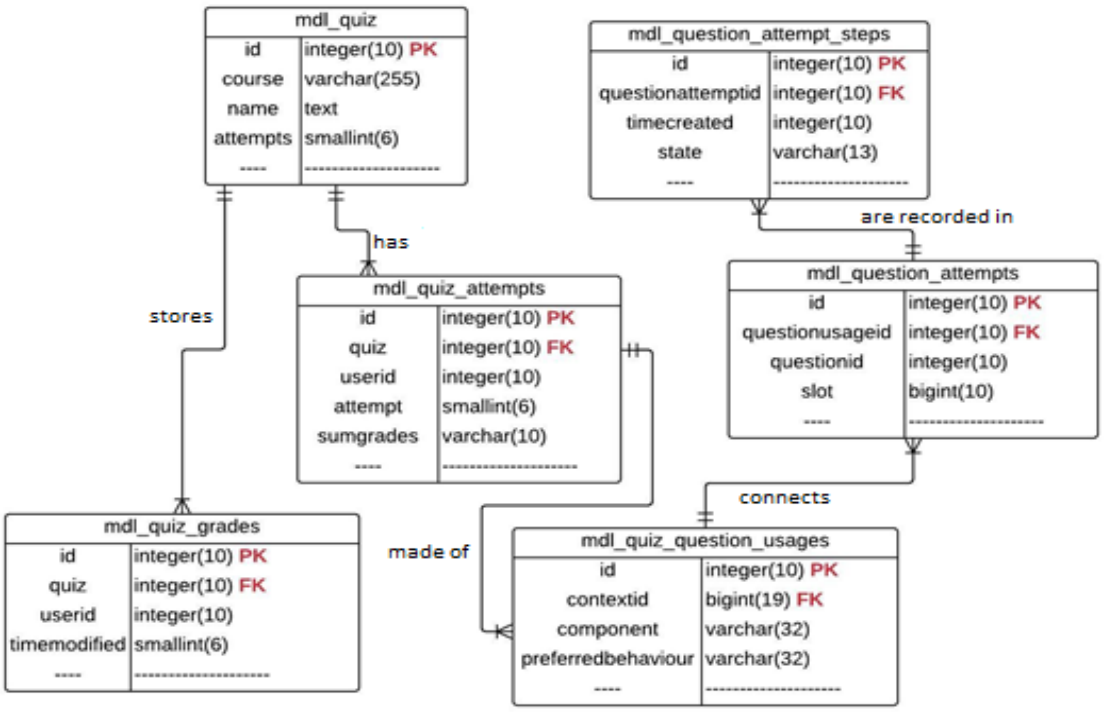


Figure 6.3: Entity relationship diagram of Moodle tables that are related to quiz-module.

records information about an attempt at one question, which is further connected to question-attempt-steps table. Question and its state with the timestamp are stored in question-attempt-steps table.

2. **Selection of process instance**

   This step is for selection of process instance and to make decisions about the granularity level. The selection of the process instance and granularity level have major impact on the quality of event log. Decision is to be made by considering the parent-child relationship between tables.

   Main activities involved in quiz-taking process are start a quiz, view/review a quiz, view a question, view feedback(optional) and submit a quiz. There are five tables that records time stamps of events related to the quiz-taking process: quiz, quiz-grades,

quiz-attempts, question-attempts and question-attempt-steps. However, for building an event log each event should relate to a case. Therefore, we need to integrate five tables in one table with column 'Case id' and all events should be referred to at least one case. In current running example there are multiple options for selecting a process

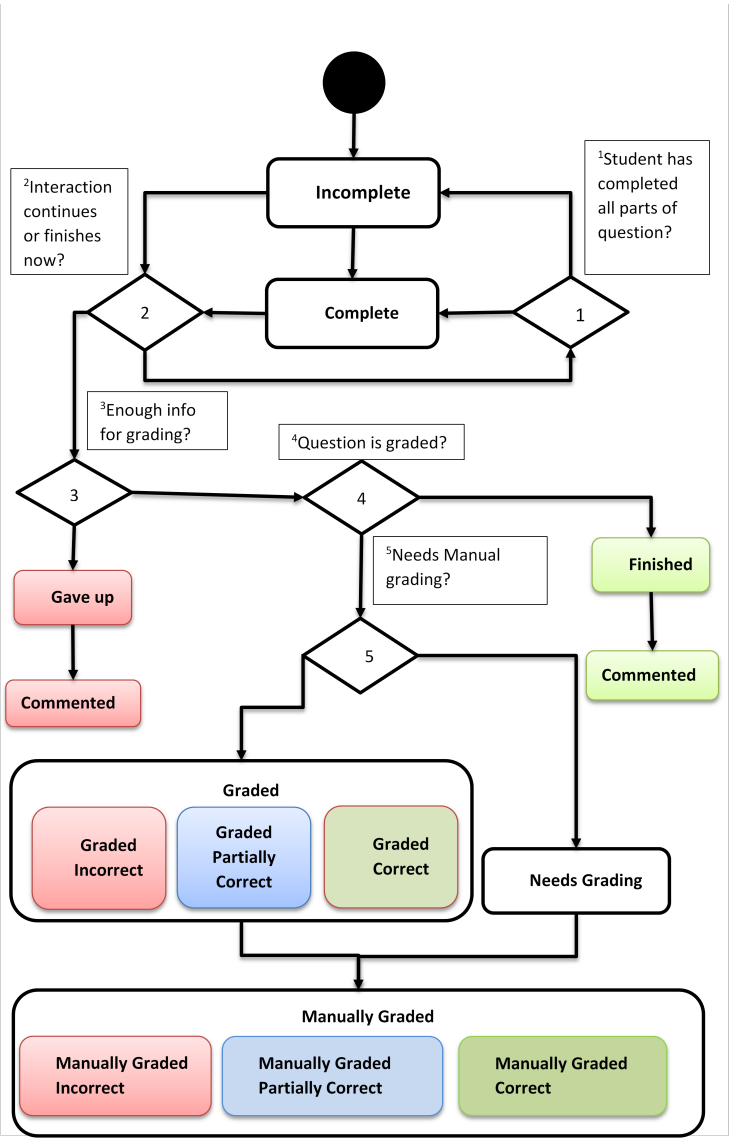


Figure 6.4: Moodle Question Engine overview.

instance (Case). For example, *Case* could be per student that allows us to obtain different patterns student followed while taking a quiz, or *Case* could be per question

to obtain a process occurred during exercises. Alternatively, it could be at quiz level, that is, at a session where quiz occurred during the course where students worked on same quiz and performed different activities.

Since the scope is limited to the quiz level, the activities performed during the quiz session have been analyzed next. Case per record in the quiz table were chosen, and all events that are associated to a Quiz (*Case*) were listed. Table 6.2 shows all the events that could be related to one *Case*.

Table 6.2: List of events associated with quiz module

Case: Quiz		
Table Name	**Activity**	**Attributes**
mdl_quiz	start-quiz view-quiz review-quiz submit-quiz	quiz-grades student-id quiz-status
mdl_quiz-attempts	start-quiz-attempt finish-quiz-attempt	quiz-attempt-state quiz-attempt-grade student-id
mdl_quiz-attempt-steps	state-to-do state-finish state-invalid state-complete state-unprocessed state-gaveup state-needsgrading state-grade-write state-grade-wrong state-grade-partial	question-grade student-id

3. **Selection of activities and related attributes**

Table quiz has four timestamps per record. If only the quiz table is considered, then there will only four events per *Case* and all information about the quiz attempts and questions states will be unused. It will result in a sequential process consisting of

following steps; *start quiz*, *view quiz*, *review quiz* and *submit quiz* (Figure: 6.2).

To develop a process model that consists of more activities other tables should be considered as well. By using domain knowledge such as cardinality, references are made to link other tables. If quiz, quiz-attempts and question-attempt-steps tables are considered then all the events or subset of events with their time stamps can be considered along their relevant activities. Table 6.2 shows list of activities and attributes that are associated with selected tables.

After selection of process instance and granularity level, all relevant activities have to be selected. Activities have to be grouped based their time stamps. In addition, those attributes that give more information about the process instance or about the activity are identified and selected. For example, activity performed by whom and where (location) or a role that is responsible for the activity.

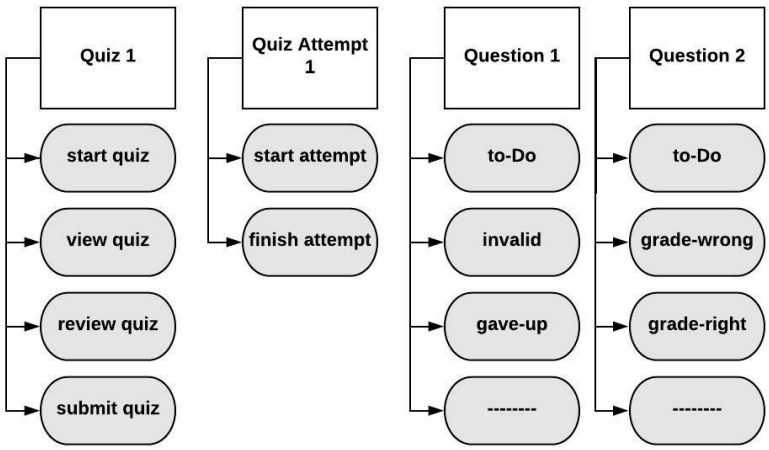


Figure 6.5: Activities associated with quiz-module tables.

6.6 Event log challenges

Educational technologies have facilitated large volumes of data from heterogeneous sources to be stored in different formats and at different granularity levels [197, 198]. Collection and integration of such voluminous data scattered over the learning platform, is not a trivial task. Moreover, precision is to be considered in filtering of event activities for process models to discover specific behaviors. Real life data extracted from business domains (e.g., LMS) comprise multiple and parallel activities that may or may not be very frequent. Learner behaviors have much variability, therefore, filtering steps to form event logs are much harder compared to other business processes (e.g., payment at checkout kiosks).

Following sub-sections highlights challenges faced in data extraction for creating a process-oriented event log.

Complex and voluminous data

Moodle database comprises more than 346 tables. Depending upon the *Case* granularity level, multiple tables have to be queried to get information about one activity. To select cases and accordingly write queries that join a number of tables requires in-depth domain knowledge. Fortunately, Moodle is open source and has plenty of online resources. The Moodle developer community is very responsive to forum questions and they regularly contribute to the online library to help users understand the complexities of the database.

- **Moodle documentation -database schema:** The technical documentation from the online Moodle resources provided an in-depth view of the underlying database schema (i.e., purpose of the table, column names and description, keys, etc.). With more than 346 tables overall, of which many are not relevant to the problem in hand, the documentation helped identify sub-schema applicable to the context being investigated.

- **Moodle online community:** An online community forum provides a platform for people having shared interest on Moodle know-how to interact with each other. The community members are quick to respond to problems posted on the forum. One forum group ADH (or the 'ad-hoc contributed my SQL queriers' group) in particular supports users to extract data More than 1000 SQL queries are posted on ADH to enable users extract information from a MySQL database. This forum group provides many resources to assist in data extractions from database.

Dealing with many-to-many relationships

Quiz-taking is a relatively small process considering the bigger picture of other activities undertaken by students during a course. It was noted that the quiz-taking process involved over 5 tables with table relationships not limited to simple one-to-one relationships. Therefore, data extractions have to be handled logically to avoid ambiguous case definitions.

Case-id	Timestampe	Activity	Grade
QuizId	2017-02-27 13:05:09	quiz_viewed	3.75
	2017-02-27 13:05:09	quiz_started	3.75
	2017-02-27 13:05:40	quiz_viewed	3.75
	2017-02-27 13:07:43	quiz_viewed	3.75
	2017-02-27 13:14:16	quiz_submitted	3.75
	2017-02-27 13:17:37	quiz_viewed	5
	2017-02-27 13:17:37	quiz_started	5
	2017-02-27 13:19:25	quiz_viewed	5
	2017-02-27 13:27:21	quiz_viewed	5
	2017-02-27 13:33:26	quiz_submitted	5

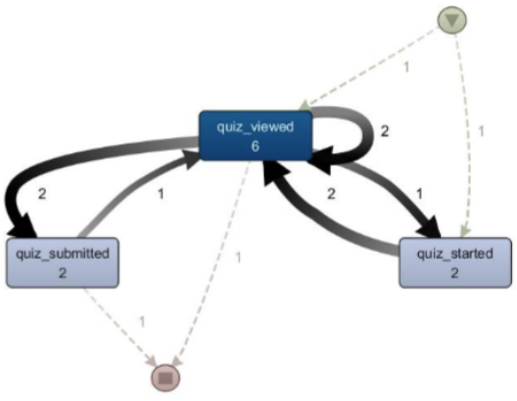


Figure 6.6: A record from event log of a student who takes a quiz along with process mining view of activities

The running example deals with activities performed by a student during the quiz-taking process. The first challenge is selection of the *Case id*, since the *Case id* must be related to end-to-end process activities. For example, quiz-taking process broadly involves three steps: *start quiz*, *view quiz* and *submit quiz*. Figure 6.6 is a simple representation of event log activities performed by a student during a quiz-taking process for a particular quiz. Activities like *quiz-started*, *quiz-submitted* and *quiz-viewed* are repeated many times for the selected quiz (which is evident from the event time stamps). Within this context, by selecting *Quiz-id* as case, all activities that belong to the quiz are transformed to a process map (Figure 6.6) after applying process mining on the event log.

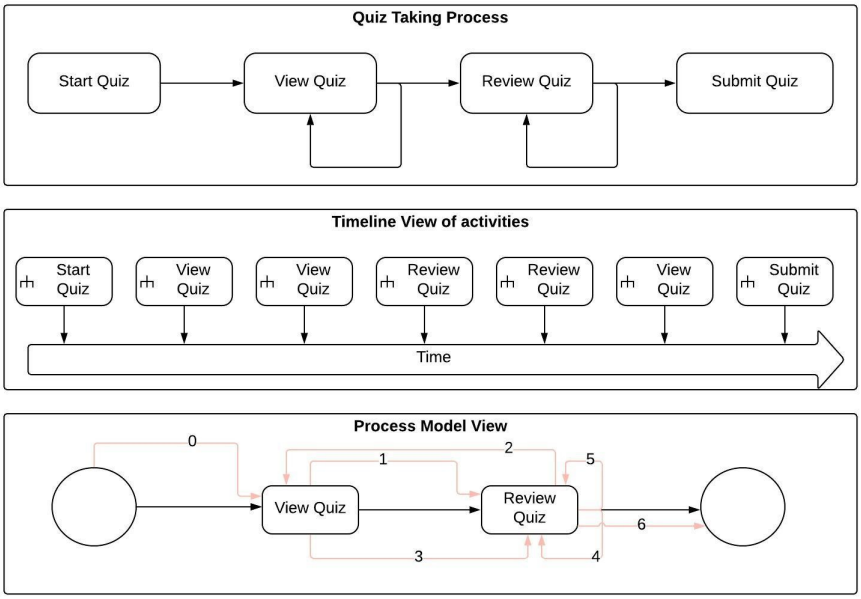


Figure 6.7: Divergences of cases visualized reality vs. process mining view of activities

The process map in Figure 6.6 shows that a student started and submitted the quiz twice; but this activity has been performed on multiple quiz-attempts. This situation is known as divergence of data [199]; it occurs as a result of many-to-many relationships (i.e., each student can attempt a quiz many times and each quiz are attempted by many students). Figure

6.7 shows three different perspectives of quiz-taking process, (1) quiz-taking as a process, (2) timeline view of the activities performed during the quiz-taking process and (3) process mining view of the quiz-taking process. Therefore, process map shown in Figure 6.6 does not fully demonstrate how the process occurred. So, the real challenge is the perspective taken in the investigation.

Divergence can be resolved by adding more granularities to activities. Like in given example, if we just consider the quiz-taking event then we can face divergence issue, but if we add quiz-attempt or add question level granularity then this problem can be resolved. By changing the perspective, we get a process model (shown in Figure 6.8) that reflects more details of the quiz-taking process. It is clear from the process map that student viewed and submitted quiz in two different attempts and this was evident by adding to the granularity level in the event log.

Case ID	Event ID	Timestamp	Activity	Grade
1	1	2017-02-27 13:05:09	quiz_viewed	3.75
1	1	2017-02-27 13:05:09	quiz_started	3.75
1	1	2017-02-27 13:05:40	quiz_viewed	3.75
1	1	2017-02-27 13:07:43	quiz_viewed	3.75
1	1	2017-02-27 13:14:16	quiz_submitted	3.75
1	2	2017-02-27 13:17:37	quiz_viewed	5
1	2	2017-02-27 13:17:37	quiz_started	5
1	2	2017-02-27 13:19:25	quiz_viewed	5
1	2	2017-02-27 13:27:21	quiz_viewed	5
1	2	2017-02-27 13:33:26	quiz_submitted	5

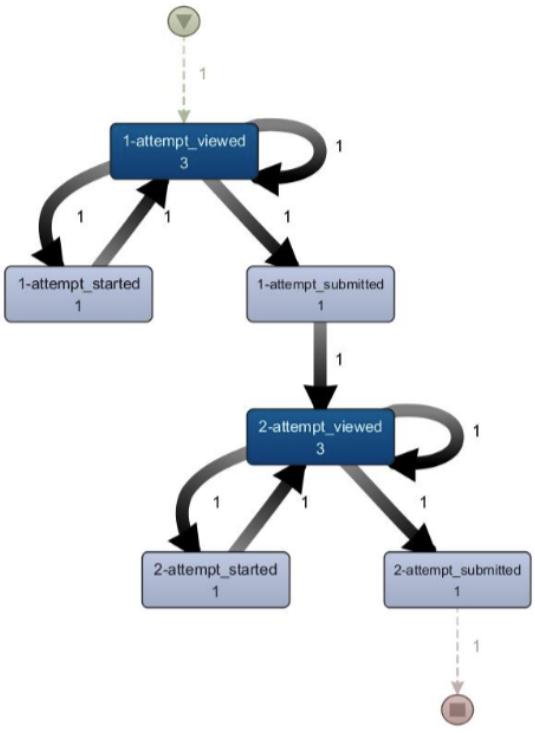


Figure 6.8: Divergences of cases visualized reality vs. process mining view of activities

Missing time stamps

In the event log, events are ordered per case. For each case we merge time stamped data from multiple tables. However, we faced missing time stamp issues in the event log. A small fragment of the quiz table (Figure 6.9) shows each quiz with two time stamp records, namely 'timeopen' (i.e., the time quizzes were available to students for attempting the quiz) and 'timeclose' (i.e., the time after which the quiz will not be available). As is evident in Figure 6.9, some quizzes have specific set time limits while others have no time limit; therefore, we have set the 'timeclose' to null value to represent no time limit. Following sub-sections detail situations having missing time stamps for events.

- **Missing quiz-open time and quiz-close time** Missing quiz-open and quiz-close times situations occur when the quiz has no predefined time bound, or the quiz is available and visible from the beginning of the course. Some instructors often setup non-compulsory practice quizzes to help students learn some subject concept; and since quizzes are not compulsory, it is up to the students to attempt them or not. Most of the time practice quizzes do not contribute towards the final mark. So, if we are interested

	quizid	name	timeopen	timeclose	grade	timelimit	overduehandling	grademethod
0	66234	Test Yourself - Lab 3	1970-01-01 12:00:00+12:00	1970-01-01 12:00:00+12:00	10.00000	0	autoabandon	1
1	66229	Mastery Test 2	2017-03-06 00:30:00+13:00	2017-03-20 19:00:00+13:00	10.00000	2	autosubmit	1
2	66228	Mastery Test 1	2017-02-27 09:00:00+13:00	2017-03-12 23:00:00+13:00	10.00000	2	autosubmit	1
3	66232	Test Yourself - Lab 1	1970-01-01 12:00:00+12:00	1970-01-01 12:00:00+12:00	10.00000	0	autoabandon	1
4	66233	Test Yourself - Lab 2	1970-01-01 12:00:00+12:00	1970-01-01 12:00:00+12:00	10.00000	0	autoabandon	1
5	66230	Mastery Test 3	2017-03-13 00:30:00+13:00	2017-03-27 03:00:00+13:00	10.00000	2	autosubmit	1
6	69389	Test Yourself--Lab 4	1970-01-01 12:00:00+12:00	1970-01-01 12:00:00+12:00	7.00000	0	autoabandon	1
7	66231	Mastery Test 4	2017-03-20 00:30:00+13:00	2017-04-03 03:00:00+12:00	10.00000	2	autosubmit	1

Figure 6.9: Some records of the quiz table.

in seeing student groups who attempted practice quizzes, we cannot get a complete view on student activities with missing quiz *timeopen* and *timeclose* on the generated

process map, we could not get a complete view on students' activities. Having common time allocations for quiz opening and its subsequent closing can help generate process maps which can be used to analyze student behaviors during the quiz sessions. But in situations where students could do practice quizzes at their convenience, we cannot analyze all students' behavior from one common session.

- **Missing quiz-close time only** This situation occurs when there are no bounds are set to quiz attempts. The quiz can be attempted at any time during the course or as long as it is visible to students.

- **Missing attempt-finish time** A quiz can be in following four different states (Figure 6.10). These are:

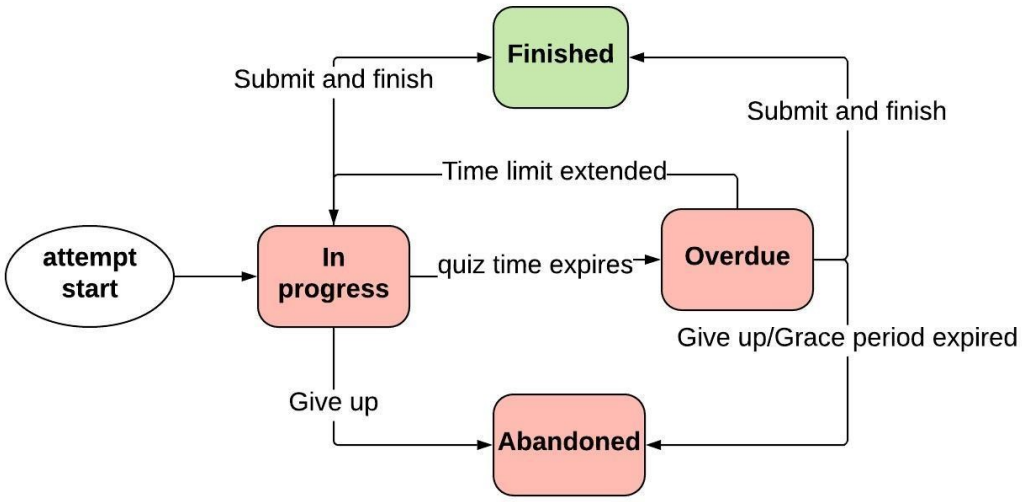


Figure 6.10: Different states of the quiz. *Attempt-finish-time* will be missing if quiz is in 'in-progress' state.

1. *In-progress:* Quiz has started but not yet finished. In this state, the 'attempt-finish-time' is null.

2. *Finished:* The quiz attempt is submitted. The *attempt-finish-time* is the timestamp

when the student submits.

3. *Abandoned:*if quiz is not submitted on time then attempt is considered as abandoned. The state is again directed to *in-progress* and *attempt-finish-time* is null.
4. *Overdue:*In some cases, students are given grace-period time to submit the quiz after the set time. If quiz is submitted within the given grace period time, it changes to 'finished' state otherwise it remains in the 'abandoned' state.

In scenarios where the finish time is missing and there is no bound on the quiz closing time, the students might attempt the quiz at a much later time. Therefore, event logs which capture these late attempts would refer to long-running processes. One possible solution is to remove those cases which are incomplete, or the finish time is missing especially when the average duration of that process is short. However, this is not an optimal solution since it will result in loss of useful information like behavior of students who did not complete quiz. To get proper picture of the process we need to involve the state of the quiz as well. For example, where finish time is missing, and state is *abandoned* or *overdue*, we can add another activity called *abandoned-quiz* for such records. Therefore, the end of quiz will be related to either of the two activities: *submit-quiz* or *abandoned-quiz*. Both end states are illustrated in Figure 6.11.

Granularity

Process with large number of activities results in a fine-granular event log. Like other high-tech systems, Moodle events are generated automatically; therefore, fine-granular events are hard to handle as it results in 'spaghetti' like process models which are difficult to interpret. It is beyond the human cognitive system to understand process models generated from fine granular event log. It is quite challenging to decide which level of granularity one should go.

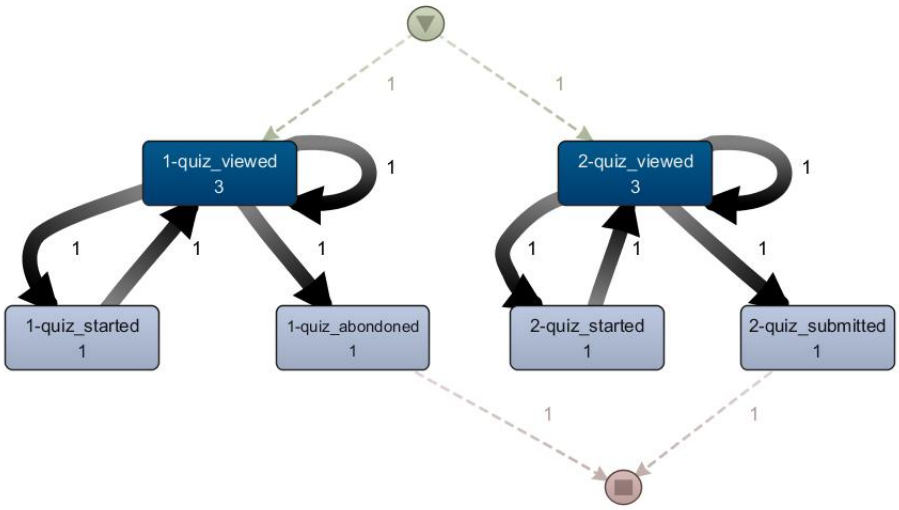


Figure 6.11: Different states of the quiz. *Attempt-finish-time* will be missing if quiz is in *in-progress* state.

If we go for higher level abstraction, we might get simple process models, but it might overlap with low level activities. One has to decide to fill the gap between high level abstraction to the low-level events that are relevant to the end user who is interested to understand or improve the processes.

From current running example it is evident that selection of the relevant events and activities was not easy. Each activity can be broken down to many sub-activities corresponding to different states. Therefore, if all activities in *question-attempt* level are selected, then a complex process model would be generated, and which would be hard to analyze. One 'question-attempt' could go in sixteen different states as previously shown in Figure 6.4. Every state could be considered as an activity as the time stamps are also recorded. If quiz is selected as process instance, then granularity level till question attempt will result in a 'spaghetti' process model. Therefore, we bound granularity level to 'quiz-attempt' to avoid

getting complex process models. An alternate solution was to aggregate the related events for example, *gradedincorrect*, *gradecorrect*, *gradepartialcorrect*, to one event that is *grade* if we are not interested in the outcome of grade but more interested in the timestamp of *grade*.

Collateral events

Collateral events are multiple events that are essentially referring to one particular step in a process within a case [124]. Moodle logs very low level of activities. For example, when student submits a quiz online there are states that capture the state of the quiz or question, also there are other states triggered based on the outcome of the grade (e.g., a 'complete' question state can further trigger events like *gradecorrect*, *gradeincorrect* or *gradepatiallycorrect* with duplication of time stamps or with difference of very short time period). Few examples are shown in Figure 6.12. These events are independent of each other with different labels but are

	id	studentid	questionattemptid	state	grade	actiontime
0	95517751	73604	33561110	todo	None	2017-03-14 16:55:14+13:00
1	95518459	73604	33561110	invalid	None	2017-03-14 17:00:00+13:00
2	95519186	73604	33561110	complete	1.0000000	2017-03-14 17:04:42+13:00
3	95524330	73604	33561110	gradedright	1.0000000	2017-03-14 17:46:02+13:00
4	95913934	73604	33698674	todo	None	2017-03-18 15:33:28+13:00

Figure 6.12: Some records of the quiz-grade table showing collateral events with difference of short time period.

repeated with difference of short time period to show that one important step of a process. These kinds of pro12 events can make the process model unnecessarily complex and does not give useful insights about process. We followed recommendation by Suriadi et al. [124] to merge such activities into a single activity and consider timestamp as either earliest or the latest.

Partial or incomplete traces

This refers to the situation where one or more events are missing in a trace. To refer to an example, there are many ways to submit assignments. Some students prefer to submit online, some submit hard copy, or some others could email their assignment to the instructor. For those who submit hard copy or email separately, the *submit-time* remains null in table assignment. Such missing events results in a process model represent partial reality, since submit assignment event has occurred in reality but has not been logged in the database. There are methods that can filter incomplete traces, but it will result in loss of information.

6.7 Recommendations

Analyzing students data and detecting interaction patterns from learning management systems have gained much attention among the learning analytics and process mining research communities. Learning analytics emphasizes on use of education data to co-relate students' online behavior with their academic performance so as to provide timely support to students; however, there is no focus on the process of learning as a whole. Research shows that process mining provides robust methods to leverage temporal data and inform us on dynamic behavioral patterns. Like any other data-driven approach, process mining takes event log as an input to produce process-related information and provide visual representation of the process for further analysis. The quality of process mining analysis depends on the accuracy and completeness of the event logs. The input data is often scattered and stored in enterprise systems that are static and not process-oriented. Therefore, identifying process related data, extraction and conversion of static data to the required format to build an event log is a complex task. This study has described multiple challenges faced in constructing event logs from process unaware LMS. Following are the lessons learned.

- Moodle database is complex and overwhelming; therefore, it is not possible to understand the entire database schema at once. It is best to focus on smaller modules or subsystems, since the corresponding sub-schema represents a smaller database and has more contextual relevancy

- Investigate different pathways, to identify all possible activities that are related to an event. This will help to find all necessary data that can be part of an event log, which can then be extracted. However, in doing any extractions, we must particularly take care of existing parent-child table relationships.

- Judiciously scope the problem in line with the motivating research questions. Select all the events and activities that are of interest to the end user who requires them rendered in a process model.

- Make a distinction on defining the process instance at parent or child level and consider the implications of this selection. If the selected process instance is at parent level, there might be low-level event at child level which needs to be aggregated or ignored (if not relevant) to avoid complex process model that is hard to comprehend.

- When selecting events or activities, focus should also be given to the other attributes that might be helpful to give more insights. For example, in this study, the 'state' attribute helped to make a new event.

- Avoid repetition of collateral events that are referring to one important event that matters in process model.

- The context of the data is necessary to be considered when interpreting the results of process mining.

Part 2: Prediction of Performance by Incorporating Process Mining Features.

This section compares effectiveness of existing popular machine learning algorithms for early identification of at-risk students (those who are likely to fail) and shows the effect of process mining features in the performance of these techniques. The significance of this study is the integration of process mining to extend features. Extended features are obtained as a result of process-conformance testing. EDM techniques are then evaluated on two kinds of datasets: one with process mining features and other without. Some widely known [200] classifiers, namely: Random Forest [87], Logistic Regression, Naive Bayes [201] and K-Nearest Neighbor [202] were used.

6.8 Research Design

Figure 6.13 shows series of steps that starts from preparation of datasets from raw data. Two types of datasets obtained from MOOC course details of data preparation are discussed in following section.

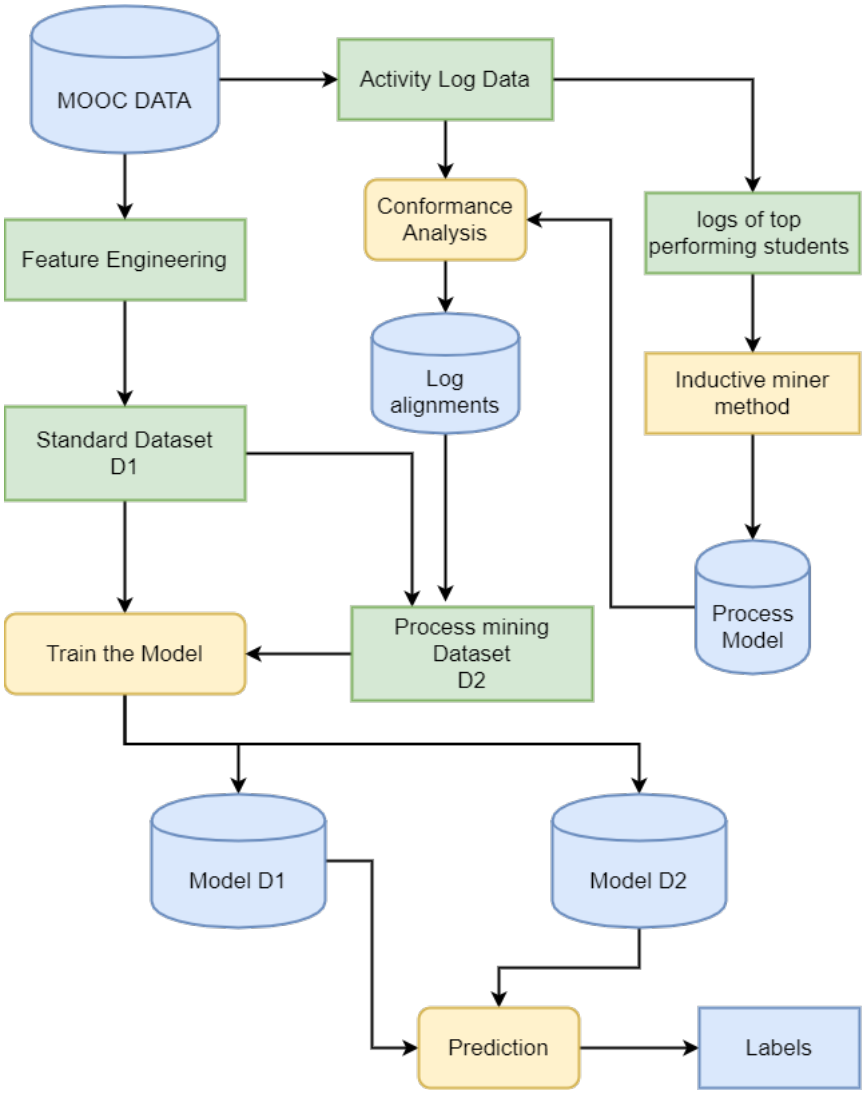


Figure 6.13: Research Design

6.8.1 Data preparation

In this study, the data were obtained from Coursera for MOOC course *Principles of Economics* offered in Summer 2014 (See Section 3.3). The dataset consisted of assessments grades, solution submission time, video lecture interaction log, participant's demographic information, time spent weekly and final grades. The course was designed as an eight-week introduction to the study of economics. Total number of students were more than 3000; however, data of students who were registered at the time the course started (i.e., on June

Table 6.3: Definition of featurs extracted from MOOC activity log.

S.No.	Feature	Explanation
1	Age	Age in years
2	Education	Highest qualification
3	Gender	Female/Male/Null
4	Average score in weekly quiz	Average score in quizzes of particular week
5	Number of quizzes attempted	Average attempt for quiz in particular week
6	Quiz lag	Duration between first and last activity of quiz
7	Lecture lag	Duration between first and last activity of Lecture
8	Total lecture attended	Total lectures attended in particular week
9	Video activity count	Activity counts during video lecture (pause, play, stop, etc.)
10	Efforts in seconds	Total time spent in a particular week

24, 2014) and whose final score was not missing are included. Total data of 167 students were extracted, out of which 40 students passed the course while rest had failed. Students with scores greater than 0.5 were considered passed. The final dataset obtained was thus imbalanced in regard to the final grade distribution (shown in Figure: 6.14).

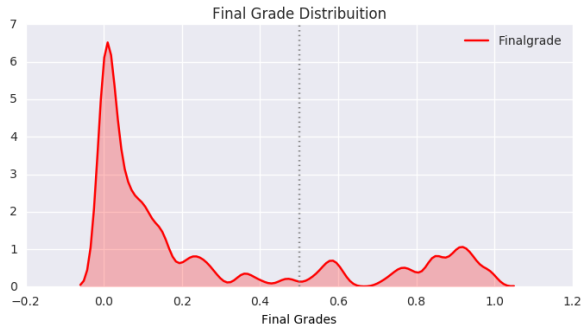


Figure 6.14: Final grade distribution

Raw data obtained from MOOC courses was divided into two datasets. First dataset (Dataset-D1) with standard features which included attributes like demographics, assessment grades, time spent on activities, video watch activities etc. This dataset will be referred as *D1* for future reference. Details of the features used in *D1* are given in Table 6.3 . Second dataset (Dataset-D2) has been generated next using logs of weekly activities during the course. It included features that reflects the differences in behavior of students with respect to the

behavior of top performing students in the course. These measures have been obtained as a result of process conformance testing [203].

In process conformance testing, given a normative model M and an event log L, difference between the process behavior and L can be explained. Conformance checking was performed using model representing top student's weekly activities and log of other student's weekly activities. The log was replayed using the model to establish a precise relationship between event and model elements and to analyze the deviation of student's from modeled behavior. Output of conformance testing is a fitness score that is assigned to each student(case). The fitness score, obtained based on weekly logs of top performing students and other students, was used as features and integrated with *standard features* and thus makes second dataset which is referred as Dataset-D2 for future references.

Following steps have performed to prepare dataset (Dataset-D2) with process mining features.

- **Step 1:** Using inductive miner method in ProM [204], process model was generated using activity logs of top performing students having grade more than 90%. The result of this step is a process model shown in Figure 6.15.

- **Step 2:** By using "Replay a log on Petri net for performance/conformance analysis" method in Prom, log model alignment was generated shown in figure 6.16. Inputs to this method were, process model of top performing students and log of activities of other students.

- **Step 3:** The log model alignment generated and exported it in CSV(comma separated value) format. Fitness scores were extracted for each student and integrated them with the *standard features* in dataset-1 (D1). This helped characterize the fitness scores as

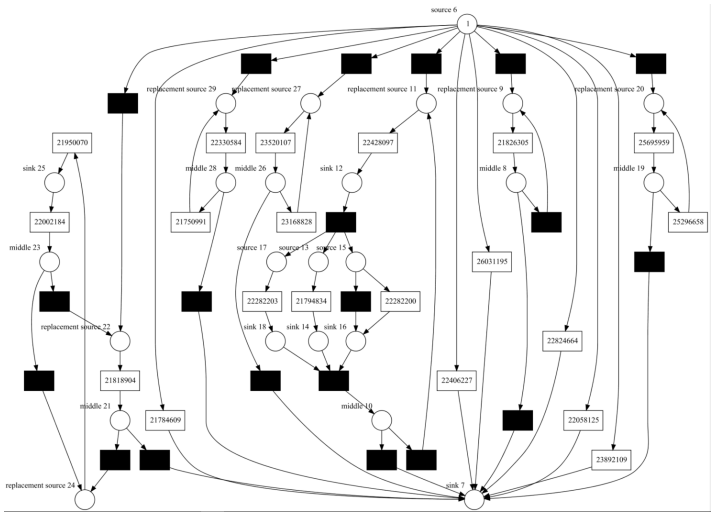


Figure 6.15: Process model of top performing students using Inductive miner method based on log of Week-1 to Week-2

Figure 6.16: Result of replaying log history on process model for conformance analysis

process mining features in dataset-2 (D2).

TRACE:concept:name	TRACEAL:IsReliable	TRACEAL:RawCost	TRACEAL:MoveLogFitness	TRACEAL:MoveModelFitnes	TRACEAL:TraceFitness
01045beeae237fc791307920f30cb80bd3eb650	TRUE	3	0	0	0
01045beeae237fc791307920f30cb80bd3eb650	TRUE	3	0	0	0
01045beeae237fc791307920f30cb80bd3eb650	TRUE	3	0	0	0
012113d6241059a1c191068a899c0396fa3f3ee	TRUE	9	0	0	0
012113d6241059a1c191068a899c0396fa3f3ee	TRUE	9	0	0	0
012113d6241059a1c191068a899c0396fa3f3ee	TRUE	9	0	0	0

Figure 6.17: Screen-shot of report generated after replaying log on process model

Same process was repeated and datasets were created using weekly-logs.

6.8.2 Experimental Design

This experiment compared effectiveness of existing popular machine learning algorithms for early identification of students, who are likely to fail and to investigate the effect of process mining features in the performance of the techniques.

To estimate the generalization capability of the model to future dataset, 10-fold cross validation technique was used. This technique splits the original dataset into 10 subsets of equal size, preserving the original ratio of minority and majority class instances. One subset is left for the validation and rest are used for training the model. This process is repeated 10 times using different subsets for training and validation each time. In the end average results across each iteration is computed. These methods are used for predicting student's final outcome in a course as Pass or Fail, based on the two datasets discussed in section 6.8.1. Prediction was based on the learners demographics and dynamic data of the previous week. Each dataset was divided in weekly basis. After each week, prediction was made based on the available data, that is of the current week and also previous weeks. For example, Week-3 dataset consists of all available data until then (i.e., week 3, week-2 and week-1). We assume that prediction accuracy will improve as more data becomes available in upcoming weeks. For instance, prediction after week-4 means that we used all available data till week-4, which might include scores of assessments and quizzes, which are part of final score and ultimately improve accuracy. Prediction accuracy at early stages is important so that timely interventions can be made to help students.

6.8.3 Evaluation Measures

In order to compare the performance of each classifier, F1-score and Area Under Curve(AUC) were used. Due to the imbalanced nature of dataset, overall accuracy might be misleading.

1. **F1-Score:** F1-Score is widely used in binary classification problems. F1-Score is the harmonic mean between Precision and Recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

True Positive is the number of positive instances correctly classified as positive.

False Positive is the number of negative instances incorrectly classified as positive.

False Negative is the number of positive instances incorrectly classified as negative.

$$F1 - Score = \frac{2 * Precision * Recall}{precision + Recall}$$

2. **Area Under Curve(AUC):** A Receiver Operating Characteristic (ROC) curve is a way to compare diagnostic tests. It is a plot of the true positive rate against the false positive rate. The Area under Curve (AUC) is a number between 0 and 1.

$$FalsePositiveRate = \frac{FP}{(FP + TN)}$$

$$TruePositiveRate = \frac{TP}{(TP + FN)}$$

Table 6.4: Machine learning algorithms' parameters

Classifier	**Training Setting**	**Implementation Source**
KNN	K=3	scikit-learn [205]
Random Forest	estimator=10	scikit-learn [205]
Naïve Bayes	Gaussian default setting	scikit-learn[205]
Logistic Regression	default settings	scikit-learn [205]

6.9 Experimental Results

The experimental results are presented and discussed next. Table 6.5 displays the result of Dataset-1 (D1) which includes features like demographic and grading scores. Table 6.6 displays the result of Dataset-2 (D2) which have enriched standard features with process mining features. Next, effectiveness of machine learning algorithms in predicting at-risk students (those who are at-risk of failure on MOOCs dataset) was analysed. Four machine learning techniques were used on the two datasets. Table 6.5 shows the results of effectiveness of machine learning algorithms using Dataset 1(using standard features only) to predict student's likely to fail. Results show that maximum F1-score obtained is 0.78 by Naive Bayes classifier after week-1. For week-2 F1-score improved to 0.89 by Naive Bayes classifier. After week-2 F1-score of all classifiers drop. In MOOC environment it is normal that students are active in first week and also the assessments are easy to score high compared to the later weeks. After week-4 we observe continuous growth in F1-score for almost all classifiers. Maximum score achieved after week-8 is 0.86 by Random forest. Different classifiers performed differently for each week data, but overall Random forest and logistic regression performed better than rest of the techniques. The performance of the models looks promising, till the mid of the course (after week-4) F1-score reaches to 80% by logistic regression.shows the results of effectiveness of machine learning algorithms using Dataset-1 (having standard features only) to predict student's likely to fail. Results show that maximum F1-score obtained is 0.78 by Naive Bayes classifier after week-1. For week-2 F1-score improved to 0.89 by Naive Bayes classifier. After week-2 F1-score of all classifiers drop. In MOOC environment it is normal that students are active in first week and also the assessments are easy to score high compared to the later weeks. After week-4, continuous growth in F1-score was observed for almost all classi-

Table 6.5: Comparative results of the effectiveness of Machine learning algorithms on the dataset using standard features and mean ranks of classifiers from highest (1) to lowest (N)

Dataset	**LR**	**RF**	**NB**	**KNN**
Week-1	0.77	0.712	0.788	0.724
Week-2	0.879	0.866	0.89	0.848
Week-3	0.794	0.803	0.618	0.678
Week-4	0.8	0.799	0.715	0.722
Week-5	0.836	0.858	0.764	0.75
Week-6	0.836	0.84	0.743	0.747
Week-7	0.85	0.842	0.503	0.75
Week-8	0.841	0.866	0.536	0.801
Rank(mean)	**1.75**	1.875	3.125	3.25

Table 6.6: Comparative results of the effectiveness of Machine learning algorithms on the dataset using process mining features and mean ranks of classifiers from highest (1) to lowest (N)

Dataset	**LR**	**RF**	**NB**	**KNN**
Week-1	0.831	0.817	0.829	0.816
Week-2	0.831	0.833	0.861	0.796
Week-3	0.842	0.852	0.872	0.808
Week-4	0.87	0.845	0.871	0.825
Week-5	0.878	0.892	0.878	0.844
Week-6	0.854	0.889	0.88	0.835
Week-7	0.865	0.868	0.879	0.828
Week-8	0.879	0.886	0.89	0.848
Rank(mean)	2.56	2.0	**1.43**	4

fiers. Maximum score achieved after week-8 is 0.86 by Random Forest. Different classifiers performed differently for each week data, but overall Random Forest and Logistic Regression performed better than rest of the techniques. The performance of the models looks promising, till the mid of the course (after week-4), and F1-score reached to 80% by Logistic Regression.

Table 6.6 shows results of classifiers using process mining features. Using process mining features F1-score shows improvement for almost all weeks. After week-5 and week-6, F1-score drops but still has a maximum score of 0.87 by Naive Bayes. Results show that overall

all classifiers performed well in predictions; however Naive Bayes method outperformed all methods by scoring maximum accuracy of 0.89 after week-8.

In order to measure the significance of above findings, we used Friedman test [206] methodology for comparison of multiple classifiers over multiple datasets. The Friedman test is a non-parametric test used to compare observations repeated on same subjects. Chi-square with k-1 degree of freedom is the test statistic for the Friedman's test, where k is the number of repeated measures. When the p-value is small ($p < 0.05$), null hypothesis is rejected. The goal of this test is to check for significance difference among the performance of machine learning techniques. Null hypothesis of our study is "There is no difference among the performance of multiple classifiers". After applying Friedman test, the p-values are 0.02 for Dataset-1 and 0.001 for Dataset-2. As p-values are less than 0.05, null hypothesis is rejected. We conclude that there is a significant difference between the performance of classifiers. The calculation of mean ranks of classifiers (from highest to lowest), shows that the Logistic Regression and Naive Bayes scored highest ranks for Dataset-1 and Dataset-2 respectively, and thus outperformed other classifiers on these datasets.

Figures 6.18 show the performance of classifiers when compared with second metric, i.e., AUC. Results are almost similar like in the case of F1-score. All classifiers performed better with process mining features compared with standard features alone.

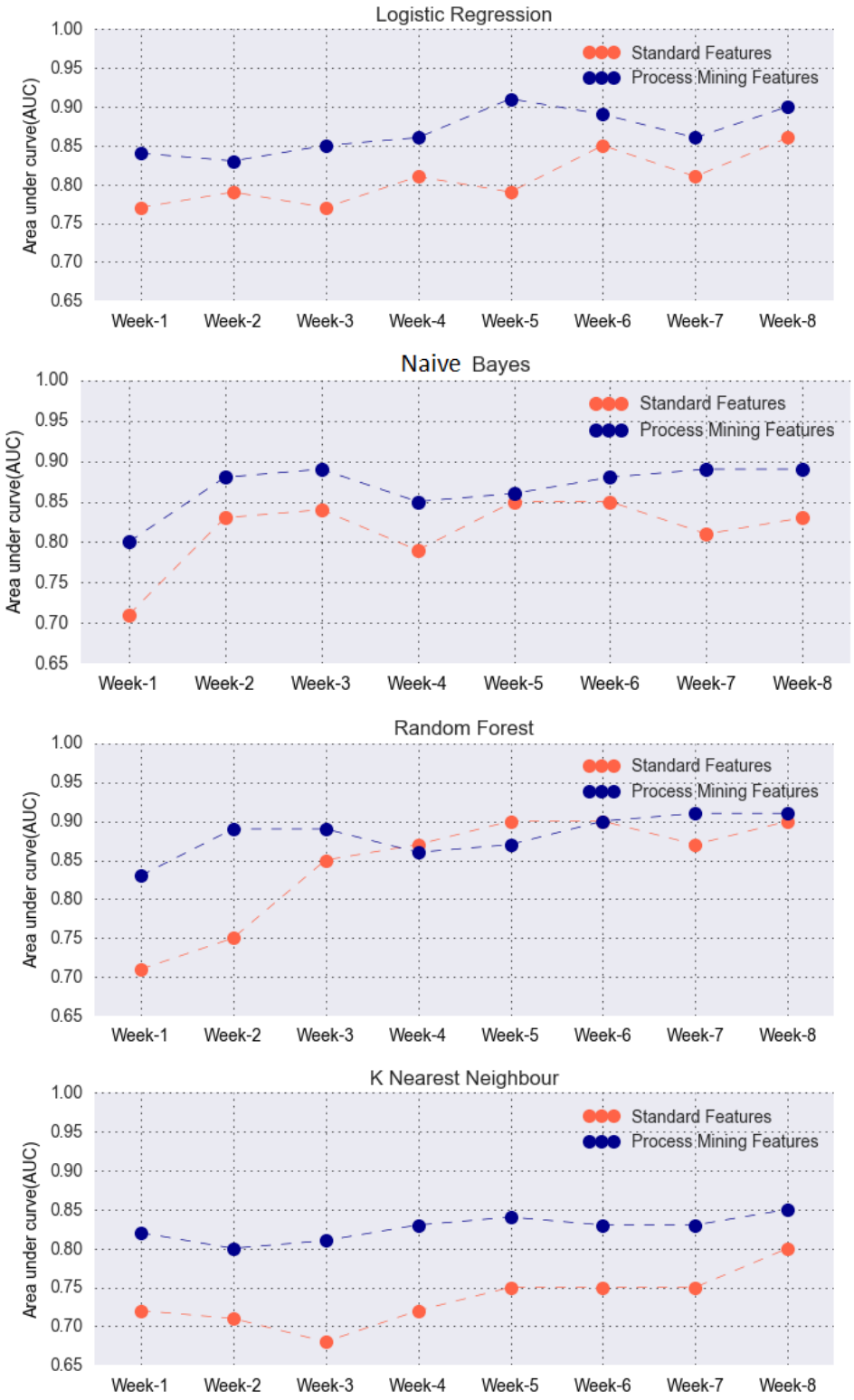


Figure 6.18: Comparative results of classification methods on the dataset consisting of process mining features and without process mining features.

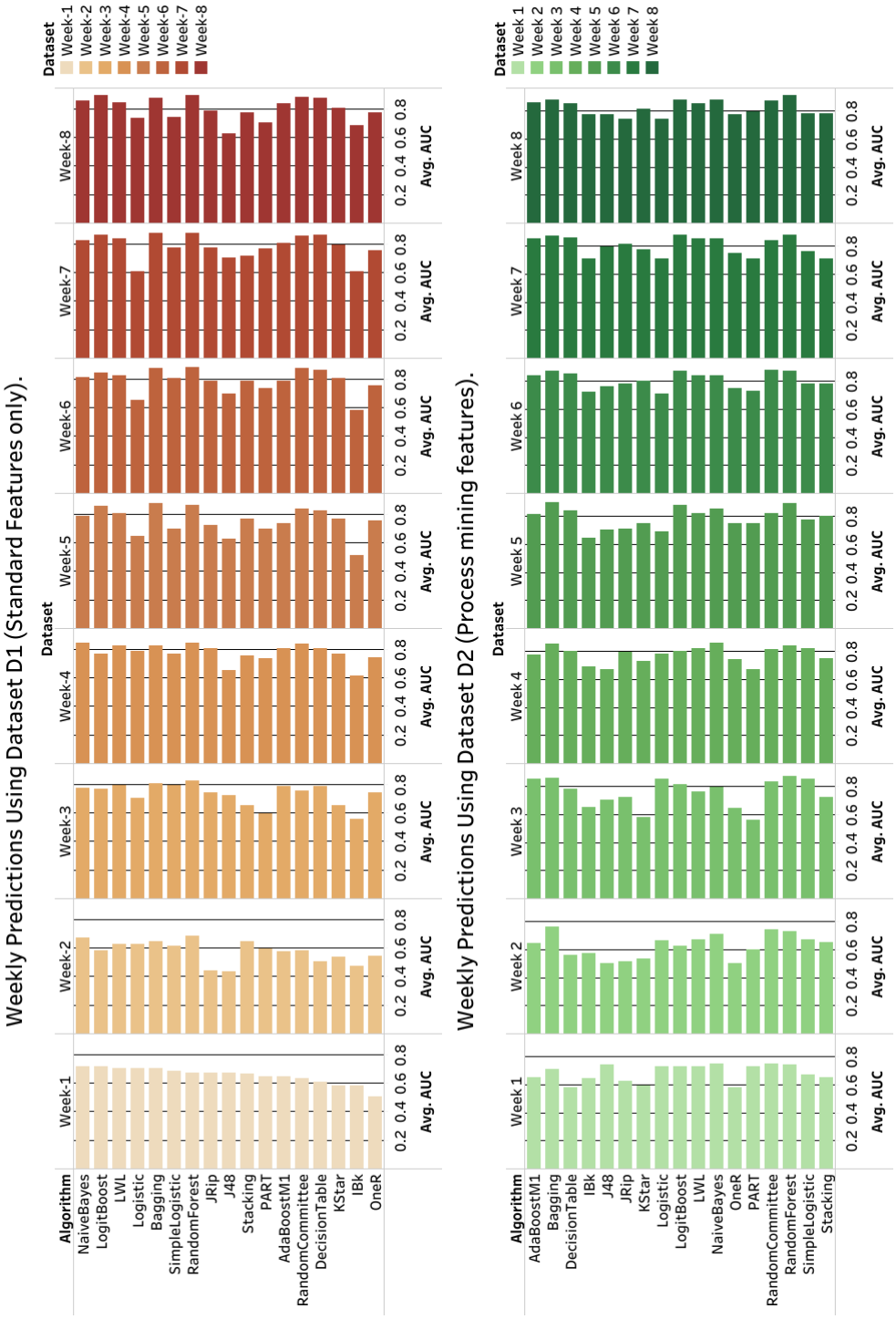


Figure 6.19: Comparative results of different classification methods on the datasets.

The study further investigated whether integration of process mining features is able to increase the effectiveness of the machine learning algorithm for the MOOCs problem domain? Same prediction experiment was conducted on Dataset-2, which consists of features used in Dataset 1 and additional features obtained from conformance testing results. Table 6.6 displays the F1-score obtained after evaluating machine learning algorithms on Dataset 2. F1-score improves after each week as expected. Figures 5.19 show the comparative results of the effectiveness of machine learning algorithms when applied on Dataset-1 with standard features and Dataset-2 which contains process mining features as additional features to the standard features. Results showed that for all weeks, F1-score improved using process mining features except for week-2 for some methods. In order to measure the significance of these results, paired t-test was applied on the results. Following p-values were obtained as a result of t-test: (Logistic Regression) $pvalue = 0.052$; (Random Forest) $pvalue = 0.02$; (K-Nearest Neighbor) $pvalue = 0.007$; (Naive Bayes) $pvalue = 0.01$. To present a significance difference, p-value should be normally less than 0.05 [143], Therefore, it can be concluded that all classifiers present a statistically significant improvement in F1-score when process mining features were integrated with standard features, except Logistic Regression.

6.9.1 Feature Importance

Next, to identify the relevant predictor variables, variable importance measure produces by the Random Forest classifiers were used. The RF model was trained with 10,000 trees on Dataset-2 to rank the 10 features by their respective importance measures. Dataset-2 was chosen as it comprises both standard features and process mining features; hence, which features are more informative to the target variable were investigated. Table 6.7 shows the top ten important features for each week. Results show that features that measure weekly time

spent and video watching activities were the most important features for all weeks. Second most important features are related to process mining.

Rank	Week-1	Week-2	Week-3	Week-4	Week-5	Week-6	Week-7	Week-8
1	LecLagw1	Vid-Act-w2	Timespentw3	VidActw4	VidActW4	VidActw4	VidActw4	VidActw4
2	VidActw1	VidActw1	VidActw2	Timespentw3	VidActw5	VidActw5	VidActw5	VidActw5
3	Timespentw1	LecLagw1	VidActw1	Timespentw4	Timespentw3	Timespentw3	Timespentw3	Timespentw8
4	Tracefitness	Timespentw1	Quizattempw3	VidActw2	Timespentw4	Timespentw6	Timespentw6	Timespentw3
5	Movelogfit	Timespentw2	LecLagw1	VidAc-w1	Queue-state	Timespentw4	Timespentw7	Timespentw6
6	Movemodelfit	Queuestate	Timespentw2	Tracefitness	VidActw2	VidActw2	Timespentw4	Timespentw7
7	Queuestate	Movemodelfit	Queuestate	Movemodelfit	Movelogfit	Tracefitness	Queuestate	Timespentw4
8	Rawfitcost	Movelogfit	Tracefitness	Movelogfit	Tracfitness	Queuestate	Tracefitness	Queuestate
9	Quizattmw1	Tracefitness	Movelogfit	Queuestate	Movemodelfit	Movemodelfit	Movemodelfit	Movemodelfit
10	Tracelength	Quizlagw2	Movemodelfit	Timespentw2	VidActw1	Movelogfit	Movelogfit	Tracefitness

Table 6.7: Feature importance by Random Forest Classifier for Dataset-2

6.9.2 Limitations of Datasets

MOOCs environment is different from traditional learning setups, which makes it challenging to analyze the data. Large amount of missing data, multiple number of attempts for assignment submission, multiple time registration, higher rate of dropout, are some of the major challenges faced during analysis of MOOCs data. In this study, a total of 167 students' data has been used of which majority of students belonged to one class and also the final dataset was imbalanced. It has been found that in a typical MOOC setting, students are active in the early weeks and become inactive or withdraw from the course in later weeks which results in a large number of missing values. Figure 6.20 shows the average number of quizzes attempted by students during each week. Students attempted most of the quizzes in week-1 only. Figure 6.21 shows the average number of lectures watched during the course. It is evident that last three weeks were the most inactive weeks. Majority of the students either did not watch lectures or withdrew from the course. These indicate unequal patterns in participation. Figure 6.22 shows average time spent during the course. The graph shows

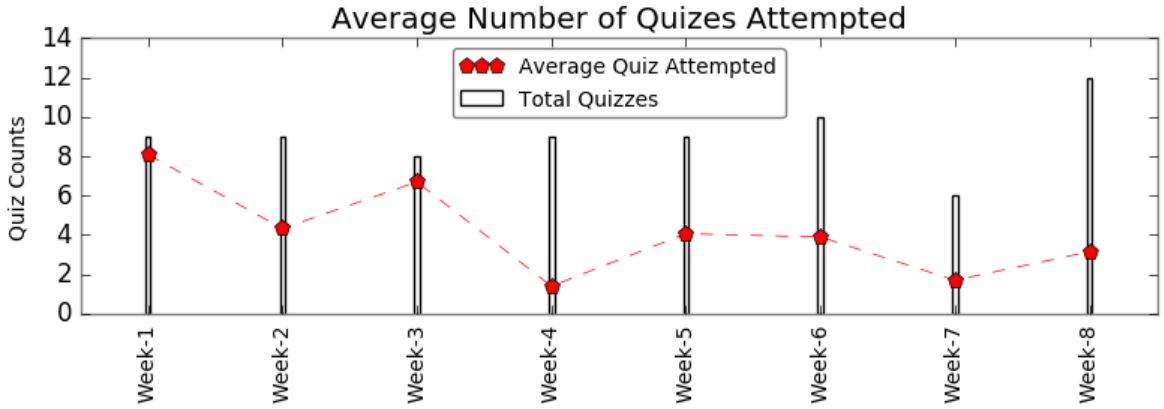


Figure 6.20: Average number of quizzes attempted(Failed or passed) by students during the course

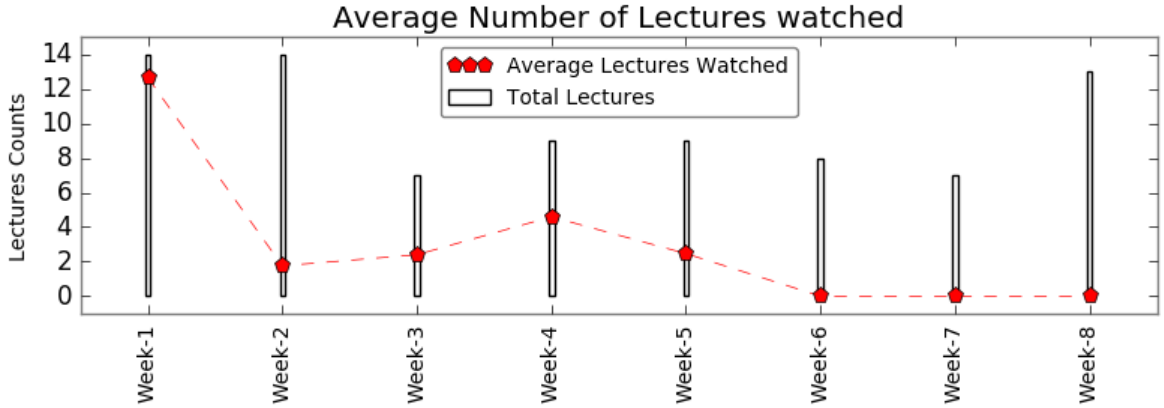


Figure 6.21: Average number of lectures watched by students during the course

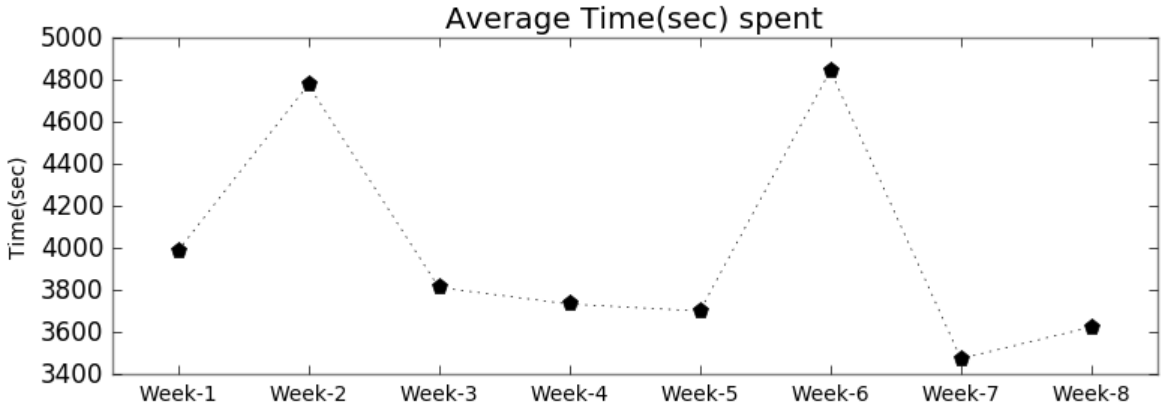


Figure 6.22: Average number of time(seconds) spent during the course

week-7 to be the least active week. However online engagement traces do not necessarily reflects all activities related to learning process. Due to this reason, measurement of success and participation in MOOC environment must be reconsidered [207] [208].

This study recognizes these limitations in the datasets; however this subset(of 167 students) is representative of student participation and completion from a real-world MOOC environment. The datsets have been used to mainly demonstrate, effectiveness of data mining techniques using process mining features.

6.10 Summary

This study has outlined innovative use of process mining techniques in education data mining to help educators gather data driven insight on student performances in enrolled courses. The first part of this chapter has demonstrated operational measures to include relevant activity data in an event log from a process-unaware database (using Moodle LMS). Challenges and lessons learned while extracting static data from non-process-oriented systems that do not follow the format required for process mining projects have been reported. Researcher further demonstrated a running example with student data extracted from activity logs when students engaged in quiz-taking process. While this study demonstrated a quiz-taking process which is only one of the learning activities performed during the course of study, there are other learning activities too which can be used. Future studies can follow a similar educational process mining approach by accommodating other learning activities like assignment submission, reading online resources, watching video lectures, etc.

In second part of the chapter, a MOOC dataset was used as a case study for predicting student's performance through the traces they leave while pursuing a course. The study has described a possibility that integrating process mining approaches can help achieve high

prediction accuracy. The use of features, obtained from process mining approach, for the purpose of prediction of students performance is novel. Data mining/machine learning algorithms were applied to weekly generated student data, as students progressed through a course, to enable prediction of students who may be at risk of not satisfying course requirements, or are rather likely to fail. Finally, a comparative analysis of four techniques (Logistic Regression, Random forest, Naive Bayes, K-Nearest Neighbor) was conducted to evaluate two datasets, one with standard features used in literature and second with features obtained from process conformance testing. By integrating process mining features with traditional features, effectiveness of some techniques have improved. Logistic Regression and Naive Bayes classifiers outperform other techniques in a statistically significant way, that is, for dataset with standard features and for dataset with extended features respectively. The importance of features using Random Forest classifier was also measured. Results show that process mining features were among top 10 important features for all datasets, however features related to weekly time spent and video watching activities were most important features among all.

The significance of our study is the use of process mining to enrich the features and results show that overall performance is statistically significantly improved using process mining features. The limitation of this study is the missing values and the small size of the data. This study recognizes these limitations in the datasets; however, this subset (of 167 students) is representative of student participation and completion from a real-world MOOC environment.

NOTE: Chapter 6 is a partial re-print of following two articles: The thesis author was the primary investigator of both these articles.

1. R. Umer,., Susnjak, T., Mathrani, A., Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. Journal of Research in Innovative

Teaching Learning, 10(2), 160–176. https://doi.org/10.1108/JRIT-09-2017-0022.

2. R. Umer,., Susnjak, T., Mathrani, A., Suriadi, S. (2019 in-press) Data Quality Challenges in Educational Process Mining: Building Process-Oriented Event Logs from Process-Unaware Online Learning Systems. Int. J. of Business Information Systems

Chapter 7

Conclusions

Government's funding to higher education providers are based upon graduates produced rather than on student enrollments. Therefore, unfinished degrees or delayed degree completions are major concerns for higher education providers since it impacts their long-term financial security and overall cost of effectiveness. Providers work towards developing strategies for improving the quality of their education and increasing their enrollment and retention rates.

In the recent past, use of predictive models for enabling real-time identification of struggling students have increasingly been recognized as a way forward to improve students' retention rate. These predictive models are developed by utilizing students' demographic data and other forms of behavioral data extracted from a range of resources within educational settings to predict performance of new students. Once those students who are struggling to engage with the course have been identified (i.e., they have been classed as at-risk of failing a course), then instructors can plan proactive measures to support these students in addressing their learning difficulties. In this manner, predictive models provide a safety net for students, who

can then be provided with timely support, thereby positively impacting their performance.

In this section all chapters of the study are summarized, and findings are discussed. The thesis has provided a synthesis of literature in the domain of education data mining, machine learning and education process mining on current methods used for prediction of students' academic performance. Additionally, this review has highlighted gaps and identified challenges in current research studies, which consequently forge a path for conducting future research studies. The systematic literature review presented includes sections which covers the methodology for review, results of review provide an overview of types of data, and machine/data mining methods used for prediction of students' performance. Further, the current status of education process mining is described and discussed. The methodology used for the conduct of this study is given in great detail. The chapter provides details of data sources and describes the three datasets that were extracted from these sources. Additionally, it describes the general methods that were common for all experiments in the research. Training procedures, evaluation methods and the description of the environment and hardware used are also described.

Based on literature findings, this study hypothesized that courses could be divided into distinct subgroups based on the characteristics of course as well as based upon associated online behaviors of the enrolled students. Therefore, classification of courses was attempted by mining the Learning Management System log data. The aim was to investigate the use of large-scale LMS data in one institution for more than 10 thousand courses and find out how various disciplines have made use of LMS tools. The investigation was based on the LMS tools usage, to see if distinct subgroups could be found that could provide futuristic ground for general and portable predictive model for similar courses.

Clustering methods were applied on more than 4000 heterogeneous courses to divide them

into smaller homogeneous groups. Using hierarchical clustering approach, 4 clusters of courses were identified. Cluster 1, comprising 32.5% of the courses is the largest cluster and demonstrated high level of online activities in the category of assignment, forum and resource. More than 50% of the courses used non-passive actions in activity item assignment specifically. Cluster1 can be titled as assignment-based activities since almost 22% courses have utilized quiz module of Moodle. Cluster 2 which is smallest in number (n = 774), is the group with most inactive courses which used Moodle; it was used only for accessing course materials and most of the time was used only for passive activities like view. Cluster 2 showed higher usage of activities like forum, resource and URL. Cluster 3 (n = 915) was most diverse in terms of online activity items utilized. More than 50% of the courses utilized seven distinct activity items such as glossary, feedback, lesson and book which other clusters rarely used. Cluster 4 also showed high levels of online activities as cluster 1, especially in the category of *interactive delivery of contents* and few of courses in the cluster used *attendance* tool which no other cluster used. It is not possible to distinguish between the clusters in terms of academic disciplines.

Next, data analytics techniques have been used in different educational domains to provide deeper insights on student progression through a course, that is, draw out predictions as to which students are at risk of not satisfying course requirements, or are likely to withdraw. The LMS data of a different types of courses have been analyzed, and the study has shown how the prediction accuracy can be improved when the LMS data was integrated with assignment scores. In first scenario, event logs of five MOOCs courses were used to predict those students that are most likely to have dropped out. Results show that techniques used in this study are able to make predictions on dropouts. However, it can be further improved by integrating more features that are directly linked to the learning process like assessments, quizzes grades

etc. Nevertheless, it is useful in cases when we need to make early predictions, such as during the first or second week of the courses when assessments do not yet exist, but instead, the event logs of students' interaction with learning management system is available. This study investigated the fact that students who engage more in the courses are less likely to dropout. Additional to this, results also show that prediction accuracy is better in courses where there is a significant difference between engagement levels of two groups of students (at-risk and not-at-risk). In second scenario, different types of courses were selected from the main dataset, in terms of their instruction format (distance vs internal) and they were compared using several machine learning methods, including Random Forest, Logistic Regression, Naıve Bayes, LDA and Ensemble method to predict student' outcomes in these courses using LMS trace data and assessment scores. The aim was to predict the outcome after every week into two classes; at-risk and not-at-risk of failing the course. Results confirm that LMS data have got discriminating power, but not more than assessment scores. Courses in which students used LMS more frequently and number of assignments are more than 3, the accuracy of predictive model for such courses were high. The thesis findings show that combination of LMS data and assessment scores can improve the accuracy of predictive models. However, it is not enough to generalize the conclusion as the data used in the study is limited to one institution and there are more data that can be used for classification. The study further proposes a multi-label regression model for recommending classification algorithms to solve a prediction problems in education domain. These regression models are trained using historical data of different courses. This study utilized data of 33 courses, and more than 20 classification algorithms were used to predict students' final outcome in the course. Meta-features were calculated for each dataset which not only used statistical features, but domain knowledge was also included. Meta-features and performance of classification algorithms were integrated to

make training set for regression model. Several regression algorithms were used to predict the performance of classification algorithm for a given dataset using the meta-features. Regression models compared significantly, and the best regression model was selected using Spearman's co-relation rank, which can then be used to rank classification algorithms for a given data set. These kinds of applications are useful for recommending set of algorithms with promising performance to the end-user who may lack domain knowledge.

Next, this study has outlined innovative use of process mining techniques in education data mining to help educators gather data driven insight on student performances in enrolled courses. Chapter 6 demonstrated a running example with student data extracted from activity logs when students engaged in quiz-taking process. While this study demonstrated a quiz-taking process which is only one of the learning activities performed during the course of study, there are other learning activities too which can be used. Future studies can follow a similar educational process mining approach by accommodating other learning activities like assignment submission, reading online resources, watching video lectures, etc. The thesis has demonstrated operational measures to include relevant activity data in an event log from a process-unaware database (using Moodle LMS). Challenges and lessons learned while extracting static data from non-process-oriented systems that do not follow the format required for process mining projects have been reported.

Next, a MOOC dataset was used as a case study for predicting student's performance through the traces they leave while pursuing a course. The study has described a possibility that integrating process mining approaches can help achieve high prediction accuracy. The use of features, obtained from process mining approach, for the purpose of prediction of students' performance is novel. Data mining/machine learning algorithms were applied to weekly generated student data, as students progressed through a course, to enable prediction

of students' performance who may be at risk of not satisfying course requirements or are rather likely to fail. Finally, a comparative analysis of four techniques – Logistic Regression, Random forest, Naive Bayes and K-Nearest Neighbor – was conducted to evaluate two datasets, one with standard features used in literature and second with features obtained from process conformance testing. By integrating process mining features with traditional features, the effectiveness of some techniques improved. Logistic Regression and Naive Bayes classifiers outperformed other techniques in a statistically significant way, that is, for dataset with standard features and for dataset with extended features respectively. The importance of features using Random Forest classifier was also measured. Results show that process mining features were among top 10 important features for all datasets, and features related to weekly time spent and video watching activities were most important features among all.

7.1 Challenges and Limitations

This section briefly lists the challenges and limitations in the domain when machine learning is applied in the prediction of students' performance.

Analysis findings performed in online course is different than offline courses and therefore cannot be generalized due to the difference in the nature of the courses. For online courses, LMS provides the only venue for interaction with peers and instructors, and for performing online activities. It is reasonable to expect that activities performed in a completely online courses would be all related to learning. There are challenges regarding the availability of open dataset that contains the learners' interaction data. While there is abundance of data, there are no benchmark datasets where researchers can test their algorithms and compare with other researchers [209]. Therefore, no specific criteria on what kind of data should be collected, the real meaning behind the collected data or what pedagogical theories align with

the findings. Therefore, there is a lack of consistency in research findings which show that data is co-related to the performance of students. Also, there is much diversity in the size of the sample and features used in literature. Also, there is no acceptable minimum requirement set for the sample size. Hence these results cannot be generalized due to the undefined size of samples that have been used. More features and bigger sample size can be helpful for accurately predicting the performance.

Prediction accuracy depends on the quality and reliability of the data provided. The data that is created by instructors for testing their teaching and assessment contents gives misinformation when added as information belonging to the learners. Such data affects the performance of the classifiers. For example, in majority of courses, the performance data collected during the semester such as quizzes or assignment scores, might not be valid if they are changed later at the end of semester, which can in turn affect the prediction accuracy.

Currently, the use of LMS logs and assessment scores as independent variables for predicting students' performance, or they rely on self-reported use of technology. However, students' engagement is not limited to the mere learning management systems. For example, use of social networking sites are very popular among students. According to the EDUCASE study 90% of the student use Facebook and 58% use it several times a day [209]. Another study confirms this results that students on an average use 1 hour 40 minutes a day on Facebook and this time spent on Facebook has significant negative impact on GPA [210]. Because students spend significant amount of time online, it makes sense to find new methods to collect student generated data [211]. One of the methods is by using monitoring software. The monitoring software was developed for employees to monitor their compute time activities, although once installed, the software cannot be removed until the monitoring time is finished. This software runs in the background and logs all activities that were performed. There might be

one limitation that once students know that they are being monitored, they might perform differently.

Dataset used for prediction are mostly imbalanced, as the number of students who have failed the course are typically less than the number who passed the course. This results in negative skew as accuracy for prediction of pass students is higher than the fail students. Whereas higher accuracy is desired for predicting students who are going to fail the course so that interventions can be made to improve their performance [67]. Advanced techniques are needed to address the imbalanced nature of datasets.

7.2 Future Direction

This section presents possible future directions to support researchers who are interested in this domain.

- Most common data used in for predictions include socio-demographic information, enrollment or registration details, previous grades achieved, forum messages posted, features related to the use of LMS tools alongside other daily activities. All these data are considered as low-cost data, which can be collected with less resources such as LMS log data or through other education databases. However, many other psychological factors such as learning style, self-efficacy, achievement goal, motivation, interest, learning and teaching environment could have been accessed using surveys, or through monitoring software, which while needs more effort and higher cost, but would be more meaningful pedagogically.

- Develop an early warning system to monitor learner's performance and provide feedback to the learners.In addition, investigate the effect of different intervention strategies

such as email which is less time-consuming vs one-to-one meetings.

- Investigate the effective use and optimal time to overcome the trade-off between higher accuracy (late predictions with no time left for help) and time for intervention (early prediction with less accuracy)

- Develop a system that automatically evaluates the content of the forum messages or student-peer reviews using text mining algorithms and assigns a score to the messages based on the content.

- Develop customized reporting tools or dash-board style visualizations that track student data according to the pedagogical intent. This will also cater to the different uses of the LMS by instructors or administrators.

- Conduct investigation of LMS tracking data and network measures in relation to student performance for supporting pedagogical designs and course delivery modalities.

- Investigate important course material, topics or learning objectives that are central for enabling student success in a course in order to support instructors in their preparation of course materials.

- Investigate the issues of portability regarding face-to-face and fully online program given how much more LMS usage takes place in the later mode of instruction.

- Investigate the difference of risk indicators for international and domestic students such as diverse ethnic group, financial status, language barrier etc.

References

[1] P. Hill, "State of the anglosphere's higher education lms market: 2013 edition," *E-Literate Blog, posted November*, vol. 9, 2013. xiii, 16

[2] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of educational research*, vol. 45, no. 1, pp. 89–125, 1975. 3

[3] W. G. Spady, "Dropouts from higher education: Toward an empirical model," *Interchange*, vol. 2, no. 3, pp. 38–62, 1971. 3

[4] A. Astin, "Preventing students from," *Dropping Out. San Francisco, Jossey-B ass Publishers*, 1975. 3

[5] A. Astin *et al.*, "Retaining and satisfying students.," *Educational Record*, vol. 68, no. 1, pp. 36–42, 1987. 3, 4

[6] V. Tinto and B. Pusser, "Moving from theory to action: Building a model of institutional action for student success," *National Postsecondary Education Cooperative*, pp. 1–51, 2006. 3

[7] A. E. Bayer, "The college drop-out: Factors affecting senior college completion," *Sociology of Education*, pp. 305–316, 1968. 3

[8] K. A. Feldman and T. M. Newcomb, "The impact of college: Epilogue," *The impact of college on students*, vol. 1, pp. 325–338, 1969. 3

[9] E. Marks, "Student perceptions of college persistence, and their intellective, personality and performance correlates.," *Journal of Educational Psychology*, vol. 58, no. 4, p. 210, 1967. 3

[10] R. J. Panos and A. W. Astin, "Attrition among college students," *American Educational Research Journal*, vol. 5, no. 1, pp. 57–72, 1968. 3

[11] W. R. Habley, J. L. Bloom, and S. Robbins, *Increasing persistence: Research-based strategies for college student success*. John Wiley & Sons, 2012. 3

[12] S. L. DesJardins, D. A. Ahlburg, and B. P. McCall, "An event history model of student departure," *Econ. Educ. Rev.*, vol. 18, no. 3, pp. 375–390, 1999. 4, 20

[13] P. T. T. Ernerst T. Pascarella, "How college affects students," *Fenxi Huaxue*, vol. 32, no. 10, pp. 1365–1367, 2004. 4, 20

[14] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The prediction of students' academic performance using classification data mining techniques," *Appl. Math. Sci.*, vol. 9, no. 129, pp. 6415–6426, 2015. 4, 21

[15] W. Laura, W. Jack, M. Helen, and T. Marjorie, "Efficacy of using a single , non-technical variable to predict the ...," pp. 41–48, 2003. 4, 21

[16] Q. Jin and W. Lafayette, "A multi-outcome hybrid model for predicting student success in engineering," *2011 Annual Conference Exposition*, 2011. 4, 21

[17] C. R. Graham, "Blended learning systems," *The handbook of blended learning: Global perspectives, local designs*, pp. 3–21, 2006. 4, 90

[18] N. Cavus and T. Zabadi, "A comparison of open source learning management systems," *Procedia-Social and Behavioral Sciences*, vol. 143, pp. 521–526, 2014. 4, 90

[19] G. Siemens, "Learning analytics: envisioning a research discipline and a domain of practice," in *Proceedings of the 2nd international conference on learning analytics and knowledge*, pp. 4–8, 2012. 4, 90

[20] R. Ferguson, A. Brasher, D. Clow, A. Cooper, G. Hillaire, J. Mittelmeier, B. Rienties, T. Ullmann, and R. Vuorikari, "Research evidence on the use of learning analytics: Implications for education policy," 2016. 4, 90

[21] N. Sclater, A. Peasgood, and J. Mullan, "Learning analytics in higher education," *London: Jisc. Accessed February*, vol. 8, no. 2017, p. 176, 2016. 4, 90

[22] O. (2019), "education-at-a-glance-2019$_f$8d7880$d-en$,″ 0.6

[23] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, "A reference model for learning analytics," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5/6, pp. 318–331, 2012. 14

[24] J. Cole and H. Foster, *Using Moodle: Teaching with the popular open source course management system.* " O'Reilly Media, Inc.", 2007. 15

[25] E. Dahlstrom, D. C. Brooks, and J. Bichsel, "The current ecosystem of learning management systems in higher education: Student, faculty, and it perspectives," 2014. 15

[26] C. R. Graham, W. Woodfield, and J. B. Harrison, "A framework for institutional adoption and implementation of blended learning in higher education," *The internet and higher education*, vol. 18, pp. 4–14, 2013. 15

[27] A. Merceron, "Educational data mining/learning analytics: Methods, tasks and current trends," *CEUR Workshop Proc.*, vol. 1443, no. DeLFI, pp. 101–109, 2015. 16, 25

[28] R. Baker *et al.*, "Data mining for education," *International encyclopedia of education*, vol. 7, no. 3, pp. 112–118, 2010. 16, 64

[29] "A Data Model to Ease Analysis and Mining of Educational Data 1," no. June, 2010. 19

[30] "A Java Desktop Tool for Mining Moodle Data .," pp. (pp. 319–320)., In Proceedings of the 4th international conference on educational data mining, 2011. 19

[31] "Automatic Generation of Proof Problems in Deductive Logic," 2011. 19

[32] M. Rodrigo, R. Baker, B. McLaren, A. Jayme, and T. Dy, "Development of a workbench to address the educational data mining bottleneck," *Proc. 5th Int. Conf. Educ. Data Min.*, pp. 152–155, 2012. 19

[33] M. Johnson, M. Eagle, L. Joseph, and T. Barnes, "The EDM Vis Tool," *Proc. 4th Int. Conf. Educ. Data Min.*, pp. 349–350, 2011. 19

[34] P. Sorenson and L. P. Macfadyen, "Learner Interaction Monitoring System (LiMS): Capturing the Behaviors of Online Learners and Evaluating Online Training Courses Context : The need to demonstrate return on investment in online education and training Beyond the LMS : The Learner Interact," no. 203, 2010. 19

[35] K. E. Maull, M. G. Saldivar, and T. Sumner, "Online Curriculum Planning Behavior of Teachers," *Proc. 3rd Int. Conf. Educ. Data Min.*, no. October, pp. 121–130, 2010. 19

[36] R. Rabbany, M. Takaffoli, and O. R. Za, "Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques," *Proc. Educ. data Min.*, pp. 21–30, 2011. 19

[37] M. E. Z. PANTALEÓN, D. GARCÍA-SAIZ, and M. E. Z. PANTALEÓN, "E-learning Web Miner: A data mining application to help instructors involved in virtual courses," *Proc. 4th Int. Conf. Educ. Data Min.*, no. February 2016, pp. 2010–2011, 2011. 19

[38] A. Cohen and R. Nachmias, "What can instructors and policy makers learn about web-supported learning through web-usage mining," *The Internet and Higher Education*, vol. 14, no. 2, pp. 67–76, 2011. 19

[39] E. García, C. Romero, S. Ventura, and C. De Castro, "A collaborative educational association rule mining tool," *Internet High. Educ.*, vol. 14, no. 2, pp. 77–88, 2011. 19

[40] J. Fritz, "Classroom walls that talk: Using online course activity data of successful students to raise self-awareness of underperforming peers," *Internet High. Educ.*, vol. 14, no. 2, pp. 89–97, 2011. 19

[41] A. Anjewierden, H. Gijlers, N. Saab, and R. De Hoog, "Brick : Mining Pedagogically Interesting Sequential Patterns," *Proc. 4th Int. Conf. Educ. Data Min.*, pp. 341–342, 2011. 19

[42] L. Moreno, C. González, R. Estevez, and B. Pesquet-Popescu, "Intelligent Evaluation of Social Knowledge Building Using Conceptual Maps with MLN," *Proc. 4th Int. Conf. Educ. Data Min.*, pp. 343–344, 2011. 19

[43] A. L. Dyckhoff, D. Zielke, M. A. Chatti, and U. Schroeder, "eLAT : An Exploratory Learning Analytics Tool for Reflection and Iterative Improvement of Technology Enhanced Learning," *Proc. 4th Int. Conf. Educ. Data Min.*, pp. 355–356, 2011. 19

[44] T. Devine, M. Hossain, E. Harvey, and A. Baur, "Improving Pedagogy by Analyzing Relevance and Dependency of Course Learning Outcomes," *Computer (Long. Beach. Calif).*, 2011. 19

[45] M. Pechenizkiy, N. Trcka, P. De Bra, and P. Toledo, "CurriM : Curriculum Mining," *Proc. 5th Int. Conf. Educ. Data Min.*, no. i, pp. 1–4, 2012. 19

[46] K. Nelson, J. Clarke, S. Kift, and T. Creagh, *Trends in policies, programs and practices in the Australasian First Year Experience literature 2000–2010 The First Year in Higher Education Research Series on Evidence-based Practice.* 2011. 20

[47] S. McDonald and H. M. Edwards, "Who should test whom?," *Communications of the ACM*, vol. 50, no. 1, pp. 66–71, 2007. 20

[48] OECD, *Education at a Glance 2019.* 2019. 20

[49] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *Journal of clinical epidemiology*, 2009. 22

[50] Y. H. Hu, C. L. Lo, and S. P. Shih, "Developing early warning systems to predict students' online learning performance," *Comput. Human Behav.*, vol. 36, pp. 469–478, 2014. 25, 27, 31, 41

[51] L. P. Macfadyen and S. Dawson, "Mining lms data to develop an "early warning system" for educators: A proof of concept," *Computers and Education*, vol. 54, pp. 588–599, 2 2010. 25, 27, 28, 29, 41

[52] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm," *Computers in Human Behavior*, pp. 0–1, 6 2018. 25, 29, 41

[53] S. Young and M. McSporran, "Confident men-successful women: Gender differences in online learning," pp. 2110–2112, 2001. 25

[54] E. V. W. v. d. A. P. D. B. Mykola Pechenizkiy, Nikola Trčka, "Process mining online assessment data," *Proceedings of the 2nd International Conference on Educational Data Mining*, pp. 279–288, 2009. 25

[55] N. Johannes, "The relationship between demographics and the academic achievement of engineering students," pp. 347–355, 2017. 25

[56] L. Green and G. Celkan, "Student demographic characteristics and how they relate to student achievement," *Procedia - Social and Behavioral Sciences*, vol. 15, pp. 341–345, 2011. 25

[57] M. Nasir and R. Masrur, "An exploration of emotional intelligence of the students of iiui in relation to gender, age and academic achievement," *Bull. Educ. Res.*, vol. 32, no. 1, pp. 37–51, 2010. 25

[58] L. Ali, M. Asadi, D. Gašević, J. Jovanović, and M. Hatala, "Factors influencing beliefs for adoption of a learning analytics tool: An empirical study," *Comput. Educ.*, vol. 62, pp. 130–148, 3 2013. 25

[59] D. T. Tempelaar, B. Rienties, and B. Giesbers, "In search for the most informative data for feedback generation: Learning analytics in a data-rich context," *Comput. Human Behav.*, vol. 47, pp. 157–167, 6 2015. 25

[60] S. B. Shum and R. D. Crick, "Learning dispositions and transferable competencies," *Proc. 2nd Int. Conf. Learn. Anal. Knowl. - LAK '12*, no. May, p. 92, 2012. 25

[61] E. Aguiar, G. A. A. Ambrose, V. N. Chawla, V. Goodrich, and J. Brockman, "Engagement vs performance: Using electronic portfolios to predict first semester engineering student persistence," *J. Learn. Anal.*, vol. 1, no. 3, pp. 7–33, 2014. 27, 41

[62] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Human Behav.*, vol. 73, pp. 247–256, 2017. 5 Start. 27, 31, 32, 41

[63] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "The open university ' s repository of research publications improving retention : predicting at-risk students by analysin clicking behaviour in a virtual learning environment conference item analysing clicking behaviour in a virtual learning," *Third Conference on Learning Analytics and Knowledge (LAK 2013)*, 2013. 27

[64] V. L. Miguéis, A. Freitas, V. P. J. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decis. Support Syst.*, vol. 115, pp. 36–51, 11 2018. 27, 31, 41

[65] A. Sandoval, C. Gonzalez, R. Alarcon, K. Pichara, and M. Montenegro, "Centralized student performance prediction in large courses based on low-cost variables in an institutional context," *Internet High. Educ.*, vol. 37, pp. 76–89, 4 2018. 27, 34, 41

[66] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Computers and Education*, vol. 61, no. 1, pp. 133–145, 2013. 27, 29, 34, 41

[67] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Computers Education*, vol. 103, pp. 1–15, 12 2016. 27, 28, 29, 31, 32, 41, 174

[68] M. Hlosta, Z. Zdrahal, and J. Zendulka, "Ouroboros: early identification of at-risk students without models based on legacy data," *LAK17 - Seventh International Learning Analytics Knowledge Conference*, pp. 6–15, 2017. 28, 29, 33, 42

[69] C. Burgos, M. L. Campanario, D. De, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students ' performance : A tutoring action plan to prevent academic dropout r," vol. 66, pp. 541–556, 2018. 28, 32, 41

[70] R. Asif, A. Merceron, S. Abbas, and N. Ghani, "Computers education analyzing undergraduate students ' performance using educational data mining," *Computers Education*, vol. 113, pp. 177–194, 2017. 28, 29, 30, 33, 42

[71] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004. 28

[72] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," *2014 International Conference on Communication and Network Technologies, ICCNT 2014*, vol. 2015-March, pp. 113–118, 2015. 28, 42

[73] G. M. V. N. Ioanna Lykourentzou, Ioannis Giannoukos and V. Loumos, "Early and dynamic student achievement prediction in e-learning courses using neural networks," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, vol. 60, no. 2, p. 372–380, 2009. 28, 34, 42

[74] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from lms data: A comparison of 17 blended courses using moodle lms," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, 2017. 28, 32, 41

[75] M. Hall and E. Frank, "Combining naive bayes and decision tables," in *Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS)*, pp. 318–319, AAAI press, 2008. 28

[76] W. W. Cohen, "Fast effective rule induction," in *Twelfth International Conference on Machine Learning*, pp. 115–123, Morgan Kaufmann, 1995. 28, 59, 115

[77] B. Martin, "Instance-based learning: Nearest neighbor with generalization," Master's thesis, University of Waikato, Hamilton, New Zealand, 1995. 28

[78] S. Roy, "Nearest neighbor with generalization." 2002. 28

[79] B. R. Gaines and P. Compton, "Induction of ripple-down rules applied to modeling large databases," *Journal of Intelligent Information Systems*, vol. 5, no. 3, pp. 211–228, 1995. 28

[80] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long, "Predicting academic performance by considering student heterogeneity," *Knowledge-Based Systems*, vol. 161, no. July, pp. 134–146, 2018. 28, 41

[81] C. Romero, M. I. López, J. M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Computers and Education*, vol. 68, no. October, pp. 458–472, 2013. 28, 32, 41, 124

[82] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: A case study with high school students," *Expert Systems*, vol. 33, no. 1, pp. 107–124, 2016. 28, 30, 32, 42

[83] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993. 29, 59, 115

[84] G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, and M. Hall, "Multiclass alternating decision trees," in *ECML*, pp. 161–172, Springer, 2001. 29

[85] Y. Freund and R. R. E. Schapire, "Experiments with a new boosting algorithm," *International Conference on Machine Learning*, pp. 148–156, 1996. 29, 31

[86] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, California: Wadsworth International Group, 1984. 29

[87] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. 29, 59, 115, 148

[88] Z. Alharbi, J. Cornford, L. Dolder, and B. De La Iglesia, "Using data mining techniques to predict students at risk of poor performance," *Proceedings of 2016 SAI Computing Conference, SAI 2016*, pp. 523–531, 2016. 29, 42

[89] R. Duarte, A. Ramos-Pires, and H. Gonçalves, "Identifying at-risk students in higher education," *Total Quality Management and Business Excellence*, 2014. 29, 42

[90] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013. 29

[91] E. Howard, M. Meehan, and A. Parnell, "Contrasting prediction methods for early warning systems at undergraduate level," *The Internet and Higher Education*, vol. 37, no. January, pp. 66–75, 2018. 29, 34, 41

[92] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauría, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative," *Journal of Learning Analytics*, vol. 1, no. 1, pp. 6–47, 2014. 29, 41

[93] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*, (San Mateo), pp. 338–345, Morgan Kaufmann, 1995. 29, 59, 115

[94] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Human Behav.*, vol. 73, pp. 247–256, 2017. 29

[95] D. Herrmannova, M. Hlosta, J. Kuzilek, and Z. Zdrahal, "Evaluating weekly predictions of at-risk students at the open university: Results and issues," *EDEN 2015 Annu. Conf. Expand. Learn. Scenar. Open. Out Educ. Landsc.*, pp. 9–12, 2015. 29, 30, 42

[96] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, pp. 498–506, 2010. 29, 42

[97] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991. 30

[98] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use moodle courses," *Comput. Appl. Eng. Educ.*, vol. 21, no. 1, pp. 135–146, 2013. 30, 42

[99] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013. 31

[100] S. Jarvela and P. Hakkinen, "The levels of web-based discussions: using perspective-taking theory as an analytical tool," *Cognition in a digital world*, pp. 77–95, 2003. 32

[101] C. E. Lopez Guarin, E. L. Guzman, and F. A. Gonzalez, "A model to predict low academic performance at a specific enrollment using data mining," *Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 10, no. 3, pp. 119–125, 2015. 32, 42

[102] F. Agudo-Peregrina, S. Iglesias-Pradas, M. Conde-González, and Hernández-García, "Can we predict success from log data in vles? classification of interactions for learning analytics and their relation with performance in vle-supported f2f and online learning," *Computers in Human Behavior*, vol. 31, pp. 542–550, 2 2014. 34, 42

[103] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? -arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014. 34

[104] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades," *Artif. Intell. Rev.*, vol. 37, no. 4, pp. 331–344, 2012. 34, 42

[105] W. van der Aalst, "Process Mining: Discovery, Conformance and Enhancement of Business Processes," *Media*, vol. 136, no. 2, p. 352, 2011. 34

[106] Process Mining Group, "ProM - the leading process mining toolkit," 2014. 34

[107] N. T??cka and M. Pechenizkiy, "From local patterns to global models: Towards domain driven educational process mining," *ISDA 2009 - 9th International Conference on Intelligent Systems Design and Applications*, pp. 1114–1119, 2009. 34

[108] N. Trcka, M. Pechenizkiy, and W. van der Aalst, *Process mining from educational data.* Chapman & Hall/CRC, 2010. 34

[109] N. Trcka and M. Pechenizkiy, "From local patterns to global models: Towards domain driven educational process mining," in *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, pp. 1114–1119, IEEE, 2009. 35

[110] "Process Mining Online Assessment Data," *Proceedings of the 2nd International Conference on Educational Data Mining*, pp. 279–288, 2009. 35

[111] M. Bannert, P. Reimann, and C. Sonnenberg, "Process mining techniques for analysing patterns and strategies in students' self-regulated learning," *Metacognition and learning*, vol. 9, no. 2, pp. 161–185, 2014. 35

[112] W. M. van der Aalst, S. Guo, and P. Gorissen, "Comparative process mining in education: An approach based on process cubes," in *International Symposium on Data-Driven Process Discovery and Analysis*, pp. 110–134, Springer, 2013. 36

[113] A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-Santillán, "Clustering for improving educational process mining," in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, pp. 11–15, ACM, 2014. 36

[114] W. Van Der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. Van Den Brand, R. Brandtjen, J. Buijs, *et al.*, "Process mining manifesto," in *International Conference on Business Process Management*, pp. 169–194, Springer, 2011. 36, 129

[115] R. J. C. Bose, R. S. Mans, and W. M. van der Aalst, "Wanna improve process mining results?," in *2013 IEEE symposium on computational intelligence and data mining (CIDM)*, pp. 127–134, IEEE, 2013. 37

[116] D. Calvanese, M. Montali, A. Syamsiyah, and W. M. van der Aalst, "Ontology-driven extraction of event logs from relational databases," in *International Conference on Business Process Management*, pp. 140–153, Springer, 2016. 37

[117] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati, "Linking data to ontologies," in *Journal on data semantics X*, pp. 133–173, Springer, 2008. 37

[118] W. M. van der Aalst, "Extracting event data from databases to unleash process mining," in *BPM-Driving innovation in a digital world*, pp. 105–128, Springer, 2015. 37

[119] R. Pérez-Castillo, B. Weber, I. G.-R. de Guzmán, M. Piattini, and J. Pinggera, "Assessing event correlation in non-process-aware information systems," *Software & Systems Modeling*, vol. 13, no. 3, pp. 1117–1139, 2014. 37

[120] M. Jans and P. Soffer, "From relational database to event log: decisions with quality impact," in *International Conference on Business Process Management*, pp. 588–599, Springer, 2017. 38, 126

[121] H. Selig, "Continuous event log extraction for process mining," 2017. 38

[122] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee, "A taxonomy of dirty data," *Data mining and knowledge discovery*, vol. 7, no. 1, pp. 81–99, 2003. 38

[123] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000. 38

[124] S. Suriadi, R. Andrews, A. H. ter Hofstede, and M. T. Wynn, "Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs," *Information Systems*, vol. 64, pp. 132–150, 2017. 38, 145

[125] H. Verbeek, J. C. Buijs, B. F. Van Dongen, and W. M. Van Der Aalst, "Xes, xesame, and prom 6," in *International Conference on Advanced Information Systems Engineering*, pp. 60–75, Springer, 2010. 39

[126] B. F. van Dongen and S. Shabani, "Relational xes: Data management for process mining.," in *CAiSE Forum*, pp. 169–176, 2015. 39

[127] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, 1996. 39, 42

[128] S. Nunn, J. T. Avella, T. Kanai, and M. Kebritchi, "Learning analytics methods, benefits, and challenges in higher education: A systematic literature review," *Online Learning*, vol. 20, no. 2, pp. 13–29, 2016. 39

[129] J. W. You, "Identifying significant indicators using lms data to predict course achievement in online learning," *The Internet and Higher Education*, vol. 29, pp. 23–30, 4 2016. 41

[130] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *The Internet and Higher Education*, vol. 28, pp. 68–84, 1 2016. 41

[131] C. Wladis, A. C. Hachey, and K. Conway, "An investigation of course-level factors as predictors of online stem course outcomes," *Computers and Education*, vol. 77, pp. 145–150, 2014. 42

[132] R. Cerezo, S. Miguel, and J. C. Nú, "Computers education students ' lms interaction patterns and their relationship with achievement : A case study in higher education," vol. 96, pp. 42–54, 2016. 42

[133] A.-s. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," *Decision Support Systems*, vol. 101, pp. 1–11, 2017. 42

[134] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv preprint arXiv:1201.3417*, 2012. 42

[135] M. M. A. Tair and A. M. El-halees, "Edm-06: Mining educational data to improve students ' performance: A case study," *Palestin - chapter III*, vol. 2, no. 2, pp. 140–146, 2012. 42

[136] R. J. Waddington, S. Nam, S. Lonn, and S. D. Teasley, "improving early warning systems with categorized course resource usage," *Journal of Learning Analytics*, vol. 3, no. 3, pp. 263–290, 2016. 42

[137] O. Corrigan, A. F. Smeaton, M. Glynn, and S. Smyth, "Using educational analytics to improve test performance," in *Design for Teaching and Learning in a Networked World*, pp. 42–55, Springer, 2015. 42

[138] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "A learning analytics approach : Using online weekly student engagement data to make predictions on student performance .," *2018 Int. Conf. Comput. Electron. Electr. Eng. (ICE Cube)*, pp. 1–5, 2018. 42

[139] P. Jia and T. Maloney, "Using predictive modelling to identify students at risk of poor university outcomes," *Higher Education*, vol. 70, no. 1, pp. 127–149, 2015. 42

[140] "data_exports.dvi." https://spark-public.s3.amazonaws.com/mooc/data_exports.pdf. (Accessed on 07/20/2020). 52

[141] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific data*, vol. 4, p. 170171, 2017. 53

[142] A. M. Shahiri, W. Husain, *et al.*, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015. 58

[143] F. Gorunescu, *Data Mining: Concepts and Techniques*, vol. 12. 2011. 58, 160

[144] W. Hoeffding, "Probability inequalities for sums of bounded random variables," in *The Collected Works of Wassily Hoeffding*, pp. 409–426, Springer, 1994. 58

[145] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *Fifteenth International Conference on Machine Learning* (J. Shavlik, ed.), pp. 144–151, Morgan Kaufmann, 1998. 59, 115

[146] R. Kohavi, "The power of decision tables," in *8th European Conference on Machine Learning*, pp. 174–189, Springer, 1995. 59, 115

[147] R. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993. 59, 115

[148] D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991. 59, 115

[149] J. G. Cleary and L. E. Trigg, "K*: An instance-based learner using an entropic distance measure," in *12th International Conference on Machine Learning*, pp. 108–114, 1995. 59, 115

[150] E. Frank, M. Hall, and B. Pfahringer, "Locally weighted naive bayes," in *19th Conference in Uncertainty in Artificial Intelligence*, pp. 249–256, Morgan Kaufmann, 2003. 59, 115

[151] C. Atkeson, A. Moore, and S. Schaal, "Locally weighted learning," *AI Review*, 1996. 59, 115

[152] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Thirteenth International Conference on Machine Learning*, (San Francisco), pp. 148–156, Morgan Kaufmann, 1996. 59, 115

[153] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996. 59, 115

[154] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992. 59, 115

[155] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," tech. rep., Stanford University, 1998. 59, 115

[156] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms.* John Wiley and Sons, Inc., 2004. 59, 115

[157] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998. 59, 115

[158] S. le Cessie and J. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992. 59, 115

[159] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," vol. 95, no. 1-2, pp. 161–205, 2005. 59, 115

[160] M. Sumner, E. Frank, and M. Hall, "Speeding up logistic model tree induction," in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 675–683, Springer, 2005. 59, 115

[161] J. Hunter, "Matplotlib: A 2d graphics environment. computing in science engineering 9, 3. 90-95," 2007. 61

[162] T. Lachev and E. Price, *Applied Microsoft Power BI: Bring your data to life!* Prologika Press, 2018. 61

[163] D. Gašević, S. Dawson, and G. Siemens, "Let's not forget: Learning analytics are about learning," *TechTrends*, vol. 59, no. 1, pp. 64–71, 2015. 64

[164] G. Siemens and D. Gašević, "Special issue on learning and knowledge analytics," *Educational Technology & Society*, vol. 15, no. 3, pp. 1–163, 2012. 64

[165] B. Xu and M. Recker, "Teaching analytics: A clustering and triangulation study of digital library user data.," *Educational Technology & Society*, vol. 15, no. 3, pp. 103–115, 2012. 66

[166] K. Nagi and P. Suesawaluk, "Research analysis of moodle reports to gauge the level of interactivity in elearning courses at assumption university, thailand," in *2008 International Conference on Computer and Communication Engineering*, pp. 772–776, IEEE, 2008. 67

[167] W. J. Krzanowski and Y. Lai, "A criterion for determining the number of groups in a data set using sum-of-squares clustering," *Biometrics*, pp. 23–34, 1988. 80

[168] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974. 80

[169] J. A. Hartigan, *Clustering algorithms*. John Wiley & Sons, Inc., 1975. 80

[170] W. Sarle, "Sas technical report a-108," *SAS Institute Inc*, 1983. 80

[171] A. J. Scott and M. J. Symons, "Clustering methods based on likelihood ratio criteria," *Biometrics*, pp. 387–397, 1971. 80

[172] F. Marriott, "Practical problems in a method of cluster analysis," *Biometrics*, pp. 501–514, 1971. 80

[173] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985. 80

[174] H. P. Friedman and J. Rubin, "On some invariant criteria for grouping data," *Journal of the American Statistical Association*, vol. 62, no. 320, pp. 1159–1178, 1967. 80

[175] L. J. Hubert and J. R. Levin, "A general statistical framework for assessing categorical clustering in free recall.," *Psychological bulletin*, vol. 83, no. 6, p. 1072, 1976. 80

[176] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979. 80

[177] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987. 80

[178] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012. 80

[179] R. O. Duda, P. E. Hart, *et al.*, *Pattern classification and scene analysis*, vol. 3. Wiley New York, 1973. 80

[180] E. Beale, *Euclidean cluster analysis*. Scientific Control Systems Limited, 1969. 80

[181] D. Ratkowsky and G. Lance, "Criterion for determining the number of groups in a classification," 1978. 80

[182] G. H. Ball and D. J. Hall, "Isodata, a novel method of data analysis and pattern classification," tech. rep., Stanford research inst Menlo Park CA, 1965. 80

[183] G. W. Milligan, "A monte carlo study of thirty internal criterion measures for cluster analysis," *Psychometrika*, vol. 46, no. 2, pp. 187–199, 1981. 80

[184] T. Frey and H. Van Groenewoud, "A cluster analysis of the d2 matrix of white spruce stands in saskatchewan based on the maximum-minimum principle," *The Journal of Ecology*, pp. 873–886, 1972. 80

[185] J. O. McClain and V. R. Rao, "Clustisz: A program to test for the quality of clustering of a set of objects," *Journal of Marketing Research*, pp. 456–460, 1975. 80

[186] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974. 80

[187] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a data set," in *Proceedings 2001 IEEE International Conference on Data Mining*, pp. 187–194, IEEE, 2001. 80

[188] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, "Quality scheme assessment in the clustering process," in *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 265–276, Springer, 2000. 80

[189] O. Indicators, "Education at a glance 2016," *Editions OECD*, 2012. 90

[190] J. R. Rice, "The algorithm selection problem," in *Advances in computers*, vol. 15, pp. 65–118, Elsevier, 1976. 110

[191] P. Brazdil, C. G. Carrier, C. Soares, and R. Vilalta, *Metalearning: Applications to data mining*. Springer Science & Business Media, 2008. 110

[192] K. A. Smith-Miles, "Towards insightful algorithm selection for optimisation using meta-learning concepts," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 4118–4124, ieee, 2008. 110

[193] P. Ifinedo, J. Pyke, and A. Anwar, "Business undergraduates' perceived use outcomes of moodle in a blended learning environment: The roles of usability factors and external support," *Telematics and Informatics*, vol. 35, no. 1, pp. 93–102, 2018. 125

[194] B. Baesens, R. Bapna, J. R. Marsden, J. Vanthienen, and J. L. Zhao, "Transformational issues of big data and analytics in networked business.," *MIS quarterly*, vol. 40, no. 4, 2016. 126

[195] J. Lismont, J. Vanthienen, B. Baesens, and W. Lemahieu, "Defining analytics maturity indicators: A survey approach," *International Journal of Information Management*, vol. 37, no. 3, pp. 114–124, 2017. 128

[196] W. H. Rice and H. William, *Moodle*. Packt Publishing Birmingham, 2006. 130

[197] T. Calders and M. Pechenizkiy, "Introduction to the special section on educational data mining," *Acm Sigkdd Explorations Newsletter*, vol. 13, no. 2, pp. 3–6, 2012. 137

[198] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker, *Handbook of educational data mining*. CRC press, 2010. 137

[199] X. Lu, "Artifact-centric log extraction and process discovery," *Unpublished master's thesis, Eindhoven*, 2013. 139

[200] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, pp. 1–37, Jan 2008. 148

[201] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. 148

[202] K. Hechenbichler and K. Schliep, "Weighted k-Nearest-Neighbor Techniques and Ordinal Classification," *Mol. Ecol.*, vol. 399, p. 17, 2004. 148

[203] W. Van der Aalst, A. Adriansyah, and B. Van Dongen, "Replaying history on process models for conformance checking and performance analysis," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 182–192, 2012. 151

[204] W. M. P. Van Der Aalst, B. F. Van Dongen, C. G??nther, A. Rozinat, H. M. W. Verbeek, and A. J. M. M. Weijters, "Prom: The process mining toolkit," *CEUR Workshop Proc.*, vol. 489, 2009. 151

[205] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," vol. 12, pp. 2825–2830, 2012. 154

[206] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006. 157

[207] D. Clow, "MOOCs and the funnel of participation," *Proc. Third Int. Conf. Learn. Anal. Knowl. - LAK '13*, p. 185, 2013. 163

[208] Y. Bergner, D. Kerr, and D. E. Pritchard, "Methodological Challenges in the Analysis of MOOC Data for Exploring the Relationship between Discussion Forum Views and Learning Outcomes," *Proc. 8th Int. Conf. Educ. Data Min.*, pp. 234–241, 2015. 163

[209] E. Dahlstrom, P. Grunwald, T. de Boor, and M. Vockley, "Ecar national study of students and information technology in higher education, 2011," *EDUCUASE Center for Applied Research. Retrieved from http://net. educause. edu/ir/library/pdf/ERS1103/ERS1103W. pdf*, 2011. 172, 173

[210] R. Junco, "Too much face and not enough books: The relationship between multiple indices of facebook use and academic performance," *Computers in human behavior*, vol. 28, no. 1, pp. 187–198, 2012. 173

[211] R. Junco, "ispy: Seeing what students really do online," *Learning, Media and Technology*, vol. 39, no. 1, pp. 75–89, 2014. 173

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
On predicting academic performance with process mining in learning analytics. Journal of Research in Innovative Teaching & Learning, 10(2), 160–176. https://doi.org/10.1108/JRIT-09-2017-0022	
In which Chapter is the Manuscript /Published work:	Chapter 6
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	75%
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
Candidate's Signature:	
Date:	26-Mar-20
Primary Supervisor's Signature:	Teo Susnjak
Date:	26-Mar-20

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis) GRS Version 4– January 2019

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
Prediction of Students' Dropout in MOOC Environment, International Journal of Knowledge Engineering, Vol. 3, No. 2, December 2017 http://www.ijke.org/vol3/85KD015.pdf (ISSN: 2382-6185)	
In which Chapter is the Manuscript /Published work:	Chapter 5
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	80%
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
Candidate's Signature:	
Date:	26-Mar-20
Primary Supervisor's Signature:	Teo Susnjak
Date:	26-Mar-20

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
Data Quality Challenges in Educational Process Mining: Building Process-Oriented Event Logs from Process-Unaware Online Learning, Int. J. of Business Information Systems (in press 2020)	
In which Chapter is the Manuscript /Published work:	Chapter 6 Chapter 4
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	75%
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
Candidate's Signature:	
Date:	26-Mar-20
Primary Supervisor's Signature:	Teo Susnjak
Date:	26-Mar-20

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
A learning analytic approach: Using online weekly student engagement data to make predictions on student performance. 2018 International Conference on Computing, Electronic and Electrical Engineering (IEEE), 1–5.	
In which Chapter is the Manuscript /Published work:	Chapter 5
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	80%
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
Candidate's Signature:	
Date:	26-Mar-20
Primary Supervisor's Signature:	Teo Susnjak
Date:	26-Mar-20

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
Predicting Student's Academic Performance in a MOOC Environment, 11th International Conference on Data Mining, Computers, Communication and Industrial Applications (DMCCIA-2017)	
In which Chapter is the Manuscript /Published work:	Chapter 5
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	85%
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
NA	
Candidate's Signature:	
Date:	26-Mar-20
Primary Supervisor's Signature:	Teo Susnjak
Date:	26-Mar-20

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
Mining Activity Log Data to Predict Student's Outcome in a Course ICBDE'19 Proceedings of the 2019 International Conference on Big Data and Education (ACM) Pages 52-58	
In which Chapter is the Manuscript /Published work:	Chapter 4 Chapter 5
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	85%
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
Candidate's Signature:	
Date:	26-Mar-20
Primary Supervisor's Signature:	Teo Susnjak
Date:	26-Mar-20

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
Prediction of Students' Failure using VLE and Demographic data: Case study Open University Data.	
In which Chapter is the Manuscript /Published work:	Chapter 5
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	90%
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
In. J. Business Intelligence and Data Mining (Accepted)	
Candidate's Signature:	
Date:	27-Jul-2020
Primary Supervisor's Signature:	Teo Susnjak
Date:	27-Jul-2020

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
Preparation for online education: Exploring students' engagement with LMS tools.	
In which Chapter is the Manuscript /Published work:	Chapter 4
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
Computers & Education (Ready for submission)	
Candidate's Signature:	
Date:	27-Jul-2020
Primary Supervisor's Signature:	Teo Susnjak
Date:	27-Jul-2020

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
Application of Meta-learning in Selection of Classification algorithms for prediction of students' performance.	
In which Chapter is the Manuscript /Published work:	Chapter 5
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
Interactive learning Environment (Ready for submission).	
Candidate's Signature:	
Date:	27-Jul-2020
Primary Supervisor's Signature:	
Date:	27-Jul-2020

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Rahila Umer
Name/title of Primary Supervisor:	Dr. Teo Susnjak
Name of Research Output and full reference:	
Current Stance on Predictive Analytics in Higher Education: Opportunities, Challenges and Future Directions.	
In which Chapter is the Manuscript /Published work:	Chapter 2
Please indicate:	
☐ The percentage of the manuscript/Published Work that was contributed by the candidate:	80%
and	
☐ Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Designing study, experiments, writing manuscript and addressing the reviewers' comments.	
For manuscripts intended for publication please indicate target journal:	
Interactive learning Environment (Accepted)	
Candidate's Signature:	
Date:	27-Jul-2020
Primary Supervisor's Signature:	Teo Susnjak
Date:	27-Jul-2020

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)